Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Programa de Pós-Graduação em Estatística

Uma Nova Estatística para a Formação de Redes de Correlação entre Variantes Genéticas

Janaína Pacheco Jaeger

Porto Alegre, março de 2022.

Dissertação submetida por Janaína Pacheco Jaeger como requisito parcial para a obtenção do título de Mestre em Estatística pelo Programa de Pós-Graduação em Estatística da Universidade Federal do Rio Grande do Sul (UFRGS).

Orientadora:

Dra. Gabriela B. Cybis (Departamento de Estatística - UFRGS)

Coorientadora:

Dra. Silvana Schneider (Departamento de Estatística - UFRGS)

Comissão Examinadora:

Dra. Hildete P. Pinheiro (Departamento de Estatística - UNICAMP)

Dr. Claiton H. D. Bau (Departamento de Genética - UFRGS)

Dr. Marcio Valk (Departamento de Estatística - UFRGS)

Data de Defesa: 17 de março de 2022

Agradecimentos

Certamente a realização desse trabalho não teria sido possível sem o apoio de algumas pessoas. Entretanto, começo aqui meu agradecimento à Universidade Federal do Rio Grande do Sul (UFRGS), a qual tenho como minha segunda casa desde 1998 e que é responsável por toda a minha formação e trajetória acadêmica. Instituição pública essa, assim como tantas outras em nosso País, que mantém alto nível de excelência mesmo sem apoio, tanto por parte dos nossos governantes quanto, por vezes, da sociedade. Faço uma crítica em especial ao atual governo, que não reconhece a educação e a ciência como áreas fundamentais para o avanço de uma sociedade e para o desenvolvimento de um País. O que levamos mais de uma década para avançar, retrocedemos em menos de três anos. Mas, parafraseando Paulo Freire com sua afirmação de que "Educação não transforma o mundo. Educação muda pessoas. Pessoas transformam o mundo", sigo meus agradecimentos com admiração a elas e a eles: minhas professoras e meus professores.

Agradeço ao corpo docente do Programa de Pós-Graduação em Estatística (PPGEst) da UFRGS pelo compartilhamento de seus conhecimentos ao longo desses últimos anos. Meu muito obrigada em especial à minha orientadora, Gabriela Cybis, por aceitar participar desse desafio junto comigo. Agradeço à Silvana pela coorientação do trabalho e ao professor Eduardo Horta pela atenção de sempre, parceria e ajuda, principalmente nessa fase final do trabalho.

E o que teria sido desse Mestrado sem meus colegas? Certamente muito mais difícil e com bem menos graça. Que honra ter participado da Primeira Turma dessa Pós-Graduação junto com vocês! Obrigada Arturito e Rafa, pela amizade que construímos, pelas infinitas risadas e pelas melhores figurinhas do WhatsApp.

Certamente cabe aqui um parágrafo somente para agradecer à Pós por também ter me dado um irmão de alma, meu diplo, Felipe Grillo. Minha gratidão e admiração por ti não cabem em palavras. De todos os presentes que o PPGEst me deu, nossa amizade foi o maior e o melhor de todos. Amo tu, 100Hora!

Por fim, agradeço à minha família. Obrigada, Guilherme Porcher, por mais uma vez não ter largado minha mão ao longo desse percurso e, principalmente, por, nesse período, ter feito junto comigo o meu maior sonho, dona de um amor tão grande que nem cabe no peito: nossa filha, Alice. Dedico a ela esse trabalho. Amo vocês!

Resumo

A relação causal entre polimorfismos genéticos e diferentes fenótipos tem fundamental interesse em diversas áreas biológicas. Os Estudos de Associação Genômica Ampla (GWAS) testam milhares de variantes do genoma em busca de marcadores genéticos associados a traços de interesse, auxiliando a compreensão do mapa genótipo-fenótipo para determinada característica. Entretanto, o interesse não está somente na testagem dessas variantes de forma independente, mas também nas interações existentes entre elas. Nesse sentido, metodologias que propõem montagem de redes interligando marcadores correlacionados representam uma estratégia interessante. Climer et al. (2014) propuseram um método que, através do cálculo do Coeficiente de Correlação Personalizado (CCC), calcula correlações entre pares de SNPs para formação de redes alélicas, que são posteriormente testadas entre indivíduos caso e controle em estudos de associação. No entanto, a distribuição de probabilidade e as propriedades estatísticas desse coeficiente não foram estudadas, já que o CCC foi proposto com base em heurísticas e simulações. O presente estudo obteve propriedades estatísticas do CCC sob a hipótese nula de independência entre variantes de diferentes *loci* bialélicos. Em particular, sua esperança sugeriu forte viés de seleção dependente de frequências alélicas. Com a finalidade de eliminar esse viés, propusemos uma nova estatística de correlação, a *Standardized Average Weighted Biallelic Statistic* (SAWB), que denotamos por S_{ij}, calculada a partir da mesma matriz de pesos utilizada no CCC. Para a S_{ij}, foi demonstrada a normalidade assintótica e definido um teste estatístico correspondente. As propriedades estatísticas do CCC e da S_{ij}, assim como de suas estatísticas relacionadas, foram comparadas por estudos de simulação. Da mesma forma, para comparar as redes formadas pelos dois métodos, realizamos uma aplicação em um banco de dados para o Transtorno de Déficit de Atenção e Hiperatividade (TDAH). Tanto os estudos de simulação quanto a aplicação demonstraram os efeitos da seleção dependente de frequência do CCC e verificaram que a S_{ij} corrige esse viés. Além disso, a S_{ij}, com distribuição e propriedades teóricas conhecidas, foi capaz de identificar pares de SNPs correlacionados através de um teste estatístico com Erro Tipo I controlado e maior poder do que o teste baseado na estatística CCC. Portanto, a estatística SAWB mostrou ser uma ferramenta com potencial aplicação em GWAS para formação de redes através de correlações entre pares de SNPs bialélicos.

Palavras-chave: redes de SNPs, *estatística CCC*, *estatística SAWB*, *GWAS*

Abstract

The causal relationship between genetic polymorphisms and different phenotypes is of fundamental interest in several biological areas. The Genome Wide Association Studies (GWAS) test thousands of genome variants searching for genetic markers associated with characteristics of interest, helping improve the understanding of the genotype-phenotype map for a given trait. However, the interest lies not only in testing these variants independently, but also in the interactions between them. In this context, methodologies that propose construction of networks connecting correlated markers are an interesting strategy. Climer et al. (2014) proposed a method that, through the Custom Correlation Coefficient (CCC), computes correlations between pairs of SNPs to build allelic networks, which are subsequently tested between case and control individuals in association studies. However, the probability distribution and statistical properties of this coefficient have not been studied, since the CCC was proposed based on heuristics and simulations. The present study derives statistical properties of the CCC under the null hypothesis of independence between variants of different biallelic *loci*. In particular, its expectation value suggested strong frequency-dependent selection. In order to eliminate this bias, we proposed a new correlation statistic, the Standardized Average Weighted Biallelic Statistic (SAWB), which we denoted by S_{ij}, calculated from the same weight matrix used in the CCC. For S_{ij}, asymptotic normality was demonstrated and a corresponding statistical test was defined. The statistical properties of the CCC and S_{ij}, as well as of their related statistics, were compared by simulation studies. Additionally, to compare the networks constructed by the two methods, we performed an application on a database for Attention Deficit Hyperactivity Disorder (ADHD). Both the simulation studies and the application demonstrated the frequency-dependent selection effects of CCC and corroborated that S_{ij} corrects this bias. Furthermore, the S_{ij} statistic, with known distribution and theoretical properties, was able to identify pairs of correlated SNPs through a statistical test with controlled Type I Error and more power than the test based on the CCC. Therefore, the SAWB statistic was shown to be a tool with interesting potential for application in GWAS through network construction by correlating pairs of biallelic SNPs.

Keywords: SNP networks, *CCC statistic*, *SAWB statistic*, *GWAS*.

Índice

Capítulo 1

Introdução

1.1 Conceitos Básicos em Genética e Biologia Molecular

Um dos obstáculos que profissionais de diferentes áreas precisam superar ao entrar em um novo campo científico é a compreensão da terminologia. A dificuldade de entendimento do conjunto de termos próprios não é diferente na área da Genética e Biologia Molecular. Nesse sentido, serão abordados aqui alguns conceitos básicos e terminologias comumente utilizados nessa área. Certamente, o objetivo dessa seção não é esgotar o assunto, uma vez que é vasto e complexo, mas sim auxiliar o leitor na compreensão do estudo apresentado nessa dissertação, tanto em relação aos termos empregados no texto, quanto à metodologia e à aplicação descritas.

1.1.1 Terminologias em Genética e Biologia Molecular

Genes são seções de uma molécula helicoidal filamentar chamada ácido desoxirribonucleico (DNA), os quais ditam as propriedades inerentes a uma determinada espécie. Os genes são reconhecidos como unidades moleculares hereditárias e são formados por uma sequência ordenada de moléculas chamadas nucleotídeos. Na maioria das vezes, os genes codificam uma sequência de aminoácidos de uma proteína específica, mas também podem ser transcritos em moléculas de ácido ribonucléico (RNA) que atuam na regulação e na expressão de outros genes (GRIFFITHS et al., 2006). Os seres humanos têm aproximadamente 20.000 genes em seus 23 pares de cromossomos (NHGRI, 2022). **Cromossomos** são moléculas de DNA organizadas e condensadas no interior das células (ALBERTS et al., 2002) e o conjunto completo de informações genéticas encontradas em uma célula é chamado de **genoma** (NHGRI, 2022).

Tomando ainda o genoma humano como exemplo, uma pessoa tem duas cópias de cada gene, uma herdada de sua mãe e outra de seu pai. A maioria das formas gênicas é a mesma em todas as pessoas, mas em um pequeno número de genes encontramos formas diferentes na população. Chamamos de **alelos** as formas do mesmo gene com pequenas diferenças em sua sequência de nucleotídeos do DNA (GRIFFITHS et al., 2006). Essas pequenas diferenças contribuem para que cada pessoa apresente traços únicos, isto é, alelos diferentes de um mesmo gene produzem características diferentes na população. Ainda, se os dois alelos forem os mesmos, dizemos que o indivíduo é homozigoto para esse gene. Já se os alelos forem diferentes, o indivíduo é heterozigoto. Embora o termo alelo seja originalmente usado para descrever a variação entre os genes, atualmente também se refere à variação entre sequências

não codificantes do DNA, ou seja, que não comandam a síntese de proteínas (NHGRI, 2022). A Figura 1.1 representa diferentes pares de alelos em cromossomos humanos, evidenciando sua estrutura dada por moléculas de DNA condensadas e exemplificando a variabilidade alélica em decorrência de quatro principais bases nitrogenadas (adenina, citosina, timina e guanina), moléculas constituintes dos nucleotídeos.

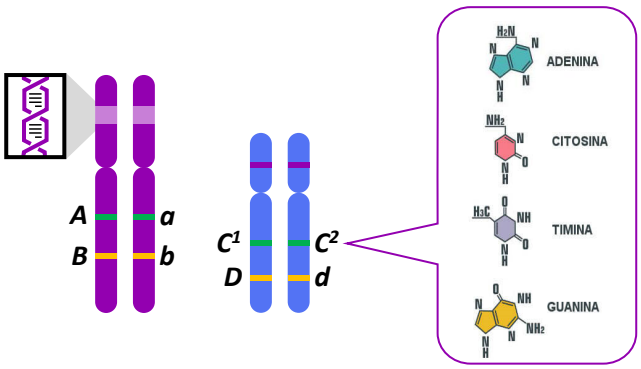

Figura 1.1: Identificação de genes alelos nos cromossomos humanos.

A constituição alélica de um organismo é o seu **genótipo**, que é o correspondente hereditário do **fenótipo** (GRIFFITHS et al., 2006). Tomando o albinismo como exemplo, os genótipos podem ser AA, Aa ou aa. Nesse caso, o fenótipo é pigmentado para os genótipos AA e Aa e albino para aa. A habilidade em sintetizar o pigmento predomina em relação à inabilidade, portanto, o alelo A é dominante em relação ao alelo a, dito recessivo. Assim, o genótipo é a informação hereditária de um organismo, enquanto o fenótipo é a expressão da informação genética sobre a qual a seleção natural atua (ALBERTS et al., 2002). Os pares de genes, ou alelos, ocorrem nos cromossomos em determinadas posições, ou ***loci*** (singular: **locus**). Portanto, alelos são variantes gênicas que ocorrem em um mesmo locus (ELSTON, 2000).

Em uma população natural, a existência de duas ou mais variantes comuns (frequência alélica maior do que 1%) é chamada de **polimorfismo**, enquanto que as várias formas são chamadas de **morfos** (GRIFFITHS et al., 2006). Quando temos um locus, no qual seus diferentes alelos determinam o fenótipo de uma característica quantitativa, o chamamos de **QTL** (*quantitative trait locus*). Tipicamente, a palavra 'quantitativa' é usada quando nos referimos a variações 'contínuas' de um traço. No entanto, o fenótipo de uma característica quantitativa pode ser discreto. Chamamos de característica quantitativa traços fenotípicos determinados, juntamente com o meio ambiente, por muitos genes de pequeno efeito (HEINO, 2013). Para a medição do efeito de um QTL no fenótipo observado, os alelos têm um efeito aditivo se a distribuição fenotípica do heterozigoto for a média das duas distribuições dos homozigotos correspondentes. Qualquer variação das médias fenotípicas dos genótipos acima ou abaixo desse efeito aditivo é chamado de efeito de dominância (ELSTON, 2000).

Quando o polimorfismo envolve a variação de um único nucleotídeo em uma posição específica do genoma, ele é chamado de Polimorfismo de Nucleotídeo Único (*Single Nucleotide Polymorphism* - SNP) (NHGRI, 2022). Comparações de sequências do DNA entre diferentes indivíduos da população revelam que essa variabilidade pode ser representada por dois nucleotídeos diferentes (SNP bialélico), ou mais (polialelia), mesmo que indivíduos diploides (com dois conjuntos cromossômicos) apresentem apenas duas cópias desses alelos. Como exemplo de SNP bialélico, podemos considerar um determinado locus que apresente

os alelos B e b como variantes e os indivíduos para esse SNP podem apresentar os genótipos BB, Bb ou bb, sendo uma variante herdada de cada um dos seus genitores. Dependendo de onde está localizado um SNP, diferentes consequências podem ocorrer em nível fenotípico. Aqueles presentes em regiões codificantes dos genes podem alterar a função e a estrutura de proteínas e, portanto, são analisados para fins de diagnóstico de doenças (SYVANEN, 2002). Já os SNPs que alteram a estrutura primária de proteínas envolvidas no metabolismo de drogas são alvos de estudos farmacogenéticos (EVANS; RELLING, 1999). Além disso, é provável que SNPs presentes em regiões reguladoras de genes possam influenciar o risco de desenvolvimento de doenças comuns. Entretanto, a maioria desses polimorfismos está localizada em regiões não codificadoras do genoma, não tendo, portanto, impacto sobre o fenótipo de um indivíduo. Contudo, tais variantes podem ser úteis como marcadores genéticos populacionais em estudos evolutivos (JORDE et al., 2000; GIOVANETTI et al., 2020).

1.2 Estudos de Associação Genômica Ampla (GWAS) e Métodos Estatísticos de Correlação para Análise de GWAS

A relação causal entre polimorfismos genéticos intraespecíficos e diferenças fenotípicas observadas entre indivíduos é de fundamental interesse em diversas áreas biológicas (KORTE; FARLOW, 2013). A capacidade de prever fatores genéticos de risco para doenças e características complexas requer a identificação dos *loci* correspondentes e a compreensão da arquitetura genética subjacente a um determinado fenótipo (REICH; LANDER, 2001). Diante do efeito poligênico (muitos genes de pequeno efeito) sob essas características, torna-se cada vez mais evidente a necessidade de estudos de associação de maior dimensão (grande conjunto de dados de SNPs) e poder estatístico adequado (LANDER; SCHORK, 1994; NJ; MERIKANGAS, 1996). Os estudos de associações genômica ampla (GWAS) visam identificar associações de genótipos com fenótipos, testando as diferenças nas frequências alélicas de variantes genéticas entre indivíduos com fenótipos diferentes (CANTOR; LANGE; SINSHEIMER, 2010; KORTE; FARLOW, 2013; UFFELMANN et al., 2021). Nessa perspectiva, a identificação de marcadores ligados a QTL por meio da análise de GWAS auxilia a compreensão do mapa genótipo-fenótipo de uma característica de interesse. Por essa abordagem, são testadas milhares de variantes, a fim de encontrar aquelas estatisticamente associadas a uma característica específica, ou doença. Entretanto, o interesse não está somente em testar alelos de forma independente, mas também interações existentes entre eles. Portanto, explorar todo o genoma na busca de associações é ainda um desafio.

São diversos os métodos que analisam correlações e interações entre genótipos e alelos. As ferramentas PLINK (PURCELL et al., 2007) e LD Matrix (CROSSLIN; QIN; HAUSER, 2010), por exemplo, são amplamente utilizadas em análises de GWAS com diferentes propósitos, incluindo esses supracitados. Dessa forma, podem ser identificadas correlações entre alelos para posterior utilização de agrupamentos dessas variantes pré-selecionadas que serão testados em estudos de associação. Nessa subseção serão apresentadas as seguintes ferramentas estatísticas para análise de GWAS por correlação: Coeficiente de Correlação de Pearson (PCC), Medidas de Desequilíbrio de Ligação (LD), Coeficiente de Correlação Customizado (CCC) e Estatística Maestro. As propriedades matemáticas, aplicações e comparações entre essas estatísticas podem ser acessadas em outros trabalhos (CANZONIERO; ROSENBERG, 2008; KANG; ROSENBERG, 2020; LEWONTIN, 1988).

