#### 1 A genome-wide association study of mammographic texture variation

2 Yuxi Liu<sup>1,2</sup>, Hongjie Chen<sup>3</sup>, John Heine<sup>4</sup>, Sara Lindstroem<sup>3,5</sup>, Constance Turman<sup>1</sup>, Erica T.

Warner<sup>6</sup>, Stacey J. Winham<sup>7</sup>, Celine M. Vachon<sup>8</sup>, Rulla M. Tamimi<sup>9,10</sup>, Peter Kraft<sup>1,2,11,\*</sup>, Xia
Jiang<sup>12,13,\*</sup>

- 5 1. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA,
- 6 USA
- 7 2. Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of
- 8 Public Health, Boston, MA, USA
- 9 3. Department of Epidemiology, University of Washington, Seattle, WA, USA
- 10 4. Division of Population Sciences, H. Lee Moffitt Cancer Center & Research Institute, Tampa,
- 11 FL, USA
- 12 5. Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA,
- 13 USA
- 14 6. Clinical and Translational Epidemiology Unit, Department of Medicine, Mongan Institute,
- 15 Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
- 16 7. Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA
- 17 8. Division of Epidemiology, Department of Quantitative Health Sciences, Mayo Clinic,
- 18 Rochester, MN, USA
- 19 9. Channing Division of Network Medicine, Department of Medicine, Brigham and Women's
- 20 Hospital and Harvard Medical School, Boston, MA, USA
- 21 10. Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA
- 22 11. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

- 23 12. Department of Clinical Neuroscience, Centre for Molecular Medicine, Karolinska Institute,
- 24 Stockholm, Sweden
- 25 13. West China School of Public Health and West China Fourth Hospital, Sichuan University,
- 26 Chengdu, China
- 27 \* Corresponding authors:
- 28 Peter Kraft, PhD. Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan
- 29 School of Public Health, 655 Huntington Avenue, Building 2-249A, Boston, MA, 02115. Email:
- 30 <u>pkraft@hsph.harvard.edu</u>
- 31 Xia Jiang, PhD. Department of Clinical Neuroscience, Center for Molecular Medicine,
- 32 Karolinska Institutet, Visionsgatan 18, Solna 171 77, Stockholm, Sweden. E-mail:
- 33 <u>xia.jiang@ki.se</u>
- 34
  35
  36
  37
  38
  39
  40
  41

# 42 Abstract

Background: Breast parenchymal texture features, including gray scale variation (V), capture
the patterns of texture variation on a mammogram and are associated with breast cancer risk,
independent of mammographic density (MD). However, our knowledge on the genetic basis of
these texture features is limited.

47 **Methods:** We conducted a genome-wide association study of V in 7,040 European-ancestry 48 women. Four V assessments representing different amounts of breast edge erosion and image 49 resolutions were generated from digitized film mammograms. We used linear regression to test 50 the single-nucleotide polymorphism (SNP)-phenotype associations adjusting for age, body mass 51 index (BMI), MD phenotypes, and the top four genetic principal components. Multivariate 52 phenotype association tests combining all four V assessments were performed. We further 53 calculated genetic correlations and performed SNP-set tests of V with MD, breast cancer risk, and other breast cancer risk factors. 54

55 Results: We identified three genome-wide significant loci associated with V: rs138141444 56 (6q24.1) in ECT2L, rs79670367 (8q24.22) in LINC01591, and rs113174754 (12q22) near 57 PGAM1P5. 6q24.1 and 8q24.22 have not previously been associated with MD phenotypes or 58 breast cancer risk, whilst 12q22 is a known locus for both MD and breast cancer risk. Among 59 known MD and breast cancer risk SNPs, we identified four variants that were associated with V 60 at the Bonferroni-corrected thresholds accounting for the number of SNPs tested: rs335189 61 (5q23.2) in *PRDM6*, rs13256025 (8p21.2) in *EBF2*, rs11836164 (12p12.1) near SSPN, and 62 rs17817449 (16q12.2) in FTO. We observed significant genetic correlations between V and mammographic dense area ( $r_g = 0.79$ ,  $P = 5.91 \times 10^{-5}$ ), percent density ( $r_g = 0.73$ ,  $P = 1.00 \times 10^{-5}$ ) 63

64	$10^{-4}$ ), and adult BMI ( $r_g = -0.36$ , $P = 3.88 \times 10^{-7}$ ). Additional significant relationships were
65	observed for nondense area (z = $-4.14$ , $P = 3.42 \times 10^{-5}$ ), estrogen receptor-positive breast cancer
66	$(z = 3.41, P = 6.41 \times 10^{-4})$ , and childhood body fatness $(z = -4.91, P = 9.05 \times 10^{-7})$ from the
67	SNP-set tests.
68	Conclusions: These findings provide new insights into the genetic basis of mammographic
69	texture variation and their associations with MD, breast cancer risk, and other breast cancer risk
70	factors.
71	
72	Keywords: Breast cancer, Breast parenchymal texture feature, Texture variation, V measure,
73	Mammographic density, GWAS, Genetic correlation
74	
75	
76	
77	
78	
79	
80	
81	

# 82 Background

83 Mammographic density (MD) phenotypes reflect the amount of dense or nondense tissue on a 84 mammogram and are well-established risk factors for breast cancer [1-3]. MD phenotypes are highly heritable with  $h^2 = 60-70\%$  from twin studies [4, 5]. Genome-wide association studies 85 86 (GWAS) have identified 55 loci that are associated with MD phenotypes [6-8], including 32 loci 87 for dense area (DA), which reflects the amount of fibroglandular tissue in the breast, 18 loci for 88 nondense area (NDA), which reflects the amount of fatty tissue in the breast, and 24 loci for 89 percent density (PD), defined as the percentage of area on a mammogram that is occupied by 90 dense tissue [9].

91 Yet, MD is a global metric that ignores local patterns of variability in breast density [10]. 92 Women with the same level of PD may have substantial heterogeneity in the structural patterns 93 of breast parenchyma, which are assessed as texture features. Compared to MD phenotypes, 94 breast parenchymal texture features are more refined and localized, and are fully automated 95 measures of the variation in parenchymal patterns on a mammogram [11]. Growing evidence 96 suggest that texture features are independent breast cancer risk factors [12-16]. Heine et al. 97 developed a summary measure of texture features called V, which captures the gray scale 98 variation on a mammogram [12]. Recent studies have shown that a higher value of V, reflecting 99 greater texture variation, is associated with an increased risk of breast cancer, independent of 100 MD [12, 16]. Understanding the mechanisms underlying texture variation and breast cancer risk, 101 especially the role of genetic variants, would provide additional insights into the development of 102 breast cancer. However, to date, no GWAS has been conducted on breast parenchymal texture 103 features.

104	In the present study, we performed a GWAS of mammographic texture variation, including four
105	different assessments of V, within the Nurses' Health Studies and Mayo Mammography Health
106	Study cohorts. We also leveraged summary statistics of breast cancer risk and MD phenotypes
107	from previous GWAS to identify shared susceptibility loci for V, MD, and breast cancer risk. We
108	further assessed the genetic relationships of V with MD phenotypes, breast cancer risk, and other
109	breast cancer risk factors by estimating genetic correlations and performing single-nucleotide
110	polymorphism (SNP)-set tests.
111	
112	
113	
114	
115	
116	
117	
118	
119	
120	
121	

# 122 Methods

#### 123 Study population

124	The Nurses'	Health Study	(NHS	) is a p	ospective c	ohort study	y established in	1976. A total
-----	-------------	--------------	------	----------	-------------	-------------	------------------	---------------

- number of 121,700 female registered nurses aged 30 to 55 residing in 11 states within the United
- 126 States completed an initial questionnaire at that time. NHSII was established in 1989 when
- 127 116,671 female registered nurses aged 25 to 42 residing in 14 states completed an initial
- 128 questionnaire. Blood samples were collected from 32,826 women in NHS cohort from 1989 to
- 129 1990 and 29,611 women in NHSII cohort from 1996 to 1999, which form the blood subcohorts.
- 130 Women in each cohort have been followed by self-administered questionnaires to collect updated
- 131 exposure and newly diagnosed disease information every two years.

