Polygenic risk scores for prediction of breast cancer risk in Asian populations

Weang-Kee Ho[1,2*], Mei-Chee Tai[2*], Joe Dennis[3], Xiang Shu[4,5], Jingmei Li[6,7], Peh Joo Ho[7], Iona Y Millwood[8,9], Kuang Lin[8], Yon-Ho Jee[10], Su-Hyun Lee[11], Nasim Mavaddat[3], Manjeet K. Bolla[3], Qin Wang[3], Kyriaki Michailidou[3,12,13], Jirong Long[4], Eldarina Azfar Wijaya[2], Tiara Hassan[2], Kartini Rahmat[14], Veronique Kiak Mien Tan[15,16], Benita Kiat Tee Tan[15,16,17], Su Ming Tan[18], Ern Yu Tan[19], Swee Ho Lim[20], Yu-Tang Gao[21], Ying Zheng[22], Daehee Kang[23,24], Ji-Yeob Choi[24,25,26], Wonshik Han[24,27], Han-Byoel Lee[24,27], Michiki Kubo[28], Yukinori Okada[29-31], Shinichi Namba[29], The Biobank Japan Project[32] , Sue K. Park[23,24,33], Sung-Won Kim[34], Chen-Yang Shen[35], Pei-Ei Wu[35], Boyoung Park[36], Kenneth Muir[37], Artitaya Lophatananon[37], Anna H. Wu[38], Chiu-Chen Tseng[38], Keitaro Matsuo[39,40], Hidemi Ito[41,42], Ava Kwong[43,44,45], Tsun L. Chan[43,46], Esther M. John[47,48], Allison W. Kurian[47,48], Motoki Iwasaki[49], Taiki Yamaji[49], Sun-Seog Kweon[50,51],Kristan J. Aronson[52], Rachel A Murphy[53,54], Woon-Puay Koh[55,56], Chiea-Chuen Khor[57], Jian-Min Yuan[58,59], Rajkumar Dorajoo[57,60], Robin G. Walters[8,9], Zhengming Chen[8,9], Liming Li[61], Jun Lv[61], Keum-Ji Jung[62], Peter Kraft[10,63], Paul D.B. Pharoah[3,64], Alison M. Dunning[64], Jacques Simard[65], Xiao Ou Shu[4], Cheng-Har Yip[66], Aishah Mohd Taib[67], Antonis C. Antoniou[3], Wei Zheng[4], Mikael Hartman[6,68,69], Douglas F. Easton[3,66†], Soo-Hwang Teo[2,67†]

* Joint first authors

† These authors jointly supervised this work

Correspondents: Weang-Kee Ho (WeangKee.Ho@nottingham.edu.my) and Soo-Hwang Teo (soohwang.teo@cancerresearch.my)

[1]School of Mathematical Sciences, Faculty of Science and Engineering, University of Nottingham Malaysia, Jalan Broga, Semenyih, 43500 Selangor, Malaysia.

[2]Cancer Research Malaysia, 1 Jalan SS12/1A, Subang Jaya, 47500 Selangor, Malaysia.

[3]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, CB1 8RN Cambridge, UK.

[4]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University, Medical Center, Nashville, TN37232, USA.

[5]Department of Epidemiology & Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY USA

[6]Department of Surgery, University Surgical Cluster, National University Hospital, 1E Kent Ridge Rd, 119228 Singapore.

[7]Genome Institute of Singapore, Laboratory of Women's Health and Genetics, 60 Biopolis St, 138672 Singapore.

[8]Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK.

[9]MRC Population Health Research Unit, University of Oxford, Oxford OX3 7LF, UK

[10]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

[11]Graduate School of Public Health, Yonsei University, Seoul, Korea

[12]Biostatistics Unit, The Cyprus Institute of Neurology & Genetics, 6 Iroon Avenue, 2371 Ayios Dometios, Cyprus

[13]Cyprus School of Molecular Medicine, The Cyprus Institute of Neurology & Genetics, 6 Iroon Avenue, 2371 Ayios Dometios, Cyprus. 43

[14]Biomedical Imaging Department, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia

[15]Department of Breast Surgery, Singapore General Hospital, Singapore, Singapore

[16]Division of Surgical Oncology, National Cancer Centre Singapore, Singapore, Singapore

[17]Department of General Surgery, Sengkang General Hospital, Singapore, Singapore

[18]Division of Breast Surgery, Changi General Hospital, Singapore, Singapore

[19]Department of General Surgery, Tan Tock Seng Hospital, Singapore 308433, Singapore

[20]KK Breast Department, KK Women's and Children's Hospital, Singapore 229899, Singapore

[21] State Key Laboratory of Oncogene and Related Genes & Department of Epidemiology, Shanghai Cancer Institute, Renji Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China.

[22]Shanghai Municipal Center for Disease Control and Prevention, Shanghai, China.

[23]Department of Preventive Medicine, Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, 03080 Seoul, Korea.

[24]Cancer Research Institute, Seoul National University, 103 Daehak-ro, Jongno-gu, Seoul 03080, Korea.

[25]Department of Biomedical Sciences, Seoul National University Graduate School, 103 Daehak-ro, Jongno-gu, 03080 Seoul, Korea.

[26]Institute of Health Policy and Management, Seoul National University, Medical Research Center, 103 Daehak-ro, Jongno-gu, Seoul, 03080 Korea.

[27]Department of Surgery, Seoul National University College of Medicine, Seoul 03080, South Korea

[28]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan.

[29]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan.

[30]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita 565-0871, Japan.

[31]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita 565-0871, Japan.

[32]Institute of Medical Science, The University of Tokyo, Tokyo, Japan.
[33]Integrated Major in Innovative Medical Science, Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, 03080 Seoul, Korea

[34]Department of Surgery, Daerim Saint Mary's Hospital, 657 Siheung-daero, Daerim-dong, Yeongdeungpo-gu, 07442 Seoul, Korea.

[35]Institute of Biomedical Sciences, Academia Sinica, 115128, Section 2, Academia Rd., Taipei, Taiwan.

[36]Department of Medicine, Hanyang University College of Medicine, Seoul, Korea.

[37]Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, The University of Manchester, Oxford Road, M13 9PL Manchester, UK.

[38]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, 1975 Zonal Ave, Los Angeles, 90033 CA, USA.

[39]Division of Cancer Epidemiology and Prevention, Aichi Cancer Center Research Institute, 1-1 Kanokoden, Chikusa-ku, 464-8681 Nagoya, Japan.

[40]Division of Cancer Epidemiology, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, 466-8550 Nagoya, Japan.

[41]Division of Cancer Information and Control, Aichi Cancer Center, Japan

[42]Division of Descriptive Cancer Epidemiology, Nagoya University Graduate School of Medicine, Nagoya University, Japan.

[43]Hong Kong Hereditary Breast Cancer Family Registry, Cancer Genetics Centre, 4 A Kung Ngam Village Road, Happy Valley, Hong Kong.

[44]Department of Surgery, The University of Hong Kong, 102 Pokfulam Road, Pok Fu Lam, Hong Kong.

[45]Department of Surgery, Hong Kong Sanatorium and Hospital, 2 Village Rd, Happy Valley, Hong Kong.

[46]Department of Pathology, Hong Kong Sanatorium and Hospital, 2 Village Rd, Happy Valley, Hong Kong.

[47]Department of Medicine, Division of Oncology, Stanford Cancer Institute, Stanford University School of Medicine, 780 Welch Road, Suite CJ250C, Stanford, 94304 CA, USA.