1.2.1 Coeficiente de Correlação de Pearson (PCC)

O Coeficiente de Correlação de Pearson (PCC) é uma medida de correlação amplamente utilizada em análises de dados genéticos, principalmente em estudos de desequilíbrio de ligação (LD). O LD refere-se a uma associação na ocorrência de alelos em dois *loci* distintos (HUDSON, 2004; PRITCHARD; PRZEWORSKI, 2001; SLATKIN, 2008) e suas medições são tipicamente baseadas na comparação das frequências observadas com as frequências esperadas desses alelos em conjunto (haplótipos), supondo independência entre eles (LEWONTIN, 1988).

Consideremos i um possível estado ou variante (alelo A ou a, por exemplo) associada a um dado SNP, digamos SNP_1. Semelhantemente, consideremos j um possível estado ou variante (alelo B ou b, por exemplo) associada a outro SNP, digamos SNP_2, onde SNP_1 e SNP_2 ocupam *loci* distintos no genoma. Sejam p_i e p_j as frequências populacionais dos alelos i e j, respectivamente, e $p_{i^c} = 1 - p_i$ e $p_{j^c} = 1 - p_j$ as frequências populacionais dos alelos i^c e j^c, respectivamente, onde $a^c = A$ e $b^c = B$. Seja ainda p_{ij} a frequência populacional de indivíduos com ambos os alelos i e j, isto é, do haplótipo ij. A medida de correlação entre dois SNPs através do PCC é dada por

$$r = \frac{p_{ij} - p_i p_j}{\sqrt{p_i p_{i^c} p_j p_{j^c}}},$$

em que $p_{ij} - p_i p_j =: D$ é a medida padrão de LD que calcula o desvio entre as frequências observadas e esperadas dos haplótipos. Se ambos os alelos estão em equilíbrio de ligação, então $D = 0$. Caso estejam em desequilíbrio de ligação, $D \neq 0$. Ainda, quando há um número maior de haplótipos do que a frequência esperada sob independência alélica, observam-se desvios positivos (D positivo). Já um número menor de haplótipos gera desvios negativos (D negativo) (LEWONTIN, 1988).

1.2.2 Medidas de Desequilíbrio de Ligação (LD)

Desequilíbrio de Ligação (LD) é um conceito fundamental em genética populacional, sendo utilizado em uma ampla variedade de contextos, tais como mapeamento de associações e detecção de seleção natural em diferentes populações (KANG; ROSENBERG, 2020). Como mencionado anteriormente, a medida padrão de LD é $D = p_{ij} - p_i p_j$. Tomando $i \in \{A, a\}$ e $j \in \{B, b\}$, a expressão para D pode ser formulada utilizando cada uma das combinações possíveis de alelos nos dois *loci* (AB, Ab, aB, e ab). Essas quatro formulações de D devolvem valores idênticos em módulo, havendo troca de sinal em relação aos desvios positivos e negativos. Normalizações dessa medida padrão permitem a avaliação da magnitude do desequilíbrio de ligação sem considerar seu sinal, ou seja, tomar um valor relativo de D (LEWONTIN, 1964). A normalização também proporciona a diminuição da influência das frequências alélicas sobre o valor da medida de LD (ZAPATA, 2000). Existem diferentes medidas de LD baseadas na normalização da medida D, as quais seguem descritas abaixo.

A medida D' (LEWONTIN, 1964), utilizada para medir LD entre pares de *loci* multialélicos (ZAPATA, 2000), é obtida pela normalização de D pela sua magnitude máxima, dado seu respectivo sinal,

$$D' = \frac{D}{D_{max}},$$

onde

$$D_{max} = \begin{cases} \min[p_i p_j^c, p_i^c p_j] \text{ se } D > 0 \\ \min[p_i p_j, p_i^c p_j^c] \text{ se } D < 0. \end{cases}$$

Portanto, se D for positivo, o valor de D_{max} irá corresponder ao menor valor obtido pelo produto das frequências dos alelos não associados, isto é, que não compõem o haplótipo. Já se o valor de D for negativo, o valor de D_{max} irá corresponder ao menor valor obtido pelo produto das frequências dos alelos que compõem os haplótipos em associação.

Já a medida r^2 (CANZONIERO; ROSENBERG, 2008), calculada a partir do quadrado do PCC (r), é definida como a medida D^2 normalizada pelo produto das frequências dos quatro alelos,

$$r^2 = \frac{D^2}{p_i p_i^c p_j p_j^c}.$$

O r^2, além de ser uma das medidas de LD mais utilizadas para locus bialélicos (MANGIN et al., 2011), também permite cálculos de tamanho efetivo populacional, parâmetro utilizado em estudos de genética de populações (WAPLES; ENGLAND, 2011).

1.2.3 Coeficiente de Correlação Customizado (CCC)

O Coeficiente de Correlação Customizado (CCC) é uma estatística proposta por Climer et al. (2014) que calcula correlações entre pares de SNPs e forma redes entre alelos. Estas redes podem, posteriormente, serem utilizadas em estudos de associação entre grupos caso e controle para uma determinada característica. O objetivo dessa análise é a identificação de padrões multi-SNPs associados a um fenótipo complexo de interesse.

Tomando $i \in \{A, a\}$ e $j \in \{B, b\}$ como variantes de dois SNPs bialélicos distintos, o CCC entre os alelos i e j é definido por

$$CCC_{ij} = \frac{9}{2} R_{ij} F_i F_j,$$

sendo R_{ij} um escore bialélico médio ponderado e F_i e F_j fatores de frequência, definidos abaixo.

O escore R_{ij}, que mede o relacionamento entre os alelos i e j em toda a amostra, é definido como

$$R_{ij} = \frac{1}{n} \sum_{k=1}^{n} r_{ij,k},$$

onde o escore bialélico ponderado $r_{ij,k}$ para o indivíduo k é atribuído de acordo com seus genótipos, conforme a Figura 1.2.

O Fator de Frequência F_i é definido por

$$F_i = 1 - \frac{f_i}{q},$$

		SNP_2					
		BB		**Bb**		**bb**	
SNP_1	**AA**	$r_{AB,k}$=1 $r_{aB,k}$=0	$r_{Ab,k}$=0 $r_{ab,k}$=0	$r_{AB,k}$=1/2 $r_{aB,k}$=0	$r_{Ab,k}$=1/2 $r_{ab,k}$=0	$r_{AB,k}$=0 $r_{aB,k}$=0	$r_{Ab,k}$=1 $r_{ab,k}$=0
	Aa	$r_{AB,k}$=1/2 $r_{aB,k}$=1/2	$r_{Ab,k}$=0 $r_{ab,k}$=0	$r_{AB,k}$=1/4 $r_{aB,k}$=1/4	$r_{Ab,k}$=1/4 $r_{ab,k}$=1/4	$r_{AB,k}$=0 $r_{aB,k}$=0	$r_{Ab,k}$=1/2 $r_{ab,k}$=1/2
	aa	$r_{AB,k}$=0 $r_{aB,k}$=1	$r_{Ab,k}$=0 $r_{ab,k}$=0	$r_{AB,k}$=0 $r_{aB,k}$=1/2	$r_{Ab,k}$=0 $r_{ab,k}$=1/2	$r_{AB,k}$=0 $r_{aB,k}$=0	$r_{Ab,k}$=0 $r_{ab,k}$=1

Figura 1.2: Escores atribuídos para as quatro relações possíveis entre um par de SNP bialélico.

em que f_i a frequência amostral do alelo i e q um parâmetro de ajuste igual a 1,5. O Fator de Frequência F_j é calculado de forma análoga para o alelo j.

Para cada interação entre um par de SNPs, quatro valores de CCC_{ij} são calculados, isto é $CCC = (CCC_{AB}, CCC_{Ab}, CCC_{aB}, CCC_{ab})$. Após, o valor máximo desses quatro valores (CCC_{max}) é comparado com um limiar para a seleção dos alelos que irão compor as redes. Climer et al. (2014) utilizaram um método *ad hoc*, que comparou por QQ-plot desvios das distribuições de valores de CCC_{max} calculados em dados reais e simulados, para o estabelecimento deste limiar. No estudo que compõe essa dissertação foram utilizados testes de permutações para essa finalidade (método descrito no artigo a seguir).

Assim, a partir do grupo caso para uma característica de interesse, os valores de CCC_{ij} são calculados e os CCC_{max} comparados ao limiar definido, para todos os pares de SNPs bialélicos. As correlações selecionadas entre SNPs ($CCC_{max} >$ limiar) são utilizadas para a formação de redes, com vértices representando os alelos e arestas as relações entre eles. Por fim, as frequências dessas redes são testadas entre grupos caso e controle para a realização de estudos de associação. Cabe ressaltar que as propriedades estatísticas do CCC_{ij} e descrições mais detalhadas do método serão apresentadas no Capítulo 2.

1.2.4 Método Maestro

Similar ao CCC, o método Maestro (CLIMER et al., 2020) é uma ferramenta de formação de redes, que liga genes correlacionados em relação ao seu padrão de expressão. Essa metodologia considera dois vértices para cada gene, um representando alta (H) e outro representando baixa (L) expressão gênica e, a partir dessa relação, identifica correlações entre tais padrões. Assim, o método Maestro identifica esses padrões sem levar em conta o fenótipo da característica de interesse, já que utiliza toda a amostra (indivíduos casos e controles) para a formação dos clusters. Somente após as redes serem formadas é que os estudos de associação são realizados.

O método Maestro utiliza a estatística *Duo* para o cálculo da correlação entre um par de genes, considerando quatro tipos de relações relevantes entre eles: ambos com alta expressão (HH), ambos com baixa expressão (LL), ou genes com expressão diferenciada e anticorrelacionadas (HL ou LH). Seja $t \in \{HH, LL, HL, LH\}$. Considerando k e l dois genes presentes no genoma, a porcentagem de indivíduos que apresentam uma das quatro relações citadas acima é designada por $R_{kl}[t]$. A estatística *Duo*, portanto, é definida pela

expressão

$$Duo_{kl}[t] = 4R_{kl}[t]F_{kt}F_{lt},$$

sendo F_{kt} o Fator de Frequência para o gene k, calculado por $F_{kt} = 1 - \frac{f_{kt}}{q}$ e f_{kt} correspondente à frequência da relação t no grupo de indivíduos (por exemplo, ambos os genes com alta expressão). O Fator de Frequência F_{lt} é calculado de forma análoga para o gene l. O parâmetro de ajuste q é igual a 1,5, como definido para o CCC.

No método Maestro, os 1.000 valores mais altos dos coeficientes de correlação Duo são selecionados para comporem as redes. Após, as frequências das redes de cada padrão de expressão são comparadas entre grupos caso e controle em estudos de associação para uma dada característica.

1.3 Escopo deste trabalho

Nessa dissertação, será apresentado o artigo "*Standardized Average Weighted Biallelic statistic (SAWB): a new method for identifying genetic correlation networks*", o qual define propriedades estatísticas do CCC e propõe uma nova estatística de teste a partir de um coeficiente de correlação denominado *Standardized Average Weighted Biallelic statistic (SAWB)*. Essa estatística, denotada por S_{ij}, tem como base a mesma matriz de pesos utilizada no CCC e calcula correlações entre pares de SNPs para formação de redes, que posteriormente são testadas entre grupos caso e controle em estudos de associação. Além disso, através de simulações, foram comparadas algumas propriedades estatísticas do CCC e S_{ij}, tais como seleção dependente de frequência, Erro Tipo I e poder. Por fim, o método proposto foi aplicado para a análise de GWAS em um banco de dados com genótipos de indivíduos casos e controles para o Transtorno de Déficit de Atenção e Hiperatividade (TDAH).

Capítulo 2

Artigo

Nesse capítulo será apresentado o artigo *Standardized Average Weighted Biallelic statistic (SAWB): a new method for identifying genetic correlation networks*, que contém os resultados do estudo desenvolvido nessa dissertação.

Standardized Average Weighted Biallelic statistic (SAWB): a new method for identifying genetic correlation networks

Janaína P. Jaeger[1,2], Felipe G. Pinheiro[1], Silvana Schneider[1], Eduardo Horta[1], and Gabriela B. Cybis[1]

[1]*Department of Statistics, Federal University of Rio Grande do Sul, RS, Brazil*

[2]*Federal Institute of Education, Science and Technology Sul-rio-grandense, RS, Brazil*

Abstract

The causal relationship between genetic polymorphisms and different phenotypes is of fundamental interest in several biological areas. The Genome Wide Association Studies (GWAS) test thousands of genome variants searching for genetic markers associated with characteristics of interest, helping improve the understanding of the genotype-phenotype map for a given trait. However, the interest lies not only in testing these variants independently, but also in the interactions between them. In this context, methodologies that propose construction of networks connecting correlated markers are an interesting strategy. Climer et al. (2014b) proposed a method that, through the Custom Correlation Coefficient (CCC), computes correlations between pairs of SNPs to build allelic networks, which are subsequently tested between case and control individuals in association studies. However, the probability distribution and statistical properties of this coefficient have not been studied, since the *CCC* was proposed based on heuristics and simulations. The present study derives statistical properties of the *CCC* under the null hypothesis of independence between variants of different biallelic *loci*. In particular, its expectation value suggested strong frequency-dependent selection. In order to eliminate this bias, we proposed a new correlation statistic, the Standardized Average Weighted Biallelic Statistic (SAWB), which we denoted by S_{ij}, calculated from the same weight matrix used in the *CCC*.

For S_{ij}, asymptotic normality was demonstrated and a corresponding statistical test was defined. The statistical properties of the CCC and S_{ij}, as well as of their related statistics, were compared by simulation studies. Additionally, to compare the networks constructed by the two methods, we performed an application on a database for Attention Deficit Hyperactivity Disorder (ADHD). Both the simulation studies and the application demonstrated the frequency-dependent selection effects of CCC and corroborated that S_{ij} corrects this bias. Furthermore, the S_{ij} statistic, with known distribution and theoretical properties, was able to identify pairs of correlated SNPs through a statistical test with controlled Type I Error and more power than the test based on the CCC. Therefore, the SAWB statistic was shown to be a tool with interesting potential for application in GWAS through network construction by correlating pairs of biallelic SNPs.
Keywords: SNP networks, *CCC statistic*, *SAWB statistic*, *GWAS*.

1 Introduction

Genome-Wide Association Studies (GWAS) scan the genome searching for genetic markers that are associated to traits of interest. These studies have provided numerous gains for determining the genetic architecture of several complex diseases, whose manifestations depend on exposure to multiple environmental, social and genetic factors (Peprah et al., 2014; Park et al., 2017; Ishigaki et al., 2020). Understanding the genetic architecture of complex diseases brings prospects of advances leading to better comprehension of these diseases through the identification of specific targets for prevention and risk reduction.

Because only a few gene variants are expected to be associated with disorder outcomes, GWAS is the standard approach for the identification of disease genes. The premise of GWAS is that common genome variation, such as single nucleotide polymorphisms (SNPs) with frequencies greater than 1%, is responsible for the risk of most genetically complex disorders (Cantor et al., 2010). However, as the effects of these gene variants do not necessarily occur independently of each other, the study of interactions between alleles from different SNPs is instrumental to identifying genetic panels underlying these disorders.

GWAS can be performed using any variation of the genome, although it is most commonly used with SNPs. The human genome has around 3.1 billion base pairs (NIH, 2022), of which around 4 to 5 million sites are SNPs (Consortium et al., 2015). Genotyping of individuals can be performed by microarrays for common variants, or by whole-genome sequencing (WGS) or whole-exome sequencing (WES), which include rare variants. In order to analyze these data, the most common approach in GWAS is to test each SNP individually, whether through

additive, non-additive, or by linear or logistic regression model (Uffelmann et al., 2021). Additionally, because of the abovementioned allelic interactions, it becomes important to evaluate the effect of combinations of the variants, but due to their combinatorial nature, testing for higher level interactions is computationally intractable. Thus, calculating pairwise interactions for network assembly and then applying a single association test for each cluster appears as an interesting alternative strategy.

Since the components of a human cell exhibit functional interdependencies, a trait is rarely a consequence of an abnormality in a single gene. In this context, network medicine is a tool that helps explore not only the complexity of a given disease, but also helps identify the molecular relationships between distinct phenotypes. Advances in this area are critical for identifying new genes associated with pathologies, for understanding the biological significance of associated diseases, and for identifying drug targets and biomarkers for complex diseases (Cano-Gamez and Trynka, 2020). While much of our understanding of biological systems is based on studies with model organisms, attention is currently focused on molecular network approach. In such system networks, vertices represent proteins, metabolites, transcription factors or genes, while edges represent their relationships, such as physical interactions, biochemical reactions, regulatory relationships, or correlations (Barabási et al., 2011). Accordingly, different studies have focused on building SNP and gene networks, which may have a direct or indirect functional relationship with a trait of interest (Liu et al., 2011; McCarter et al., 2020; Grimes and Datta, 2021).

Climer et al. (2014b) proposed a method for network construction using the Custom Correlation Coefficient (CCC), which aims to test multi-SNP association with complex traits in genome-wide studies. The approach is based on calculating pairwise correlations between (SNP-SNP) alleles to form networks (or clusters) that are subsequently tested between case and control individuals. This represents a new approach in genetic studies, since the networks connect alleles and not genotypes. The authors claim that the CCC accommodates genetic heterogeneity, in which different subsets of individuals develop a given disease due to different sets of genetic factors, and that the method is able to identify multi-allelic networks for use in association studies, even when the variants are rare. In addition, the CCC was compared with two other correlation measures widely used in genetic analyses: Pearson's correlation coefficient (PCC) and linkage disequilibrium r^2 measure (for further details see (Hayes, 2013; Gonzalez-Recio et al., 2014; Waldmann, 2019)). The CCC was more sensitive to capturing significant correlations between SNPs than the other approaches, and responded smoothly to small changes in the genotypes. Furthermore, the authors provide a software package named BlocBuster to perform the analysis and show that their method demands less computational time than usual methods of genome-wide analysis considering SNP interactions. Thus, the

CCC is an alternative for evaluating the relationship between alleles of different SNPs in a population. However, its probability distribution and theoretical properties have not yet been studied, as the statistic was proposed based on heuristics and simulations, without the derivation of its theoretical properties.