132 The Mayo Mammography Health Study (MMHS) is a prospective cohort study of 19,924 women

133 who had a screening mammogram from 2003 to 2006 at the Mayo Clinic in Rochester, MN and

agreed to participate in the study. Participants were at least 35 years old, residents of Minnesota,

135 Iowa, or Wisconsin, and had no personal history of breast cancer. Participants completed a

136 baseline questionnaire and provided consent to access any residual blood samples from clinical

137 tests over the time period. Breast cancer diagnostic information were obtained through linkage to

- 138 state-wide cancer registry data and mailed questionnaires.
- 139 Mammogram collection and processing

The mammogram collection and processing procedure in NHS and NHSII is described elsewhere
[16, 17] and is briefly summarized here. Pre-diagnostic screening mammograms were collected
within NHS and NHSII breast cancer case-control studies nested in the blood subcohorts [18]. A

143 total number of 6.258 film mammograms obtained close to the blood draw date were initially 144 collected. The study protocol was approved by the institutional review boards of the Brigham 145 and Women's Hospital and Harvard T.H. Chan School of Public Health. Film mammogram 146 craniocaudal views of both breasts were digitized using a Lumysis 85 laser film scanner or a 147 VIDAR CAD PRO Advantage scanner (VIDAR Systems Corporation, Herndon, VA, USA). 148 Digitized images were grouped based on resolution (mean resolution =  $171 \mu m$ ,  $232 \mu m$ ,  $300 \mu m$ , 149 and images with isolated resolutions). Here, we evaluated the groups of images with average resolutions of 171µm (high resolution) and 300µm (low resolution). Images with isolated 150 151 resolutions were down-sampled to 300µm and added to the low resolution group. All 171µm 152 images were further adjusted to 300µm to form a larger dataset of low-resolution images. 153 Details of mammogram acquisition, retrieval, and digitization for MMHS are described 154 elsewhere [12, 19]. Briefly, women in MMHS who agreed to participate provided written 155 informed consent to access their mammograms. A total number of 19,924 women were followed 156 up for incident cancer events. We used a case-cohort design with a random sample of 2,300 157 women from the entire MMHS cohort as the subcohort. We collected film mammograms from 158 1,194 breast cancer cases identified through August 2019, excluding women who were 159 diagnosed within 60 days of the enrollment mammogram and women with a digital mammogram. 160 We further collected mammograms from 2,167 control women from the subcohort. The study 161 protocol was approved by the Mayo Clinic institutional review board. Film mammograms of 162 both craniocaudal views were digitized on the Array 2905 laser digitizer (Array Corporation, 163 Roden, The Netherlands) with 50µm (limiting) pixel spacing and further downsampled to 200µm. 164 Both the original 50µm images (high resolution) and the down-sampled 200µm images (low 165 resolution) were used for calculation of V.

#### 166 Assessment of V

167 V is an automated measure of the gray scale variation on a mammogram. The algorithm for 168 generating V has been described previously by Heine et al. [12, 20, 21]. Briefly, there are three 169 main steps: segmentation, erosion, and calculation of variation. First, the breast is segmented 170 from the background. Then, the segmented breast area is eroded by 25% or 35% along a radial 171 direction to retain the regions where the breast was in contact with the compression paddle. 172 Finally, the V is calculated as the standard deviation of the pixel values within the eroded breast 173 region. Normalization processes, including spatial normalization, feature distribution 174 normalization, and resolution estimation, were applied to the images before calculation of V to 175 account for resolution and intensity scale differences [17]. 176 We generated four assessments of V with different proportions of erosion and image resolutions: 177 V with 35% erosion and low resolution (V65L), V with 25% erosion and low resolution (V75L), 178 V with 35% erosion and high resolution (V65H), and V with 25% erosion and high resolution 179 (V75H). These four V assessments were highly correlated with each other (Fig. 1a). We used 180 V65L as our primary univariate outcome, as it had the largest sample size.

# 181 MD phenotypes and other covariates

MD phenotypes were assessed from digitized film mammograms using Cumulus [22], a semiautomated software, by a single trained reader [12, 23]. DA and NDA were generated for each mammogram; PD was calculated as DA divided by the total breast area. DA, NDA, and PD measures in the left and right breasts were averaged. Fig. 1b shows the scatter plots and correlations of V65L and the three MD phenotypes. Body mass index (BMI) was measured at mammogram collection for all participants. Women were considered as breast cancer cases if

they were diagnosed with breast cancer after blood or mammogram collection but before June 1,
2004 (NHS), June 1, 2007 (NHSII), or August 2019 (MMHS). Age at mammogram collection
was also retrieved.

#### 191 Genotyping, quality control, and imputation

192 The full genotyping and quality control pipeline for NHS and NHSII is described elsewhere [24].

193 In the present study, we used genotype data from four platforms:  $A \square$  ymetrix 6.0, Illumina

194 HumanHap, Illumina OmniExpress, and Illumina OncoArray. Variants with call rate < 95% or

Hardy-Weinberg equilibrium  $P < 1 \times 10^{-6}$  were excluded. European ancestry principal

196 component (PC) outliers or samples with call rate < 90%, gender discordance, or extreme

197 heterozygosity were excluded.

198 The full genotyping and quality control pipeline for MMHS is also described elsewhere [25].

199 Here, we used genotype data from iCOGS and OncoArray platforms. Variants with a call rate <

200 95% or not in Hardy-Weinberg equilibrium were excluded. Samples with a call rate < 95%,

201 extreme heterozygosity, or of non-European ancestry based on genetic PCs were further

excluded.

All genotype data were imputed to the 1000 Genomes Phase 3 version 5 reference panel

separately by study and platform [26]. Number of individuals included in our GWAS by study

and platform can be found in Additional file 1: Table S1.

#### 206 Association test

All four V assessments (V65L, V75L, V65H, and V75H) were standardized to have mean zero
and unit standard deviation before analysis. SNP association analyses were performed within

each study by platform for each of the four V assessments using linear regression assuming an 209 210 additive dosage effect. RVtests [27] was used for NHS/NHSII cohorts (A vmetrix 6.0, Illumina 211 HumanHap, Illumina OmniExpress, and Illumina OncoArray) and variants were removed from 212 individual platform results if the expected minor allele counts were below 10. PLINK 2.0 [28] 213 was used for MMHS cohorts (iCOGS and OncoArray). We ran six models adjusting for different 214 covariates: Model 0 was the base model adjusting for age and the top four genetic PCs to account 215 for population structure. Model 1 further adjusted for BMI. In addition to the covariates in Model 216 1, Model 2 further adjusted for PD, Model 3 further adjusted for DA, and Model 4 further 217 adjusted for NDA. Model 5 was the fully adjusted model with age, BMI, genetic PCs, PD, DA, 218 and NDA as covariates. Fixed effect meta-analyses across studies and platforms were conducted 219 for each V assessment and model using METAL [29]. Cochran's Q statistic was used to check 220 for heterogeneity between studies and platforms. Quantile-quantile plots and genomic inflation 221 factors were used to assess systematic inflation in test statistics due to population substructure. 222 Manhattan plots were generated to visualize the overall GWAS results. LocusZoom plots [30] of 223 the 1Mb region centered around the identified lead SNPs were generated to visualize the regional 224 association results and nearby genes.

Given that the four V assessments were highly correlated with each other and might be proxies for an underlying latent phenotype, we performed multivariate phenotype association tests to pool association evidence across the four V assessments and get a single summary test statistic for each variant. We used R package MPAT [31] to obtain the summary *P* values and corresponding Z scores using test statistics from the meta-analysis results for each V assessment and model, accounting for sample overlaps of the four V assessments. We referred to this summary phenotype as VSUM, which was used as our primary multivariate outcome. SNPs with

- 232  $P < 5 \times 10^{-8}$  in any of the six models for any of the four univariate V assessments or the
- 233 multivariate VSUM were considered genome-wide significant.

#### 234 V, MD phenotypes, and breast cancer susceptibility variants

- 235 We evaluated whether the identified V loci were also associated with MD phenotypes or breast
- cancer risk using GWAS results from Breast Cancer Association Consortium (BCAC) [8, 25, 32].
- 237 To further identify shared susceptibility SNPs between V, MD phenotypes, and breast cancer
- risk, we conducted *in silico* lookups of 72 genome-wide significant MD phenotype SNPs
- identified by Sieh et al. [6] and Chen et al. [8], and 195 genome-wide significant breast cancer
- risk SNPs identified by Michailidou et al. [25] and Zhang et al. [32] in our GWAS of V. These
- 241 candidate SNPs were considered significant for V if they passed the Bonferroni-corrected
- thresholds accounting for the number of MD (P < 0.05/72) or breast cancer (P < 0.05/195) SNPs
- tested in Model 0 for any V assessment.

# 244 Genetic correlation and SNP-set test

245 Genetic correlations of the four V assessments and VSUM with MD phenotypes, breast cancer

risk, overall and stratified by estrogen receptor (ER) status, adult BMI, childhood body fatness,

- age at menarche, and age at natural menopause were estimated using linkage disequilibrium (LD)
- score regression [33, 34]. Sources of summary statistics of these traits for estimating genetic
- correlations are summarized in Additional file 2: Table S2.

While genetic correlation quantifies the shared genetic contribution to two traits on genome-wide scale, it may also capture the contribution of other traits due to pleiotropy (e.g., the effect of BMI on the correlation between V and PD). Therefore, we further performed SNP-set tests to assess

the genetic relationship between V and the above-mentioned traits using only reported genome-

254 wide significant SNPs for those traits. SNPs for each trait were collected from published GWAS

followed by LD clumping to remove any SNPs in LD ( $r^2 > 0.1$ ) with SNPs of smaller *P* value

256 (see Additional file 2: Table S3). The test statistic for V and each trait was

$$Z = \frac{\sum_{n} \operatorname{sgn}(Z_{\mathrm{V}} \cdot Z_{\mathrm{Trait}}) \cdot |Z_{\mathrm{V}}|}{\sqrt{n}}$$

where  $Z_V$  is the Z score from the SNP-specific association with V and  $Z_{Trait}$  is the Z score from the SNP-specific association with the trait of interest, and *n* is the total number of tested genomewide significant SNPs for that trait.