[48]Department of Epidemiology and Population Health, Stanford University School of Medicine, 259 Campus Drive, Stanford, 94305 CA, USA.

[49]Division of Epidemiology, Center for Public Health Sciences, National Cancer Center, 5-1-1 Tsukiji, Chuo-ku, 104-0045 Tokyo, Japan.

[50]Department of Preventive Medicine, Chonnam National University Medical School, Hwasun, Korea.

[51]Jeonnam Regional Cancer Center, Chonnam National University Hwasun Hospital, Hwasun, Korea.

[52]Department of Public Health Sciences, and Cancer Research Institute, Queen's University, 10 Stuart Street, Kingston, K7L 3N6 ON, Canada.

[53]Cancer Control Research, BC Cancer, 675 West 10th Avenue, Vancouver, BC, Canada.

[54]School of Population and Public Health, University of British Columbia, 2329 West Mall, Vancouver, BC, Canada.

[55]Healthy Longevity Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117545, Singapore.

[56]Singapore Institute for Clinical Sciences, Agency for Science Technology and Research (A*STAR), Singapore 117609, Singapore

[57]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore.

[58]Division of Cancer Control and Population Sciences, UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA, USA

[59]Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA.

[60]Health Services and Systems Research, Duke-NUS Medical School.

[61]Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing, China

[62]Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Korea.

[63]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

[64]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, 2 Worts' Causeway, CB1 8RN, Cambridge, UK.

[65]Genomics Center, CHU de Québec-Université Laval Research 2705 Blvd Laurier Québec (Québec) G1V 4G2, Canada.

[66]Sime Darby Medical Centre, 1 Jalan SS12/1A, Subang Jaya, 47500 Selangor, Malaysia.

[67]Department of Surgery, Faculty of Medicine, University of Malaya, Jalan Universiti, Kuala Lumpur 50630 Kuala Lumpur.

[68]Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore.

[69]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, 12 Science Drive 2, #10-01, 117549 Singapore.

Abstract

Purpose: Non-European populations are under-represented in genetics studies, hindering clinical implementation of breast cancer polygenic risk scores (PRS). We aimed to develop PRSs using the largest available studies of Asian ancestry and to assess the transferability of PRS across ethnic subgroups. **Methods:** The development dataset comprised 138,309 women from 17 case-control studies. PRSs were generated using a clumping+thresholding method, lasso penalized regression, Empirical Bayes approach, a Bayesian polygenic prediction approach or linear combinations of multiple PRSs. These PRSs were evaluated in 89,898 women from three prospective studies (1,592 incident cases). **Results:** The best performing PRS (genome-wide set of single-nucleotide polymorphism (SNPs)) had a hazard ratio (HR) per unit standard deviation (SD) of 1.62 (95% CI = 1.46-1.80), and an area under the receiver operating curve (AUC) of 0.635 (95%CI = 0.622–0.649). Combined Asian and European PRSs (333 SNPs) had a HR per SD of 1.53 (95%CI: 1.37-1.71) and AUC of 0.621 (95%CI: 0.608-0.635). The distribution of the latter PRS was different across ethnic subgroups, confirming the importance of population-specific calibration for valid estimation of breast cancer risk. **Conclusion:** PRSs developed in this study, from association data from multiple ancestries, can enhance risk stratification for women of Asian ancestry.

Introduction

Genetic inheritance is an important risk factor for breast cancer[1]. Rare pathogenic variants in several susceptibility genes, including *BRCA1*, *BRCA2* and *PALB2*, confer increased risks of breast cancer[2]; however, much of the genetic variation in risk is polygenic, due to a combination of large numbers of genetic variants each conferring a small increase in risk. The

effects of these variants can be summarized as polygenic risk scores (PRS)[3,4]. Mavaddat *et al*[3] developed and validated a 313 variant breast cancer PRS (PRS-313), using data from European-ancestry women in the Breast Cancer Association Consortium (BCAC)[4,5]. The lifetime risk of breast cancer was estimated to be 2.6% for women in the lowest 1% of the PRS-313 distribution and ~32% for women in the highest 1%; the latter group would be classified as at high-risk of developing breast cancer according to the National Institute for Health and Care Excellence (NICE) and other clinical management guidelines.[3] This demonstrates the potential of PRS to improve quantification of risk and consequently optimize breast cancer screening and prevention strategies[6].

Non-European populations are under-represented in genetic studies and this could limit PRS adoption and applicability[7-9], and exacerbate health disparities[10]. This is important for ethnic minorities in high income countries, where clinical evaluation of the European 313 variants PRS is already underway, but perhaps more so in low- and middle-income countries, where there is an urgent need to develop breast cancer screening strategies to address rapidly rising breast cancer incidence and high breast cancer mortality[11].

Asians constitute more than half of the world's population and are facing a dramatic increase in breast cancer incidence[12,13], but make up only 15% of participants in the breast cancer genome-wide association studies (GWAS). Efforts to develop breast cancer PRS specifically for Asian populations have so far been limited. In our previous work, we showed that PRS-313, developed in Europeans, was predictive of breast cancer risk in Asian populations, although the effect size was somewhat smaller than that reported in European populations[14]. However, an important outstanding question is whether a more predictive PRS utilizing Asian data can be developed. Thus far, the largest study to attempt this involved 23,372 women of

Asian ancestry. This study evaluated previously published breast cancer risk single-nucleotide polymorphisms (SNPs) and took forward SNPs that were significantly associated with breast cancer risk in Asians (p-value < 0.05) for PRS derivation, resulting in a 44-SNP PRS (PRS-44)[15]. Although predictive, we have shown in our previous work that the discriminatory power of PRS-44 (area under the receiver operating curve (AUC) = 0.586) was much lower than PRS-313 (AUC = 0.617), derived from European ancestry women, for predicting breast cancer risk in Asian women[14].

In this study, our objectives were twofold: (1) to develop improved breast cancer PRSs utilizing data from Asian populations and to validate their performance in prospective cohorts using the largest available breast cancer genetic study of Asian ancestry; (2) to assess the transferability of PRSs across Asian ethnic subgroups.

Materials and Methods

Study populations

The study population was divided into training, validation and testing datasets. The training datasets included (a) set 1 which comprised of 22,013 invasive cases and 22,114 controls of East Asian ancestry from studies participating in Breast Cancer Association Consortium (BCAC) and Asia Breast Cancer Consortium (ABCC) (where GWAS summary statistics of SNPs significant up to p-value < 0.0001 were available), (b) set 2 which comprised of 16,680 invasive cases and 83,414 controls of East Asian ancestry from studies participating in BCAC together with Biobank Japan (BBJ) (where GWAS summary statistics were available), and (c) set 3 which comprised of 122,977 invasive cases and 105,974 controls of European ancestry participating in BCAC[4] (where GWAS summary statistics were available). The validation dataset comprised

of (a) 6,392 cases invasive cases and 6,638 controls of Chinese or Malay ancestry, and (b) 585 invasive cases and 1,018 controls of Indian ancestry participating in two multi-ethnic case-control studies: the Malaysian Breast Cancer Genetics study and the Singapore Breast Cancer Cohort study. The testing dataset comprising 89,898 women (1,595 incident cases) from three prospective cohorts of East Asian ancestry: the Singapore Chinese Health Study[16] , the China Kadoorie Biobank[17] and the Korean Cancer Prevention Study Biobank[18]. Table S1 summarises the study design, genotyping arrays and the sample size in each study. Genotype calling, quality control procedures and imputation methods have been described previously[4,19-23]. Ancestry informative principal components (PC) were available for Asian ancestry samples in the BCAC and validation datasets, generated using methods as previously described[24]. See Supplementary Methods for more details.