The CCC statistic and accompanying methodology have been used in different studies. Its application was able to identify networks of genetic variants associated with hypertensive heart disease (Climer et al., 2014b) and psoriasis (Climer et al., 2014a). This coefficient has also been applied in evolutionary studies which investigated the genetic coadaptation between skin color genes and vitamin D receptors in different populations (Tiosano et al., 2016; Missaggia et al., 2020). Climer et al. (2020) proposed a similar statistic to CCC for gene expression-based networks and applied it in association studies for Alzheimer's disease (AD). Joubert et al. (2019) also described strategies for efficient mapping of the calculations to many-node parallel systems of CCC, avoiding performance penalty by redundant and unnecessary computations, as well as expanded this evaluation to SNP trios. These findings address the limitation of identifying markers of small effect when they are analyzed individually, because the structure of the networks allows for the identification of those small effect alleles that collectively may have a significant effect on a given phenotype.

To better understand the CCC statistic the present study derives basic statistical properties of the CCC under a null hypothesis of allele independence. In particular, we compute its expected value and find that it suggests a strong selection bias in terms of allelic frequencies. In order to eliminate this bias, we propose a new correlation statistic, the Standardized Average Weighted Biallelic Statistic (SAWB), which we denote by S_{ij}, calculated from the same weight matrix used in the CCC. We also show its asymptotic normality and define the corresponding statistical test. Both simulation studies and an application corroborate the profound effects of frequency-dependent selection on the CCC and how S_{ij} can be used to correct for such biases. Of note, we see that the S_{ij} test shows proper Type I Error control and appears consistent while the CCC presents a pathological behavior. Thus the S_{ij} statistic can be seen as an interesting alternative for performing the type of multi-allelic network selection for association studies proposed by Climer et al. (2014b) while avoiding the strong biases of the CCC statistic.

The paper is organized as follows. In section 2 the CCC statistic and its theoretical properties are presented. In this section we also define the Average Weighted Biallelic statistic S_{ij} and derive its asymptotic theory. Section 3 presents simulation studies evaluating the frequency-dependent selection of the CCC and comparing the behavior of both methods. Then, in section 4, both methods are applied to genotype data in cases and controls from a GWAS study of Attention-deficit/hyperactivity disorder (ADHD). Finally, in section 5 we

discuss the repercussions of our findings.

2 Methods

2.1 Custom Correlation Coefficient (CCC)

In this section we explore the CCC introduced by Climer et al. (2014b), a statistic that identifies correlation between alleles of SNPs.

Consider a sample of n individuals genotyped for biallelic SNPs. Here a SNP represents a specific locus in the genome, an allele represents a variant of the information contained in the corresponding SNP, and a biallelic SNP is a SNP that holds only two alleles (e.g. if SNP_1 is a biallelic SNP containing alleles A and a, the possible genotypes for any individual are AA, Aa or aa). If we let $i \in \{A, a\}$ and $j \in \{B, b\}$ denote the alleles of two distinct biallelic SNPs for which we are interested in assessing correlation, then the Custom Correlation Coefficient for alleles i and j is defined as

$$CCC_{ij} = \frac{9}{2} R_{ij} F_i F_j, \tag{1}$$

where R_{ij} is the average weighted biallelic score and F_i and F_j are frequency factors (both defined below).

In order to quantify the genetic co-occurrence of a pair of alleles (i, j), CCC_{ij} is computed using the weighted biallelic score, which is based on the expected frequency of allelic combinations, where the weights are derived assuming independence. For each individual k in the sample ($k \in \{1, ..., n\}$), a weight $r_{ij,k}$ is assigned according to the four possible relationships between the pair of alleles i in SNP_1 and j in SNP_2 (Table 1).

Table 1: CCC assigned weights for the four possible relationships between a biallelic SNPs.

		SNP_2					
		BB		**Bb**		**bb**	
SNP_1	**AA**	$r_{AB,k}$=1 $r_{aB,k}$=0	$r_{Ab,k}$=0 $r_{ab,k}$=0	$r_{AB,k}$=1/2 $r_{aB,k}$=0	$r_{Ab,k}$=1/2 $r_{ab,k}$=0	$r_{AB,k}$=0 $r_{aB,k}$=0	$r_{Ab,k}$=1 $r_{ab,k}$=0
	Aa	$r_{AB,k}$=1/2 $r_{aB,k}$=1/2	$r_{Ab,k}$=0 $r_{ab,k}$=0	$r_{AB,k}$=1/4 $r_{aB,k}$=1/4	$r_{Ab,k}$=1/4 $r_{ab,k}$=1/4	$r_{AB,k}$=0 $r_{aB,k}$=0	$r_{Ab,k}$=1/2 $r_{ab,k}$=1/2
	aa	$r_{AB,k}$=0 $r_{aB,k}$=1	$r_{Ab,k}$=0 $r_{ab,k}$=0	$r_{AB,k}$=0 $r_{aB,k}$=1/2	$r_{Ab,k}$=0 $r_{ab,k}$=1/2	$r_{AB,k}$=0 $r_{aB,k}$=0	$r_{Ab,k}$=0 $r_{ab,k}$=1

The average weighted biallelic score R_{ij}, that measures the relationship between alleles i and j in the whole sample, is defined as

$$R_{ij} = \frac{1}{n}\sum_{k=1}^{n} r_{ij,k},$$

where the weighted biallelic score $r_{ij,k}$ for individual k is assigned according to its genotypes and the scheme in Table 1 (for example, $r_{aB,k} = 1/2$ if individual k has $SNP_1 = aa$ and $SNP_2 = Bb$ or $SNP_1 = Aa$ and $SNP_2 = BB$). It follows from the definition that $R_{ij} \in [0, 1]$ and $R_{AB} + R_{Ab} + R_{aB} + R_{ab} = 1$. Also, notice that if the sample is random, then the random variables $r_{ij,1}, \ldots, r_{ij,n}$ are independent and identically distributed (iid).

The frequency factors F_i and F_j in equation (1) are introduced with the intent of adjusting CCC_{ij} for the effect of rare variants. They are defined as

$$F_i = 1 - \frac{f_i}{q}, \tag{2}$$

where f_i is the frequency of allele i in the sample and q is a tuning parameter set to 1.5, as proposed by Climer et al. (2014b). The frequency factor F_j is calculated analogously for allele j. The average weighted biallelic score R_{ij} and the frequency factors F_i and F_j are then multiplied by the constant 9/2 in order to obtain $CCC_{ij} \in [0, 1]$, thus completing the formula defined in expression (1) .

Importantly, for each interaction between a pair of SNPs, four CCC_{ij} values are calculated, i.e. $CCC = (CCC_{AB}, CCC_{Ab}, CCC_{aB}, CCC_{ab})$. Then, the maximum of these four values $CCC_{max} = \max\{CCC\}$ is compared with a threshold for identifying pairs of alleles with relevant correlations. Climer et al. (2014b) used an *ad hoc* method that compared real and simulated data for establishing the threshold value. From the selected correlations between all SNPs in the data, networks are constructed with vertices represented by alleles from different SNP and edges determining the relevant relationships between them. These clusters are then tested between case and control individuals to perform an association study with the trait of interest.

2.2 Statistical properties of the Custom Correlation Coefficient (CCC) and related statistics

In the following sessions we derive statistical properties of the CCC and related statistics under a null hypothesis of SNP independence. Before proceeding, it will be conve-

nient to summarize our modeling assumptions in terms of a collection of random vectors $(X_{ak}, X_{Ak}, X_{bk}, X_{Bk})$, $1 \leqslant k \leqslant n$, having Binomial marginals. Here the random variable X_{Ak} represents the number of alleles A, in SNP_1, for individual k; X_{bk} is the number of alleles b, in SNP_2, for the same individual, and so on. Additionally let p_ℓ represent the population frequency of allele $\ell \in \{A, a, B, b\}$ such that $0 \leqslant p_a, p_b \leqslant 1$, $p_A = 1 - p_a$, and $p_B = 1 - p_b$. The assumptions are

Assumptions.

A1 (random sampling) the random vectors $(X_{ak}, X_{Ak}, X_{bk}, X_{Bk})$, $1 \leqslant k \leqslant n$, are independent draws from $(X_a, 2 - X_a, X_b, 2 - X_b)$, with $X_\ell \sim \text{Binomial}(2, p_\ell)$ for $\ell \in \{a, b\}$.

A2 (SNP independence) for $i \in \{A, a\}$, $j \in \{B, b\}$ and $1 \leqslant k \leqslant n$, the random variables X_{ik} and X_{jk} are mutually independent.

Remark. In terms of these Binomial random variables, the allelic frequencies f_ℓ, with $\ell \in \{A, a, B, b\}$, can be written as

$$f_\ell = \frac{1}{2n} \sum_{k=1}^{n} X_{\ell k} \tag{3}$$

and similarly, as can be easily established by inspecting Table 1, the weighted biallelic score $r_{ij,k}$ is given by

$$\begin{aligned} r_{ij,k} := \tfrac{1}{4}\mathbb{I}[X_{ik} = 1, X_{jk} = 1] + \tfrac{1}{2}\mathbb{I}[X_{ik} = 1, X_{jk} = 2] \\ + \tfrac{1}{2}\mathbb{I}[X_{ik} = 2, X_{jk} = 1] + \mathbb{I}[X_{ik} = 2, X_{jk} = 2], \end{aligned} \tag{4}$$

where $\mathbb{I}[E]$ denotes the indicator function of an event E. In fact, by noticing that $\mathbb{I}[X_{ik} = 1] = X_{ik}(2 - X_{ik})$ and $\mathbb{I}[X_{ik} = 2] = \frac{1}{2}X_{ik}(X_{ik} - 1)$ and substituting in (4), we obtain $r_{ij,k} = \frac{1}{4}X_{ik}X_{jk}$.

In order to better understand the statistical properties of the CCC, we derive its expected value. First, we assess the theoretical mean and variance of the average weighted biallelic score R_{ij}.

Proposition 1. Under assumption A2 the expected value and variance of the average weighted biallelic score R_{ij} are given, respectively, by

$$\mathbf{E}(R_{ij}) = p_i p_j, \tag{5}$$

and

$$\mathbf{Var}(\sqrt{n}R_{ij}) = \frac{(p_i + p_i^2)(p_j + p_j^2)}{4} - p_i^2 p_j^2, \tag{6}$$

where the p_i and p_j are the populational frequencies of alleles i and j, respectively.

Proof. See Supplementary material S.1. ■

Note that expression (5) implies that $\mathbf{E}(R_{ij})$ is larger than the expected values of the other three allelic combinations for the pair of *loci*. For example, if $p_A > p_a$ and $p_B > p_b$, then $\mathbf{E}(R_{AB})$ is larger than $\mathbf{E}(R_{aB})$, $\mathbf{E}(R_{Ab})$ and $\mathbf{E}(R_{ab})$.

Remark. Since R_{ij} is a sample mean, we have—under the assumption of iid sampling A1—that $\sqrt{n}R_{ij}$ is asymptotically Gaussian. Under the additional assumption of independence between SNPs (Assumption A2), in fact we have

$$\sqrt{n}(R_{ij} - p_i p_j) \to \mathcal{N}\big(0, \tfrac{1}{4}(p_i + p_i^2)(p_j + p_j^2) - p_i^2 p_j^2\big).$$

In order to define the Standardized Average Weighted Biallelic Statistic (explored in the next subsection), we require estimators for the mean and variance of R_{ij} under the independence assumption (A2).

Proposition 2. Consider two biallelic SNPs under assumption A2. Then, $\widehat{\mathbf{E}}(R_{ij}) := f_i f_j$ is an unbiased and consistent estimator for $\mathbf{E}(R_{ij})$. Moreover, the estimator

$$\widehat{\mathbf{Var}}(\sqrt{n}R_{ij}) := \frac{(f_i + f_i^2)(f_j + f_j^2)}{4} - f_i^2 f_j^2 \tag{7}$$

is consistent for $\mathbf{Var}(\sqrt{n}R_{ij})$.

Proof. See Supplementary material S.1. ■

Based on the above result and exploring repeatedly the law of total expectation, we derive the expected value of the CCC, presented in Theorem 1.

Theorem 1. Consider two biallelic SNPs under assumption A2. Then, the expected value of the custom correlation coefficient is given by

$$\begin{aligned}\mathbf{E}(CCC_{ij}) = \frac{9}{2}\Bigg[p_i p_j - \frac{1}{1.5}\Bigg(p_i\left(\frac{p_j - p_j^2}{2n} + p_j^2\right) + p_j\left(\frac{p_i - p_i^2}{2n} + p_i^2\right)\Bigg) \\ + \frac{1}{2.25}\Bigg(\left(\frac{p_i - p_i^2}{2n} + p_i^2\right)\left(\frac{p_j - p_j^2}{2n} + p_j^2\right)\Bigg)\Bigg].\end{aligned} \tag{8}$$

Proof. See Supplementary material S.1. ■

Building on the results from Proposition 1, Theorem 1 and the simulation studies presented in section 3, we note the strong dependence of the four $\mathbf{E}(CCC_{ij})$ on the allele frequencies. Figures 3(A) and 2(A) present the shape of the $\mathbf{E}(R_{ij})$ and $\mathbf{E}(CCC_{ij})$ surfaces, highlighting that the maximum values of R_{ij} are higher as p_i and/or p_j approach 1. On the other hand, the maximal $\mathbf{E}(CCC_{ij})$ surface is maximized when p_i and/or p_j approach 0.75. These results suggest that the CCC statistic will tend to favor correlations between alleles corresponding to those frequencies, therefore affecting Type I Error rates.

Theorem 2. Under assumptions A1 and A2, the statistic $\sqrt{n}\big(CCC_{ij} - \frac{9}{2}f_i f_j F_i F_j\big)$ converges in distribution to a centered Normal random variable with variance

$$\left[\frac{9}{4}\left(1-\frac{p_i}{q}\right)\left(1-\frac{p_j}{q}\right)\right]^2 (p_i - p_i^2)(p_j - p_j^2).$$

Proof. See Supplementary material S.1. ■

2.3 Standardized Average Weighted Biallelic Statistic (SAWB)

In this section we propose a standardized statistic based on R_{ij} as an alternative to the CCC. We postulate that assessing correlation between SNPs through this statistic corrects for the allele frequency selection bias observed for CCC. Additionally, we demonstrate that this standardized statistic allows for the definition of a *bona fide* statistical test with Type I Error correction.

We propose a statistical test to evaluate the dependence relationship between alleles from two different SNPs. The null hypothesis states that the variants are independent, and the alternative hypothesis that the variants are not independent. Recall also that we assume altogether that the sampling scheme is iid.

We now introduce our standardized statistic, S_{ij}.

Definition 1. The *Standardized Average Weighted Biallelic* (SAWB) Statistic, S_{ij}, is defined as

$$S_{ij} := \frac{\sqrt{n}(R_{ij} - f_i f_j)}{\sqrt{\frac{(n-1)}{4n}(f_i^2 - f_i)(f_j^2 - f_j)}}. \tag{9}$$

Theorem 3. Under assumptions A1 and A2, as $n \to \infty$, the statistic S_{ij} converges in distribution to a standard normal random variable.

Proof. See Supplementary material S.2. ■

Thus we can use quantiles of the normal distribution to test for association between SNPs based on S_{ij} for sufficiently large samples.

In the course of proving Theorem 3, we use the following results, which are of interest on their own.

Lemma 1. Under assumptions A1 and A2, the statistic $\sqrt{n}(R_{ij} - f_i f_j)$ has zero mean, and variance given by

$$\mathbf{Var}\big(\sqrt{n}(R_{ij} - f_i f_j)\big) = \frac{(n-1)}{4n}\big[(p_i^2 - p_i)(p_j^2 - p_j)\big]. \tag{10}$$

Moreover, the estimator

$$\widehat{\mathbf{Var}}\big(\sqrt{n}(R_{ij} - f_i f_j)\big) = \frac{(n-1)}{4n}\big[(f_i^2 - f_i)(f_j^2 - f_j)\big] \tag{11}$$

is consistent for $\mathbf{Var}(\sqrt{n}(R_{ij} - f_i f_j))$.

Proof. See Supplementary material S.2. ■

Proposition 3. The random variables S_{ab}, S_{Ab}, S_{aB} and S_{AB} satisfy the following symmetry identity:

$$S_{ab} = S_{AB} = -S_{aB} = -S_{Ab}. \tag{12}$$

Proof. See Supplementary material S.2. ■

Note that this last result does not depend on distributional assumptions. While the CCC requires the computation of one CCC_{ij} statistic for each of the four allele combinations, this symmetry property implies that, for the S_{ij} statistic, we can use only one of the four R_{ij}, which leads to computational cost reductions.

3 Simulation Studies

In this section we present the simulation studies performed to evaluate the statistical properties of the Custom Correlation Coefficient (CCC) and the SAWB statistic (S_{ij}) proposed in Section 2.

3.1 Proprieties of CCC_{max}

In order to evaluate the sampling distribution of CCC_{max}, we investigate eight different population frequency scenarios with $p_i = p_j$, as described in Figure 1. For each of the RE=5000 replications, we simulate $n = 50$ genotypes with SNP_1 frequencies p_i and p_{i^c} and SNP_2 frequencies p_j and p_{j^c}, taking $i = A$, $i^c = a$, $j = B$, $j^c = b$ and assuming independence

between alleles from different SNPs. We calculate $CCC = (CCC_{AB}, CCC_{Ab}, CCC_{aB}, CCC_{ab})$, according to equation (1), and annotate its maximum value CCC_{max}. Figure 1 displays the density plots for these scenarios.