### 260 Sensitivity analysis

261 Our study population contains both women who developed breast cancer and women who did 262 not develop breast cancer during the follow-up period after mammogram collection. We 263 therefore further adjusted for breast cancer case-control status in Model 5 to assess its impact on 264 the genetic associations. We performed a multicollinearity check for the identified genome-wide 265 significant SNPs for Model 5, where we adjusted for all three MD phenotypes, by calculating the 266 variance inflation factor (VIF). To assess the potential impact of outliers on the association 267 results at the identified GWAS loci, we calculated the studentized residuals for all samples for 268 each genome-wide significant SNP. Samples with absolute studentized residual greater than 3 269 were considered as outliers.

#### 271 Results

272 Our GWAS meta-analysis of V comprised 7,040 women of European ancestry within the NHS, 273 NHSII, and MMHS cohorts (Table 1). Women in MMHS were older, had higher BMI and lower 274 MD compared to women in NHS and NHSII. Quantile-quantile plots and genomic inflation 275 factors indicate there was no evidence of systematic inflation of the GWAS test statistics in any 276 model for any V assessment (Additional file 2: Figure S1). Manhattan plots showing the 277  $-\log_{10}(P)$  for all tested SNPs across chromosomes are present in Additional file 2: Figure S2. 278 Quantile-quantile plots of the heterogeneity P value indicate there was limited evidence of 279 heterogeneity in the test results across studies and platforms (Additional file 2: Figure S3). 280 In total, we identified three independent loci that reached the genome-wide significant threshold of  $P < 5 \times 10^{-8}$  in any model for any V assessment: 6q24.1 (*ECT2L*), 8q24.22 (*LINC01591*), and 281 12q22 (*PGAM1P5*) (Table 2). 6q24.1 (Lead SNP: rs138141444,  $P = 1.24 \times 10^{-8}$  for V75H, 282 283 Model 0) is a novel locus that has not previously been associated with MD phenotypes or breast 284 cancer risk. Fig. 2a shows the regional association results for 6q24.1 from Model 0 (adjusting for 285 age and genetic PCs) for V75H where the association was genome-wide significant. The 286 association results were consistent across models with the same direction and similar effect sizes as well as *P* values. 8q24.22 (Lead SNP: rs79670367,  $P = 2.38 \times 10^{-8}$  for VSUM, Model 5) is 287 288 neither a MD nor breast cancer risk locus. Fig. 2b shows the regional association results for 289 8q24.22 from Model 5 (adjusting for age, BMI, DA, NDA, PD, and genetic PCs) for VSUM. 290 The association between V and rs79670367 was more significant when we adjusted for PD 291 (Model 2), DA (Model 3), or both (Model 5) and became less significant without adjustment for 292 any MD phenotypes (Model 0 and 1) or adjusting for NDA only (Model 4). The direction of association was consistent across models. 12q22 (Lead SNP: rs113174754,  $P = 4.42 \times 10^{-8}$  for 293

VSUM, Model 3) has previously been associated with NDA (rs11836367,  $P = 8.40 \times 10^{-9}$ ,  $r^2 =$ 

295 0.59 with rs113174754) [6], overall breast cancer risk (rs113174754,  $P = 1.08 \times 10^{-24}$ ), and ER+

breast cancer risk (rs113174754,  $P = 1.37 \times 10^{-18}$ ) [25]. This locus is also significantly

associated with breast size (rs17356907,  $P = 1.30 \times 10^{-13}$ ,  $r^2 = 0.47$  with rs113174754) [35]. Fig.

298 2c shows the regional association results for 12q22 from Model 3 (adjusting for age, BMI, DA,

and genetic PCs) for VSUM. The association between V and rs113174754 became non-

300 significant when we adjusted for NDA. The direction of association with V was consistent across

301 models and consistent with the association with NDA (opposite direction) and breast cancer risk

302 (same direction).

303 We identified four additional loci that had previously been associated with MD phenotypes or

304 breast cancer risk and reached the Bonferroni-corrected thresholds accounting for the number of

305 MD or breast cancer SNPs tested ( $P < 0.05/72 = 6.94 \times 10^{-4}$  for MD,  $P < 0.05/195 = 2.56 \times 10^{-4}$ 

306 for breast cancer risk) in Model 0: 5q23.2 (*PRDM6*), 8p21.2 (*EBF2*), 12p12.1 (*SSPN*), and

307 16q12.2 (*FTO*) (Table 2). 5q23.2 (Lead SNP: rs335189,  $P = 7.30 \times 10^{-5}$  for VSUM, Model 0) is

308 a known locus for DA ( $P = 2.84 \times 10^{-11}$ ) and PD ( $P = 5.78 \times 10^{-10}$ ) [8]. The associations with V

were significant in Model 0, Model 1 (adjusting for age, BMI, and genetic PCs), and Model 4

310 (adjusting for age, BMI, NDA, and genetic PCs) but not with adjustment for DA, PD, or both

311 (Model 2, 3, and 5). 8p21.2 (Lead SNP: rs13256025,  $P = 5.74 \times 10^{-5}$  for VSUM, Model 0) has

312 previously been associated with breast cancer risk ( $P = 1.40 \times 10^{-8}$ ) [32]. The associations with

V were significant in Model 0 and 1, and became non-significant when we adjusted for DA,

NDA, or PD (Model 2, 3, 4, and 5). Although this locus has not been reported as a MD locus, the

315 *P* value of the association between the lead SNP and PD was close to the genome-wide

316 significant threshold ( $P = 4.46 \times 10^{-7}$ ) [8]. 12p12.1 (Lead SNP: rs11836164,  $P = 6.69 \times 10^{-5}$  for

VSUM, Model 0) is a known locus for DA ( $P = 1.66 \times 10^{-9}$ ) [8]. The associations with V were 317 318 significant in Model 0, 1, and 4, and became non-significant when we adjusted for DA, PD, or both (Model 2, 3, and 5). 16q12.2 (Lead SNP: rs17817449,  $P = 1.12 \times 10^{-6}$  for VSUM, Model 0) 319 is a known locus for PD ( $P = 5.06 \times 10^{-9}$ ) [8], overall ( $P = 2.52 \times 10^{-21}$ ), ER+ ( $P = 5.59 \times 10^{-14}$ ), 320 and ER- breast cancer risk ( $P = 1.80 \times 10^{-10}$ ). This locus is also significantly associated with 321 BMI (rs17817449,  $P = 5.10 \times 10^{-19}$ ) [36] and breast size (rs62033406,  $P = 3.70 \times 10^{-7}$ ,  $r^2 = 0.89$ 322 323 with rs17817449) [35]. The associations with V were significant in Model 0 and became non-324 significant when we adjusted for BMI or any MD phenotype. The directions of association with 325 V were consistent with those significant associations with MD (same direction for PD and DA, 326 opposite direction for NDA) or breast cancer (same direction) for all four loci. Association 327 results of all identified V loci for all models and V assessments can be found in Additional file 1: 328 Table S4. There was no substantial difference between the results of different V assessments. 329 The full lookup results of the 72 MD phenotype SNPs and 195 breast cancer SNPs can be found 330 in Additional file 1: Table S5 and Table S6. We observed significant positive genetic correlations between V and dense area ( $r_g = 0.79$ , P =331  $5.91 \times 10^{-5}$  for VSUM, Model 0) and percent density ( $r_g = 0.73$ ,  $P = 1 \times 10^{-4}$  for VSUM, Model 332 333 0) (Fig. 3a). The correlations became non-significant using GWAS results from Model 2 334 (adjusting for age, BMI, PD, and genetic PCs). Positive correlations were also observed with overall ( $r_g = 0.20$ ,  $P = 6.90 \times 10^{-3}$  for VSUM, Model 0) and ER+ ( $r_g = 0.22$ ,  $P = 4.60 \times 10^{-3}$  for 335 336 VSUM, Model 0) breast cancer and became non-significant when adjusting for PD. We also observed a significant negative association with adult BMI ( $r_g = -0.36$ ,  $P = 3.88 \times 10^{-7}$  for 337 338 VSUM, Model 0), which became non-significant when adjusting for BMI. A strong negative correlation was observed for NDA ( $r_g = -0.60$ ,  $P = 5.20 \times 10^{-3}$  for VSUM, Model 0) before 339

adjusting for PD. Genetic correlation results were similar across V assessments; the full resultsare summarized in Additional file 1: Table S7.