All studies were approved by the relevant institutional ethics committees and review boards, and all participants provided written informed consent.

Statistical methods

Polygenic risk scores (PRS) were given by:

$$PRS = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \cdots + \beta_m x_m$$

where x_k is the allele dosage for SNP k, β_k is the corresponding weight, and m is the total number of SNPs. PRSs were standardised to have unit standard deviation (SD) in the control subjects. Logistic regression models, adjusted for the first 10 PCs and study, were used to estimate odds ratios (ORs) for association between the standardised PRSs and breast cancer

risk in the validation set. The studies in the validation set were genotyped in two batches and hence treated as different strata for the purposes of adjustment. Cox proportional hazard model, adjusted for the first 2 PCs for SCHS and KCPS-II and the first 12 PCs for CKB, was used to estimate hazard ratios per SD (HR_{perSD}) for the association between the PRS and breast cancer risk in the test set. The discrimination of PRS were assessed using AUC. The HR_{perSD} and AUC were obtained individually for each study and combined using a fixed-effect meta-analysis. Test of heterogeneity between studies were obtained using *rma()* command in the *metafor* package in R[26].

The approaches for SNPs selection to be included in PRS and the corresponding weights are described in subsequent sections. Figure 1 and Figure S1 summarises the methods and dataset. The lists of SNPs and the weights for the PRS computation are given in Table S2-4

Clumping and thresholding approach (C+T)

Training dataset 1 was used in these analyses. SNPs clumping (within 1Mb windows) was conducted to remove highly correlated SNPs (pairwise correlation $r^2 > 0.9$); the SNP with the lowest p-value for association in the correlated pairs was retained, resulting in 3,050 SNPs. SNPs were further clumped within pre-specified clumping window sizes and threshold of a correlation r^2. PRSs were then computed using the subset of SNPs that were significant at a pre-specified p-value threshold (set at $5x10^{-8}$ and then increased in steps of 10^{-10} up to 10^{-3}). The PRS with the highest AUC in the validation dataset was selected as the best PRS. The clumping and derivation of PRSs were done using PRSice v2.11[26], while the AUCs for PRSs were generated using the pROC package in *R*.

To account for the joint effect of SNPs used to derive the best PRS, we computed the optimal weight, from the summary statistics, for SNP *j* using the following formula:

$$\gamma_j = {\gamma'}_j \Big/ \sqrt{2p_j(1-p_j)} \qquad (1)$$

where $\underline{\gamma'} = R^{-1}\underline{\beta'}$, R is the correlation matrix between the SNP genotypes, β' is the predicted normalised marginal effect sizes of the SNPs, and p_j is the effect allele frequency of SNP j (see Supplementary Methods).

Lasso penalized regression

All 3,050 SNPs described in C+T section were included in these analyses, together with genotype data from Asian controls in BCAC OncoArray studies for calculating linkage disequilibrium among SNPs. The analyses were run using the package *lassosum* in R[27] across different values of the penalty and shrinkage parameters, and the PRS giving the highest correlation between PRS and disease status (default metric in the method) in the validation dataset was selected.

Linear combination of European PRS with Asian PRS

Of the 313 SNPs included in PRS developed for European women[3], only 287 SNPs with imputation info score > 0.9 in validation dataset were retained for subsequent analyses. Reported weights[3] were used to derive the European PRS (hereafter denoted as PRS_{287_EUR}). Asian PRSs generated from C+T or lasso penalized regression were linearly combined with $PRS_{287\_EUR.}$ The relative contribution of each PRS were estimated by logistic regression using the validation dataset.

Re-weighting of European-based PRS

We considered two sets of weights for PRS derivation using the 287 SNPs: (i) Asian weights estimated from the training dataset 1, taking into account the correlation between SNPs using

Equation (1) (hereafter denoted as PRS_{287_ASN}); and (ii) weights based on a combination of the Asian and European weights using an Empirical Bayes approach (hereafter denoted as PRS_{287_EB}), where the optimal weight is given by

$$\beta_{j,EB} = {\beta_{jA,EB}}\Big/{\sqrt{2p_j(1-p_j)}}.$$

Here, $\beta_{jA,EB}$ is the estimated posterior effect sizes in Asians given the data and p_j is the allele frequency for SNP j (see Supplementary Methods). Other approaches to combine European and Asian-specific weights were also explored, including fixed effect meta-analysis, but only the method that gave the best AUC is presented here.

We also considered linear combinations of the re-weighted European PRSs with Asian PRSs generated from C+T method or lasso penalized regression (as described above).

Bayesian polygenic prediction approach (PRS-CSx)

Training sets 2 and 3 were used as training datasets for PRS-CSx[28] together with Asians and Europeans in the 1000 Genomes Phase 3 project as LD reference panels[29]. PRSs generated using European- (hereafter denoted as PRS_{GW_EUR}) and Asian-specific posterior weights (hereafter denoted as PRS_{GW_ASN}) were linearly combined (hereafter denoted as PRS_{GW_EUR}+ PRS_{GW_ASN}) in the validation dataset. The analyses were repeated across a range of global shrinkage parameter (ϕ) and the ϕ that gave the linear combination of PRSs with the highest AUC in the validation dataset was selected as the optimal ϕ. Analyses were run using the published Python code-based tool in Github[27].

PRSs for the South Asian population

The predictive performance of PRSs developed for East Asian-ancestry women in Indian-ancestry women were assessed using AUC and OR per SD. Given the much smaller sample size for Indian-ancestry women, we did not attempt to generate a South Asian-specific PRS, but we considered estimating the weights in the linear combinations of multiple PRSs using the South Asian validation dataset.

Absolute risk of breast cancer by PRS percentiles

The age-specific absolute risks of developing breast cancer in each PRS percentile were obtained by constraining to the incidence of overall population breast cancer incidence (see Supplementary Methods). The details of these methods have been described previously[3]. We calculated lifetime and 10-year absolute risks using Singaporean mortality and breast cancer incidence in 2017[30,31]. For birth-cohort specific incidences, age-specific breast cancer incidences for the 1960-1969 and 1970-1979 birth cohorts were calculated using data on breast cancer incidence in Singapore from 1968 to 2017[30]. For women born between 1980-1989, incidences could only be calculated up to age 35, hence breast cancer incidences were projected by assuming an annual increase in breast cancer incidence of 3.9%[32].

Results

Genetic diversity within Asian populations

Figure 1 summarises the dataset and methods used in this study. The populations are clustered, consistent with geography and population history, with the Chinese-ancestry women (Malaysia/Singapore/mainland China/Hong Kong/Taiwan) form a distinct cluster that is genetically closer to Japanese/Koreans women than to Indian-ancestry women (Figure 2(a)). The Malay-ancestry women from Malaysia/Singapore are genetically closer to Chinese-

ancestry women than to Indian-ancestry women. Given the large genetic distance between Indian-ancestry women from the other populations, the primary validation dataset was based on Chinese-ancestry and Malay-ancestry women, and Indian-ancestry women were evaluated separately.

PRSs developed using Asian-specific SNPs

For C+T, SNPs were removed if they were within 250kb of a SNP already selected and correlated at $r^2 > 0.1$, leaving 1,326 SNPs for analysis. For East Asian-ancestry women, the best PRS was obtained at a p-value threshold of 5.74×10^{-7}, resulting in a 46-SNP PRS (PRS_{46}) (Figure S2), with OR per SD (OR_{perSD}) (95%CI) of 1.35 (1.30–1.39; AUC = 0.586) (Table 1). Other combinations of clumping size and correlation threshold r^2 did not result in PRSs that showed appreciable improvement (Figure S3).