In Figure 1 it can be noted that the CCC_{max} distributions are frequency-dependent. Higher values of CCC_{max} are found at allelic frequencies equal to 0.75 and lower values as allelic frequencies depart from 0.75. Additionally, the variability of CCC_{max} also seems to be frequency-dependent, being higher when allelic frequencies are more equiprobable inside each SNP, i.e. $p_A \approx p_a$ and $p_B \approx p_b$.

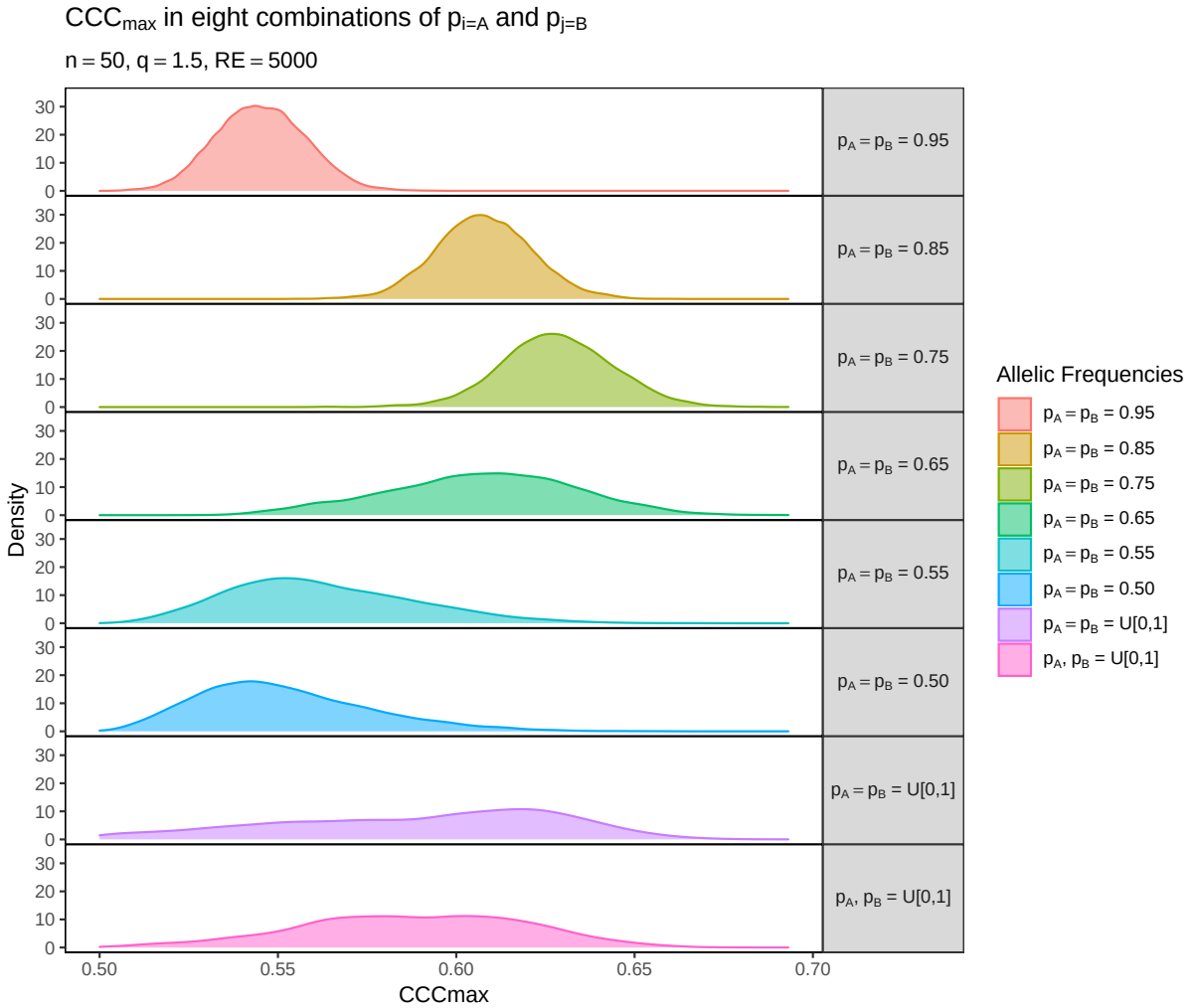

Figure 1: Distribution of CCC_{max} values from simulated data in eight different combinations of allele frequencies.

Furthemore, we evaluate the convergence of CCC_{ij} sample mean at different number of replications (RE), $\overline{CCC_{ij,RE}}$, to their correspondent theoretical expected values $\mathbf{E}(CCC_{ij})$,

as given by expression (8), via Monte Carlo simulations. We consider 25 allelic frequency combinations $p_i \times p_j$, where $p_i, p_j \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, four sample sizes $n \in \{10, 20, 30, 50\}$, and successive values of replication number $RE \in \{1, 2, \ldots, 10000\}$. In each scenario, we calculate the absolute difference between the Monte Carlo estimates $\sum_{r=1}^{RE} CCC_{ij}^{(r)}/RE$ and the theoretical expected value of CCC_{ij}. The Monte Carlo estimates converge to the theoretical expected value in all $p_i \times p_j$ combinations, increasing their accuracy as the sample size n and the number of replications RE increase (data not shown).

3.2 Frequency-dependent selection in CCC_{ij} and S_{ij}

To visualize the expectation $\mathbf{E}(CCC_{ij})$ as a function of the allelic frequencies, we draw the corresponding surfaces by calculating the theoretical values of $\mathbf{E}(CCC_{ij})$ for $n = 1000$ in 400 different allelic frequency combinations, with p_i and p_j ranging from 0.01 to 0.99 in grid increments of 0.05. We then generate $RE = 100$ replications of CCC_{max} for each scenario and register which alleles i and j correspond to CCC_{max}, according to the same simulation scheme described in subsection 3.1. The CCC_{max} are calculated based on SNP data simulated in the same conditions as the $\mathbf{E}(CCC_{ij})$, that is, same sample size and allelic frequencies for p_i and p_j, assuming independence between alleles i and j.

Figure 2 combines the theoretical $\mathbf{E}(CCC_{ij})$ surfaces and the respective simulated mean CCC_{max} for each scenario. One can easily see that the four $\mathbf{E}(CCC_{ij})$ surfaces are symmetric and each one dominates as the highest curve in different regions of the $p_i \times p_j$ grid (Figure 2A). The colored points indicate that in all the $RE = 100$ replications for CCC_{max}, the i and j alleles are equivalent to the corresponding alleles of the highest $\mathbf{E}(CCC_{ij})$. The black points indicate that not all i and j alleles in CCC_{max} are consistent with respective highest theoretical $\mathbf{E}(CCC_{ij})$ surface. These misidentifications occur near the limits between $\mathbf{E}(CCC_{ij})$ curves, where values of CCC_{ij} are similar for different i and j (Figure 2C).

The same simulation scenario described for the $\mathbf{E}(CCC_{ij})$ function is performed for $\mathbf{E}(R_{ij})$ and R_{max} (Figure 3) and for $\mathbf{E}(S_{ij})$ and S_{ij} (Figure 4). It is easy to see that for both CCC_{ij} and R_{ij}, higher values of the maxima and theoretical expected values are found at specific combinations of p_i and p_j. As one can see in expression (5), the highest R_{ij} corresponds to the alleles i and j with the highest frequencies. A similar phenomenon is observed in CCC where the maximum is found approaching allele frequency 0.75 (see curve with purple dots for $E(CCC_{AB})$). Additionally, when considering the overlay between the four surfaces, one can see that the surface corresponding to the higher allele frequencies is always maximal. This becomes explicit noting that these surfaces are plotted in relation to the allelic frequencies of p_A and p_B, and, for example, when $p_A > p_a$ and $p_B < p_b$, the surface

that dominates is the red colored one ($\mathbf{E}(CCC_{Ab})$ and $\mathbf{E}(R_{Ab})$), once alleles A and b have the highest frequencies at their respective *loci*. Combining this observation with the fact that only for a few points in the intersection of curves the maximum for the simulated values do not always coincide with the maximal curve, it becomes apparent that, under independence, the allele combination that yields the CCC_{max} will consistently be the one with higher allele frequencies. Regarding S_{ij}, the points are close and randomly distributed along the surface of $\mathbf{E}(S_{ij})$ and and no frequency selection is observed.

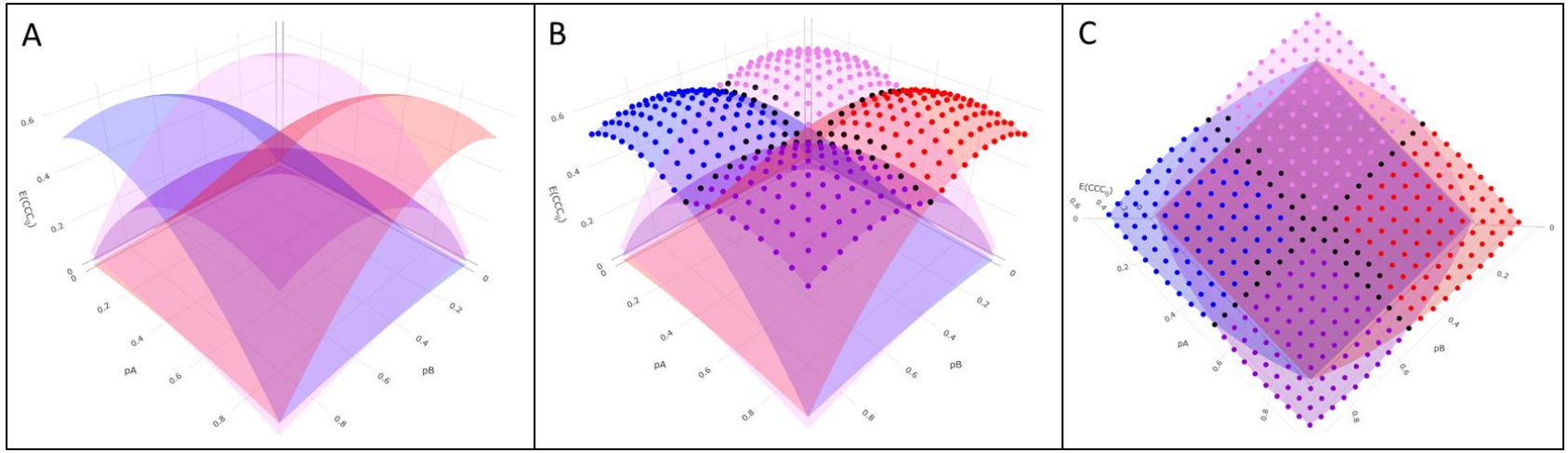

Figure 2: A. $\mathbf{E}(CCC_{ij})$ surfaces; B Simulated CCC_{max} for each p_A and p_B combination; C. Black points show the misidentifications near the limits between $\mathbf{E}(CCC_{ij})$ curves. Purple surface corresponds to $\mathbf{E}(CCC_{AB})$; pink surface corresponds to $\mathbf{E}(CCC_{ab})$; red surface corresponds to $\mathbf{E}(CCC_{Ab})$; and blue surface corresponds to $\mathbf{E}(CCC_{aB})$.

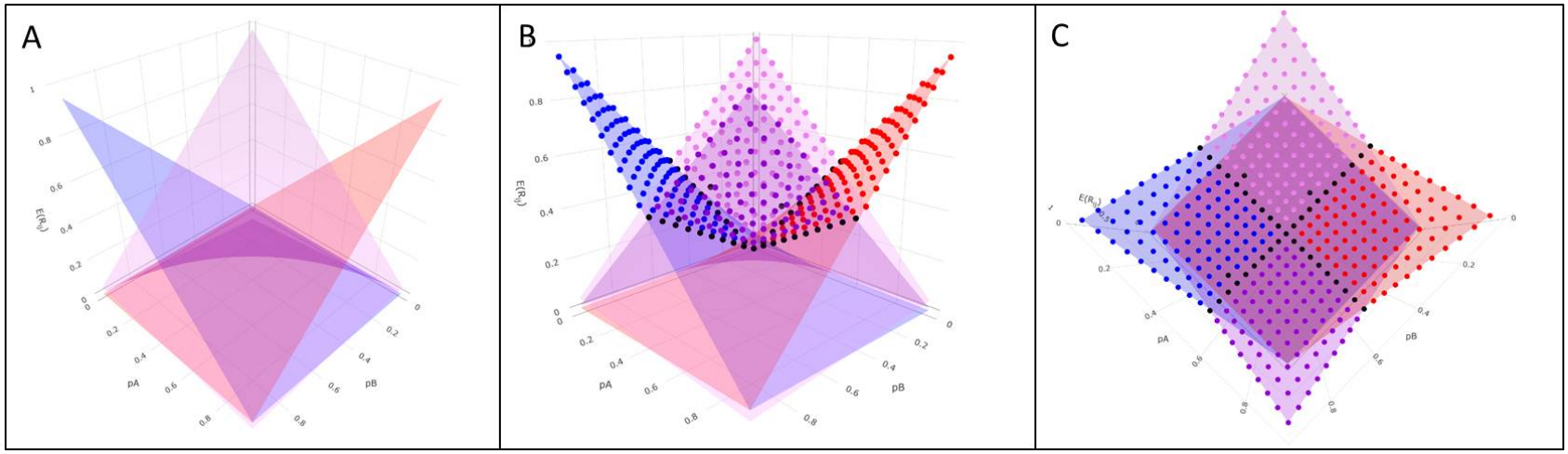

Figure 3: A. $\mathbf{E}(R_{ij})$ surface; B. Simulated R_{max} for each p_A and p_B combination; C. Black points show the misidentifications near the limits between $\mathbf{E}(R_{ij})$ curves. Purple surface corresponds to $\mathbf{E}(R_{AB})$; pink surface corresponds to $\mathbf{E}(R_{ab})$; red surface corresponds to $\mathbf{E}(R_{Ab})$; and blue surface corresponds to $\mathbf{E}(R_{aB})$.

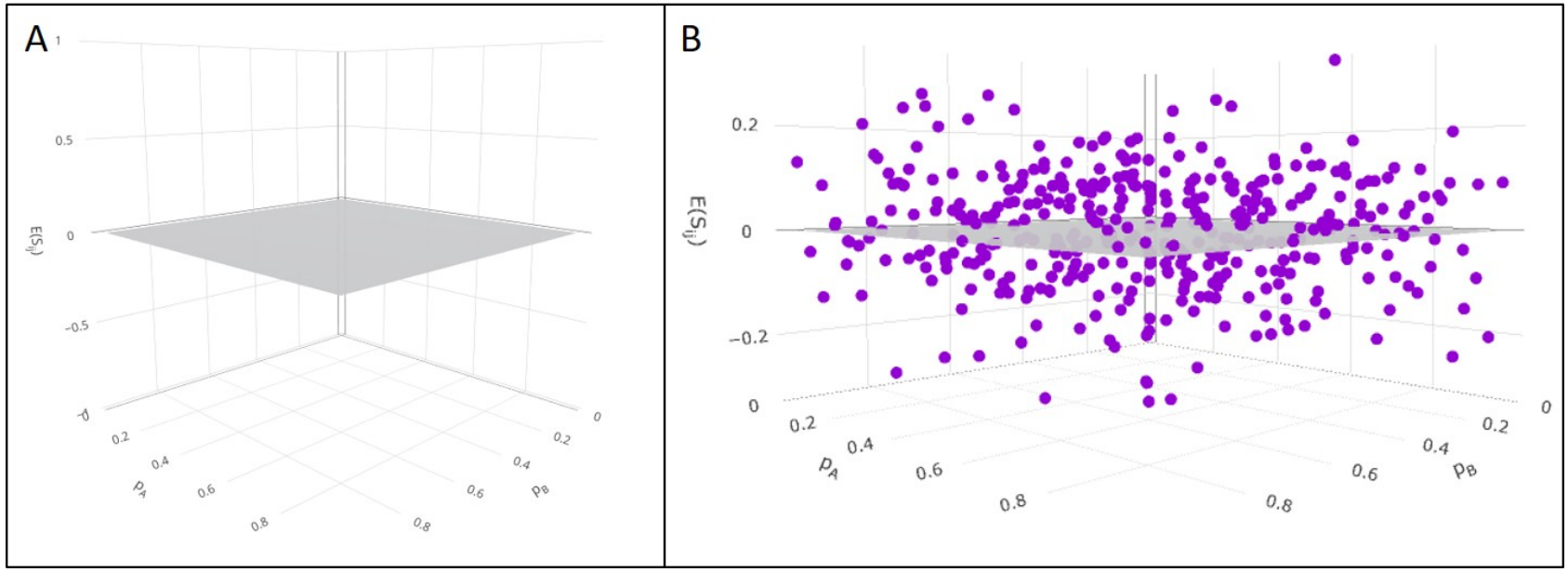

Figure 4: A. $\mathbf{E}(S_{ij})$ surface; B. Simulated S_{ij} values for each p_A and p_B combination.

To further evaluate the existence of frequency-dependent selection in the CCC_{max} and S_{ij} statistics we perform another simulation study. We simulate $RE = 10000$ samples of size $n = 1000$, with $p_A, p_B \overset{\text{iid}}{\sim} U[0,1]$, with the same simulation scheme described in subsection 3.1. We select the CCC_{max} above the threshold and significant S_{ij} values and plot their distributions according to the respective allelic frequencies (Figure 5). For the CCC_{max} statistic the threshold value used to single out correlations that compose the graph is the 0.95 quantile of the simulated CCC_{max}, and all CCC_{max} above the threshold are selected. For the S_{ij} statistic, we use the 0.025 and 0.975 quantiles of the standard normal distribution to establish significance. Figure 5 shows that selected CCC_{max} are more frequent when p_i (Figure 5C) or p_j (Figure 5E) are close to 0.75. By contrast, significant S_{ij} values are uniformly distributed along p_i (Figure 5D) and p_j (Figure 5F) range. Note that in Figure 5C and E CCC_{max} above threshold correspond to the most frequent alleles at each *loci*, particularly when i and j allelic frequencies are close to 0.75. Therefore, it is easy to see that the CCC_{max} distribution is influenced by frequency-dependent selection, whereas S_{ij} does not suffer from such issue (Figure 5D and F).