342 In addition to the genetic relationships of V with DA, NDA, PD, and breast cancer risk identified 343 by genetic correlations, we further identified a significant positive association between V and ER+ breast cancer (z = 3.41,  $P = 6.41 \times 10^{-4}$  for VSUM, Model 0) and a significant negative 344 345 association between V and childhood body fatness from the SNP-set test using genome-wide significant SNPs for childhood body fatness (z = -4.91,  $P = 9.05 \times 10^{-7}$  for VSUM, Model 0) 346 347 (Fig. 3b). The overall pattern of the associations was similar for genetic correlation and SNP-set 348 test. It is worth noting that for MD phenotypes and childhood body fatness, the associations with 349 V remained nominally significant (P < 0.05) if we further adjust for BMI and PD in the SNP-set 350 test. Plots showing Z scores from GWAS of V and GWAS of MD phenotypes [8] and breast 351 cancer [25] for the SNPs included in the SNP-set tests are present in Fig. 4. SNP-set test results 352 across all models and V assessments can be found in Additional file 1: Table S8. 353 No substantial change on the top findings was observed after including breast cancer case-354 control status as a covariate (Additional file 1: Table S4). There was no multicollinearity issue 355 for the effect estimates of the genome-wide significant SNPs in Model 5 (VIFs all close to 1). 356 There were 10 outliers with absolute studentized residual greater than 3 for rs79670367 at 357 8q24.22 from Model 5 for V65L. The effect estimates for the effect allele increased by 24% after 358 removing those outliers. No substantial impact of outliers was found for other identified V SNPs. 359

# 361 Discussion

362 While MD continues to be one of the most well-established and widely used mammographic risk 363 factors for breast cancer, there are gaps in our knowledge of mammographic features themselves 364 and their relationship with breast cancer risk. Current MD measures do not capture the 365 heterogeneity in the distribution of dense breast tissue on a mammogram, known as texture 366 variation. Increasing evidence have shown that the performance of texture variation on 367 discriminating breast cancer outcomes is either comparable or even higher than the performance of MD measures [12, 16, 37, 38]. Understanding the contributing mechanisms of texture 368 369 variation on breast cancer risk, especially the involved genetic components, would expand our 370 knowledge on breast cancer development. In this study, we performed the first GWAS meta-371 analysis of mammographic texture variation, focusing on a summary measure of gray scale 372 variation on mammograms (V). We identified three genome-wide significant V loci: 6q24.1 373 (ECT2L), 8q24.22 (LINC01591), and 12q22 (PGAM1P5), the first two of which have not 374 previously been associated with MD or breast cancer risk. Four additional loci for MD or breast 375 cancer risk, 5q23.2 (PRDM6), 8p21.2 (EBF2), 12p12.1 (SSPN), and 16q12.2 (FTO), were also 376 found associated with V.

Different models of the SNP-V association were fit to capture different effects. Model 0 with only age and genetic PCs as covariates can capture both the effect of genetic variants on V and the effect that was mediated by BMI or MD phenotypes. We also fit Model 5 adjusting for all MD phenotypes together to assess the variant effect that was independent of all adjusted covariates. Although PD can be calculated from DA and NDA, previous GWAS of MD still identified different loci and genetic effects for different MD measures. We therefore fit the fully adjusted Model 5 to minimize the effect of MD phenotypes on the V associations. Collinearity

384 issue in Model 5 did not have an impact on the effect estimates of the variants. Comparing the 385 results from different models may also provide evidence for the underlying relationships between 386 the genetic variants, V, and other adjusted covariates as well as boost power to detect V SNPs. 387 For example, if we observed a SNP-V association in models with and without adjustment for 388 MD, then it is likely that the SNP influences V through other pathways that are independent of 389 density; if the SNP-V association was only observed in model without adjusting for density, then 390 it indicates that the SNP effect on V might be largely mediated by density. Downstream analyses 391 need to be performed to confirm the relationships. Both V65L and the calculated summary 392 statistics of the four V assessments, VSUM, were used as our primary outcomes. We have a 393 larger sample size thus a greater power for low resolution V assessments compared to high 394 resolution assessments (sample size for V65L and V75L = 7,040; sample size for V65H and 395 V75H = 4,763). Although a previous study looking at the relationship between V and breast 396 cancer risk in NHS/NHSII used a different assessment, V75L, as the outcome [16], these two 397 low resolution V assessments were highly correlated with each other ( $\rho = 0.98$ , Fig. 1) and there 398 was no substantial difference in the GWAS results of these two assessments (Additional file 1: 399 Table S4). Using VSUM also has the advantage of boosting power given that the SNP 400 associations were similar across different V assessments.

Among the three genome-wide significant V loci, **12q22** is also associated with NDA and breast cancer risk in consistent direction, suggesting that at least part of its genetic effect on V is mediated by NDA or the genetic effect on NDA is mediated by V, and there are potential shared biological pathways between these three traits. These hypotheses are further supported by the fact that 12q22 is also associated with total breast size and its association with V was most significant when adjusting for DA and became non-significant when adjusting for NDA. The

407 lead variant rs113174754 at 12q22 is an indel near pseudogene *PGAM1P5* and is 30kb upstream 408 of protein coding gene NTN4 (see Fig. 2c). NTN4 encodes a member of the netrin family of 409 proteins, which involved in axon guidance, tumorigenesis, and angiogenesis. NDA SNP at 12q22 410 (rs11836367-C, correlated with the effect allele of rs113174754) has been found to downregulate 411 NTN4 in mammary tissue [6]. NTN4 has also been identified as a candidate breast cancer risk gene by colocalization analysis, where the C allele of SNP rs61938093 ( $r^2 = 0.48$  with the effect 412 413 allele of rs113174754) at this region reduced NTN4 promoter activity and knockdown of NTN4 414 promoted breast cell proliferation and tumor growth [39]. These findings suggest a shared 415 genetic basis and potential biological mechanisms for mammographic risk factors, especially 416 breast adipose tissue (represented by NDA), and breast cancer risk at this locus, and may also 417 explain the observed association between V and breast cancer risk. 6q24.1 and 8q24.22 are V 418 loci that have not been seen associated with MD phenotypes or breast cancer risk. The lead 419 variant rs138141444 at 6q24.1 is an intronic indel in ECT2L. The lead variant rs79670367 at 420 8q24.22 is an intronic SNP in *LINC01591*. Neither these two genes nor nearby genes have been 421 associated with breast cancer risk. The genetic effects of these two loci on V are therefore likely 422 through mechanisms not mediated by MD. It should also be noted that the effect allele frequency 423 for rs79670367 is less than 5% and the outlier analysis indicated that the association results 424 might be influenced by influential outliers. Moreover, only about half of the samples have 425 genotype data on this variant (available in NHS/NHSII Illumina HumanHap and MMHS 426 OncoArray). Further studies are needed to confirm the findings at these two loci. 427 Four additional V loci have previously been associated with breast cancer risk or MD phenotypes. 428 The lead variant rs13256025 at **8p21.2** is an intronic SNP in protein coding gene *EBF2*. *EBF2* 429 encodes well conserved DNA-binding helix-loop-helix transcription factors, which involved in

430	differentiation of osteoblasts. Although little is known about the role of EBF2 in breast cancer
431	development, studies have shown that inactivation of EBF genes can lead to tumorigenesis via
432	accumulation and expansion of undifferentiated progenitor cells [40]. 16q12.2 is associated with
433	both PD and breast cancer risk in the same direction with its lead SNP rs17817449 located in
434	FTO. FTO is a well-established susceptibility gene for obesity [41]. In our analysis, the
435	association was only significant in the base model and became non-significant when adjusting
436	for BMI, suggesting that its genetic effect on V might be mediated by BMI. FTO is
437	overexpressed in breast cancer cells, which affects the energy metabolism of the cells [42].
438	<b>5q23.2</b> is a known locus for DA and PD. The lead variant rs335189 is an intronic SNP in
439	PRDM6. PRDM6 encodes a transcriptional repressor involved in the regulation of endothelial
440	cell proliferation, survival, and differentiation, and may play a role in breast cancer
441	tumorigenesis [7, 43]. The lead variant rs11836164 at <b>12p12.1</b> is an intronic SNP near SSPN and
442	is only associated with DA. Functional analysis needs to be performed to further investigate the
443	role of identified V SNPs in mammary development and breast cancer etiology.
444	Consistent with the phenotypic relationships we observed for V and MD measures, there were
445	strong positive genetic correlations of V with DA and PD, and negative genetic correlations with
446	NDA. The positive genetic correlations between V and breast cancer risk (overall and ER+
447	specific) were also nominally significant, further supporting that the observed phenotypic
448	association between V and breast cancer risk can at least be partially explained by shared genetic
449	components. The magnitude of these genetic correlations is comparable to those between MD
450	and breast cancer risk [6]. A genetic variant can be associated with multiple traits, which is
451	known as pleiotropy. Studies have shown that jointly analyzing GWAS data of multiple traits
452	can boost power to detect genetic associations for each trait and improve the prediction

453 performance [44, 45]. In our analysis, we observed significant genetic correlations of V with MD 454 phenotypes and BMI using genome-wide association results. It is therefore very likely that a 455 substantial number of variants are associated with both MD phenotypes, especially NDA, and 456 BMI, which would dilute the correlations we observed for any pair of the traits. SNP-set tests 457 may provide more evidence for the shared mechanism underlying two traits using only 458 susceptibility variants. Here, we found that even if we adjust for PD in the model, there were still 459 significant correlations between V and PD based on genome-wide significant SNPs for PD, 460 indicating that the genetic contribution of V cannot be fully explained by PD and PD is either a 461 mediator or collider of the association between the genetic variants and V (Fig. 4a). Correlations 462 of V with breast cancer and childhood body fatness were also stronger at the susceptibility 463 variants. There were still correlations, though not significant, after adjusting for PD, providing 464 evidence for the genetic relationship between V and these traits that were not mediated by MD 465 (Fig. 4b).

