

# 1 A Breast Cancer Polygenic Risk Score Validation in 15,490 Brazilians 2 using Exome Sequencing

3

4 Flávia Eichenberger Rius<sup>1,2</sup>, Rodrigo Guindalini<sup>3</sup>, Danilo Viana<sup>1</sup>, Júlia Salomão<sup>1</sup>,  
5 Laila Gallo<sup>1</sup>, Renata Freitas<sup>1</sup>, Cláudia Bertolacini<sup>1</sup>, Lucas Taniguti<sup>1</sup>, Danilo  
6 Imparato<sup>1</sup>, Flávia Antunes<sup>1</sup>, Gabriel Sousa<sup>1</sup>, Renan Achjian<sup>1</sup>, Eric Fukuyama<sup>1</sup>,  
7 Cleandra Gregório<sup>1</sup>, Iuri Ventura<sup>1</sup>, Juliana Gomes<sup>1</sup>, Nathália Taniguti<sup>1</sup>, Simone  
8 Maistro<sup>2</sup>, José Eduardo Krieger<sup>4</sup>, Yonglan Zheng<sup>5</sup>, Dezheng Huo<sup>6</sup>, Olufunmilayo I.  
9 Olopade<sup>5</sup>, Maria Aparecida Koike<sup>2</sup>, David Schlesinger<sup>1</sup>

10

11 1. Mendelics, São Paulo, Brazil.

12 2. Comprehensive Center for Precision Oncology - C2PO, Centro de Investigação  
13 Translacional em Oncologia (CTO), Departamento de Radiologia e Oncologia, Instituto  
14 do Cancer do Estado de Sao Paulo (ICESP), Hospital das Clinicas HCFMUSP,  
15 Faculdade de Medicina, Universidade de Sao Paulo, Sao Paulo, Brazil.

16 3. Instituto D'Or de Pesquisa e Ensino (IDOR), São Paulo, Brazil.

17 4. Instituto do Coração, Hospital das Clínicas da Faculdade de Medicina da  
18 Universidade de São Paulo - FMUSP, São Paulo, Brazil.

19 5. Medicine and Human Genetics, Center for Clinical Cancer Genetics and Global  
20 Health, University of Chicago Medical Center, Chicago, USA.

21 6. Department of Public Health Sciences, University of Chicago, Chicago, USA.

22

## 23 Abstract

### 24 Purpose

25 Brazil has a highly admixed population. Polygenic Risk Scores (PRS) have been mostly  
26 developed from European population studies and applying them to other populations is  
27 challenging. To assess the use of PRS for breast cancer (BC) risk in Brazil, we validated  
28 four PRSs developed in the Brazilian population.

### 29 Patients and Methods

30 We analyzed 6,362 women with a history of breast cancer and 9,128 unphenotyped  
31 adults as controls in a sample obtained from a clinical laboratory. Genomic variants  
32 were imputed from exomes and scores were calculated for all samples.

### 33 Results

34 After excluding individuals with known pathogenic or likely pathogenic variants in  
35 *BRCA1*, *BRCA2*, *PALB2*, *PTEN*, or *TP53*, and first-degree relatives of the probands,  
36 5,730 cases and 8,847 controls remained. Four PRS models were compared, and PRS  
37 3820 from Mavaddat *et al.* 2019 performed best, with an Odds Ratio (OR) of 1.41 per  
38 standard deviation (SD) increase (p-value: < 0.0001) and an OR of 1.94 (p-value: <  
39 0.0001) for the individuals in the top risk decile. PRS 3820 also performed well for  
40 different ancestry groups: East Asian majority (Group 1), Non-European majority (Group  
41 2), and European majority (Group 3), showing significant effect sizes for all groups:  
42 (Group 1: OR 1.54, p-value 0.006; Group 2: OR 1.44, p-value: <0.001; Group 3 OR:

43 1.43, p-value: <0.001). PRS 90% compares with monogenic moderate BC risk genes  
44 (PRS90 OR: 1.94; CHEK2 OR: 1.89; ATM OR: 1.99).

## 45 Conclusion

46 PRS 3820 can be accurately used in the Brazilian population. This will allow a more  
47 precise BC risk assessment of mutation-negative women in Brazil.

48

## 49 Introduction

50 Breast cancer (BC) is a critical global health concern, representing the most common  
51 cancer diagnosed among women<sup>1</sup>. In Brazil, over 70,000 women are diagnosed with BC  
52 every year, accounting for 30% of all cancers in the female population<sup>2</sup>.

53 Approximately 10% of all BC cases are attributable to germline pathogenic variants in  
54 susceptibility genes<sup>3</sup>. Rare variants in high penetrance genes (*BRCA1*, *BRCA2*, *TP53*,  
55 *PTEN*, and *PALB2*) and in moderate penetrance genes (*CHEK2* and *ATM*) are  
56 associated with a more than 4-fold and 1.5–4 fold increased risk of BC, respectively<sup>4,5</sup>.

57 Rare variants in these genes account for approximately 25% of the genetic risk. The  
58 remaining genetic risk (~75%) is derived from common, low penetrance variants that  
59 individually confer small risk, but which combined effect can be substantial<sup>4–6</sup>.

60 Genome-wide association studies (GWASs) have been predominantly carried out in  
61 European populations<sup>7–10</sup>. Evaluation of PRS across different genetic and environmental  
62 backgrounds is essential to enable the implementation of genetic risk stratification  
63 strategies for individuals from non-European populations<sup>11</sup>.

64 The Brazilian population exhibits a unique, highly admixed, genetic composition. It is  
65 mostly derived from a combination of Native Americans, Southern Europeans  
66 (Portuguese, Spanish, and Italian) that immigrated in the period 1500-1900, and  
67 Sub-Saharan Africans brought through extensive slave trading until the 1800s. More  
68 recently, from 1822 to the first half of the 1900s, other smaller waves of immigration also  
69 contributed to Brazil's remarkable diversity, including Japanese, Lebanese, German,  
70 and Eastern Europeans<sup>12</sup>. Three in every four Brazilians have multiple genetic  
71 ancestries<sup>13,14</sup>. Given Brazil's genetic diversity, any PRS developed in predominantly  
72 European populations requires validation before it can be used in clinical settings.  
73 Several laboratory methods are available for genotyping variants directly or indirectly  
74 (imputation), including microarrays, whole exome sequencing (WES), and whole  
75 genome sequencing (WGS). WES offers an affordable and scalable alternative to arrays  
76 and WGS, while allowing for simultaneous rare and common variant genotyping.  
77 In this study, we evaluate four BC PRSs<sup>7,8,15</sup> developed using WES in 15,490 Brazilians.  
78

## 79 **Methods**

### 80 **Study population**

81 A total of 15,490 individuals were selected for this study, including 6,362 women with  
82 breast cancer history, and 9,128 adult unphenotyped controls. Both clinical and genetic  
83 data were collected from a database of a College of American Pathology  
84 (CAP)-accredited laboratory (Mendelics, São Paulo, SP, Brazil). All BC and control  
85 subjects provided Informed Consentment for use of retrospective anonymized data for

86 research purposes. Samples were anonymized before analysis. Clinical records such as  
87 BC histological type and age of diagnosis were obtained from genetic test requisitions.  
88 The study was IRB-approved (CAAE: 70112423.3.0000.0068 ).