For lasso penalised regression, the best PRS was obtained at penalty parameter (λ) = 0.014 and shrinkage parameter (s) = 0.9, resulting in a PRS that included 2,985 SNPs (PRS_{2985}) (Figure S4), with OR_{perSD} (95%CI) of 1.41 (1.36–1.46; AUC = 0.596), slightly more predictive than the PRS_{46} (Table 1).

Linear combinations of European PRS and Asian PRSs

Combining PRS_{287_EUR} and PRS_{46} (OR_{perSD} (95%CI) = 1.54 (1.49-1.60); AUC = 0.623) markedly improved the predictive accuracy in East Asian-ancestry women, as compared to using the Asian-specific PRSs alone (Table 1). The improvement was marginal compared to using PRS_{287_EUR} alone (OR_{perSD} (95%CI) = 1.50 (1.45-1.56); AUC = 0.615), but relative contribution of PRS_{46} to the linear combination model was approximately 30% (Table S5). Combining PRS_{287_EUR} with PRS_{2985} further increased the OR per SD and AUC compared to PRS_{46}+PRS_{287_EUR}.

PRSs developed by integrating Asian weights into the European PRS

For East Asian-ancestry women, PRS_{287_EB} (OR_{perSD} (95%CI) = 1.53 (1.47–1.58); AUC = 0.620) was slightly more predictive than PRS_{287_ASIAN} (OR_{perSD} (95%CI) = 1.50 (1.45–1.56); AUC = 0.615) and PRS_{287EUR} (OR_{perSD} (95%CI) = 1.50 (1.45–1.56); AUC = 0.614), and markedly more predictive than PRS_{46} and PRS_{2985} (Table 1). A linear combination of PRS_{287_EB} with PRS_{46} further improved the PRS performance compared to PRS_{46}+ PRS_{287_EUR}.

Continuous shrinkage PRSs (PRS-CSx)

The best combined PRS for East Asian ancestry women was obtained at $\phi = 10^{-4}$ (Table S6), with $OR_{per\ D}$ (95%CI) of 1.62 (1.52-1.68) and AUC of 0.636 for PRS_{GW_EUR} + PRS_{GW_ASN}, markedly better than all the PRSs described thus far (Table 1). This improvement was mainly driven by the contribution of PRS_{GW_EUR} (OR_{perSD} (95%CI) = 1.59 (1.53-1.65); AUC = 0.629). The OR per SD (95%CI) and AUC for PRS_{GW_ASN} alone was 1.44 (1.39-1.49) and 0.601, respectively, only slightly better than PRS_{46} (Table S6).

PRSs for Indian-ancestry population

The East Asian-ancestry women derived PRSs (as shown in Table 1) were all predictive of risk in South Asian-ancestry women but the OR_{perSD} were reduced compared to East Asian-ancestry women. While linear combination of Asian-based and European-based PRSs improved the PRS performance compared to individual PRSs in East Asians, the improvement in women of South Asian ancestry was observed only when PRS_{2985} was considered in the linear combination (Table 2). There was no improvement in the effect sizes when European-based PRS was combined with PRS_{46}. Whereas incorporating Asian weights via the EB approach improved the performance of PRSs in East Asians, there was no improvement in

performance in women of South Asian ancestry. Re-estimating the weights of the combined models using South Asian-ancestry women in the validation dataset did not lead to an appreciable difference in predictive performance (Table S7).

Evaluation of PRSs in prospective cohorts

The predictive performance of PRSs in the East Asian-ancestry women was replicated in the prospective cohorts (Table 1). Thus, the effect sizes were smallest for PRS based on Asian data alone (HR_{perSD} (95%CI) = 1.40 (1.25-1.56) for PRS_{46} and 1.45 (1.31-1.61) for PRS_{2985}), larger for PRS based on the European PRS (HR_{perSD} (95%CI) = 1.50 (1.35-1.65) for PRS_{287_EB}) and larger still for PRS based on combining the Asian and European PRS (HR_{perSD} (95%CI) = 1.53 (1.37-1.71) for PRS_{46}+PRS_{287_EB}). As in the validation dataset, PRS generated using PRS-CSx showed the strongest association with breast cancer risk (HR_{perSD} (95%CI) = 1.62 (1.46-1.80)) and highest AUC (0.635). There was no evidence of heterogeneity in the HRs among studies for any PRS (Figure S5).

Absolute breast cancer risk predictions

We used PRS_{46}+PRS_{287_EB} to demonstrate the potential of translating PRS into clinical tool for Asian population. Based on East Asian-ancestry women in the validation dataset, the estimated breast cancer ORs (95%CI) for women in the lowest and highest 1% of the PRS distribution were 0.53 (0.33-0.82) and 3.01 (2.25-4.06), respectively, compared to middle quintile. The estimated ORs did not differ from those predicted under a theoretical polygenic model in which the log OR increases linearly with the PRS (Table S8). The corresponding lifetime risks of developing breast cancer by age 80 years, on current incidence rates, were ~2% and ~19% respectively (Figure 3(a)). Assuming that a 10-year absolute risk threshold of 2.3%[35] is used to define women at sufficient risk to justify screening, approximately 12% of

Chinese women would reach the risk threshold before or at age 40 (Figure 3(b)). Figure S6 shows the distribution of the 10-year absolute risk at age 40 for women who were born between 1980-1989 using projected incidence rates (see Methods). It is projected that the proportion of women who would reach the risk threshold would rise to 29%.

Generalisability of PRS across Asian ethnic subgroups

We demonstrate the generalisability of PRS across Asian ancestry population using the three ethnic groups in the validation set and PRS_{46}+PRS_{287_EB} as an example. This combined PRS was predictive of risk in all ethnic groups, with the effect size higher in Chinese-ancestry women compared to Malay and Indian-ancestry women (OR_{perSD} (95%CI) = 1.56 (1.50-1.63) for Chinese versus 1.51 (1.39-1.64) for Malays and 1.49 (1.33-1.66) for Indians, heterogeneity p-value = 0.983) (Figure S7 and Table S9). The PRS distribution was, however, different among the three ethnic groups. While there was only a marginal difference in the SD, the means differed markedly, being highest in Chinese and lowest in Indians (mean (SD) in Chinese, Malay, and Indian controls were -0.118 (0.439), -0.197 (0.556), and -0.328 (0.455), respectively, p-values for pair-wise comparison of means < 0.0001) (Table S9). Figure 3(c) showed that if the Chinese PRS distribution was applied to Indians without adjustment, the 95th percentile in Indians corresponds, approximately, to the 90th percentile in the Chinese population, resulting in underestimation of risk in Indian women. The difference in the PRS distributions is even more apparent when women of European ancestry is used as reference (Figure 3(d)).

The patterns of PRS distribution by population (Figure 2b) are mirrored in the genetic clusters shown in Figure 2(a). The largest differences in the means of the standardised PRS_{46}+PRS_{287_EB}

were observed between the Indian-ancestry women and Japanese/Korean women (with Indians being the biggest outlier).

Discussion

Personalised risk stratification for prevention and early detection of breast cancer has gained increasing interest; however, it is important to recognize the need to study women representing diverse ancestries, to lessen health disparities. Our study provides essential information about the utility of PRSs for breast cancer risk prediction in women of Asian ancestry. We developed and validated different PRSs for East Asian ancestry women: the key observations were (a) PRSs generated by integrating information from European ancestry and Asian ancestry GWAS datasets performed better than PRSs based purely on weights derived from single-ancestry GWAS data, and (b) there were substantial differences in PRS distributions across ethnic groups.