466 Our study focuses on a summary texture measure, V, but there are also many other texture 467 features. For example, Manduca et al. systematically evaluated 1,443 textural features and 468 identified six independently validated strongest features [13]. Malkov et al. identified 15 texture 469 features that were significantly associated with breast cancer risk, several of which were only 470 weakly correlated with PD [46]. Studying the genetics of these features or their combinations 471 may provide additional information for the genetic architecture of breast parenchymal texture 472 variation. Our study included breast cancer cases, which might be concerning since V has been 473 associated with breast cancer risk. However, both theoretical [47] and empirical [48] evidence 474 suggest that including cases of a rare outcome does not bias the association estimates in GWAS 475 of a secondary outcome, except when both the genetic variant being analyzed and the secondary

476	outcome are very strong risk factors-stronger than those exhibited by breast cancer risk SNPs,
477	V, or BMI. Indeed, we did not observe any substantial changes on the top findings after further
478	adjusting for breast cancer case-control status in the model. Moreover, the direction of the
479	associations we observed—e.g., a breast cancer risk allele was positively associated with V—are
480	opposite of those expected if the SNP-V association is solely an artefact due to collider bias.
481	Multiple testing issue caused by studying four V assessments may also be a concern, we
482	therefore estimated a single summary test statistic, VSUM, to minimize the impact of multiple
483	testing and to boost power. Studying the computerized automated texture feature can also reduce
484	the potential bias caused by measurement error that studies on semi-automated MD measures are
485	usually susceptible to.
486	
487	
488	
489	
400	
490	
491	
492	
102	
-133	
494	

# 495 Conclusions

496	In conclusion, we performed a GWAS of breast parenchymal texture variation, V, and identified
497	three independent loci at genome-wide significance, including 12q22 (PGAM1P5) that are
498	associated with MD phenotypes and breast cancer risk, and 6q24.1 (ECT2L) and 8q24.22
499	(LINC01591) that are novel V susceptibility loci. Four additional V loci were identified from
500	looking up MD and breast cancer susceptibility SNPs in GWAS of V, including 5q23.2
501	(PRDM6), 8p21.2 (EBF2), 12p12.1 (SSPN), and 16q12.2 (FTO). These findings provide the first
502	evidence of the genetic basis of V and shared genetic components between V, MD, and breast
503	cancer risk. Future studies are needed to confirm our findings and further improve our
504	understanding of the mechanisms underlying the relationship between texture features, MD, and
505	breast cancer development.
506	
507	
508	
509	
510	
511	
512	
513	

# 514 List of abbreviations

<b>DUAL.</b> DIEast Calleel Association Consolution, Divit. Douy mass muex, DA. dense area,	515	BCAC: Breast (	Cancer Association	n Consortium;	BMI: body	y mass index:	: DA: dense	area; ER:
---	-----	----------------	--------------------	---------------	-----------	---------------	-------------	-----------

- 516 estrogen receptor; GWAS: genome-wide association study; LD: linkage disequilibrium; MD:
- 517 mammographic density; MMHS: Mayo Mammography Health Study; NDA: nondense area;
- 518 NHS: Nurses' Health Study; PC: principal component; PD: percent density; SNP: single-
- 519 nucleotide polymorphism; VIF: variance inflation factor; V75H: V with 25% erosion and high
- resolution; V75L: V with 25% erosion and low resolution; V65H: V with 35% erosion and high
- resolution; V65L: V with 35% erosion and low resolution.
- 522
- 523
- 524
- 525
- 526
- 527
- 528
- 529
- 530
- 531

# 532 **Declarations**

#### 533 Ethics approval and consent to participate

- 534 The study protocol was approved by the institutional review boards of the Brigham and
- 535 Women's Hospital, Harvard T.H. Chan School of Public Health, and the Mayo Clinic, and was
- 536 in accordance with the 1964 Declaration of Helsinki and its later amendments or comparable
- 537 ethical standards. All participants provided written informed consent.

# 538 **Consent for publication**

539 Not applicable

#### 540 Availability of data and materials

- 541 NHS/NHSII: The data that support the findings of this study are available from the Nurses'
- 542 Health Studies, however they are not publicly available. Investigators interested in using the data
- 543 can request access, and feasibility will be discussed at an investigators' meeting. Limits are not
- 544 placed on scientific questions or methods, and there is no requirement for co-authorship.
- 545 Additional data sharing information and policy details can be accessed
- 546 at http://www.nurseshealthstudy.org/researchers.
- 547 MMHS: The summary statistics generated from the current study are available from the
- 548 corresponding author on reasonable request.

# 549 Competing interests

550 The authors declare that they have no competing interests

## 551 Funding

- 552 This work is supported by the National Cancer Institute (R01CA175080 and R01CA131332 to
- 553 R.M.T., R01CA244670 to S.L., and R03CA224196 to X.J.), Avon Foundation for Women,
- 554 Susan G. Komen for the Cure, and Breast Cancer Research Foundation. The Nurses' Health
- 555 Study is supported by the National Cancer Institute (UM1CA186107, P01CA87969, and
- 556 R01CA49449). The Nurses' Health Study II is supported by the National Cancer Institute
- 557 (U01CA176726 and R01CA67262). The Mayo Mammography Health Study is supported by the
- 558 National Cancer Institute (R01CA128931 and R01CA97396).

#### 559 Authors' contributions

- 560 XJ, PK, RMT, and CVM conceived and designed the study. RMT and ETW prepared the
- 561 mammographic texture variation data for NHS/NHSII. CMV and JH prepared the
- 562 mammographic texture variation data for MMHS. CT prepared the genotype data for
- 563 NHS/NHSII. SJW prepared the genotype data for MMHS. HC and SL prepared the
- 564 mammographic density GWAS data. YL analyzed and interpreted the data, and was a major
- 565 contributor in writing the manuscript. All authors read and approved the final manuscript.

# 566 Acknowledgements

- 567 We would like to thank the participants and staff of the NHS and NHSII for their valuable
- 568 contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO,
- 569 CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH,
- 570 OK, OR, PA, RI, SC, TN, TX, VA, WA, WY.

571	The BCAC MD GWAS was supported by CA244670 and CA194393. BCAC is funded by the
572	European Union's Horizon 2020 Research and Innovation Programme (grant numbers 634935
573	and 633784 for BRIDGES and B-CAST respectively), and the PERSPECTIVE I&I project,
574	funded by the Government of Canada through Genome Canada and the Canadian Institutes of
575	Health Research, the Ministère de l'Économie et de l'Innovation du Québec through Genome
576	Québec, the Quebec Breast Cancer Foundation. The EU Horizon 2020 Research and Innovation
577	Programme funding source had no role in study design, data collection, data analysis, data
578	interpretation or writing of the report. Additional funding for BCAC is provided via the
579	Confluence project which is funded with intramural funds from the National Cancer Institute
580	Intramural Research Program, National Institutes of Health.
581	Genotyping of the OncoArray was funded by the NIH Grant U19 CA148065, and Cancer UK
582	Grant C1287/A16563 and the PERSPECTIVE project supported by the Government of Canada
583	through Genome Canada and the Canadian Institutes of Health Research (grant GPH-129344)
584	and, the Ministère de l'Économie, Science et Innovation du Québec through Genome Québec
585	and the PSRSIIRI-701 grant, and the Quebec Breast Cancer Foundation. Funding for iCOGS
586	came from: the European Community's Seventh Framework Programme under grant agreement
587	n° 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (C1287/A10118,
588	C1287/A10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007,
589	C5047/A10692, C8197/A16565), the National Institutes of Health (CA128978) and Post-Cancer
590	GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 - the GAME-ON
591	initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health
592	Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer, and Komen

- 593 Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer
- 594 Research Fund.