89

## 90 **Relatedness calculation and data filtering**

91 Relatedness of individuals was obtained from the exomes using somalier software<sup>16</sup>,  
92 following the standard protocol required for a VCF file  
93 (<https://github.com/brentp/somalier#readme>). Concerning related individuals removal, if  
94 two individuals had a first-degree relationship, one of them was randomly selected to be  
95 included in the dataset. However, if individuals had two or more first-degree  
96 relationships, all related individuals were excluded from the dataset. This process  
97 resulted in a total of 211 removals. Furthermore, 73 individuals were removed from the  
98 sample due to unavailability of files necessary for genome imputation.

99 PRS analyses were performed after filtering out cases and controls with pathogenic or  
100 likely-pathogenic (P/LP) variants in BC genes *BRCA1*, *BRCA2*, *TP53*, *PALB2*, and *PTEN*.

101

## 102 **Exome sequencing and imputation**

103 Exome sequencing data were generated from buccal swab or venous blood samples  
104 with standard protocol for Illumina Flex Exome Prep, using a custom probe set from  
105 Twist Biosciences. Sequencing was conducted in Illumina sequencers and the  
106 bioinformatics pipeline for data analysis followed Broad Institute's GATK best practices

107 (<https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflo>  
108 ws), with alignment to GRCh38.

109 Imputation of exomes was based on a panel of 2,504 individuals of all ancestries from  
110 the 1000 Genomes Project (1KGP)<sup>17</sup> on GRCh38 (2017 release)  
111 (<https://www.internationalgenome.org/data-portal/data-collection/grch38>). All regions  
112 captured from the exome sequencing comprehending at least 1x coverage, as well as  
113 off-target regions, were considered for the imputation, performed using Glimpse (v1.1.0)  
114 software<sup>18</sup>.

115

#### 116 **Polygenic Risk Score calculation**

117 Four BC PRSs with publicly available summary statistics, from three different studies,  
118 were evaluated in this work: Khera *et al.* 2018<sup>7</sup>, with 5,218 variants; Mavaddat *et al.*  
119 2019<sup>8</sup> PRSs (with 313 and 3,820 variants); and UK Biobank<sup>15</sup> (UKBB) PRS obtained  
120 from a variant thresholding ( $p$ -value < 10e-5) on summary statistics for phenotype code  
121 20001\_1002, with 7,538 variants.

122 The PRS variants were selected based on exome bed kit distance and minor allele  
123 frequency (MAF). Additionally, the PRS from Mavaddat study, originally with 3,820  
124 variants, had a pathogenic variant of moderate-penetrance in *CHEK2* gene (*CHEK2*  
125 p.Ile157Thr - Clinvar: RCV000144596) that was removed to avoid conflation with  
126 monogenic risk.

127 PRS calculation was performed using a software developed by Mendelics, evaluating  
128 the weighted sum of beta values, in which weights are based on the number of the

129 individual's alleles containing the variant of the PRS file. The sum is normalized by all  
130 beta positive and negative values so the final value can be between zero and one.

131

### 132 **Genetic Principal Component Analyses (PCA) Assessment**

133 PCA was calculated for exomes from a projection in 1KGP<sup>17</sup> and Human Genome  
134 Diversity Project (HGDP)<sup>19</sup> samples. Only variants with MAF > 1% and that could have  
135 been directly genotyped using WES were included for the PCA analysis in 1KGP and  
136 HGDP samples using plink2<sup>20</sup>. Exomes were converted to plink bfile format (bed, bim,  
137 and fam files) and had duplicated variants removed. PCA projection for 10 PCs was  
138 calculated using plink2 –score method, with allele frequencies from the breast cancer  
139 case-control sample.

140

### 141 **Ancestry evaluation**

142 Admixture<sup>21</sup> was used to extract continental ancestries from all non-related and data  
143 completed exomes. The analysis was supervised by the 1KGP samples, after removal  
144 of South Asian, Oceania, and admixed Americans from the GRCh38 1KGP release of  
145 2017. South Asian and Oceania ancestries were removed because they are not a  
146 significant part of Brazilian ancestral composition. Latin American admixed populations  
147 (Colombian, Peruvian, Puerto Rican, and Mexican) were removed to avoid confounding  
148 with the native americans belonging to the same population label. Continents evaluated  
149 were: Africa - AFR, America - AMR, East Asia - EAS, and Europe - EUR. Ancestry

150 results were further used for splitting individuals into groups according to their ancestry  
151 composition, to further analyze the effect size of PRS on each group.

152

### 153 Paired imputed and sequenced genomes analysis

154 Exome-imputed variants and directly sequenced variants from WGS were compared  
155 using 1001 samples from an independent Brazilian population dataset  
156 (<http://elsabrasil.org/>) that had both WES and WGS available. The WES were  
157 sequenced and imputed also using the same method previously described. BC  
158 PRS-3820 from Mavaddat *et al.* study was calculated for both imputed and sequenced  
159 genomes, and their Spearman correlation was calculated using R software base  
160 function *cor.test*.

161

### 162 Statistical analyses

163 PRS values were standardized according to the control values prior to all statistical  
164 analyses. PCs were Z-scored prior to analyses. To assess the effect size of PRS on  
165 breast cancer status (0 = control, 1 = case) corrected for PCs, Odds Ratio (OR) per  
166 standard deviation of PRS was calculated by performing a logistic regression of BC  
167 status with PRS and PCs 1 to 10 as predictors. AUC for the full dataset evaluation was  
168 obtained using the yardstick R package ([yardstick.tidymodels.org/](https://yardstick.tidymodels.org/)) *roc\_auc* function, in  
169 the testing data split (25%). In order to find segmentation effect-sizes, individuals were  
170 classified into deciles or percentiles following the left-open and right-closed intervals.  
171 OR for deciles was calculated by first selecting only the decile analyzed and the interval



172 from 40-60% individuals as the control section, and binarizing it (0 = belongs to the  
173 control interval 40-60%, 1 = belongs to the decile analyzed, for example, 10%); and  
174 performing a logistic regression analysis on the binarized decile information with  
175 correction for PCs 1 to 10. A similar approach was conducted for calculating the OR on  
176 percentiles for comparison with Mavaddat's<sup>8</sup> PRS validation. For each ancestry  
177 proportion group, AUC was estimated using 10-fold cross-validation with the R package  
178 *caret*<sup>22</sup>. All PRS 95% confidence intervals (CI) were obtained from the logistic  
179 regression output from the R function *glm* (stats package<sup>23</sup>). OR and CI for genes  
180 *BRCA1*, *BRCA2*, *PALB2*, *TP53*, *ATM* and *CHEK2* were obtained using *epitools* R  
181 package<sup>24</sup>. All statistical tests performed were two-tailed.