Based on the largest available breast cancer GWAS datasets, the best PRS for East Asian-ancestry women was based on PRS-CSx approach[28] (PRS_{GW_ASN}+PRS_{GW_EUR}). This PRS had a notably larger effect size than the European PRS (PRS_{287_EUR}) that we had previously shown to be the best breast cancer PRS for women of Asian ancestry[14] (HR_{perSD} in prospective cohorts: 1.62 versus 1.46; Table 1). It is noteworthy that the predictive performance of this PRS was similar to that achieved in European populations (HR_{perSD} (95%CI) of 313-SNP PRS: 1.59 (1.54-1.64) as reported in *Mavaddat et al*[3]). However, despite the rapid drop in cost associated with next-generation sequencing, implementation of PRS comprising ~1 million SNPs can be practically more challenging compared to the implementation of the European PRS that included only 313 variants.

We showed that adaptions based on the European 313-SNP PRS can improve risk prediction in women of East Asian ancestry. First, incorporating SNPs in identified in the Asian populations (PRS46) improved predictive power. This approach of linearly combining PRSs may reduce the gap in prediction accuracy between European and non-European populations as described previously[33]. Second, incorporating Asian weights further improved predictive power (PRS_{46}+PRS_{287_EB}), but to a lesser extent. The 313-SNP PRS is being used in several clinical studies in European populations, including the MyPeBs[8] and WISDOM[7] trials, and the PRS_{46}+PRS_{287_EB} PRS would be relatively easy to implement in clinical settings.

The PRS generated for women of East Asian ancestry were also predictive for women South Asian ancestry, but the effect sizes were smaller. When combining East Asian-derived genome-wide PRS with European-derived genome-wide PRS in women of South Asian ancestry using the PRS-CSx approach, it was noticeable that the East Asian component made a smaller contribution to the linear combination (relative contribution ~ 14%, Table S5). These results demonstrate the need for larger studies of women of South Asian ancestry both to optimize the PRS and validate in prospective cohorts.

One of the challenges of moving PRS into clinical implementation is transferability across different ethnic groups. Several studies have evaluated the population-level applicability of European PRSs to non-European populations for various diseases[34-37]. Similar to these studies, we showed that the mean of the PRS distribution differ substantially between European and Asian ethnic subgroups. We showed that if the European PRS ($PRS\text{-}287_{EUR}$) was applied to an Asian population without adjustment, the 60th percentile in Chinese ancestry and Malay-ancestry women and 80th percentile in Indian-ancestry women corresponds, approximately, to the 90th percentile in the European population, resulting in overestimation of risk in these

women (Figure 3(d)). To our knowledge, no studies thus far have looked at the transferability of breast cancer PRS within diverse Asian ethnic subgroups. Our results showed that while the effect sizes appeared to be similar across ethnic groups (Table S8), the mean PRS distribution differed substantially across Asian populations (Table S7 and Figure 2(b)). For example, even though Japanese, Koreans and Han Chinese are conventionally classified as East Asians in genetic analyses, the means PRS were markedly different between these ethnic groups (Figure 2(b)). The differences are sufficiently large to affect risk classification, so comparing the PRS for an individual woman with the correctly calibrated ethnic-specific distribution is crucial for valid risk prediction. This however can be problematic for admixed individuals, where the genomes composed from multiple ancestries that may be closely or distantly related to the reference population. As more samples of Asian ancestry become available, it may be possible to combine ethnic-specific PRSs with ancestry components to derive better multi-ethnic PRSs[32].

Our work is subject to several limitations. Firstly, although we have demonstrated that the predictive performance of European PRS can be improved by integrating weights from Asians using an empirical Bayes approach, the absolute increase in predictive accuracies is marginal. Secondly, our studies focus on developing PRS without using individual-level training data. When such data is available, it may be possible to develop PRS with higher accuracy using methods that fit all variants simultaneously, such as the step-wise hard-thresholding method as described in Mavaddat *et al*[3], or considering subtype-specific disease analyses to retain more informative variants. Thirdly, our results showed that PRSs developed using Asian-derived GWAS dataset had significantly poorer performance compared to the European PRS indicating that further improvement is likely to require much larger Asian discovery dataset. Finally, PRSs were linearly combined using the validation dataset and hence the reported

performance is likely subject to overfitting. Although we have shown that performance of the combined PRSs in East Asians were replicated in the prospective cohorts, we did not have a similar independent dataset for South Asian women for such replication.

In summary, we have shown that genome-wide PRS derived from trans-ancestry method had significantly higher predictive accuracy for women of Asian ancestry than existing breast cancer PRSs. We also showed that European-based PRS can be improved for use in Asian populations by integrating population-specific weights and combined with Asian-specific PRS. Importantly, the differences in distribution of the same PRS across different ethnic groups (among Asians, and between Asian and Europeans) emphasise the need for ethnic-specific calibration before translating PRS into practice for diverse Asian populations.

Data availability

Summary statistics (odds ratios and confidence limits) for all SNPs used in derivation of various PRSs are provided in Supplementary Table S2-S4 of the manuscript. Summary statistics of European breast cancer GWAS analysis used in this study can be accessed via BCAC website (http://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/oncoarray-and-combined-summary-result/). Summary statistics of GWAS analyses from The Biobank Japan Project can be accessed via The BioBank Japan Project website (https://pheweb.jp/pheno/BrC). Request for access to individual level data from BCAC studies can be made via the Data Access Coordinating Committee of BCAC (BCAC Coordinator: BCAC@medschl.cam.ac.uk). Request for access to the ABCC data could be requested by submission of an inquiry to Dr. Wei Zheng (wei.zheng@vanderbilt.edu).