# 596 **References**

597 1. Byrne C, Schairer C, Wolfe J, Parekh N, Salane M, Brinton LA, Hoover R, Haile R: 598 Mammographic features and breast cancer risk: effects with time, age, and 599 menopause status. J Natl Cancer Inst 1995, 87(21):1622-1629. 600 2. Boyd NF, Martin LJ, Yaffe MJ, Minkin S: Mammographic density and breast cancer risk: 601 current understanding and future prospects. Breast Cancer Res 2011, 13(6):223. 602 Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, Jong RA, Hislop G, Chiarelli A, Minkin 3. 603 S et al: Mammographic density and the risk and detection of breast cancer. N Engl J 604 Med 2007, 356(3):227-236. 605 Boyd NF, Dite GS, Stone J, Gunasekara A, English DR, McCredie MR, Giles GG, Tritchler D, 4. Chiarelli A, Yaffe MJ et al: Heritability of mammographic density, a risk factor for 606 607 breast cancer. N Engl J Med 2002, 347(12):886-894. 608 5. Stone J, Dite GS, Gunasekara A, English DR, McCredie MR, Giles GG, Cawson JN, Hegele 609 RA, Chiarelli AM, Yaffe MJ et al: The heritability of mammographically dense and 610 nondense breast tissue. Cancer Epidemiol Biomarkers Prev 2006, 15(4):612-617. Sieh W, Rothstein JH, Klein RJ, Alexeeff SE, Sakoda LC, Jorgenson E, McBride RB, Graff RE, 611 6. 612 McGuire V, Achacoso N et al: Identification of 31 loci for mammographic density 613 phenotypes and their associations with breast cancer risk. Nat Commun 2020, 614 **11**(1):5116. 615 Lindstrom S, Thompson DJ, Paterson AD, Li J, Gierach GL, Scott C, Stone J, Douglas JA, 7. 616 dos-Santos-Silva I, Fernandez-Navarro P et al: Genome-wide association study 617 identifies multiple loci associated with both mammographic density and breast cancer 618 risk. Nat Commun 2014, 5:5303. 619 Chen H, Fan S, Stone J, Thompson DJ, Douglas J, Li S, Scott C, Bolla MK, Wang Q, Dennis J 8. 620 et al: Genome-wide and transcriptome-wide association studies of mammographic 621 density phenotypes reveal novel loci. Breast Cancer Res 2022, 24(1):27. 622 9. Boyd NF, Rommens JM, Vogt K, Lee V, Hopper JL, Yaffe MJ, Paterson AD: 623 Mammographic breast density as an intermediate phenotype for breast cancer. Lancet 624 Oncol 2005, 6(10):798-808. 625 Byng JW, Boyd NF, Fishell E, Jong RA, Yaffe MJ: The quantitative analysis of 10. 626 mammographic densities. Phys Med Biol 1994, 39(10):1629-1638. 627 11. Gastounioti A, Conant EF, Kontos D: Beyond breast density: a review on the advancing 628 role of parenchymal texture analysis in breast cancer risk assessment. Breast Cancer 629 *Res* 2016. **18**(1):91. 630 Heine JJ, Scott CG, Sellers TA, Brandt KR, Serie DJ, Wu FF, Morton MJ, Schueler BA, 12. 631 Couch FJ, Olson JE et al: A novel automated mammographic density measure and 632 breast cancer risk. J Natl Cancer Inst 2012, 104(13):1028-1037. 633 Manduca A, Carston MJ, Heine JJ, Scott CG, Pankratz VS, Brandt KR, Sellers TA, Vachon 13. 634 CM, Cerhan JR: Texture features from mammographic images and risk of breast cancer. 635 Cancer Epidemiol Biomarkers Prev 2009, 18(3):837-845. 636 14. Nielsen M, Karemore G, Loog M, Raundahl J, Karssemeijer N, Otten JD, Karsdal MA, 637 Vachon CM, Christiansen C: A novel and automatic mammographic texture

638 639		resemblance marker is an independent risk factor for breast cancer. <i>Cancer Epidemiol</i>
640	15	Wanders IOP van Gils CH. Karssemeijer N. Holland K. Kallenberg M. Peeters PHM
641	13.	Nielsen M, Lillholm M: The combined effect of mammographic texture and density on
642		breast cancer risk: a cohort study. Breast Cancer Res 2018, <b>20</b> (1):36.
643	16.	Warner ET, Rice MS, Zeleznik OA, Fowler EE, Murthy D, Vachon CM, Bertrand KA, Rosner
644		BA, Heine J, Tamimi RM: Automated percent mammographic density, mammographic
645		texture variation, and risk of breast cancer: a nested case-control study. NPJ Breast
646		Cancer 2021, <b>7</b> (1):68.
647	17.	Oh H, Rice MS, Warner ET, Bertrand KA, Fowler EE, Eliassen AH, Rosner BA, Heine JJ,
648		Tamimi RM: Early-Life and Adult Anthropometrics in Relation to Mammographic Image
649		Intensity Variation in the Nurses' Health Studies. Cancer Epidemiol Biomarkers Prev
650		2020, <b>29</b> (2):343-351.
651	18.	Tworoger SS, Missmer SA, Eliassen AH, Spiegelman D, Folkerd E, Dowsett M, Barbieri RL,
652		Hankinson SE: The association of plasma DHEA and DHEA sulfate with breast cancer
653		risk in predominantly premenopausal women. Cancer Epidemiol Biomarkers Prev 2006,
654		<b>15</b> (5):967-971.
655	19.	Olson JE, Sellers TA, Scott CG, Schueler BA, Brandt KR, Serie DJ, Jensen MR, Wu FF,
656		Morton MJ, Heine JJ et al: The influence of mammogram acquisition on the
657		mammographic density and breast cancer association in the Mayo Mammography
658		Health Study cohort. Breast Cancer Res 2012, 14(6):R147.
659	20.	Heine JJ, Cao K, Rollison DE: Calibrated measures for breast density estimation. Acad
660		Radiol 2011, <b>18</b> (5):547-555.
661	21.	Heine JJ, Cao K, Rollison DE, Tiffenberg G, Thomas JA: A quantitative description of the
662		percentage of breast density measurement using full-field digital mammography. Acad
663		<i>Radiol</i> 2011, <b>18</b> (5):556-564.
664	22.	Boyd NF, Stone J, Martin LJ, Jong R, Fishell E, Yaffe M, Hammond G, Minkin S: <b>The</b>
665		association of breast mitogens with mammographic densities. Br J Cancer 2002,
666		<b>87</b> (8):876-882.
667	23.	Yaghjyan L, Pettersson A, Colditz GA, Collins LC, Schnitt SJ, Beck AH, Rosner B, Vachon C,
668		Tamimi RM: Postmenopausal mammographic breast density and subsequent breast
669		cancer risk according to selected tissue markers. Br J Cancer 2015, 113(7):1104-1113.
670	24.	Lindstrom S, Loomis S, Turman C, Huang H, Huang J, Aschard H, Chan AT, Choi H,
671		Cornelis M, Curhan G et al: A comprehensive survey of genetic variation in 20,691
672		subjects from four large cohorts. PLoS One 2017, 12(3):e0173997.
673	25.	Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, Lemacon A, Soucy P, Glubb
674		D, Rostamianfar A et al: Association analysis identifies 65 new breast cancer risk loci.
675		Nature 2017, <b>551</b> (7678):92-94.
676	26.	Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K,
677		Jun G, Fritz MH et al: An integrated map of structural variation in 2,504 human
678		genomes. Nature 2015, <b>526</b> (7571):75-81.
679	27.	Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ: RVTESTS: an efficient and comprehensive tool
680		for rare variant association analysis using sequence data. Bioinformatics 2016,
681		<b>32</b> (9):1423-1426.

682	28.	Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: Second-generation PLINK:
683		rising to the challenge of larger and richer datasets. <i>Gigascience</i> 2015, 4:7.
684	29.	Willer CJ, Li Y, Abecasis GR: METAL: fast and efficient meta-analysis of genomewide
685		association scans. Bioinformatics 2010, <b>26</b> (17):2190-2191.
686	30.	Boughton AP, Welch RP, Flickinger M, VandeHaar P, Taliun D, Abecasis GR, Boehnke M:
687		LocusZoom.js: Interactive and embeddable visualization of genetic association study
688		results. Bioinformatics 2021.
689	31.	Liu Z, Lin X: A Geometric Perspective on the Power of Principal Component Association
690		Tests in Multiple Phenotype Studies. J Am Stat Assoc 2019, 114(527):975-990.
691	32.	Zhang H, Ahearn TU, Lecarpentier J, Barnes D, Beesley J, Qi G, Jiang X, O'Mara TA, Zhao
692		N, Bolla MK et al: Genome-wide association study identifies 32 novel breast cancer
693		susceptibility loci from overall and subtype-specific analyses. Nat Genet 2020,
694		<b>52</b> (6):572-581.
695	33.	Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, ReproGen C,
696		Psychiatric Genomics C, Genetic Consortium for Anorexia Nervosa of the Wellcome
697		Trust Case Control C, Duncan L et al: An atlas of genetic correlations across human
698		diseases and traits. <i>Nat Genet</i> 2015, <b>47</b> (11):1236-1241.
699	34.	Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of
700		the Psychiatric Genomics C, Patterson N, Daly MJ, Price AL, Neale BM: LD Score
701		regression distinguishes confounding from polygenicity in genome-wide association
702		studies. Nat Genet 2015, 47(3):291-295.
703	35.	Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA: Detection and interpretation of
704		shared genetic influences on 42 human traits. Nat Genet 2016, 48(7):709-717.
705	36.	Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, Highland HM, Patel YM,
706		Sorokin EP, Avery CL et al: Genetic analyses of diverse populations improves discovery
707		for complex traits. <i>Nature</i> 2019, <b>570</b> (7762):514-518.
708	37.	Gierach GL, Li H, Loud JT, Greene MH, Chow CK, Lan L, Prindiville SA, Eng-Wong J,
709		Soballe PW, Giambartolomei C et al: Relationships between computer-extracted
710		mammographic texture pattern features and BRCA1/2 mutation status: a cross-
711		sectional study. Breast Cancer Res 2014, 16(4):424.
712	38.	Li H, Giger ML, Olopade Ol, Margolis A, Lan L, Chinander MR: Computerized texture
713		analysis of mammographic parenchymal patterns of digitized mammograms. Acad
714		Radiol 2005, <b>12</b> (7):863-873.
715	39.	Beesley J, Sivakumaran H, Moradi Marjaneh M, Shi W, Hillman KM, Kaufmann S, Hussein
716		N, Kar S, Lima LG, Ham S et al: <b>eQTL Colocalization Analyses Identify NTN4 as a</b>
717		Candidate Breast Cancer Risk Gene. Am J Hum Genet 2020, 107(4):778-787.
718	40.	Liao D: Emerging roles of the EBF family of transcription factors in tumor suppression.
719		Mol Cancer Res 2009, <b>7</b> (12):1893-1901.
720	41.	Loos RJ, Yeo GS: The bigger picture of FTO: the first GWAS-identified obesity gene. Nat
721		<i>Rev Endocrinol</i> 2014, <b>10</b> (1):51-61.
722	42.	Liu Y, Wang R, Zhang L, Li J, Lou K, Shi B: The lipid metabolism gene FTO influences
723		breast cancer cell energy metabolism via the PI3K/AKT signaling pathway. Oncol Lett
724		2017 <i>,</i> <b>13</b> (6):4685-4690.