182

## 183 Results

### 184 Case-control sample selection and characteristics

185 After removal of 211 subjects with a first-degree relationship and 73 with missing files  
186 necessary for imputation, a total of 15,206 subjects remained (**Supplementary Table**  
187 **1**). Four percent of all cases and controls were removed from the analysis due to their  
188 presence of pathogenic or likely-pathogenic (P/LP) variants in high penetrance genes  
189 with OR > 5 for breast cancer: *BRCA1*, *BRCA2*, *TP53*, *PALB2*, and *PTEN* (n = 629).  
190 Therefore, the sample used for PRS evaluation consisted of 5,730 women with a BC  
191 history, and 8,847 unphenotyped controls, both with known sex and age (**Table 1**).

192

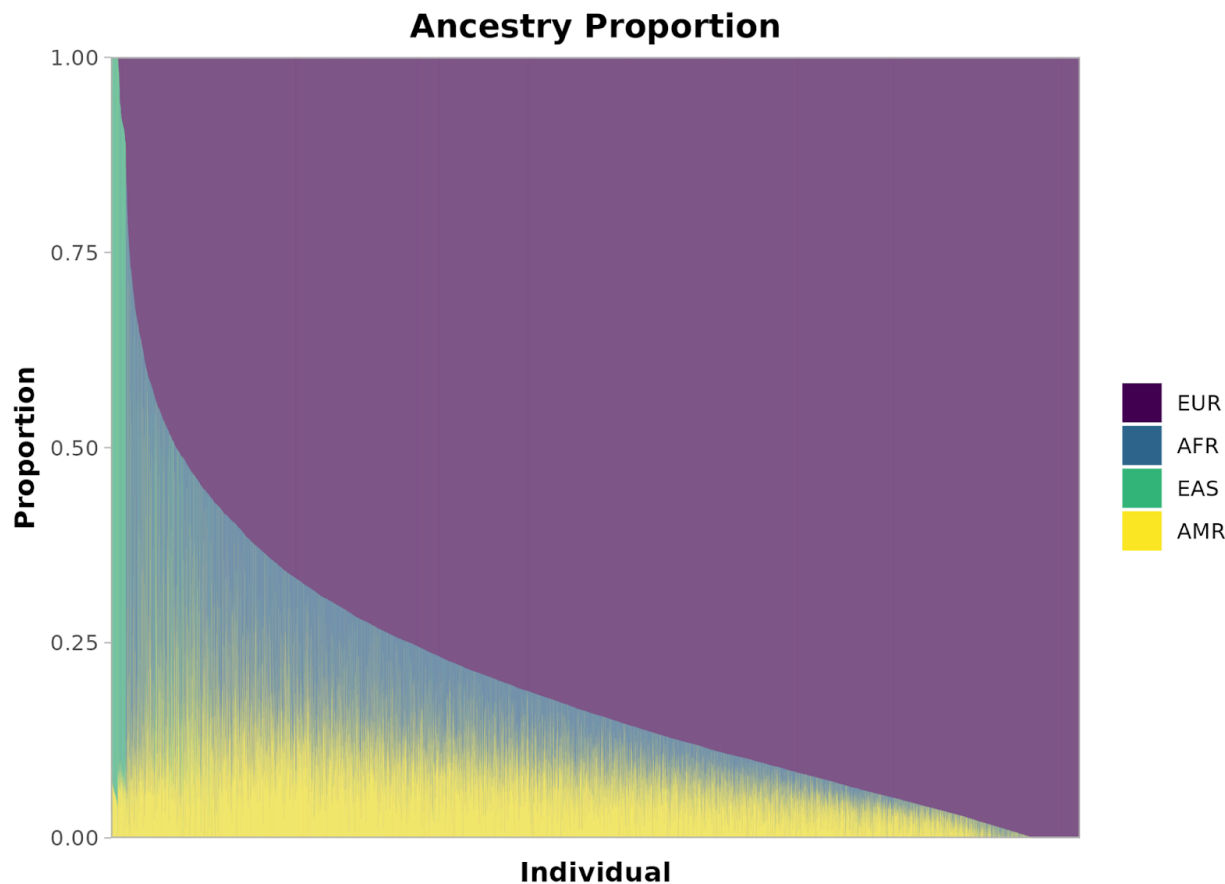
193 **Table 1.** Demographics of cases and controls in BC dataset used for PRS evaluation

		<b>Case</b>	<b>Control</b>	<b>Total</b>	<b>p value</b>
	Total	5,730	8,847	14,577	-
<b>Sex</b>	F	4,225	5,730	9,955	-
	M	-	4,662	4,662	-
<b>Age</b>	Total	49.5 (11.7)	41.6 (13.3)	44.8 (13.3)	0.000
	F	49.5 (11.7)	42 (13.7)	46.4 (13.1)	0.000
	M	-	41.3 (12.9)	41.3 (12.9)	-

194 p-values obtained from two-tailed t-tests.

195

196 Ancestry composition of our sample was obtained using ADMIXTURE analysis<sup>21</sup>,  
197 supervised by EUR, EAS, AFR and non-admixed AMR populations of 1KGP and HGDP.  
198 The results show that the majority of individuals have EUR as their greatest ancestry  
199 proportion (median 84%, SD 18%). Besides that, a significant portion of AFR (median  
200 6%, std. dev. 12%) and AMR (median 8%, SD 7%) ancestries are present,  
201 complemented with a variety of EUR proportions. A small quota of EAS is also observed  
202 (median < 1%, SD 12%), composed by 214 individuals with over 70% of this ancestry.



204 **Figure 1. Ancestry composition of our Brazilian cohort.** Estimated ancestries are  
205 shown as proportions per individual. Each thin bar represents one individual and their  
206 ancestry proportion. Europe (EUR) in purple, Africa (AFR) in blue, East Asia (EAS) in  
207 green and America (AMR) in yellow.

208

### 209 **Effect sizes of four different PRSs in the Brazilian population**

210 Four PRS files from three studies were selected for initial effect size investigation in our  
211 cohort (**Supplementary Table 2**). All four PRS files had their variants further filtered to  
212 address only variants covered by the imputation of our exomes.

213

214 PCA was performed on the exomes to capture the population genetic structure. PRSs  
215 were calculated for the imputed genomes (details described in the **Methods**) and  
216 standardized by z-score to improve interpretability. To avoid confounding from P/LP  
217 variants on PRS effect, we have evaluated only individuals without those rare variants  
218 ( $n = 14,577$ ). Effects were corrected for the ten first PCs, and results are all reported in  
219 **Supplementary Table 3**.