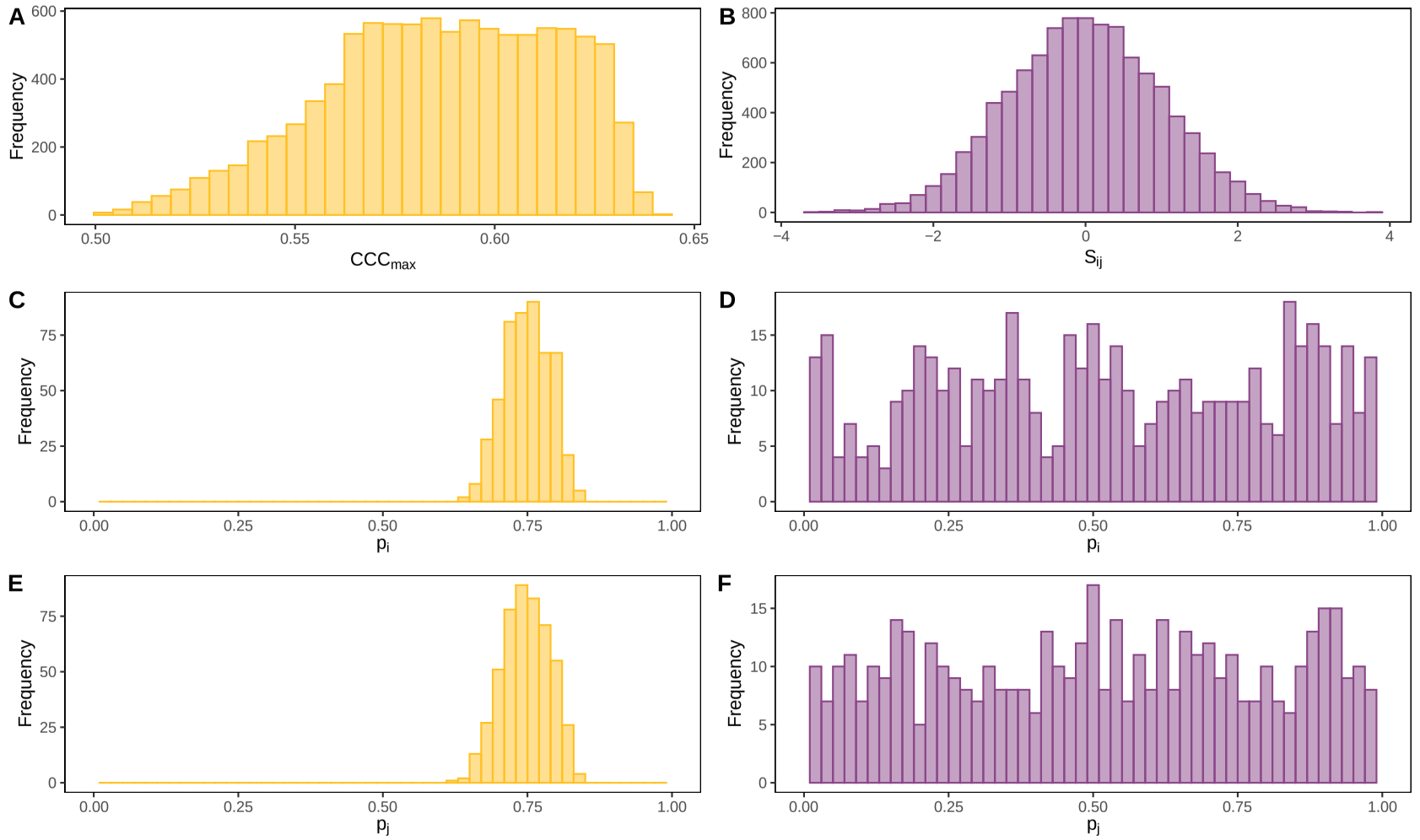

Figure 5: Distribution of CCC_{max} and S_{ij} and Frequency-dependent selection of CCC_{max} distribution. **A** Distribution of CCC_{max}. **B** Distribution of S_{ij}. **C** Allelic frequency distribution of p_i for CCC_{max} above threshold. **D** Allelic frequency distribution of p_i for significant S_{ij}. **E** Allelic frequency distribution of p_j for CCC_{max} above threshold. **F** Allelic frequency distribution of p_j for significant S_{ij}.

3.3 Inference properties of CCC_{ij} and S_{ij}

3.3.1 Convergence of S_{ij} to the standard normal distribution

To investigate the convergence of the four S_{ij}, with $i \in \{A, a\}$ and $j \in \{B, b\}$, to the standard normal distribution as n approaches infinity, we simulate S_{ij} values for $RE = 1000$ replications of 49 different scenarios with $p_A, p_B \in \{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$ for a sample size $n = 1000$, assuming independence between alleles i and j.

In Figure 6 one can note that, under the null hypothesis of independence between alleles, the distributions of the S_{ij} seems to approach the standard normal distribution, except for rare variants. Because of the symmetric relationship between the four S_{ij}, i.e. $S_{AB} = S_{ab} = -S_{Ab} = -S_{aB}$, proven in subsection S.2, there are two overlapping curves in the density plots. It is important to observe that this convergence to the standard normal distribution is slower when both p_A and p_B frequencies are extreme, and these distributions are still far from Gaussian for $n = 1000$. These results can also be visualized in the violin plot for the same simulation (Figure S1). However, understanding the impact of less common (MAF $1 - 5\%$) or rare (MAF $< 1\%$) variations in complex human diseases and traits is still a challenge, as there is a clear need to increase the statistical power of these studies to target their usually small or modest effects (Bomba et al., 2017).

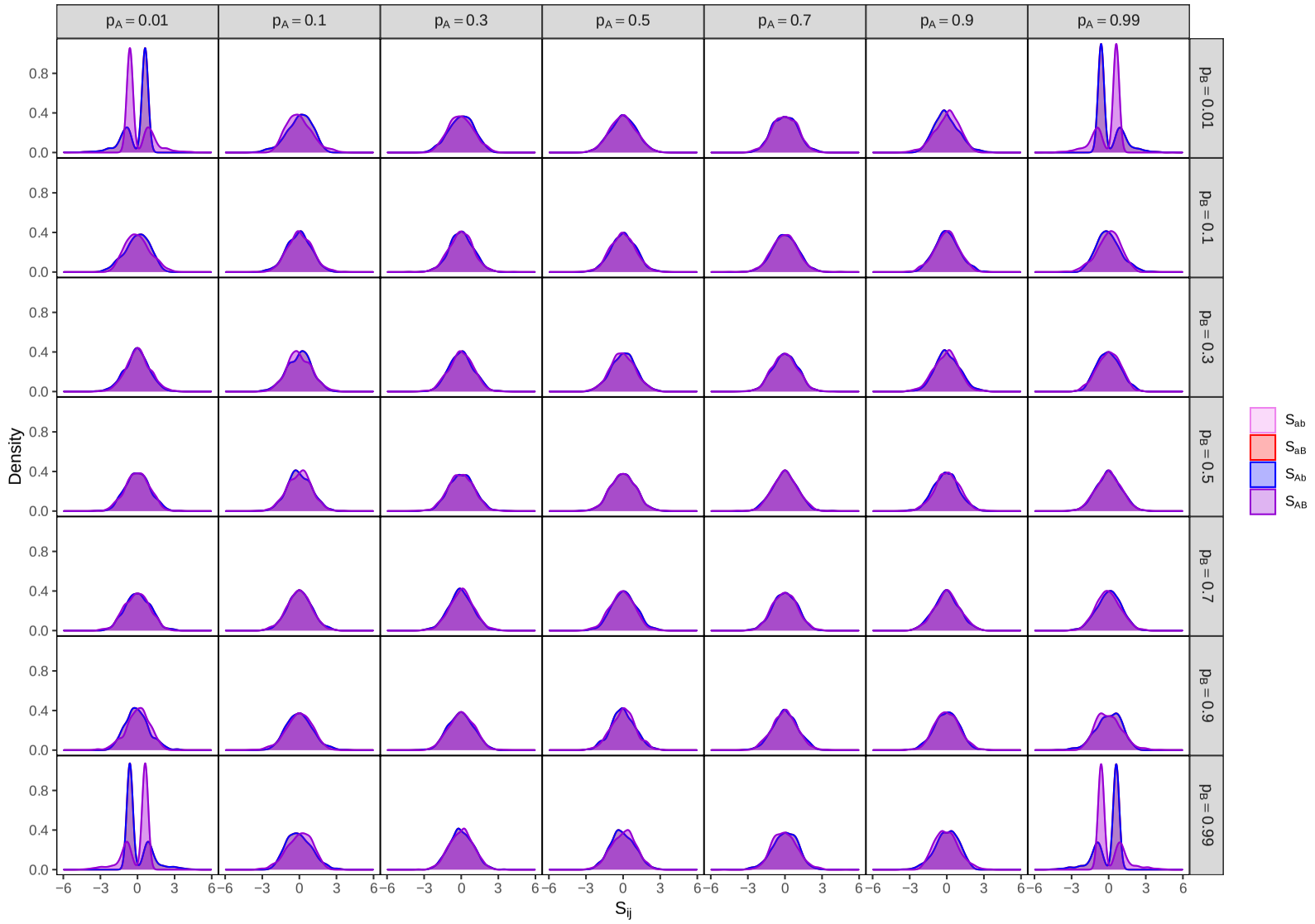

Figure 6: S_{ij} distributions in different allelic frequency combinations for a sample size $n = 1000$.

3.3.2 Type I Error of CCC_{ij} and S_{ij} statistics

In order to compare the Type I Error rates of the CCC_{ij} and S_{ij} statistical tests, we perform a simulation study that accommodates different allelic frequencies with $p_A, p_B \in \{0.01, 0.05, 0.1, 0.10, 0.25, 0.5\}$, assuming independence between alleles i and j. We generate SNP_1 and SNP_2 genotypes considering incremental sample sizes $n = (25, 50, 100, 1000)$. Given each scenario above, we perform $RE = 5000$ replications calculating both CCC_{max} and S_{ij}.

The CCC_{max} threshold for each sample size is obtained from a subset of the 1000 Genomes Project dataset (Consortium et al., 2015), containing 2.176 biallelic SNPs from NTAD gene cluster of chromosome 11 (Mota et al., 2012). We randomly select 10,000 pairs of SNPs from the dataset, and sample $n = 1000$ individuals (largest n). For each pair of SNPs, we shuffle the genotypes of the first SNP and keep the second one intact to simulate independence and then compute the $CCC_{max}^{H_0}$, the CCC_{max} under the null hypothesis H_0. We use the 0.95 quantile of the $CCC_{max}^{H_0}$ to determine the threshold values for the simulated CCC_{max}. This procedure is repeated for each sample size n.

For each sample size, the Type I Errors are then calculated by the proportion of selected CCC_{max} (above threshold) and significant S_{ij} (under -1.96 or above the 1.96, respectively). Table 2 displays the Type I Error for both statistics.

One can see that the Type I Error for CCC_{ij} is extremely small in most scenarios and approaches zero as the sample size n increases. However, the Type I Error rate tends to increase when p_A and/or p_B approach 0.25. When the sample size is $n = 1000$ and $p_A = p_B = 0.25$, all CCC_{max} calculated are selected (P(Type I Error) = 1). In this case, as observed in Figure 5, the CCC_{max} correspond to the alleles with the highest frequencies (CCC_{ab}). Regarding the S_{ij} statistic, we can observe Type I Error rates close to 0.05 in all sample sizes. A small deviation from this pattern (with higher values) is observed when sample sizes are small (n=25 and n=100) and at least one of the alleles is a rare variant. However, with n=1000 individuals, the Type I Error rate is adequate for all allele frequency combinations tested. This control of Type I Error presented by the S_{ij} statistic also highlights the absence of frequency-dependent selection already mentioned in subsection 3.2.

Table 2: Type I Error of CCC_{ij} and S_{ij} statistics.

			CCC_{ij}					S_{ij}				
			p_B					p_B				
			0.01	**0.05**	**0.10**	**0.25**	**0.50**	**0.01**	**0.05**	**0.10**	**0.25**	**0.50**
$n = 25$	p_A	**0.01**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0490	0.0855	0.0555	0.0510	0.0175
		0.05	0.0000	0.0000	0.0006	0.0116	0.0010	0.0775	0.0655	0.0485	0.0505	0.0455
		0.10	0.0000	0.0006	0.0116	0.1096	0.0146	0.0560	0.0555	0.0480	0.0430	0.0580
		0.25	0.0002	0.0086	0.1034	0.4728	0.1312	0.0565	0.0490	0.0510	0.0495	0.0525
		0.50	0.0000	0.0006	0.0104	0.1182	0.0606	0.0155	0.0410	0.0565	0.0555	0.0550
$n = 100$	p_A	**0.01**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0455	0.0665	0.0520	0.0545	0.0440
		0.05	0.0000	0.0000	0.0000	0.0000	0.0000	0.0705	0.0390	0.0385	0.0385	0.0555
		0.10	0.0000	0.0000	0.0002	0.0798	0.0000	0.0535	0.0470	0.0605	0.0445	0.0455
		0.25	0.0000	0.0004	0.0684	0.8618	0.0120	0.0430	0.0560	0.0470	0.0510	0.0500
		0.50	0.0000	0.0000	0.0000	0.0100	0.0004	0.0355	0.0490	0.0490	0.0515	0.0545
$n = 500$	p_A	**0.01**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0615	0.0410	0.0420	0.0465	0.0415
		0.05	0.0000	0.0000	0.0000	0.0000	0.0000	0.0425	0.0410	0.0465	0.0490	0.0515
		0.10	0.0000	0.0000	0.0000	0.0012	0.0000	0.0435	0.0400	0.0520	0.0530	0.0440
		0.25	0.0000	0.0000	0.0012	0.9960	0.0000	0.0365	0.0525	0.0605	0.0480	0.0445
		0.50	0.0000	0.0000	0.0000	0.0000	0.0000	0.0530	0.0445	0.0450	0.0435	0.0530
$n = 1000$	p_A	**0.01**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0540	0.0460	0.0460	0.0425	0.0465
		0.05	0.0000	0.0000	0.0000	0.0000	0.0000	0.0395	0.0490	0.0550	0.0480	0.0505
		0.10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0405	0.0525	0.0545	0.0510	0.0500
		0.25	0.0000	0.0000	0.0000	1.0000	0.0000	0.0410	0.0525	0.0525	0.0440	0.0495
		0.50	0.0000	0.0000	0.0000	0.0000	0.0000	0.0515	0.0470	0.0465	0.0445	0.0505

3.3.3 Power of CCC_{ij} and S_{ij} statistics

We also investigate the power of both statistics, CCC_{ij} and S_{ij}, accounting for different degrees of association between simulated SNPs. We choose Clayton's Copula (Clayton, 1978) in order to generate dependent samples for the SNP data and set different values for Kendall's tau correlation coefficient $\tau \in \{0.15, 0.3, 0.45, 0.6, 0.85, 0.99\}$. The relationship between Clayton's Copula parameter α and Kendall's tau is given by $\alpha = 2\tau/(1-\tau)$. Several allelic frequencies combinations $p_A, p_B \in \{0.01, 0.05, 0.1, 0.25, 0.5\}$ and sample sizes $n \in \{25, 50, 100, 200, 500, 1000\}$ are also considered in this simulation. We obtain SNP_1 and SNP_2 samples according to the following procedure: 1) first we generate $n = 1000$ (largest sample size) uniform random values u_1 and v; 2) we generate the uniform random value u_2 from u_1 and v according to Clayton's Copula; 3) we obtain SNP_1 and SNP_2, respectively, through the probability integral transform, by mapping the values u_1 and u_2 onto the genotype

cumulative probability intervals, e.g.

$$SNP_1 = \begin{cases} AA, & \text{if} \quad u_1 \leqslant p_A; \\ Aa, & \text{if} \quad p_A < u_1 \leqslant p_A + 2p_A p_a; \\ aa, & \text{if} \quad u_1 > p_A + 2p_A p_a; \end{cases}$$

4) we then calculate CCC_{max} and S_{ij} from SNP_1 and SNP_2; and 5) we finally obtain the proportion of selected CCC_{max} and significant S_{ij} for each sample size n. To select correlations and define significance, we rely on the same thresholds as in the Type I Error simulations for CCC_{ij} and the same critical values of the standard normal distribution for S_{ij}.

Figure 7 shows the behavior of the power of both the CCC_{ij} and S_{ij} statistics. One can see that for any degree of association, i.e., Kendall's τ, the power of S_{ij} statistic is better behaved than the power of CCC_{ij}. For S_{ij} statistics, the power increases as the sample size n increases, which is not always observed in CCC_{ij}. In Figure 8, it is easy to see the effect of frequency-dependent selection on the power of CCC_{ij}. For this statistic, the power is equal or close to zero when at least one variant has low frequency. However, the power approaches 1 as the sample size increases when the p_A and p_B are close to 0.25. Reasonable power values are attained in frequencies close to 0.5, only for Kendall's τ equal or above 0.45. The S_{ij} statistic appears to be consistent, with all power curves approaching 1 as the sample size increases, for all combinations of p_i and p_j. It is important to note that in these simulations few intermediate frequencies were considered.