References

1. Shiovitz, S. & Korde, L. A. Genetics of breast cancer: a topic in evolution. *Ann Oncol* **26**, 1291-1299, doi:10.1093/annonc/mdv022 (2015).
2. Dorling, L. *et al*. Breast Cancer Risk Genes - Association Analysis in More than 113,000 Women. *N Engl J Med*. **4**;384(5):428-439(2021).
3. Mavaddat, N. *et al*. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet* **104**, 21-34, doi:10.1016/j.ajhg.2018.11.002 (2019).
4. Michailidou, K. *et al*. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92-94, doi:10.1038/nature24284 (2017).
5. Milne, R. L. *et al*. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat Genet* **49**, 1767-1778, doi:10.1038/ng.3785 (2017).
6. Pashayan, N. *et al*. Publisher Correction: Personalized early detection and prevention of breast cancer: ENVISION consensus statement. *Nat Rev Clin Oncol* **17**, 716, doi:10.1038/s41571-020-0412-0 (2020).
7. Esserman, L. J., Study, W. & Athena, I. The WISDOM Study: breaking the deadlock in the breast cancer screening debate. *NPJ Breast Cancer* **3**, 34, doi:10.1038/s41523-017-0035-5 (2017).
8. MyPeBS Personalizing Breast Screening, <*https://mypebs.eu/the-project/*>
9. Personalized risk assessment for prevention and early detection of breast cancer: Integration and Implementation, <*https://www.genomecanada.ca/en/personalized-risk-assessment-prevention-and-early-detection-breast-cancer-integration-and*>
10. Martin, A. R. *et al*. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584-591, doi:10.1038/s41588-019-0379-x (2019).
11. Youlden, D. R., Cramb, S. M., Yip, C. H. & Baade, P. D. Incidence and mortality of female breast cancer in the Asia-Pacific region. *Cancer Biol Med* **11**, 101-115, doi:10.7497/j.issn.2095-3941.2014.02.005 (2014).
12. Bhoo-Pathy, N. et al. Breast cancer research in Asia: adopt or adapt Western knowledge? *Eur J Cancer* **49**, 703-709, doi:10.1016/j.ejca.2012.09.014 (2013).
13. Heer, E. et al. Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study. *Lancet Glob Health* **8**, e1027-e1037, doi:10.1016/S2214-109X(20)30215-1 (2020).
14. Ho, W. K. et al. European polygenic risk score for prediction of breast cancer shows similar performance in Asian women. *Nat Commun* **11**, 3833, doi:10.1038/s41467-020-17680-w (2020).
15. Wen, W. *et al.* Prediction of breast cancer risk based on common genetic variants in women of East Asian ancestry. *Breast Cancer Res* **18**, 124, doi:10.1186/s13058-016-0786-1 (2016).
16. Hankin, J. H. *et al.* Singapore Chinese Health Study: development, validation, and calibration of the quantitative food frequency questionnaire*. Nutr Cancer* **39**, 187-195, doi:10.1207/S15327914nc392_5 (2001).
17. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol **40**, 1652-1666, doi:10.1093/ije/dyr120 (2011).
18. Jee, Y. H. *et al*. Cohort Profile: The Korean Cancer Prevention Study-II (KCPS-II) Biobank. *Int J Epidemiol* 47, 385-386f, doi:10.1093/ije/dyx226 (2018).
19. Michailidou, K. *et al*. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45**, 353-361, 361e351-352, doi:10.1038/ng.2563 (2013).
20. Shu, X. *et al.* Identification of novel breast cancer susceptibility loci in meta-analyses conducted among Asian and European descendants. *Nat Commun* **11**, 1217, doi:10.1038/s41467-020-15046-w (2020).

21. Dorajoo, R. *et al.* Loci for human leukocyte telomere length in the Singaporean Chinese population and trans-ethnic genetic studies. *Nat Commun* **10**, 2491, doi:10.1038/s41467-019-10443-2 (2019).
22. Gan, W. *et al*. Evaluation of type 2 diabetes genetic risk variants in Chinese adults: findings from 93,000 individuals from the China Kadoorie Biobank. *Diabetologia* **59**, 1446-1457, doi:10.1007/s00125-016-3920-9 (2016).
23. The BioBank Japan Project website <https://pheweb.jp/pheno/BrC>
24. Amos, C. I. *et al*. The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev* **26**, 126-135, doi:10.1158/1055-9965.EPI-16-0106 (2017).
25. Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1–48. https://doi.org/10.18637/jss.v036.i03
26. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* 8, doi:10.1093/gigascience/giz082 (2019).
27. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genet Epidemiol* **41**, 469-480, doi:10.1002/gepi.22050 (2017).
28. Github repository: PRS-CSx. <https://github.com/getian107/PRScsx>
29. Genomes Project, C. *et al*. A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
30. *Age-Specific Death Rates, Annual* (Government of Singapore, 2017).
31. Cancer Incidence in Five Continents. (International Agency for Research on Cancer, Lyon, (2014).
32. Jara-Lazaro, A.R., S. Thilagaratnam, and P.H. Tan, *Breast cancer in Singapore: some perspectives.* Breast Cancer, 2010. **17**(1): p. 23-8.
33. Márquez-Luna, C., et al. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet Epidemiol*;**41**(8):811-823 (2017).
34. Martin, A.R., *et al*. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*;**51(4)**:584-591 (2019).
35. Martin, A.R., *et al*. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet*; **6**;100(4):635-649 (2017).
36. Reisberg, S., Iljasenko, T., Lall, K., Fischer, K., Vilo, J. Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. *PLoS One* ;**12(7)**:e0179238 (2017).
37. Isgut, M., Sun, J., Quyyumi, A.A., Gibson, G. Highly elevated polygenic risk scores are better predictors of myocardial infarction risk early in life than later. *Genome Med*;**13**(1):13 (2021).

Funding and acknowledgements

Funding

This study was supported by grants from Newton-Ungku Omar Fund [grant no: MR/P012930/1] and Wellcome Trust [grant no: v203477/Z/16/Z]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The Malaysian Breast Cancer Genetic Study was established using funds from the Malaysian Ministry of Science, and the Malaysian Ministry of Higher Education High Impact Research Grant [grant no: UM.C/HIR/MOHE/06]. The Malaysian Mammographic Density Study was established using funds raised through the Sime Darby LPGA tournament and the High Impact Research Grant. Additional funding was received from Yayasan Sime Darby, PETRONAS, Estee Lauder Group of Companies and other donors of Cancer Research Malaysia.

WKH is the recipient of L'Oreal-UNESCO For Women in Science National Fellowship. JL is the recipient of a National Research Foundation Singapore Fellowship (NRF-NRFF2017-02). ACA is supported through Cancer Research - UK (C12292/A20861). JS holds a Canada Research Chair in Oncogenetics.

The PERSPECTIVE I&I project is funded by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the Ministère de l'Économie et de l'Innovation du Québec through Genome Québec, the Quebec Breast Cancer Foundation, the CHU de Quebec Foundation and the Ontario Research Fund.

BCAC is funded by Cancer Research UK [C1287/A16563, C1287/A10118], the European Union's Horizon 2020 Research and Innovation Programme (grant numbers 634935 and 633784 for BRIDGES and B-CAST respectively), and by the European Community´s Seventh Framework Programme under grant agreement number 223175 (grant number HEALTH-F2-2009-223175) (COGS). The EU Horizon 2020 Research and Innovation Programme funding source had no role in study design, data collection, data analysis, data interpretation or writing of the report.

Genotyping of the OncoArray was funded by the NIH Grant U19 CA148065, and Cancer UK Grant C1287/A16563 and the PERSPECTIVE project supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research (grant GPH-129344) and, the Ministère de l'Économie, Science et Innovation du Québec through Genome Québec and the PSRSIIRI-701 grant, and the Quebec Breast Cancer Foundation. Funding for the iCOGS infrastructure came from: the European Community's Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (C1287/A10118, C1287/A10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565), the National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 - the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer, and Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund. The DRIVE Consortium was funded by U19 CA148065.

MYBRCA is funded by research grants from the Malaysian Ministry of Higher Education (UM.C/HlR/MOHE/06) and Cancer Research Malaysia. MYMAMMO is supported by research grants from Yayasan Sime Darby LPGA Tournament and Malaysian Ministry of Higher Education (RP046B-15HTM). SGBCC is funded by NUS Start Up Grant, National University Cancer Institute Singapore (NCIS) Centre Grant, NMRC Clinical Scientist Award, NMRC Clinician Scientist Award-Senior Investigator, Asian Breast Cancer Research Fund and Breast Cancer Prevention Programme under Saw Swee Hock School of Public Health. Recruitment of controls were funded by the Biomedical Research Council, (05/1/21/19/425).