220

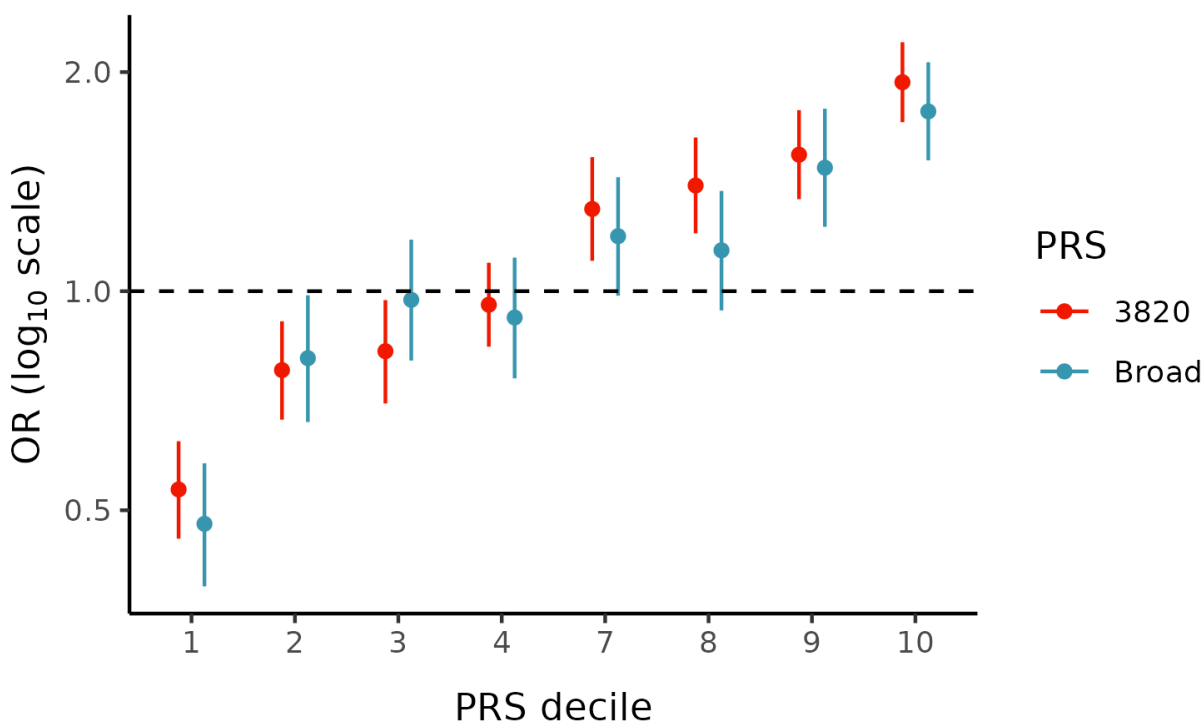
221 Both  $PRS_{\text{Broad}}$  and  $PRS_{3820}$  performed well, with very significant effect sizes (both  
222 p-values  $< 0.0001$ ) following the direction of risk rise as the PRS increases ( $OR_{\text{Broad}}$ :  
223 1.52;  $OR_{3820}$ : 1.41).  $PRS_{313}$  and  $PRS_{\text{UKBB}}$  have not reached significance level for their OR  
224 results (p-value<sub>313</sub>: 0.315 and p-value<sub>UKBB</sub>: 0.985). Goodness of fit of the model is also  
225 greater for  $PRS_{3820}$  (Nagelkerke pseudo- $R^2$ : 0.061) and  $PRS_{\text{Broad}}$  (Nagelkerke  
226 pseudo- $R^2$ : 0.051). Note that pseudo- $R^2$  values should not be interpreted as a linear  
227 regression  $R^2$  value, but as a metric of improvement from null model to fitted model,  
228 which has its value mainly by being compared between different PRS models in which a  
229 greater pseudo- $R^2$  indicates a better goodness of fit to the data.

230

231 Since  $PRS_{\text{Broad}}$  and  $PRS_{3820}$  showed significant results per standard deviation, they were  
232 used to split the data into deciles to evaluate BC risk conferred by PRS in each strata.  
233 These analyses were also corrected for the first ten PCs. Interestingly, shorter  
234 confidence intervals and a better “staircase” shape can be seen for  $PRS_{3820}$  plot in  
235 comparison to  $PRS_{\text{Broad}}$  (**Figure 2**). Moreover, especially the top 10% (90-100% interval)  
236 present a much greater effect for  $PRS_{3820}$  ( $OR_{90-100}$ : 1.94; CI: 1.71 - 2.20) compared to

237 PRS<sub>Broad</sub> (OR<sub>90-100</sub>: 1.77; CI: 1.51 - 2.10) (**Supplementary Table 4**), indicating a better  
238 performance of the former in identifying women with increased risk of BC. Therefore we  
239 decided to focus our next analyses on PRS<sub>3820</sub>, which was the best PRS to identify BC  
240 risk in our Brazilian population.

241



242

243 **Figure 2. Effect sizes by decile of PRS<sub>3820</sub> and PRS<sub>Broad</sub>.** Odds Ratios (OR) and  
244 Confidence Intervals (CI) for PRS<sub>3820</sub> (red) and PRS<sub>Broad</sub> (blue). ORs for both PRS  
245 deciles were corrected for the first ten PCs.