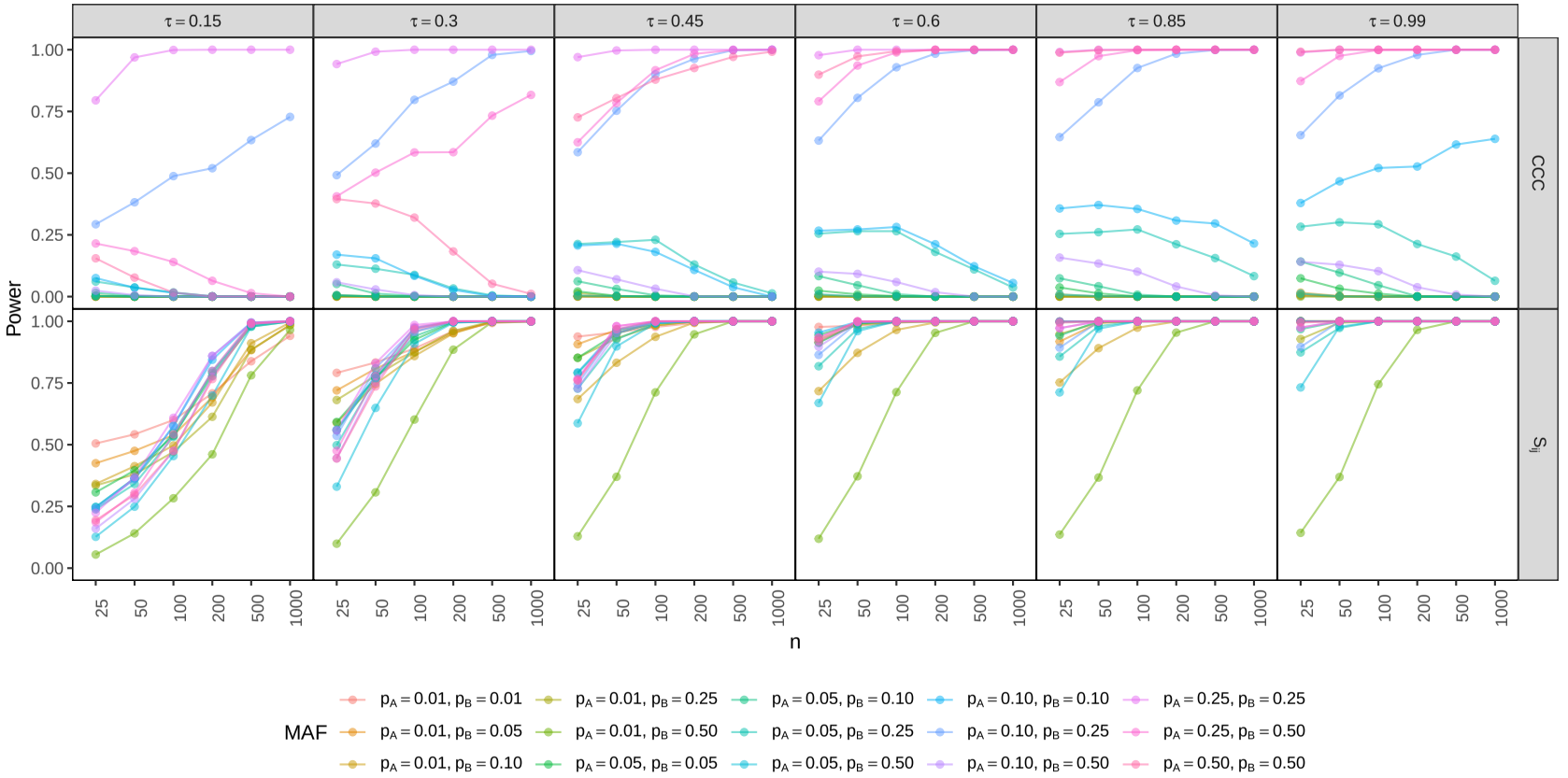

Figure 7: Power of CCC_{ij} and S_{ij} in different τ combinations

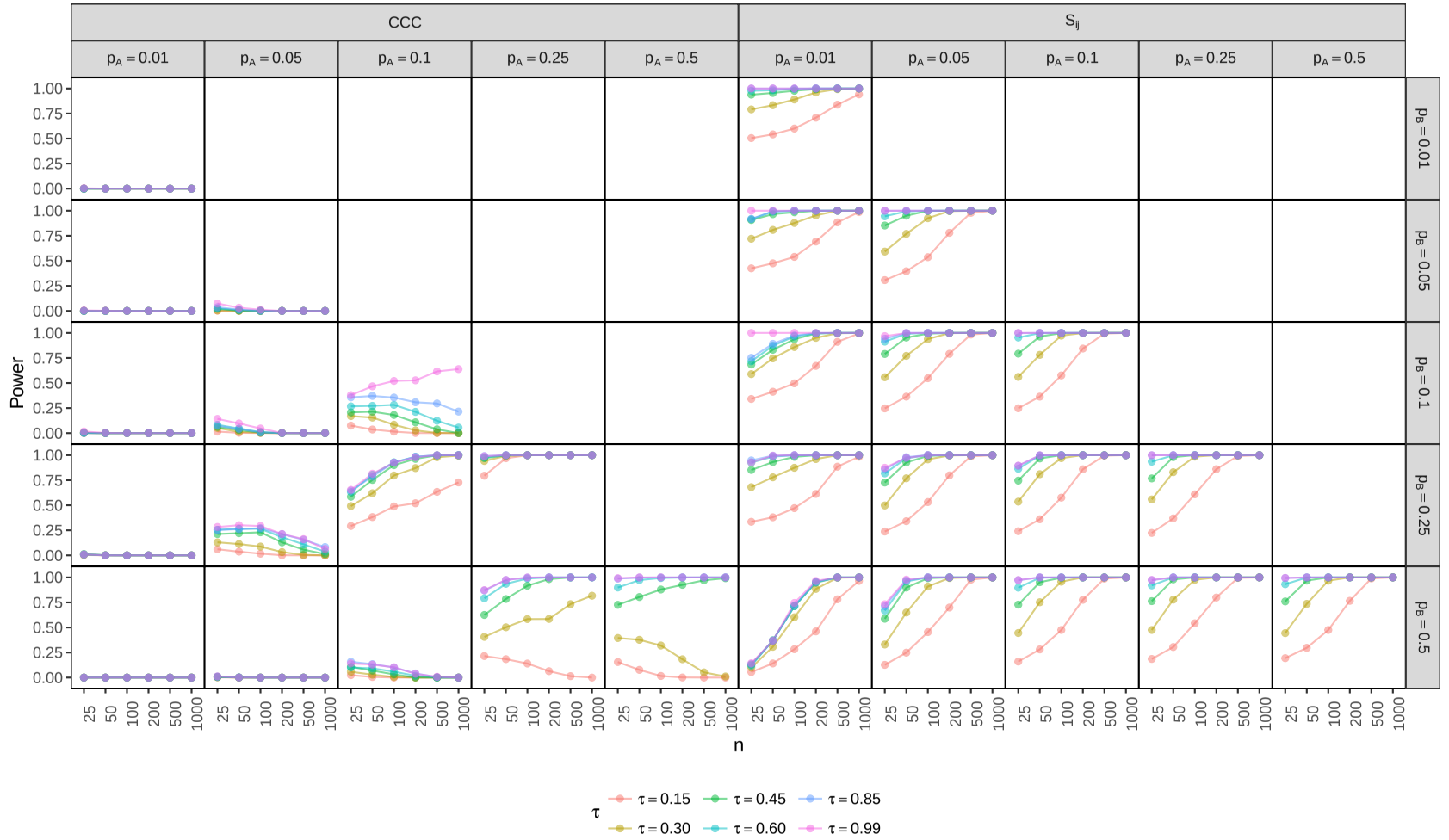

Figure 8: CCC_{ij} and S_{ij} power in different allelic frequency combinations.

4 Application

Network studies in complex human diseases have many clinical and biological applications (Barabási et al., 2011). To compare the networks formed by the CCC_{ij} and S_{ij} statistics, we analyze an Attention Deficit Hyperactivity Disorder (ADHD) database. The dataset is composed of 417 ADHD cases from the ADHD Outpatient Program (ProDAH-A) and 463 blood donor controls, all of them from Hospital de Clínicas de Porto Alegre (HCPA). The sample consists of 880 adult white Brazilians (18 years or older), 445 males and 435 females. After genotyping and post imputation quality control (QC), a total of 771 biallelic variants from the NTAD cluster of the chromosome 11 (Mota et al., 2012) are retained, with Minor Allele Frequency (MAF) greater than or equal to 0.05. This sample is a single site of a larger dataset, described in Rovira et al. (2020), that is part of the Psychiatric Genomics Consortium (PGC) and the International Multi-centre persistent ADHD CollaboraTion (IMpACT).

For each pair of alleles, the values of CCC_{ij} and S_{ij} are calculated according to equations 1 and 9. To select the CCC_{max} that will correspond to edges in the network, the threshold is computed from the case subset following the same procedure of subsection 3.3.2. Of the 296,835 SNP pairs considered in this study, we found that 2.597% are selected for the CCC networks (CCC_{max} values equal to or greater than threshold of 0.65). The same proportion of allele pairs with the highest S_{ij} values is used for network building. However, the S_{ij} statistic does not select any combination of alleles, since it is symmetric (either all four S_{ij} are significant, or they are not significant). It is important to note that when CCC_{max} is annotated, indirectly R_{max} is being selected as well up to the influence of the frequency factors F_i and F_j. Slight deviations from this observation can be found in the overlays of the $\mathbf{E}(CCC_{ij})$ surfaces (presented in subsection 3.2, Figure 2). Thus, for this application the i and j alleles associated with R_{max} are selected for the formation of the S_{ij} networks for the purpose of comparison with the CCC_{ij} allelic networks.

The heuristic described above is chosen because when we use standard normal distribution quantiles for the selection of significant S_{ij}, few networks are generated and all of which containing a massive amount of vertices. This is due to the S_{ij} statistic being very sensitive in finding significant correlations between SNPs (high statistical power). However, the same conformation of networks with extensive number of vertices are difficult to find in case and control individuals, making traditional association studies unfeasible. Since the goal of this application is to compare the two statistics (CCC_{ij} and S_{ij}), the same proportion of significant values is used.

We find 12 clusters in the CCC_{ij} network, two with a large number of vertices and edges (Clusters #1 and #2), one of intermediate size (Cluster #4) and nine small ones (Table

3). When we apply the S_{ij} statistic to the data, the amount of clusters is larger than the amount formed from the CCC_{ij}. Two large clusters are obtained (Clusters #1 and #5), six clusters of intermediate size (Clusters #3, #4, #7, #8, #12 and #14) and nine small ones (Table 4). Of note, Cluster #1 created from S_{ij} contains the same set of alleles found in Cluster #1 from CCC_{ij}. However, 120 alleles are added to Cluster #1 from S_{ij}, as well as 1,023 additional interactions. Therefore, when we apply the heuristic of analyzing the same proportion of selected CCC_{max} and significant S_{ij}, the S_{ij} statistic generates more diversified networks (with higher number of alleles) than CCC_{ij}. Besides the higher number of clusters obtained, the S_{ij} statistic yields a total of 806 vertices (alleles), while a total of 596 vertices (alleles) are found from CCC_{ij}. The number of edges (allelic interactions) is similar between both statistics, since the same proportion of selected CCC_{max} and significant S_{ij} is considered for network construction.

Tables 3 and 4 also show the number and percentage of individuals with and without ADHD in each cluster. We apply Fisher's exact test to verify possible associations between cases and controls for the disorder, however, no significant associations are found considering all clusters in both statistics (CCC_{ij} and S_{ij}). This result suggests that this absence of association is not due to a possible lack of representativeness in case and control groups as we observe when using standard normal distribution quantiles to select significant S_{ij}.

Table 3: Number of vertices and edges of clusters in the CCC network and test of association between cases and controls for Attention Deficit Hyperactivity Disorder (ADHD).

Clusters	**Number of Vertices**	**Number of Edges**	**Cases(%)**	**Controls(%)**	**Fisher p-value**
Cluster 1	328	24275	156(37.41)	174(37.58)	1.000
Cluster 2	195	6113	46(11.03)	41(8.86)	0.309
Cluster 3	9	31	327(78.42)	366(79.05)	0.869
Cluster 4	32	379	269(64.51)	294(63.50)	0.779
Cluster 5	5	10	396(94.96)	434(93.74)	0.468
Cluster 6	4	3	314(75.30)	332(71.71)	0.252
Cluster 7	4	6	322(77.22)	382(82.51)	0.053
Cluster 8	3	3	356(85.37)	389(84.02)	0.640
Cluster 9	2	1	385(92.33)	415(89.63)	0,196
Cluster 10	5	10	396(94.96)	426(92.01)	0.102
Cluster 11	3	3	404(96.88)	450(97.19)	0.844
Cluster 12	3	3	405(97.12)	451(97.41)	0.838

Table 4: Number of vertices and edges of clusters in the S_{ij} network and test of association between cases and controls for Attention Deficit Hyperactivity Disorder (ADHD).

Clusters	**Number of Vertices**	**Number of Edges**	**Cases(%)**	**Controls(%)**	**Fisher p-value**
Cluster 1	448	25298	173(41.49)	205(44.28)	0.414
Cluster 2	8	28	410(98.32)	454(98.06)	0.806
Cluster 3	23	217	389(93.29)	428(92.44)	0.695
Cluster 4	34	263	90(21.58)	107(23.11)	0.627
Cluster 5	185	4289	231(55.40)	231(49.89)	0.105
Cluster 6	6	15	387(92.81)	429(92.66)	1.000
Cluster 7	32	379	269(64.51)	294(63.50)	0.779
Cluster 8	11	55	403(96.64)	445(96.11)	0.721
Cluster 9	5	10	396(94.96)	434(93.74)	0.468
Cluster 10	4	3	314(75.30)	332(71.71)	0.252
Cluster 11	4	6	322(77.22)	382(82.51)	0.053
Cluster 12	17	136	406(97.36)	457(98.70)	0.219
Cluster 13	3	3	356(85.37)	389(84.02)	0.640
Cluster 14	16	120	307(73.62)	327(70.63)	0.329
Cluster 15	5	10	396(94.96)	426(92.01)	0.102
Cluster 16	3	3	404(96.88)	450(97.19)	0.844
Cluster 17	2	1	395(94.72)	431(93.09)	0.328

In subsection 3.2 we show the frequency-dependent selection of the CCC_{ij} in simulation studies. To assess frequency-dependent selection of both CCC_{ij} and S_{ij} statistics in the ADHD data, we compare allele frequencies of all SNPs to that of the SNPs present in clusters created from CCC_{ij} and S_{ij}. Figure 9 shows that the curve of allelic frequencies of SNPs present in clusters formed from both S_{ij} and CCC_{ij} contain almost exclusively alleles with frequency larger than 0.5. This striking manifestation of the property seen in the simulations under H_0, indicates that also in real data the CCC almost always selects the higher frequency alleles within a SNP pair. It is easy to see that when we use R_{max} as the criteria for allele selection in the S_{ij} networks, we import the same property for the S_{ij} network. Furthermore, the CCC curve presents high spikes and a conspicuous absence of SNPs with higher frequencies, highlighting additional frequency-dependent selection effect in this statistic. On the other hand the curve of allelic frequencies of those SNPs present in clusters formed from S_{ij} is closer to the curve of frequencies of all SNPs.

In this application, we used a database with some expectation of association between allelic networks and ADHD. However, with no change in the allele selection strategy, we observe that none of the statistics used (CCC_{ij} and S_{ij}) are able to find association. We postulate that it occurs due to the frequency-dependent selection does not pick the best

possible combination of alleles, but rather, generates networks with the most frequent alleles of the *loci*, especially when these frequencies are close to 0.75. Therefore, the CCC_{ij} statistic is a biased tool in selecting correlated alleles to compose the clusters.

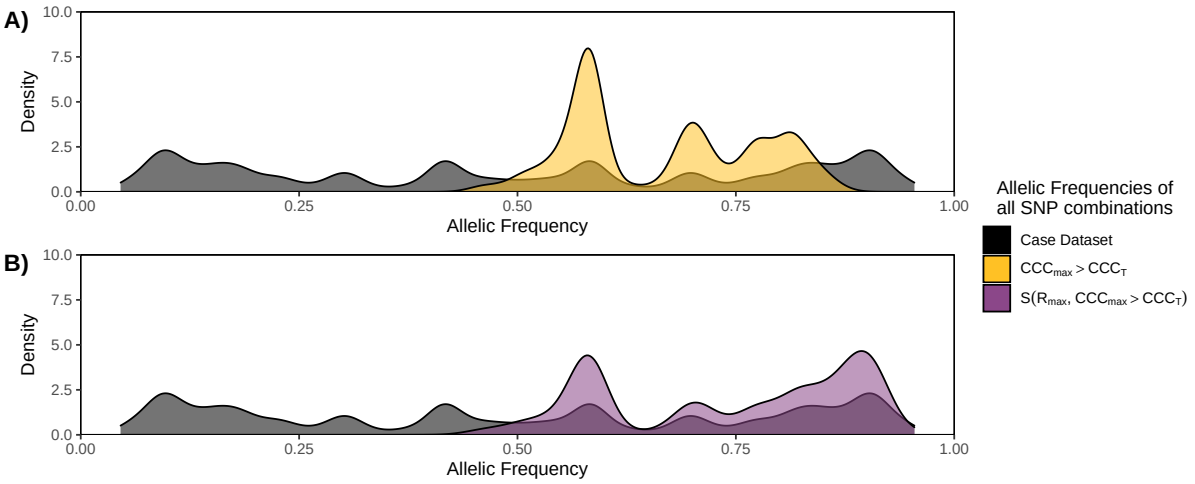

Figure 9: Comparison of allelic frequencies of the CCC_{ij} and S_{ij} networks to all SNPs evaluated in individuals with Attention Deficit Hyperactivity Disorder (ADHD).