725	43.	Casamassimi A, Rienzo M, Di Zazzo E, Sorrentino A, Fiore D, Proto MC, Moncharmont B,
726		Gazzerro P, Bifulco M, Abbondanza C: Multifaceted Role of PRDM Proteins in Human
727		<b>Cancer</b> . Int J Mol Sci 2020, <b>21</b> (7).
728	44.	Aguirre M, Tanigawa Y, Venkataraman GR, Tibshirani R, Hastie T, Rivas MA: Polygenic
729		risk modeling with latent trait-related genetic components. Eur J Hum Genet 2021,
730		<b>29</b> (7):1071-1081.
731	45.	Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, Nguyen-Viet TA,
732		Wedow R, Zacher M, Furlotte NA et al: Multi-trait analysis of genome-wide association
733		summary statistics using MTAG. Nat Genet 2018, 50(2):229-237.
734	46.	Malkov S, Shepherd JA, Scott CG, Tamimi RM, Ma L, Bertrand KA, Couch F, Jensen MR,
735		Mahmoudzadeh AP, Fan B et al: Mammographic texture and risk of breast cancer by
736		tumor type and estrogen receptor status. Breast Cancer Res 2016, 18(1):122.
737	47.	Monsees GM, Tamimi RM, Kraft P: Genome-wide association scans for secondary traits
738		using case-control samples. Genet Epidemiol 2009, 33(8):717-728.
739	48.	Lindstrom S, Vachon CM, Li J, Varghese J, Thompson D, Warren R, Brown J, Leyland J,
740		Audley T, Wareham NJ et al: Common variants in ZNF365 are associated with both
741		mammographic density and breast cancer risk. Nat Genet 2011, 43(3):185-187.
742		
743		
744		
745		
746		
747		
748		
749		
750		
751		
752		
753		
754		
755		
756		
757		
758		
759		
760		
761		
762		
763		
764 765		
705		
/00 707		
/0/ 760		
/0ŏ		

# Tables

# Table 1 Characteristics of NHS/NHSII and MMHS study population

	NHS/NHSII (n = 4831)	<b>MMHS</b> (n = 2209)
	Mean (SD)	Mean (SD)
Age (years)	53.8 (9.2)	58.9 (11.9)
BMI (kg/m <sup>2</sup> )	25.9 (5.3)	28.0 (6.2)
Dense area	43.4 (29.3)	23.8 (17.0)
Nondense area	109.3 (73.7)	130 (67.1)
Percent density	32.8 (19.7)	17.9 (12.9)

Abbreviations: NHS Nurses' Health Study, MMHS Mayo Mammography Health Study, SD standard deviation, BMI body mass index

<b>Table 2</b> Novel V loci and their associations with mammographic density phenotypes and breast cancer risk, overall and stratified by
estrogen receptor status

-						V65L			VSUM			MD phenoty	pes <sup>b</sup>		Breast cance	r risk <sup>b</sup>	
Region	Lead SNP	Position	Gene <sup>a</sup>	EA/OA	EAF	Model	β (SE)	P value	Model	Z score	P value	Phenotype	Z score	P value	Phenotype	β (SE)	P value
Genome-wide significant loc <sup>*</sup>																	
	0.0																
6q24.1 <sup>d</sup>	rs138141444	139157426	ECT2L	CT/C	0.58	0	-0.08 (0.02)	1.02E-05	0	-5.32	1.04E-07	DA	-0.31	7.55E-01	Overall	0.00 (0.01)	8.09E-01
						1	-0.07 (0.02)	3.06E-05	1	-5.06	4.10E-07	NDA	-0.01	9.90E-01	ER+	0.00 (0.01)	9.29E-01
						2	-0.06 (0.01)	6.43E-05	2	-4.99	6.00E-07	PD	-0.12	9.05E-01	ER-	0.02 (0.01)	2.32E-01
						3	-0.06 (0.01)	2.28E-04	3	-4.62	3.87E-06						
						4	-0.06 (0.02)	1.14E-04	4	-4.96	7.12E-07						
	70/702/7	126200666	I DICOLCOL		0.00	5	-0.05 (0.01)	1.95E-04	5	-4.69	2.78E-06	D.	0.12	0.005.01	0 11	0.00 (0.02)	1.025.02
8q24.22	rs/96/036/	136280666	LINC01591	A/C	0.02	0	-0.34 (0.14)	1./1E-02	0	-2.55	1.09E-02	DA	0.13	8.99E-01	Overall	-0.08 (0.03)	1.92E-02
						1	-0.41 (0.14)	2.75E-03	1	-3.08	2.05E-03	NDA	0.45	6.5/E-01	ER+	-0.08 (0.04)	3.58E-02
						2	-0.57 (0.11)	4.63E-07	2	-5.34	9.32E-08	PD	-0.64	5.25E-01	ER-	-0.10 (0.06)	1.25E-01
						3	-0.55 (0.12)	2.41E-06	3	-5.07	3.9/E-0/						
						4	-0.39 (0.13)	2.23E-03	4	-3.30	9.72E-04						
10.00	110174754	0.00010.07	DOLLARDS		0.57	5	-0.56 (0.11)	3.01E-07	5	-5.58	2.38E-08	D.		1.055.02	0 11	0.07 (0.01)	1.005.04
12q22	rs1131/4/54	96021847	PGAMIPS	A/ACCIGIAGI	0.57	0	0.07 (0.02)	1.25E-04	0	4.53	5.91E-06	DA	-2.34	1.95E-02	Overall	0.07 (0.01)	1.08E-24
						1	0.07 (0.02)	8.74E-05	1	4.61	4.05E-06	NDA	-3.97	7.26E-05	EK+	0.07 (0.01)	1.3/E-18
						2	0.06 (0.02)	3.39E-05	2	4.62	3.82E-06	PD	0.81	4.1/E-01	ER-	0.06 (0.01)	3.44E-06
						3	0.08 (0.02)	3.40E-07	3	5.47	4.42E-08						
						4	0.03 (0.02)	1.95E-05	4	5.03	2.03E-04						
						5	0.06 (0.01)	2.65E-05	5	4.44	8.90E-06						
MD or bro	east cancer risk lo	ci <sup>e</sup>															
5-22.2	ro225190	122446856	DDDM6	C/C	0.70	0	0.07 (0.02)	2.15E.04	0	2.07	7 20E 05	DA	6.66	2.84E 11	Overall	0.02 (0.01)	0.55E.05
5425.2	18555169	122440850	TKDM0	0	0.70	1	-0.07(0.02)	2.13E-04 2.67E-04	1	-3.97	1.24E-04	NDA	-0.00	0.30E-03	ER+	-0.03(0.01)	1.54E-03
						2	-0.00 (0.02)	6.26E.02	2	-5.04	2.25E 02	PD	6.20	5.78E 10	ED	-0.03 (0.01)	2 25E 01
						3	-0.03(0.01)	6.42E-02	23	-2.20	1.73E-02	FD	-0.20	5.781-10	LK-	-0.01 (0.01)	5.25E-01
						4	-0.04 (0.02)	4.65E.04	4	-5.15	4.50E.04						
						4	-0.03(0.02)	2.61E-02	4	-2.58	9.85E-03						
8n21 2	re13256025	25831778	FRF?	T/C	0.20	0	-0.03 (0.01)	6.23E-05	0	-2.58	5.74E-05	DA	3 3 2	9.17E-04	Overall	0.04 (0.01)	2 12E-07
0021.2	1315250025	25051770	LDF2	1/C	0.20	1	0.09(0.02)	7.58E-05	1	3.04	8.01E-05	NDA	_4.26	2.01E-05	EP+	0.03 (0.01)	8 77E-04
						2	0.03(0.02)	1.60E-02	2	2.54	1.11E-02	PD	-4.20	2.01E-05	ER-	0.05 (0.01)	1.78E-04
						3	0.04(0.02)	5.86E-03	3	2.94	3.47E-03	TD	5.05	4.401-07	LK-	0.00 (0.01)	1.761-04
						4	0.05 (0.02)	1 39E-03	4	3 23	1.22E-03						
						5	0.04(0.02)	3 31E-02	5	2 37	1.79E-02						
12n12.1	rs11836164	26446625	SSPN	T/C	0.76	0	-0.09(0.02)	8.22E-06	0	-3.99	6.69E-05	DA	-6.03	1.66E-09	Overall	0.01 (0.01)	6 95E-02
1-21-211	1011000101	20110020	55111	1/0	0.70	1	-0.08(0.02)	7 20E-06	1	-4.05	5 19E-05	NDA	-0.73	4 67E-01	ER+	0.00(0.01)	7 35E-01
						2	-0.04(0.02)	7 39E-03	2	-2.37	1 77E-02	PD	-5.43	5 55E-08	ER-	0.03 (0.01)	5 72E-02
						3	-0.05(0.02)	3 18E-03	3	-2.55	1.09E-02	12	5.15	0.002.00	210	0.05 (0.01)	0.122.02
						4	-0.08(0.02)	1.14E-05	4	-4.17	3.03E-05						
						5	-0.04(0.02)	6.93E-03	5	-2.47	1.34E-02						
16a12.2	rs17817449	53813367	FTO	T/G	0.61	0	0.09 (0.02)	1.63E-07	0	4.87	1.12E-06	DA	4.64	3.49E-06	Overall	0.06 (0.01)	2.52E-21
10412.2		22220000				1	0.06 (0.02)	2.46E-04	1	3.32	8.93E-04	NDA	-3.10	1.94E-03	ER+	0.06 (0.01)	5.59E-14
						2	0.04 (0.01)	1.50E-03	2	2.85	4.36E-03	PD	5.85	5.06E-09	ER-	0.07 (0.01)	1.80E-10
						3	0.05 (0.01)	1.51E-04	3	3.35	8.20E-04		2.55				
						4	0.05 (0.02)	2.42E-03	4	2.77	5.66E-03						
						5	0.04 (0.01)	1.46E-03	5	2.80	5.14E-03						