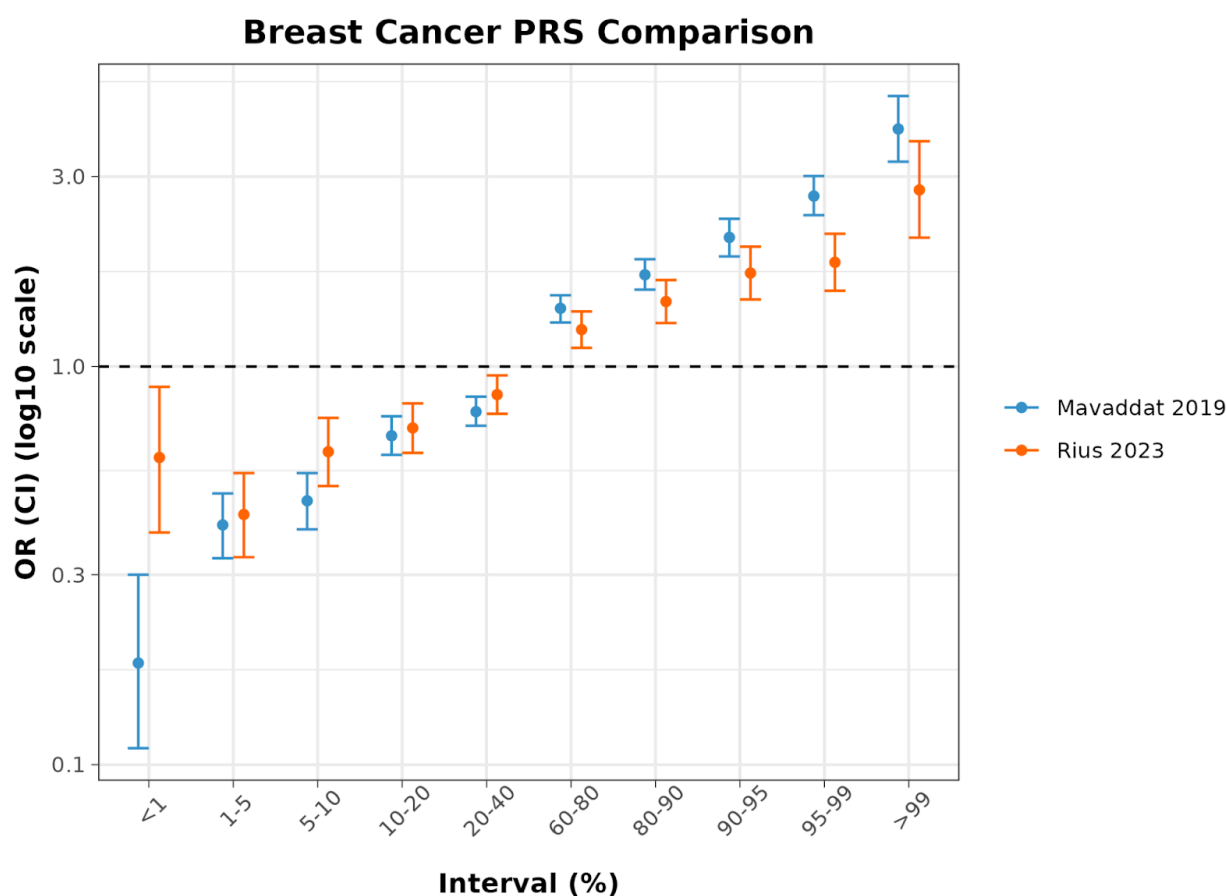
246

#### 247 PRS<sub>3820</sub> performance compared with the original study

248 As seen in the previous results, the PRS<sub>3820</sub> showed a positive association with  
249 increased risk of BC (OR per standard deviation: 1.41; CI: 1.36 - 1.47) after correction

250 for the first 10 principal components (PCs). This association was slightly lower when  
251 compared to the original study test set, composed of only Europeans (OR: 1.66; CI:  
252 1.61 - 1.70). Besides that, performance of our model with PRS<sub>3820</sub> in identifying BC  
253 cases was very similar to the original study ( $AUC_{\text{Brazilians}}: 0.610$  vs.  $AUC_{\text{Europeans}}: 0.636$ ).  
254 After calculating OR per percentiles, we observed that the PRS<sub>3820</sub> exhibited an  
255 expressive risk increase for our admixed population, although the increase was smaller  
256 than the original study, which applied the PRS<sub>3820</sub> to a population with the same  
257 ancestry it was originated from ( $OR_{\text{Brazilians}} >99: 2.72$ ;  $OR_{\text{Europeans}} >99: 3.95$ ).

258



259

260 **Figure 3. Comparison of PRS<sub>3820</sub> performance for Europeans and Brazilians.** The  
261 plot shows the PRS<sub>3820</sub> adapted in this study (orange), with 2,892 variants, compared  
262 with the original from Mavaddat *et al.* study (blue), with 3,820 variants.

263

264 The lower interval, comprehending the lowest 1% of PRS values, showed a smaller  
265 decrease in BC risk compared to the original study. This result is probably related to the  
266 small sample size of this section, with only 31 cases and 88 controls available to  
267 calculate OR. In addition, the 95th to 99th percentile interval exhibited marginal growth  
268 in odds ratio (OR) when contrasted with the interval immediately below (OR 90th-95th:  
269 1.75, OR 95th-99th: 1.83). Besides that, both effect sizes show an expressive increase  
270 in BC risk due to PRS results. This might be partly due to the cohort sample size. Our  
271 study evaluated a total of 14,577 individuals, while Mavaddat's evaluated twice this  
272 number in their test dataset composed of joined cohorts (n = 29,751).

273

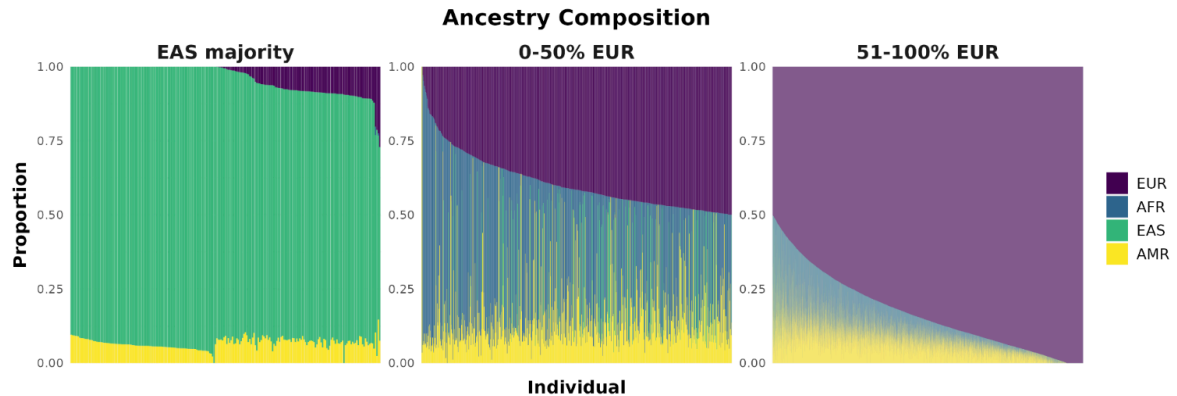
### 274 **PRS evaluation by ancestry composition**

275 Since our sample contains a great majority of EUR ancestry proportion, we decided to  
276 evaluate the PRS effect size in different ancestry compositions. We have created three  
277 groups: EAS majority (> 50% EAS, n = 217), 0 - 50% EUR (n = 763) and 51 - 100%  
278 EUR (n = 13,597). All three bins had statistically significant (p < 0.001) ORs above 1.40  
279 (1.54, 1.44 and 1.43, respectively) per PRS standard deviation, showing a positive  
280 association of the PRS value with increased BC risk. The EAS majority group shows a  
281 wider confidence interval due to the small sample size (cases = 64, controls = 153).  
282 Besides that, the lower tail of the 95% confidence interval has an OR of 1.14

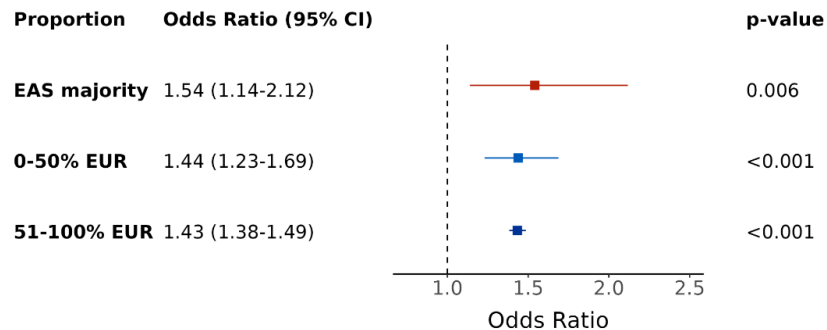
283 (Supplementary Table 5), which means at least 14% risk rise for each unit of  
 284 standardized PRS increase.