5 Discussion

In this paper we proposed a statistical test to evaluate the dependency relationship between alleles from two different SNPs, named Standardized Average Weighted Biallelic Statistic (SAWB). This new statistic is based on the average weighted biallelic score (R_{ij}) and is introduced as an alternative to the CCC_{ij}.

Climer et al. (2014b) claimed that the CCC_{ij} accommodates genetic heterogeneity and introduced a network model applied in identifying association of a given phenotype in different subgroups of individuals. In addition, they proposed allelic (instead of genotypic) networking, which they claim is able to more accurately identify which variants of each SNP have an effect on the observed trait, even if these variants are rare.

In order to verify the advantages of using the CCC, we studied the properties of this coefficient and related statistics, and compare them to S_{ij}. In simulation studies, we analyzed the $E(CCC_{ij})$, $E(R_{ij})$ and $E(S_{ij})$ surfaces in different combinations of p_i and p_j. The four surfaces were symmetric for CCC_{ij} and R_{ij}, with highest points along the p_i x p_j grid when these frequencies are close to 0.75 (or 0.25 when considering the complementary frequency axis). Regarding S_{ij}, a random dispersion pattern with no clear tendency for the allelic frequency combinations was found. In another simulation study in which the significant

values for CCC_{max} and S_{ij} were selected, we found that the significant CCC_{max} values were more frequent when p_i or p_j were close to 0.75. In contrast, the S_{ij} values were uniformly distributed along the p_i and p_j range, showing that this statistic does not suffer from frequency-dependent selection.

In comparing the Type I Error rate for the CCC_{ij} and S_{ij} statistical tests, we observed that the Type I Error rate for the test based on CCC_{ij} was extremely small in most of the scenarios evaluated, approaching zero as the sample size n increased. However, due to frequency-dependent selection, the Type I Error rate increased when p_A and/or p_B approached 0.25. At sample size $n = 1000$ it was equal to 1, indicating that all the CCC_{max} calculated for $p_a = p_b = 0.75$ were selected to compose clusters. For the S_{ij} statistic, the Type I Error rates were close to 0.05 in the different allelic frequencies and sample sizes. Small deviations were found in small samples and when at least one of the variants was rare. Note that for larger samples ($n = 500$ or $n = 1000$) no selection bias was observed, and the Type I Error was controlled for the different allelic frequencies tested.

Concerning the power of the CCC_{ij} and S_{ij} statistical tests, once again the S_{ij} curves showed better behavior than the CCC_{ij} ones at different degrees of departure from the null hypothesis tested. We found that the CCC_{ij} test was not consistent for rare variants, where the power was equal or close to zero even for increasing sample sizes, in contrast to the claim of Climer et al. (2014b) that the CCC captures correlation for these variants. On the other hand, the power approached 1 as sample size increased for p_A and p_B allelic frequencies close to 0.25, preferentially selecting CCC_{ab} to compose the clusters and showing frequency selection for CCC_{ij}. In contrast, for the S_{ij} statistic, the power curves indicated consistency, with the power increasing as sample size n increased, in different τ and combinations of allelic frequencies.

In the application to the ADHD database, S_{ij} statistic tended to find larger more loosely connected networks, identifying more candidate alleles than CCC_{ij}. Furthermore, the clusters constructed by the CCC_{ij} and S_{ij} statistics (when we used R_{max} as a criterion for allele selection in S_{ij}) were made up by the most frequent alleles in the *loci*. Therefore, the CCC does not appear to be an effective network selection strategy. This does not mean that the associations found using this statistic in previous studies were invalid. However, it could be noted that the construction of these networks was occurring in a suboptimal way, due to the biases showed in this study.

It is worth mentioning that this application should be seen as a preliminary result focused on the comparison of networks formed by both statistics, and that the method could potentially be followed by alternative procedures of network analysis. Further studies will be required to consider association tests that do not simply consider absolute presence or

absence of clusters but take in to account more nuanced approaches to network testing. The S_{ij} was shown to be a statistical test that corrects frequency-dependent selection presented from CCC_{max} with power to detect association. However, we still need to determine an improved allelic selection strategy, or to devise a methodology for association testing with SNP networks.

In summary, we had extensively shown profound effects that frequency-dependent selection has on the CCC, biasing the corresponding networks to disproportionately include alleles with certain frequencies. In this context, we have proposed a new statistical test able to identify sets of correlated SNPs, with known distribution and properties, such as expectation and variance values, Type I Error and power. In addition, this test has controlled Type I Error and was, for most allelic frequencies, more powerful than the test based on the CCC statistic.

References

(2022). Genome. https://www.genome.gov/genetics-glossary/Genome. Accessed: 2022-02-14.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56–68.

Bomba, L., Walter, K., and Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biology*, 18.

Cano-Gamez, E. and Trynka, G. (2020). From gwas to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in Genetics*, 11:424.

Cantor, R. M., Lange, K., and Sinsheimer, J. S. (2010). Prioritizing gwas results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1):6–22.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65:141–151.

Climer, S., Templeton, A., and Zhang, W. (2014a). Allele-specific network reveals combinatorial interaction that transcends small effects in psoriasis gwas. *PLoS computational biology*, 10:e1003766.

Climer, S., Templeton, A. R., Garvin, M., Jacobson, D., Lane, M., Hulver, S., Scheid, B., Chen, Z., Cruchaga, C., and Zhang, W. (2020). Synchronized genetic activities in alzheimer's brains revealed by heterogeneity-capturing network analysis. *bioRxiv*.

Climer, S., Yang, W., de las Fuentes, L., Dávila-Román, V. G., and Gu, C. C. (2014b). A custom correlation coefficient (ccc) approach for fast identification of multi-snp association patterns in genome-wide snps data. *Genetic epidemiology*, 38(7):610–621.

Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68.

Gonzalez-Recio, O., Rosa, G., and Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Science*, 166.

Grimes, T. and Datta, S. (2021). A novel probabilistic generator for large-scale gene association networks. *PLOS ONE*, 16:e0259193.

Hayes, B. (2013). Overview of statistical methods for genome-wide association studies (gwas). *Methods in molecular biology (Clifton, N.J.)*, 1019:149–69.

Ishigaki, K., Akiyama, M., Kanai, M., Takahashi, A., Kawakami, E., Sugishita, H., Sakaue, S., Matoba, N., Low, S.-K., Okada, Y., Terao, C., Amariuta, T., Gazal, S., Kochi, Y., Horikoshi, M., Suzuki, K., Ito, K., Koyama, S., Ozaki, K., and Kamatani, Y. (2020). Large-scale genome-wide association study in a japanese population identifies novel susceptibility loci across different diseases. *Nature Genetics*, 52.

Joubert, W., Nance, J., Climer, S., Weighill, D., and Jacobson, D. (2019). Parallel accelerated custom correlation coefficient calculations for genomics applications. *Parallel Computing*, 84.

Liu, Y., Lee, Y. F., and Ng, M. (2011). Snp and gene networks construction and analysis from classification of copy number variations data. *BMC bioinformatics*, 12 Suppl 5:S4.

McCarter, C., Howrylak, J., and Kim, S. (2020). Learning gene networks underlying clinical phenotypes using snp perturbation. *PLOS Computational Biology*, 16:e1007940.

Missaggia, B., Reales, G., Cybis, G., Hünemeier, T., and Bortolini, M. (2020). Adaptation and co-adaptation of skin pigmentation and vitamin d genes in native americans. *American Journal of Medical Genetics*, 184.

Mota, N., Araujo-Jnr, E., Paixão-Côrtes, V., Bortolini, M., Henrique, C., and Bau, C. (2012). Linking dopamine neurotransmission and neurogenesis: the evolutionary history of the ntad (ncam1-ttc12-ankk1-drd2) gene cluster. *Genetics and Molecular Biology*, 35:912–918.

Park, S., Cheng, I., and Haiman, C. (2017). Genome-wide association studies of cancer in diverse populations. *Cancer Epidemiology Biomarkers & Prevention*, 27:cebp.0169.2017.

Peprah, E., Xu, H., Tekola-Ayele, F., and Royal, C. (2014). Genome-wide association studies in africans and african americans: Expanding the framework of the genomics of human traits and disease. *Public Health Genomics*, 18.

Rovira, P., Demontis, D., Sánchez-Mora, C., Zayats, T., Klein, M., Mota, N. R., Weber, H., Garcia-Martínez, I., Pagerols, M., Vilar-Ribó, L., et al. (2020). Shared genetic back-

ground between children and adults with attention deficit/hyperactivity disorder. *Neuropsychopharmacology*, 45(10):1617–1626.

Tiosano, D., Audi, L., Climer, S., Zhang, W., Templeton, A., Fernández-Cancio, M., Gershoni-Baruch, R., Sánchez-Muro, J., Kholy, M., and Hochberg, Z. (2016). Latitudinal clines of the human vitamin d receptor and skin color-genes. *G3 (Bethesda, Md.)*, 6.

Uffelmann, E., Huang, Q., Munung, N. S., Vries, J., Okada, Y., Martin, A., Martin, H., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1.

van der Vaart, A. (2000). *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press.

Waldmann, P. (2019). On the use of the pearson correlation coefficient for model evaluation in genome-wide prediction. *Frontiers in Genetics*, 10.

Williams, D. (1991). *Probability with Martingales*. Cambridge mathematical textbooks. Cambridge University Press.

Supplementary Material

S.1 Statistical properties of the Custom Correlation Coefficient (CCC) and related statistics

Proof to Proposition 1. It is immediate that $\mathbf{E}(R_{ij}) = \mathbf{E}(r_{ij})$, where we write $r_{ij} = r_{ij,1}$ for simplicity. It is also clear from the definition that

$$r_{ij} = \begin{cases} 1 & \text{with probability } p_i^2 p_j^2 \\ \frac{1}{2} & \text{with probability } 2p_i(1-p_i)p_j^2 + 2p_i^2 p_j(1-p_j) \\ \frac{1}{4} & \text{with probability } 4p_i(1-p_i)p_j(1-p_j) \\ 0 & \text{with probability 1 minus the sum of the above probabilities.} \end{cases}$$

From this, a direct computation yields $\mathbf{E}(r_{ij}) = p_i p_j$.

For the variance of R_{ij}, assumptions A1 and A2 tell us that $\mathbf{Var}(\sqrt{n}R_{ij}) = \mathbf{Var}(r_{ij}) = \mathbf{E}(r_{ij}^2) - p_i^2 p_j^2$, with

$$\mathbf{E}(r_{ij}^2) = p_i^2 p_j^2 + \tfrac{1}{4} \times 2(p_i(1-p_i)p_j^2 + p_i^2 p_j(1-p_j)) + \tfrac{1}{16} \times 4p_i(1-p_i)p_j(1-p_j).$$

Thus, $\mathbf{Var}(r_{ij}) = \frac{1}{4}(p_i + p_i^2)(p_j + p_j^2) - p_i^2 p_j^2$. This completes the proof. ■

Proof to Proposition 2. Recall that f_i and f_j are the sample frequencies of alleles i and j, respectively. Thus, for instance, we can write

$$f_i = \frac{1}{2n} \sum_{k=1}^{n} X_{ik}, \tag{S1}$$

with $X_{ik} \sim \mathsf{Bin}(2, p_i)$ representing the amount of i alleles in individual k. It is important to note that, in relation to f_i, the sum of X_{ik} is divided by $2n$ because the individuals in the sample have two alleles in each SNP. Therefore, the X_{ik} variable can assume three different values, namely 0, 1, or 2. From this it follows that $\mathbf{E}(f_i) = 2np_i/(2n) = p_i$. Assuming independence between the allele i in SNP_1 and allele j in SNP_2, we obtain $\mathbf{E}(f_i f_j) = p_i p_j$, establishing the claim of unbiasedness. Consistency follows from the Law of Large Numbers (Williams, 1991) applied to f_i and f_j individually, together with the continuous mapping theorem.

We now turn to the variance estimator. Clearly,

$$n \times \left(\widehat{\mathbf{Var}}(R_{ij}) - \mathbf{Var}(R_{ij})\right) = \frac{(f_i + f_i^2)(f_j + f_j^2) - (p_i + p_i^2)(p_j + p_j^2)}{4} - (f_i^2 f_j^2 - p_i^2 p_j^2).$$

Again, since $f_i \to p_i$ almost surely by the Law of Large Numbers, and similarly for f_j, a few applications of the Continuous Mapping Theorem, together with Slutsky's Lemma (van der Vaart, 2000), tells us that the above expression converges to zero almost surely, which completes the proof. ■

Proof to Theorem 1. In order to calculate the expected value of the custom correlation coefficient ($\mathbf{E}(CCC_{ij})$), we need to access some intermediary properties.

Let n_{AA} be the number of individuals with genotype AA in the sample, and define $n_{Aa}, n_{aa}, n_{BB}, n_{Bb}, n_{bb}$ analogously. Additionally let $\nu = (n_{AA}, n_{Aa}, n_{aa}, n_{BB}, n_{Bb}, n_{bb})$. The expected value of the weighted biallelic score given the sample size of each possible genotypic combination, is

$$\mathbf{E}(r_{ij}|\nu) = f_i f_j \tag{S2}$$

In order to prove equation (S2), without loss of generality, let $i = A$ and $j = B$. Let w be the genotypes of individual k for both SNPs, such that $w \in \eta = \{$AABB, AaBB, aaBB, AABb,...,aabb$\}$ and let $r_{ij}(w)$ be the value of the r_{ij} for an individual of genotype w (according to Table 1). Multiplying the probability of the genotypes, given the sample subtotals ν, to their weighted biallelic scores r_{ij}, we have

$$\begin{aligned}
\mathbf{E}(r_{AB}|\nu) &= \sum_{w \in \eta} r_{AB}(w) P(r_{AB}(w)|\nu) \\
&= P(AABB|\nu) r_{AB}(AABB) + P(AABb|\nu) r_{AB}(AABb) + P(AaBB|\nu) r_{AB}(AaBB) + \\
&\quad P(AaBb|\nu) r_{AB}(AaBb) \\
&= \left(\frac{n_{AA}}{n}\frac{n_{BB}}{n}\right) + \left(\frac{n_{AA}}{n}\frac{n_{Bb}}{n}\frac{1}{2}\right) + \left(\frac{n_{Aa}}{n}\frac{n_{BB}}{n}\frac{1}{2}\right) + \left(\frac{n_{Aa}}{n}\frac{n_{Bb}}{n}\frac{1}{4}\right) \\
&= \left(\frac{2n_{AA} + n_{Aa}}{2n}\right)\left(\frac{2n_{BB} + n_{Bb}}{2n}\right) \\
&= f_A f_B.
\end{aligned}$$

In general, this result is valid for all i and j, because we can exchange the labels.With this result, we can compute the expected value of the weighted biallelic score r_{ij} given the allelic frequencies. Note that

$$\mathbf{E}(r_{ij}|f_i, f_j) = \mathbf{E}(\mathbf{E}(r_{ij}|\nu, f_i, f_j)|f_i, f_j)$$

And as f_i and f_j are deterministic functions of ν, we have $\mathbf{E}(r_{ij}|\nu, f_i, f_j) = \mathbf{E}(r_{ij}|\nu)$. Thus, it readily follows from expression (S2) that $\mathbf{E}(r_{ij}|f_i, f_j) = f_i f_j$.