The ACP study is funded by the Breast Cancer Research Trust, UK. KM and AL are supported by the NIHR Manchester Biomedical Research Centre, the Allan Turing Institute, and, by the ICEP (Cancer Research UK (C18281/A19169). CBCS is funded by the Canadian Cancer Society (grant # 313404) and the Canadian Institutes of Health Research. The HERPACC was supported by MEXT Kakenhi (No. 170150181 and 26253041) from the Ministry of Education, Science, Sports, Culture and Technology of Japan, by a Grant-in-Aid for the Third Term Comprehensive 10-Year Strategy for Cancer Control from Ministry Health, Labour and Welfare of Japan, by Health and Labour Sciences Research Grants for Research on Applying Health Technology from Ministry Health, Labour and Welfare of Japan, by National Cancer Center Research and Development Fund, and "Practical Research for Innovative Cancer Control (15ck0106177h0001)" from Japan Agency for Medical Research and development, AMED, and Cancer Bio Bank Aichi. The KOHBRA study was partially supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), and the National R&D Program for Cancer Control, Ministry of Health & Welfare,

Republic of Korea (HI16C1127; 1020350; 1420190). LAABC is supported by grants (1RB-0287, 3PB-0102, 5PB-0018, 10PB-0098) from the California Breast Cancer Research Program. Incident breast cancer cases were collected by the USC Cancer Surveillance Program (CSP) which is supported under subcontract by the California Department of Health. The CSP is also part of the National Cancer Institute's Division of Cancer Prevention and Control Surveillance, Epidemiology, and End Results Program, under contract number N01CN25403. The Northern California Breast Cancer Family Registry (NC-BCFR) were supported by grant U01CA164920 from the USA National Cancer Institute of the National Institutes of Health. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR. The NGOBCS was supported by the National Cancer Center Research and Development Fund (Japan). The SBCGS was supported primarily by NIH grants R01CA64277, R01CA148667, UMCA182910, and R37CA70867. Biological sample preparation was conducted the Survey and Biospecimen Shared Resource, which is supported by P30 CA68485. The scientific development and funding of this project were, in part, supported by the Genetic Associations and Mechanisms in Oncology (GAME-ON) Network U19 CA148065. SEBCS was supported by the BRL (Basic Research Laboratory) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2012-0000347). The TWBCS is supported by the Taiwan Biobank project of the Institute of Biomedical Sciences, Academia Sinica, Taiwan. The BioBank Japan Project (BBJ) is supported by the Ministry of Education, Culture, Sports, Sciences and Technology from the Japanese Government); the Hwasun Cancer Epidemiology Study-Breast (HCES-Br) supported by the

Biobank of Chonnam National University Hwasun Hospital, a member of the Korea Biobank Network.

The SCHS study was supported by grants from the National Medical Research Council, Singapore (NMRC/CIRG/1456/2016) and the National Institutes of Health, USA (R01 CA144034 and UM1 CA182876). China Kadoorie Biobank was supported as follows: Baseline survey and first re-survey: Hong Kong Kadoorie Charitable Foundation; long-term follow-up: UK Wellcome Trust (212946/Z/18/Z, 202922/Z/16/Z, 104085/Z/14/Z, 088158/Z/09/Z), National Natural Science Foundation of China (91843302), and National Key Research and Development Program of China (2016YFC 0900500, 0900501, 0900504, 1303904). DNA extraction and genotyping: GlaxoSmithKline, UK Medical Research Council (MC_PC_13049, MC-PC-14135). The UK Medical Research Council (MC_UU_00017/1, MC_UU_12026/2 MC_U137686851), Cancer Research UK (C16077/A29186; C500/A16896) and the British Heart Foundation (CH/1996001/9454), provide core funding to the Clinical Trial Service Unit and Epidemiological Studies Unit at Oxford University for the project.

Acknowledgements

We thank all the individuals who took part in these studies and all the researchers, clinicians, technicians and administrative staff who have enabled this work to be carried out. The BCAC study would not have been possible without the contributions of the staff of the McGill University and Génome Québec Innovation Centre, Stig E. Bojesen, Sune F. Nielsen, Borge G. Nordestgaard, and the staff of the Copenhagen DNA laboratory, and Julie M. Cunningham, and the staff of Mayo Clinic Genotyping Core Facility.

MYBRCA thanks study participants and all research staff at Cancer Research Malaysia, University Malaya, and Sime Darby Medical Centre who assisted in recruitment and

interviews (particularly Siti Norhidayu Hassan, Patsy Pei-Sze Ng, Sook-Yee Yoon, Shivaani Mariapun, and Joanna Lim) for their contributions and commitment to this study.

For SGBCC, we want to thank the program manager Jenny Liu, clinical research coordinators/research assistants Siew-Li Tan, Siok-Hoon Yeo, Ting-Ting Koh, Amanda Ong, Jin-Yee Lee, Michelle Mok, Ying-Jia Chew, Jing-Jing Hong and Hui-Min Lau for their contributions in recruitment, Yen-Shing Yeoh for data preparation and Alexis Khng for processing the DNA samples. We also want to thank all the participants' support to SGBCC.

The ACP study wishes to thank the participants in the Thai Breast Cancer study. Special Thanks also go to the Thai Ministry of Public Health (MOPH), doctors and nurses who helped with the data collection process. CBCS thanks study participants, co-investigators, collaborators and staff of the Canadian Breast Cancer Study, and project coordinators Agnes Lai and Celine Morissette. HKBCS thanks Hong Kong Sanatorium and Hospital, Dr Ellen Li Charitable Foundation, The Kerry Group Kuok Foundation, National Institute of Health 1R03CA130065 and the North California Cancer Center for support. We thank all investigators of the KOHBRA (Korean Hereditary Breast Cancer) Study. LAABC thanks all the study participants and the entire data collection team. SBCGS thank study participants and research staff for their contributions and commitment to the studies.

SCHS thank the Singapore Cancer Registry for identification of cancer cases within their cohort. CKB acknowledges the study participants and members of the survey teams in each of the 10 regional centres, and the project development and management teams based at Beijing, Oxford and the 10 regional centres. China Kadoorie Biobank acknowledges the contribution of participants, project staff, and China National Centre for Disease Control and

Prevention (CDC) and its regional offices. China's National Health Insurance provided electronic linkage to all hospital treatments.

Authors contribution

Conceptualization: W.K.H., J.Simard, A.C.A., D.F.E., S.H.T.; Project administration: M.C.T., M.K.B, Q.W., E.A.W.; Data curation: J.D., N.M., K.Michailidou, J.Long; Formal analysis: W.K.H., M.C.T., X.S., J.Li, P.J.H., I.Y.M., K.L., Y.H.J., S.H.Lee, D.F.E; Resources: T.H., K.R., V.K.M.T, B.K.T.T., S.M.T., E.Y.T., S.H.Lim, Y.T.G., Y.Z., D.K., J.Y.C., W.H., H.B.L., M.K., Y.O., N.M., B.B.J., S.K.P., S.W.K., C.Y.S., P.E.W., B.P., K.Muir, A.L., A.H.W., C.C.T., K.Matsuo, H.I., A.K., T.L.C., E.M.J., A.W.K., M.I., T.Y., S-S.K., K.J.A., R.A.M., W.P.K., C.C.K., J.M.Y, R.D., R.G.W., Z.C., L.L., J.Lv, K.J.J, P.K, P.D.P.P., A.M.D., X.O.S., C.H.Y., N.A.M.T., W.Z., M.H., S.H.T. Supervision: D.F.E., S.H.T.; Writing – original draft: W.K.H., M.C.T., A.C.A., D.F.E., S.H.T.; Writing – review & editing: all authors.

Conflict of Interest

The authors confirm that they have no conflict of interests.

Ethics statement

The Malaysian Breast Cancer Genetic Study (MyBrCa) was approved by the Independent Ethics Committee, Ramsay Sime Darby Health Care (reference no: 201109.4 and 201208.1), and the Medical Ethics Committee, University Malaya Medical Centre (reference no: 842.9). Analyses using CKB data were conducted under research approval 2020-0047. Each study listed was approved by the local institutional ethics committees and review boards, and all participants provided written informed consent.