285

A.



B.



286

287

288 **Figure 4. Breast cancer Odds Ratio by ancestry proportion.** The cohort was split  
 289 into three groups based on main ancestry: EAS majority (>50% EAS), 0 - 50% EUR and  
 290 51 - 100% EUR (A) Ancestry composition of each group, with colors representing  
 291 continental ancestries for each subject. (B) Breast cancer ORs by PRS<sub>3820</sub> standard  
 292 deviation for the three groups. p-values displayed were corrected for  
 293 multiple-hypothesis testing using Bonferroni method.



294

### 295 **Comparison of PRS derived from genomes imputed from exomes with WGS**

296 A correlation of 0.76 (p value < 2.2e-16) was obtained between BC PRS<sub>3820</sub> values  
297 calculated from imputed genomes and WGS, showing a consistent concordance  
298 between both methods (**Supplementary Figure 1A**).

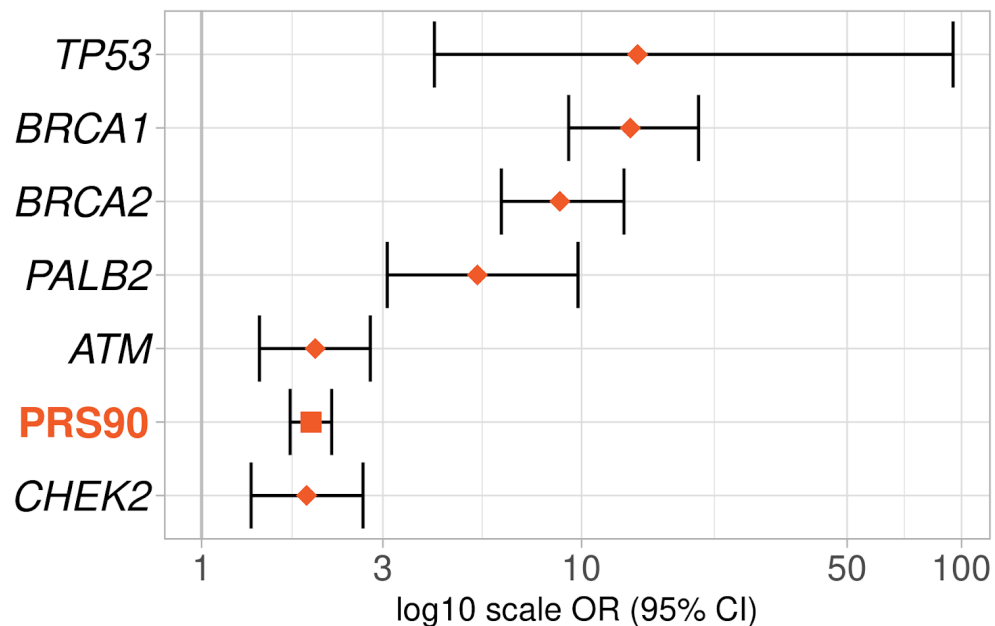
299

300 When we compared imputed (exome) and sequenced genomes (WGS), most of the  
301 extreme PRS<sub>3820</sub> values were concordant (decile 1: 56%; decile 10: 60%)  
302 (**Supplementary Figure 1B**). Furthermore, most of the proportion which is not in the  
303 same decile is in the surrounding deciles, which indicates a low deviation from the  
304 purpose of predicting risk.

305

### 306 **Comparison of PRS and breast cancer genes effect size**

307 For the purpose of understanding how the PRS<sub>3820</sub> effect size compare to known high  
308 and moderate risk genes for BC, we have compared OR of the top PRS<sub>3820</sub> decile  
309 (PRS90) with all pathogenic variants located in *TP53*, *BRCA1*, *BRCA2*, *PALB2*, *ATM*  
310 and *CHEK2* genes (**Figure 5**) in this cohort of individuals.



311

312 **Figure 5. Effect sizes of 90th percentile of PRS and BC genes in BC risk.** Effect  
313 sizes (OR and CI) were obtained according to the presence of pathogenic variants in  
314 the genes *TP53*, *BRCA1*, *BRCA2*, *PALB2*, *ATM* and *CHEK2*, or belonging to the 90th to  
315 100th percentiles of PRS<sub>3820</sub>.

316

317 As expected, *TP53*, *BRCA1* and *BRCA2* present the most extreme BC risks (OR: 14.05,  
318 CI: 4.1-95.05; OR: 13.43, CI: 9.25-20.32; and OR: 8.77, CI: 6.15-12.93, respectively).  
319 PRS90 risk (OR: 1.94, CI:1.71-2.2) can be compared with moderate risk BC genes *ATM*  
320 (OR: 1.99, CI: 1.42-2.78) and *CHEK2* (OR: 1.89, CI: 1.35-2.66). This result indicates  
321 how an increased risk for BC due to PRS90 could be interpreted in the clinical context,  
322 potentially following the same care protocols as for a moderate risk monogenic variant  
323 for BC.

324

## 325 Discussion

326 In the present study we have validated a breast cancer PRS developed from Europeans  
327 in the highly admixed Brazilian population. The PRS adapted from Mavaddat *et al.* study  
328 with 2,892 variants<sup>8</sup> showed a statistically significant risk prediction value (OR: 1.41 per  
329 SD). Furthermore, individuals classified in the top decile had an expressive effect size  
330 (OR: 1.94; CI: 1.71 - 2.20) of almost one-fold increased risk of BC compared to the  
331 middle percentiles (40-60%). This PRS highest decile risk is comparable with the  
332 previously reported risks for moderate-penetrance monogenic variants in *ATM*, *NF1*,  
333 and *CHEK2* genes (1.82, 1.93, and 2.47 OR, respectively)<sup>25</sup>, and also with risks in *ATM*  
334 and *CHEK2* calculated in our sample (1.99 and 1.89 OR, respectively).