<sup>a</sup>Nearest gene within 500kb of the lead SNP

<sup>b</sup>Summary statistics were from Breast Cancer Association Consortium [8, 25]

<sup>c</sup>Genome-wide significant loci ( $P < 5 \times 10^{-8}$ ) in any model for any V assessment

# <sup>d</sup>6q24.1 (rs138141444) is genome-wide significant for V75L (Additional file 1: Table S4)

<sup>e</sup>Significant V loci at the Bonferroni-corrected threshold of  $P < 6.94 \times 10^{-4}$  accounting for the 72 MD phenotype SNPs tested or  $P < 2.56 \times 10^{-4}$  accounting for the 195 breast cancer SNPs tested based on Model 0.

*Abbreviations: SNP* single-nucleotide polymorphism, *EA* effect allele, *OA* other allele, *EAF* effect allele frequency, *MD* mammographic density, *DA* dense area, *NDA* nondense area, *PD* percent density, *ER* estrogen receptor

# Figures

**Fig. 1** Scatter plots of four V assessments and V65L by three mammographic density phenotypes.  $\mathbf{a} = \text{scatter plots of four V assessments; } \mathbf{b} = \text{scatter plots of V65L by dense area (DA), nondense area (NDA), and percent density (PD). Spearman correlation between the two measures is shown on the upper left or right corner of each plot. Red lines on the plots are the diagonal lines.$ 

**Fig. 2** Regional association plots for the three genome-wide significant V loci.  $\mathbf{a} = 6q24.1$  (rs138141444; V75H, Model 0);  $\mathbf{b} = 8q24.22$  (rs79670367; VSUM, Model 5);  $\mathbf{c} = 12q22$  (rs113174754; VSUM, Model 3). Model and V assessment with the most significant results for each locus are shown. Each plot is centered around the lead SNP of each locus. SNPs in the 95% credible set at each locus are shown in color. Physical positions are based on NCBI Genome Reference Consortium Human Build 37. Plots were generated using LocusZoom [30].

Abbreviations: SNP single-nucleotide polymorphism

**Fig. 3** Genetic correlation and SNP-set test results of V with mammographic density phenotypes, breast cancer risk, and other breast cancer risk factors. **a** = genetic correlations between V and other traits; **b** = SNP-set test results of the relationship of V and other traits. Results of Model 0 and 2 for V65L and VSUM are shown. Estimates passed the Bonferroni threshold ( $P < 0.05/40 = 1.25 \times 10^{-3}$ ) are marked with triple asterisk (\*\*\*); estimates with P < 0.01 are marked with double asterisk (\*\*); estimates with nominal significance (P < 0.05) are marked with single asterisk (\*). Genetic correlations between VSUM (Model 2) and MD phenotypes were not estimated due to the out of bounds heritability of V.

Abbreviations: SNP single-nucleotide polymorphism, DA dense area, NDA nondense area, PD percent density, ER estrogen receptor, BMI body mass index

**Fig. 4** Z scores from GWAS of V, mammographic density phenotypes, and breast cancer risk for SNPs included in SNP-set test.  $\mathbf{a}$  = scatter plots of Z scores from GWAS of V by Z scores from GWAS of percent density (PD), dense area (DA), and nondense area (NDA) for mammographic density SNPs;  $\mathbf{b}$  = scatter plots of Z scores from GWAS of V by Z scores from GWAS of overall breast cancer risk and stratified by estrogen receptor (ER) status for breast cancer SNPs. For each

SNP, GWAS results from Model 0 and 2 for VSUM are shown with gray and red dots, respectively. RS number for some SNPs are not shown on the plots. Gray line is the fitted linear regression line of Z scores for results from Model 0; red line is the fitted linear regression line of Z scores for results from Model 2. Note that some of the overall breast cancer risk SNPs are not genome-wide significant because we obtained the Z scores from one study and those SNPs were reported by other studies.

Abbreviations: GWAS genome-wide association study, SNP single-nucleotide polymorphism

# **Additional files**

**Additional file 1** (XLSX):

**Table S1.** Number of GWAS subjects by study and platform.

Table S4. GWAS results of significant V loci for all V assessments and models.

Table S5. Lookup results of 72 MD phenotype SNPs in GWAS of V.

Table S6. Lookup results of 195 breast cancer SNPs in GWAS of V.

**Table S7.** Genetic correlation results of all V assessments and models.

Table S8. SNP-set test results of all V assessments and models.

# Additional file 2 (DOCX):

**Table S2.** Sources of summary statistics of breast cancer risk and breast cancer risk factors for calculating genetic correlation.

**Table S3.** Sources of summary statistics of breast cancer risk and breast cancer risk factors for the SNP-set test.

Figure S1. Quantile-quantile plots of the GWAS meta-analysis results.

Figure S2. Manhattan plots of the GWAS meta-analysis results.

Figure S3. Quantile-quantile plots of the *P* value of heterogeneity.

Fig. 1

a





Fig. 2



100 rs113174754 (cM/Mb) Some contribution - 80 Most contribution Rate - 60 bination 40 20 95.80 96.00 96.20 96.40 Chromosome 12 (Mb)

–FGD6 ←F	RNU6-735P 	<i>←NTN4</i>	AMDHD1→
VEZT→	METAP2→ ₩₩	<i>←AC09000</i>	01.1 ←YPEL5P3
Y_RNA→ 	←USP44	<i>←RNU6-247</i> / 	P AC007298.2→ H
<i>←CBX3P5</i> 	PGAM1F	°5→ SNF ┿	RPF→
RNU6-808P→ 	Y_RNA	→ ←(	CCDC38
MIR331→ 		LINC02410 H	$\rightarrow  \leftarrow RN7SL88P$
<i>MIR</i> 3685→ ∣		AC090	0001.2→ 
<i>←AC08</i>	4879.1		←HAL
←AC08	84879.2 		←LTA4H
*	–AC018475.1		AC007298.1→

a

VEST MODELO 0.81 0.79 DA -\*\*\* \*\*\* -0.56 -0.7 -0.6 NDA -\*\* \*\* 0.58 0.73 0.73 PD -\*\*\* \*\*\* 0.04 0.04 0.29 0.2 Overall breast cancer -\*\* \*\* 0.28 0.08 0.22 0.11 ER+ breast cancer -\*\* \*\* 0.27 -0.04 0.17 -0.09 ER- breast cancer --0.47 -0.05 -0.36 -0.06 BMI -\*\*\* \*\*\* -0.14 0.17 -0.13 0.19 Childhood body fatness -0.23 0.12 0.19 0.12 Age at menarche -\*\* 0.13 0.06 0.1 0.05 Age at natural menopause -

b

1.0

0.5 0.0

-0.5

-1.0



10 5 0 -5 -10

# Fig. 4

а