Considering that the average weighted biallelic score R_{ij} is the sample mean of independent individual weights r_{ij}, we have

$$\mathbf{E}(R_{ij}|f_i, f_j) = f_i f_j. \tag{S3}$$

Furthermore, from expression (1) we have, $CCC_{ij} = \frac{9}{2} R_{ij} F_i F_j$, thus

$$\begin{aligned}\mathbf{E}(CCC_{ij}|f_i, f_j) &= \mathbf{E}\left[\left(\frac{9}{2} R_{ij}\left(1 - \frac{f_i}{1.5}\right)\left(1 - \frac{f_j}{1.5}\right)\right)\middle| f_i, f_j\right] \\ &= \frac{9}{2} f_i f_j \left(1 - \frac{f_i}{1.5}\right)\left(1 - \frac{f_j}{1.5}\right).\end{aligned}$$

Finally, once more applying the conditional expectation property, we have

$$\begin{aligned}&\mathbf{E}(CCC_{ij}) \\ &= \mathbf{E}[\mathbf{E}(CCC_{ij}|f_i, f_j)] \\ &= \mathbf{E}\left[\frac{9}{2} f_i f_j \left(1 - \frac{f_i}{1.5}\right)\left(1 - \frac{f_j}{1.5}\right)\right] \\ &= \frac{9}{2}\mathbf{E}\left[f_i f_j - \frac{f_i f_j^2}{1.5} - \frac{f_i^2 f_j}{1.5} + \frac{f_i^2 f_j^2}{2.25}\right] \\ &= \frac{9}{2}\left[\mathbf{E}(f_i)\,\mathbf{E}(f_j) - \frac{1}{1.5}(\mathbf{E}(f_i)\,\mathbf{E}(f_j^2) + \mathbf{E}(f_i^2)\,\mathbf{E}(f_j)) + \frac{1}{2.25}(\mathbf{E}(f_i^2)\,\mathbf{E}(f_j^2))\right] \\ &= \frac{9}{2}\left[p_i p_j - \frac{1}{1.5}(p_i(\frac{p_j - p_j^2}{2n} + p_j^2) + p_j(\frac{p_i - p_i^2}{2n} + p_i^2)) + \frac{1}{2.25}((\frac{p_i - p_i^2}{2n} + p_i^2)(\frac{p_j - p_j^2}{2n} + p_j^2))\right],\end{aligned}$$

thus completing the proof. ■

Proof to Theorem 2. We will write $C_{ij} := CCC_{ij}$ for simplicity. Recall that

$$C_{ij} = \frac{9}{2} R_{ij} F_i F_j,$$

where $F_i = (1 - f_i/q)$ (analogously for F_j), and where $q > 0$ is a tuning parameter. Thus,

since $\mathbb{E}(R_{ij}|f_i, f_j) = f_i f_j$, we have $\mathbb{E}(C_{ij} \mid f_i, f_j) = \frac{9}{2} f_i f_j F_i F_j$, and then

$$\sqrt{n}(C_{ij} - \tfrac{9}{2} f_i f_j F_i F_j) = \sqrt{n}(R_{ij} - f_i f_j)\big[\tfrac{9}{2} F_i F_j\big].$$

As established in the proof of Theorem 3 (see bellow), it holds that $\sqrt{n}(R_{ij} - f_i f_j) \to Z \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = \frac{1}{4}(p_i - p_i^2)(p_j - p_j^2)$. Moreover, by the Weak Law of Large Numbers,

$$F_\ell \to 1 - \frac{p_\ell}{q} \quad \text{in probability,} \quad \ell = i, j.$$

Therefore, by Slutsky's Lemma, it holds that

$$\sqrt{n}(C_{ij} - \tfrac{9}{2} f_i f_j F_i F_j) \to \frac{9}{2} Z \left(1 - \frac{p_i}{q}\right)\left(1 - \frac{p_j}{q}\right).$$

This completes the proof. ■

S.2 Standardized Average Weighted Biallelic Statistic (SAWB)

Proof to Lemma 1. From Propositions 1 and 2, we have $\mathbf{E}(R_{ij} - f_i f_j) = 0$. As for the variance of $R_{ij} - f_i f_j$, the preceding identity allows us to expand

$$\mathbf{Var}(R_{ij} - f_i f_j) = \mathbf{E}(R_{ij}^2) + \mathbf{E}(f_i^2 f_j^2) - 2\,\mathbf{E}(R_{ij} f_i f_j). \tag{S4}$$

We know that $\mathbf{E}(R_{ij} f_i f_j) = \mathbf{E}[f_i f_j\, \mathbf{E}(R_{ij} \mid f_i, f_j)] \overset{*}{=} \mathbf{E}[(f_i f_j)(f_i f_j)] = \mathbf{E}(f_i^2 f_j^2)$, where the starred equality follows from (S3), so (using independence between alleles i and j) we arrive at $\mathbf{Var}(R_{ij} - f_i f_j) = \mathbf{E}(R_{ij}^2) - \mathbf{E}(f_i^2)\,\mathbf{E}(f_j^2)$. Now, we have

$$\mathbf{E}(f_i^2) = \mathbf{Var}(f_i) + p_i^2 = \frac{1}{2n}(p_i - p_i^2) + p_i^2 \tag{S5}$$

since the assumption of iid sampling yields $\mathbf{Var}(f_i) = \frac{1}{4n^2} \times n\,\mathbf{Var}(X_{i1}) = \frac{1}{4n} 2p_i(1 - p_i) = \frac{1}{2n} p_i(1 - p_i)$. Proposition 1 in turn gives us

$$\mathbf{E}(R_{ij}^2) = \frac{1}{n}\left(\frac{(p_i + p_i^2)(p_j + p_j^2)}{4} - p_i^2 p_j^2\right) + p_i^2 p_j^2. \tag{S6}$$

We see then that

$$\begin{aligned} 4n^2\,\mathbf{Var}(R_{ij} - f_i f_j) &= n(p_i + p_i^2)(p_j + p_j^2) - 4np_i^2 p_j^2 + 4n^2 p_i^2 p_j^2 \\ &\quad - (p_i - p_i^2 + 2np_i^2)(p_j - p_j^2 + 2np_j^2) \\ &= (n-1)(p_i - p_i^2)(p_j - p_j^2). \end{aligned}$$

Consistency of the proposed estimator follows from consistency of f_i and f_j, combined with the Continuous Mapping Theorem and an application of Slutsky's Lemma. This completes the proof. ■

Proof to Theorem 3. We first show that $\sqrt{n}(R_{ij} - f_i f_j) \xrightarrow{d} N(0, \sigma^2)$, for some $\sigma^2 > 0$. In fact, we can write

$$\begin{aligned} \sqrt{n}(R_{ij} - f_i f_j) = \sqrt{n}(R_{ij} - p_i p_j) - p_i\sqrt{n}(f_j - p_j) - p_j\sqrt{n}(f_i - p_i) \\ - \sqrt{n}(f_j - p_j)(f_i - p_i). \end{aligned} \tag{S7}$$

Since f_i is a sample mean of iid random variables with expected value p_i, the Weak Law of Large Numbers tells us that $(f_i - p_i) = o_{\mathbf{P}}(1)$. Similarly, the Central Limit Theorem implies $\sqrt{n}(f_j - p_j)$ is bounded in probability ($O_{\mathbf{P}}(1)$). Thus, the term in (S7) is $O_{\mathbf{P}}(1)o_{\mathbf{P}}(1) = o_{\mathbf{P}}(1)$, and, *a fortiori*, $\sqrt{n}(f_j - p_j)(f_i - p_i) \to 0$ in distribution. Slutsky's Lemma then tells us that $\sqrt{n}(R_{ij} - f_i f_j)$ has the same limiting distribution as

$$\sqrt{n}(R_{ij} - p_i p_j) - p_i\sqrt{n}(f_j - p_j) - p_j\sqrt{n}(f_i - p_i) =: \sqrt{n}(\bar{Z}_n - \mu),$$

where $\bar{Z}_n := n^{-1}\sum_{k=1}^n Z_k$, with $Z_k := r_{ij,k} - \frac{1}{2}p_i X_{jk} - \frac{1}{2}p_j X_{ik}$ and $\mu := \mathbf{E}(Z_1) = -p_i p_j$. Since the Z_k's are iid with $\mathbf{E}(Z_k) = \mu$, the Central Limit Theorem yields $\sqrt{n}(\bar{Z}_n - \mu) \to \mathcal{N}(0, \sigma^2)$ in distribution, where $\sigma^2 = \mathbf{Var}(Z_1)$. It is not difficult to establish that $\sigma^2 = \frac{1}{4}(p_i^2 - p_i)(p_j^2 - p_j)$.

We have shown that $\sqrt{n}(R_{ij} - f_i f_j) \to Z$, where $Z \sim \mathcal{N}(0, \sigma^2)$. Now, the Weak Law of Large Numbers, together with a few applications of Slutsky's Lemma and the Continuous Mapping Theorem, tells us that

$$\sqrt{\frac{n-1}{4n}(f_i^2 - f_i)(f_j^2 - f_j)} \to \sigma \quad \text{in probability.}$$

Employing Slutsky's Lemma one last time, we get $S_{ij} \to \sigma^{-1}Z \sim \mathcal{N}(0, 1)$. ■

Remark. In the denominator of S_{ij}, we could use the consistent estimator $\widehat{\sigma}^2 = \frac{1}{4}(f_i^2 - f_i)(f_j^2 - f_j)$ in place of the slightly more intricate $\widehat{\mathbf{Var}}(\sqrt{n}(R_{ij} - f_i f_j))$, but we chose to employ the latter as it is an estimator of the correct finite-sample variance $\mathbf{Var}(\sqrt{n}(R_{ij} - f_i f_j))$.

Proof to Proposition 3. Let us first show that $S_{ab} = S_{AB}$. We have, for $1 \leqslant k \leqslant n$,

$$\begin{aligned} r_{ab,k} &= \tfrac{1}{4} X_{ak} X_{bk} = \tfrac{1}{4}(2 - X_{Ak})(2 - X_{Bk}) \\ &= \tfrac{1}{4}(4 - 2X_{Bk} - 2X_{Ak} + X_{Ak}X_{Bk}) \\ &= 1 - \tfrac{1}{2} X_{Bk} - \tfrac{1}{2} X_{Ak} + r_{AB,k} \end{aligned}$$

so, summing along k and dividing by n, we arrive at $R_{ab} = 1 - f_B - f_A + R_{AB}$. Thus

$$\begin{aligned} R_{ab} - f_a f_b &= 1 - f_B - f_A + R_{AB} - (1 - f_A)(1 - f_B) \\ &= 1 - f_B - f_A + R_{AB} - [1 - f_B - f_A + f_A f_B] \\ &= R_{AB} - f_A f_B. \end{aligned}$$

Moreover

$$f_a^2 - f_a = (1 - f_A)^2 - (1 - f_A) = f_A^2 - f_A, \tag{S8}$$

and similarly $f_b^2 - f_b = f_B^2 - f_B$. This yields the desired identity.

As for the equality $S_{ab} = -S_{Ab}$, again we can write

$$\begin{aligned} r_{ab,k} &= \tfrac{1}{4} X_{ak} X_{bk} = \tfrac{1}{4}(2 - X_{Ak}) X_{bk} \\ &= \tfrac{1}{2} X_{bk} - r_{Ab,k}. \end{aligned}$$

Taking the sample average of the above quantity yields $R_{ab} = f_b - R_{Ab}$, and then

$$\begin{aligned} R_{ab} - f_a f_b &= f_b - R_{Ab} - (1 - f_A) f_b \\ &= -(R_{Ab} - f_A f_b). \end{aligned}$$

Now it is just a matter of using the computations in (S8) to conclude that $S_{ab} = -S_{Ab}$. ■

S.3 Figures

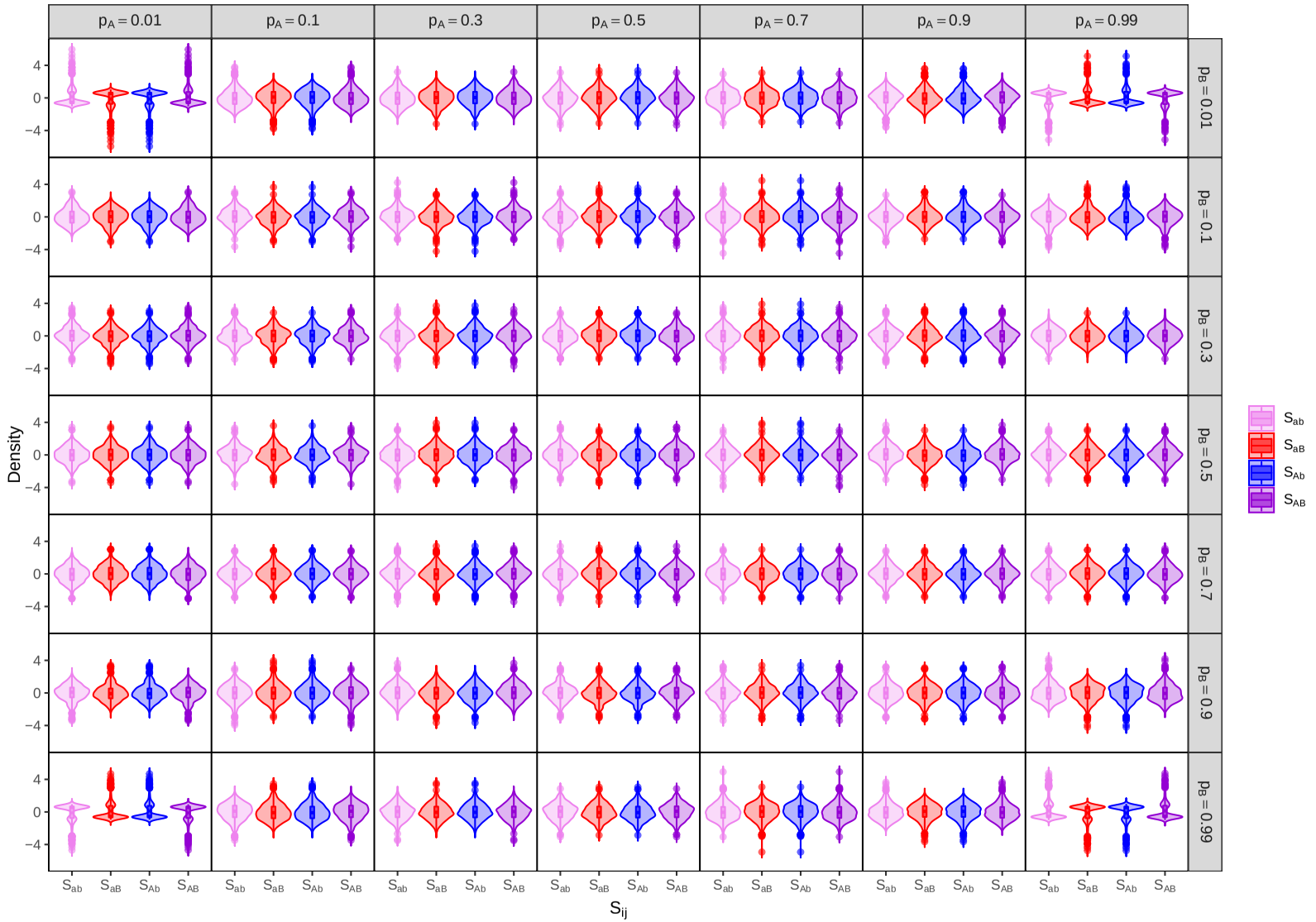

Figure S1: S_{ij} distributions under H_0 for different allelic frequency combinations.

Referências Bibliográficas

ALBERTS, B. et al. Molecular Biology of The Cell. [S.l.: s.n.], 2002. v. 1. ISBN 0-8153-3218-1.

CANTOR, R.; LANGE, K.; SINSHEIMER, J. Prioritizing gwas results: A review of statistical methods and recommendations for their application. American journal of human genetics, v. 86, p. 6–22, 01 2010.

CANZONIERO, J.; ROSENBERG, N. Mathematical properties of the measure of linkage disequilibrium. Theoretical population biology, v. 74, p. 130–7, 09 2008.

CLIMER, S. et al. Synchronized genetic activities in alzheimer's brains revealed by heterogeneity-capturing network analysis. bioRxiv, Cold Spring Harbor Laboratory, 2020.

CLIMER, S. et al. A custom correlation coefficient (ccc) approach for fast identification of multi-snp association patterns in genome-wide snps data. Genetic epidemiology, Wiley Online Library, v. 38, n. 7, p. 610–621, 2014.

CROSSLIN, D.; QIN, X.; HAUSER, E. Assessment of ld matrix measures for the analysis of biological pathway association. Statistical applications in genetics and molecular biology, v. 9, p. Article35, 01 2010.

ELSTON, R. Introduction and overview. Statistical Methods in Medical Research, v. 9, p. 527–541, 12 2000.

EVANS, W.; RELLING, M. Pharmacogenomics: Translating functional genomics into rational therapeutics. Science, v. 286, 10 1999.

GIOVANETTI, M. et al. Evolution patterns of sars-cov-2: Snapshot on its genome variants. Biochemical and Biophysical Research Communications, v. 538, 11 2020.

GRIFFITHS, A. et al. An Introduction To Genetic Analysis. [S.l.: s.n.], 2006.

HEINO, M. Quantitative traits. Stock Identification Methods: Applications in Fishery Science: Second Edition, p. 59–76, 10 2013.

HUDSON, R. Linkage disequilibrium and recombination. In: _____. [S.l.: s.n.], 2004. ISBN 9780470022627.

JORDE, L. et al. The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and y-chromosome data. American journal of human genetics, v. 66, p. 979–88, 04 2000.

KANG, J. T.; ROSENBERG, N. A. Mathematical properties of linkage disequilibrium statistics defined by normalization of the coefficient d = pab – papb. Human Heredity, v. 84, p. 1–17, 02 2020.

KORTE, A.; FARLOW, A. The advantages and limitations of trait analysis with gwas: A review. Plant methods, v. 9, p. 29, 07 2013.

LANDER, E.; SCHORK, N. Genetic dissection of complex traits. Science (New York, N.Y.), v. 265, p. 2037–48, 10 1994.

LEWONTIN, R. The interaction of selection and linkage. i. general considerations; heterotic models. Genetics, v. 49, p. 49–67, 02 1964.

LEWONTIN, R. On measures of gametic disequilibrium. Genetics, v. 120, p. 849–852, 11 1988.

MANGIN, B. et al. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity, v. 108, p. 285–91, 08 2011.

NHGRI. National Human Genome Research Institute. 2022. ⟨https://www.genome.gov/genetics-glossary/⟩. Accessed: 2022-02-14.

NJ, R.; MERIKANGAS, K. The future of genetic studies of complex human diseases. Science (New York, N.Y.), v. 273, p. 1516–7, 10 1996.

PRITCHARD, J.; PRZEWORSKI, M. Linkage disequilibrium in humans: Models and data. American journal of human genetics, v. 69, p. 1–14, 08 2001.

PURCELL, S. et al. Plink: A tool set for whole-genome association and population-based linkage analyses. American journal of human genetics, v. 81, p. 559–75, 10 2007.

REICH, D.; LANDER, E. Reich, d. e. & lander, e. s. on the allelic spectrum of human disease. trends genet. 17, 502-510. Trends in genetics : TIG, v. 17, p. 502–10, 10 2001.

SLATKIN, M. Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. Nature reviews. Genetics, v. 9, p. 477–85, 07 2008.

SYVANEN, A.-C. Accessing genetic variation: Genotyping single nucleotide polymorphisms. Nature reviews. Genetics, v. 2, p. 930–42, 01 2002.

UFFELMANN, E. et al. Genome-wide association studies. Nature Reviews Methods Primers, v. 1, 12 2021.

WAPLES, R.; ENGLAND, P. Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. Genetics, v. 189, p. 633–44, 08 2011.

ZAPATA, C. The d' measure of overall gametic disequilibrium between pairs of multiallelic loci. Evolution; international journal of organic evolution, v. 54, p. 1809–12, 11 2000.