335

336 This study is based on a previous study from Mavaddat *et al.* 2019, which developed  
337 and validated a PRS with 3,820 variants evaluating aggressive BC risk (metastatic BC).  
338 For all BC subtypes (ER+ and ER-) they found an OR of 1.71 per SD (CI: 1.64 - 1.79) in  
339 the validation set (n = 29,751; cases = 11,428), and OR 1.66 per SD (CI: 1.61 - 1.70) in  
340 the prospective set (n = 190,040; cases = 3,215). These values are even greater  
341 compared to the widely used 313 PRS (OR: 1.65 per SD; CI: 1.59 - 1.72 in validation  
342 set). However, they included a *CHEK2* gene pathogenic variant in the PRS and worked  
343 with only aggressive BC, which may have led to overestimating their OR values. A study  
344 from Liu and colleagues has evaluated another modification of the same PRS with  
345 3,820 variants developed from Mavaddat *et al.* for African, Latin, and European

346 populations<sup>26</sup>. According to the study, the effect size of this PRS to a BC risk in an  
347 European sample (n = 33,594) was 1.40 per standard deviation, a result very similar to  
348 ours for a Brazilian sample (OR 1.41 per SD; n = 14,477). They deliberately have  
349 included women with *in situ* ductal BC as well as women with metastatic BC, what they  
350 claim to be a reason for OR decline compared to the original study, which included only  
351 metastatic BC women both in their discovery and validation sets. Our study, however,  
352 does not distinguish BC types, therefore we hypothesize that both metastatic and *in situ*  
353 BC are included, which may be a factor, together with genetic population structure, that  
354 decreased the OR compared to the original study.

355

356 Furthermore, significant effect sizes per PRS standard deviation were obtained for  
357 distinct ancestry compositions within our sample. Due to the high proportion of EUR  
358 (median 84%, std. dev. 18%), we separated the sample into groups with different  
359 ancestry compositions. Despite the small sample size (n = 217) of the EAS majority  
360 group (**Supplementary Table 5**), there was a statistical significance (adjusted p-value:  
361 0.006) for the effect size in this group, which had similar magnitude (OR: 1.54, CI:  
362 1.40-2.12) of the full sample (OR: 1.41, CI: 1.36-1.47, p-value: < 0.0001). Also, PRS<sub>3820</sub>  
363 had significant and expressive effect sizes on BC risk for both EUR proportion groups  
364 (0-50% EUR OR: 1.44, CI: 1.23-1.69; 51-100% EUR OR: 1.43, CI: 1.38-1.49). These  
365 results evidence that, for individuals with a more prominent East Asian ancestry, for  
366 admixed individuals, and for predominantly Europeans, PRS<sub>3820</sub> is still effective in  
367 stratifying BC risk.

368

369 All of our PRS values were calculated according to a new methodology: the imputation  
370 of genomes from exomes. This approach has demonstrated to be very successful for  
371 PRS calculation and assessment of BC risk in our study, and could be very interesting  
372 for laboratories that already perform exome sequencing as a cost-effective methodology  
373 to identify P/LP variants for BC. A variety of studies have compared low-pass genome  
374 sequencing with arrays for different applications, such as pharmacogenetics, GWAS,  
375 CNV detection, and PRS calculation<sup>27,28,29</sup>. The study of Li *et al.*<sup>28</sup> reported improved  
376 accuracy for polygenic risk prediction of imputed low-pass genome compared to array  
377 imputation for both coronary artery disease and BC. Despite the slight difference we  
378 found between PRS values calculated from sequenced genomes and imputed genomes  
379 from exomes (Spearman correlation: 0.76), decile classification showed satisfactory  
380 concordance between both methods for the majority of results in the extreme deciles (1  
381 and 10th), which are the most important to define decreased or increased risk.  
382 Unfortunately, it was not possible to assess the predictive power of PRS values  
383 calculated from genomes of BC patients due to unavailability of paired exome and  
384 genome data.

385

386 Among familial BC cases, approximately 25% have a P/LP germline variant reported<sup>30</sup>.  
387 In the Brazilian population, a robust study with 1,663 breast cancer patients detected  
388 20.1% of P/LP germline variants using multigene panel testing<sup>4,6</sup>. A 2017 study reported  
389 that 18% of the hereditary BC can be explained by a polygenic effect of variants  
390 discovered in a GWAS<sup>31</sup>. Therefore, employing this PRS in the clinical practice might  
391 bring an elucidation to BC Brazilian families without high or moderate-effect germline

392 variants detected. Moreover, women without prior knowledge of their familial BC  
393 condition, or even those with a high PRS risk by chance, will have the possibility to be  
394 informed of their results and share them with their physicians to adopt preventive  
395 actions accordingly to their risk strata, such as intensifying surveillance adding breast  
396 magnetic resonance imaging to mammography screening<sup>32</sup>.

397

398 In conclusion, our work was able to validate a PRS developed in Europeans in the  
399 Brazilian population, using imputed genomes from exomes. The top decile of this PRS  
400 presents a risk comparable to moderate-risk monogenic variants for BC. Future studies  
401 will be required to evaluate the combination of PRS with P/LP variants and clinical  
402 factors in order to deliver more informative results to patients, thus physicians can  
403 recommend prevention strategies based on their combined polygenic and monogenic  
404 BC risk.

405

## 406 **Ethics Statement**

407 This work was approved by the Ethics Committee Comissão para análise de projeto de  
408 pesquisa of Hospital das Clínicas da FMUSP - CAPPesq under the CAAE number  
409 70112423.3.0000.0068.

410

## 411 **Acknowledgements**

412 We thank all individuals once sequenced in Mendelics laboratory who have consented  
413 to participate in this research. We also thank all UKBB participants for their contribution  
414 to the PRS hereby analyzed, and all authors from previous studies on BC PRSs in  
415 which we based our validation (Khera *et al.* 2018 and Mavaddat *et al.* 2019). Maria  
416 Aparecida Azevedo Koike Folgueira received research support from Conselho Nacional  
417 de Desenvolvimento Científico e Tecnológico, Brazil (CNPq-308052/2022-6).

418

## 419 **Data Availability**

420 All variants and betas which compose the four evaluated PRSs are available as  
421 Supplementary Information. Individual cases and controls data are not publicly available  
422 due to the confidentiality consentment agreement signed by all included in the study.

423

## 424 **Competing Interests**

425 Flávia Eichemberger Rius, Danilo Viana, Júlia Salomão, Laila Gallo, Renata Freitas,  
426 Cláudia Bertolacini, Lucas Taniguti, Danilo Imparato, Flávia Antunes, Gabriel Sousa,  
427 Renan Achjian, Eric Fukuyama, Cleandra Gregório, Iuri Ventura, Juliana Gomes,  
428 Nathália Taniguti, and David Schlesinger are currently employed by Mendelics, or were  
429 employed at the time of the study.

430 Rodrigo Guindalini acted as a consultant for AstraZeneca, Janssen Oncology,  
431 Roche/Genentech and Igenomix; received speaker honoraria from AstraZeneca, Bristol  
432 Myers Squibb, GlaxoSmithKline, Merck Sharpe & Dohme Brasil, Novartis, and Roche  
433 outside the submitted work; and has equity in Mendelics Análise Genômica.

434 Olufunmilayo I. Olopade is co-founder at CancerIQ; serves as scientific advisor at  
435 Tempus; and has received research funding from Color Genomics and  
436 Roche/Genentech.

437 José Eduardo Krieger, Yonglan Zheng, Dezheng Huo, Simone Maistro and Maria  
438 Aparecida Koike declare no competing interests.

#### 439 **Author Contributions**

440 Generated Main Data: Flávia Eichemberger Rius, Danilo Viana, Júlia Salomão, Laila  
441 Gallo, Renata Freitas, Cláudia Bertolacini, Lucas Taniguti, Danilo Imparato, Flávia  
442 Antunes, Gabriel Sousa, Renan Achjian, Eric Fukuyama, David Schlesinger.

443 Analyzed Data: Flávia Eichemberger Rius, Rodrigo Guindalini, Danilo Viana, Lucas  
444 Taniguti, Danilo Imparato, Flávia Antunes, Gabriel Sousa, Renan Achjian, Eric  
445 Fukuyama, Yonglan Zheng, Dezheng Huo, Olufunmilayo I. Olopade, Maria Aparecida  
446 Koike, David Schlesinger.

447 Other Contributions: Cleandra Gregório, Iuri Ventura, Juliana Gomes, Nathália Taniguti,  
448 Simone Maistro, José Eduardo Krieger.

449

#### 450 **References**

451 1. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence



- 452 and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**,  
453 209–249 (2021).
- 454 2. Instituto Nacional de Câncer. *Estimativa 2023 : incidência de câncer no Brasil*.  
455 (Ministério da Saúde, 2023).
- 456 3. Nielsen, F. C., van Overeem Hansen, T. & Sørensen, C. S. Hereditary breast and  
457 ovarian cancer: new genes in confined pathways. *Nat. Rev. Cancer* **16**, 599–612  
458 (2016).
- 459 4. Guindalini, R. S. C. *et al.* Detection of germline variants in Brazilian breast cancer  
460 patients using multigene panel testing. *Sci. Rep.* **12**, 4190 (2022).
- 461 5. Shiovitz, S. & Korde, L. A. Genetics of breast cancer: a topic in evolution. *Ann.*  
462 *Oncol.* **26**, 1291–1299 (2015).
- 463 6. Melchor, L. & Benítez, J. The complex genetic landscape of familial breast cancer.  
464 *Hum. Genet.* **132**, 845–863 (2013).
- 465 7. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify  
466 individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224  
467 (2018).
- 468 8. Mavaddat, N. *et al.* Polygenic risk scores for prediction of breast cancer and breast  
469 cancer subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).
- 470 9. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer  
471 susceptibility loci from overall and subtype-specific analyses. *Nat. Genet.* **52**,  
472 572–581 (2020).
- 473 10. Morra, A. *et al.* Association of germline genetic variants with breast cancer-specific  
474 survival in patient subgroups defined by clinic-pathological variables related to

- 475 tumor biology and type of systemic treatment. *Breast Cancer Res.* **23**, 86 (2021).
- 476 11. Mars, N. *et al.* Genome-wide risk prediction of common diseases across ancestries  
477 in one million people. *Cell Genomics* **2**, None (2022).
- 478 12. Salzano, Freire-Maia, F. M. N. As origens. in *Populações Brasileiras: Aspectos*  
479 *Demográficos, Genéticos e Antropológicos* (1967).
- 480 13. Souza, A. M. de, Resende, S. S., Sousa, T. N. de & Brito, C. F. A. de. A systematic  
481 scoping review of the genetic ancestry of the Brazilian population. *Genet. Mol. Biol.*  
482 **42**, 495–508 (2019).
- 483 14. Naslavsky, M. S. *et al.* Whole-genome sequencing of 1,171 elderly admixed  
484 individuals from São Paulo, Brazil. *Nat. Commun.* **13**, 1004 (2022).
- 485 15. Sudlow, C. *et al.* UK Biobank: an open access resource for identifying the causes of  
486 a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779  
487 (2015).
- 488 16. Pedersen, B. S. *et al.* Somalier: rapid relatedness estimation for cancer and  
489 germline studies using efficient genome sketches. *Genome Med.* **12**, 62 (2020).
- 490 17. 1000 Genomes Project Consortium *et al.* A global reference for human genetic  
491 variation. *Nature* **526**, 68–74 (2015).
- 492 18. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing  
493 and imputation of low-coverage sequencing data using large reference panels. *Nat.*  
494 *Genet.* **53**, 120–126 (2021).
- 495 19. Bergström, A. *et al.* Insights into human genetic variation and population history  
496 from 929 diverse genomes. *Science* **367**, (2020).
- 497 20. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and

- 498 richer datasets. *Gigascience* **4**, 7 (2015).
- 499 21. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of  
500 ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
- 501 22. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.*  
502 **28**, (2008).
- 503 23. R Foundation for Statistical Computing. *R: A Language and Environment for*  
504 *Statistical Computing*. (<https://www.R-project.org/>, 2023).
- 505 24. Aragon, T. J., Fay, M. P., Wollschlaeger, D. & Omidpanah, A. *epitools:*  
506 *Epidemiology Tools. Tools for training and practicing epidemiologists including*  
507 *methods for two-way and multi-way contingency tables*. (CRAN, 2020).
- 508 25. Hu, C. *et al.* A Population-Based Study of Genes Previously Implicated in Breast  
509 Cancer. *N. Engl. J. Med.* **384**, 440–451 (2021).
- 510 26. Liu, C. *et al.* Generalizability of polygenic risk scores for breast cancer among  
511 women with european, african, and latinx ancestry. *JAMA Netw. Open* **4**, e2119084  
512 (2021).
- 513 27. Wasik, K. *et al.* Comparing low-pass sequencing and genotyping for trait mapping  
514 in pharmacogenetics. *BMC Genomics* **22**, 197 (2021).
- 515 28. Li, J. H., Mazur, C. A., Berisa, T. & Pickrell, J. K. Low-pass sequencing increases  
516 the power of GWAS and decreases measurement error of polygenic risk scores  
517 compared to genotyping arrays. *Genome Res.* **31**, 529–537 (2021).
- 518 29. Chaubey, A. *et al.* Low-Pass Genome Sequencing: Validation and Diagnostic Utility  
519 from 409 Clinical Cases of Low-Pass Genome Sequencing for the Detection of  
520 Copy Number Variants to Replace Constitutional Microarray. *J. Mol. Diagn.* **22**,

521 823–840 (2020).

522 30. Bahcall, O. Common variation and heritability estimates for breast, ovarian and  
523 prostate cancers. *Nat. Genet.* (2019) doi:10.1038/ngicogs.1.

524 31. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci.  
525 *Nature* **551**, 92–94 (2017).

526 32. Monticciolo, D. L., Newell, M. S., Moy, L., Lee, C. S. & Destounis, S. V. Breast  
527 Cancer Screening for Women at Higher-Than-Average Risk: Updated  
528 Recommendations From the ACR. *J. Am. Coll. Radiol.* (2023)  
529 doi:10.1016/j.jacr.2023.04.002.