Figure legends

Figure 1. Overview of methods for polygenic risk scores (PRS) development. Inputs are summary statistics from meta-analysis of multiple GWAS datasets – BCAC ASN+ABCC denoted training dataset 1, BCAC ASN+BBJ denoted training dataset 2 and BCAC-EUR denoted training dataset 3 as described in the method section. LD ref: reference panel for linkage disequilibrium. LD ref: BCAC ASN denoted Oncoarray studies in BCAC Asian studies were used as reference panel; LD ref: BCAC EUR denoted BCAC studies of European ancestries were used as reference panel; 1000G ASN and 1000G EUR denoted the Asian and European samples, respectively, in 1000Genome Project. Figure 1 illustrate methods using East Asian ancestry women (Chinese and Malays) as an example, same methods were applied to South Asian ancestry women in the validation dataset.

Figure 2: Principal components analysis and mean of PRS_{46}+PRS_{287_EB} by country and ethnicity. (a) Principal component plot by country. Principal components analysis of samples genotyped with OncoArray as listed in Table S1. The samples were grouped according to country (Thailand, Taiwan, Hong Kong, China, Korea and Japan). For Malaysia and Singapore (M+S), the samples were further categorized by their self-reported ethnic origin (Chinese, Malay and Indian). (b) Mean of standardised PRS_{46}+PRS_{287_EB} in controls by country. PRS was standardised by the control SDs of each study. Error bars represent 95% CI. The mean of standardised PRS_{46}+PRS_{287_EB} in European controls were included for reference.

Figure 3: Absolute breast cancer risk by percentiles of PRS and PRS distribution by ancestry. (a) Lifetime and (b) 10-year absolute risk of developing breast cancer for Chinese women calculated using Singaporean incidence and mortality data and OR per SD of PRS_{46}+PRS_{287_EB}

in Chinese (1.56 as reported in Table S9). The grey dashed lines in the (a) and (b) represent the average lifetime risk and absolute 10-year risk, respectively, of Singaporean Chinese women. The red horizontal dashed line (2.3%) in the (b) represents the 10-year absolute risk of a 50-year old European women where screening is recommended; (c) the distribution of PRS_{46}+PRS_{287_EB} in Chinese, Indian and Malay-ancestry women, generated using ethnic-specific mean and standard deviation of controls as reported in Table S9 and the corresponding cumulative breast cancer risk by age 80, generated using calendar-specific breast cancer incidence and mortality rates for Chinese, Malay and Indian women in Singapore[36]. Area under the curves represent the percentiles of PRS_{287_EB}. The right vertical dashed line represents the 90th percentile cutoff for PRS distribution in Chinese-ancestry women; For example, the 95th percentile in Indians (lifetime risk = 11%) corresponds, approximately, to the 90th percentile in the Chinese population. If Chinese PRS distribution was used as a reference, these Indian women would be categorised as 90th percentile and hence would be told that their corresponding lifetime risk was 9% instead of 11%; (d) the distribution of European PRS (PRS_{287_EUR}) for women of European, Chinese, Malay, or Indian ancestry. The right vertical dashed line represents the 90th percentile cutoff for PRS distribution in European-ancestry women.

.

Table 1. Mean, standard deviation, and the association of polygenic risk scores (PRS) with breast cancer risk in women of East Asian ancestry

Method	PRS	Validation set[a]				Test set[b]			
		Cases Mean (SD)	Control Mean (SD)	OR per SD[†] (95% CI)	AUC	Cases Mean (SD)	Control Mean (SD)	HR per SD[*] (95% CI)	AUC[*]
[1] Clumping and Thresholding	[c]PRS_{46}	-0.387 (0.446)	-0.538 (0.443)	1.37 (1.32-1.42)	0.589	-0.299 (0.433)	-0.444 (0.438)	1.40 (1.25-1.56)	0.600
[2] Penalised regression	[c]PRS_{2985}	0.075 (0.455)	-0.082 (0.452)	1.41 (1.37-1.47)	0.598	0.107 (0.460)	-0.059 (0.458)	1.45 (1.311.61)	0.608
[3] EUR SNPs+ EUR weights	[c]PRS_{287_EUR}	0.865 (0.548)	0.640 (0.549)	1.50 (1.45-1.56)	0.615	0.876 (0.549)	0.679 (0.541)	1.46 (1.34-1.60)	0.609
[4] EUR SNPs +ASN weights	[c]PRS_{287_ASN}	-0.533 (0.445)	-0.714 (0.447)	1.50 (1.45-1.56)	0.614	-0.552 (0.448)	-0.731 (0.441)	1.49 (1.33-1.66)	0.608
[5] EUR SNPs+ EB weights	[c]PRS_{287_EB}	0.343 (0.491)	0.135 (0.492)	1.53 (1.47-1.58)	0.620	0.341 (0.493)	0.153 (0.485)	1.50 (1.35-1.65)	0.609
Combine [1] + [3]	[d]PRS_{46} + PRS_{287_EUR}	0.058 (0.440)	-0.134 (0.437)	1.54 (1.49-1.60)	0.623	0.103 (0.442)	-0.075 (0.436)	1.52 (1.36-1.70)	0.620
Combine [2] + [3]	[d]PRS_{2985} + PRS_{287_EUR}	0.062 (0.447)	-0.139 (0.444)	1.56 (1.50-1.61)	0.626	0.080 (0.454)	-0.106 (0.447)	1.54 (1.38-1.72)	0.622
Combine [1] + [4]	[d]PRS_{46} + PRS_{287_ASN}	0.052 (0.425)	-0.127 (0.423)	1.52 (1.47-1.58)	0.619	0.070 (0.425)	-0.113 (0.421)	1.52 (1.35-1.70)	0.621
Combine [2] + [4]	[d]PRS_{2985} + PRS_{287_ASN}	0.055 (0.430)	-0.130 (0.430)	1.54 (1.48-1.60)	0.621	0.057 (0.435)	-0.135 (0.427)	1.53 (1.37-1.72)	0.623
Combine [1] + [5]	[d]PRS_{46} + PRS_{287_EB}	0.061 (0.446)	-0.137 (0.443)	1.55 (1.50-1.61)	0.625	0.089 (0.447)	-0.089 (0.441)	1.53 (1.37-1.71)	0.621

Combine [2] + [5]	[d]PRS_{2985} + PRS_{287_EB}	0.063 (0.451)	-0.139 (0.449)	1.56 (1.51-1.62)	0.627	0.077 (0.455)	-0.120 (0.447)	1.55 (1.39-1.72)	0.623
[6] PRS-CSx	[d]PRS_{GW_EUR}+PRS_{GW_ASN}	0.082 (0.493)	-0.159 (0.489)	1.62 (1.52-1.68)	0.636	-0.145 (0.511)	-0.388 (0.511)	1.62 (1.46-1.80)	0.635

[a]Validation cohort which consist of 6,392 breast cancer cases and 6,638 control of Chinese- and Malay- ancestry from MyBrCa and SGBCC (Table S1).

[b]Prospective cohorts which consist of 89,898 control and 1,592 breast cancer cases from 3 prospective cohorts, Singapore Chinese Health Study (SCHS), China Kadoorie Biobank (CKB) and Korean Cancer Prevention Study-II Biobank (KCPS-II) (Table S1).

[c]PRSs were derived using 46, 2,985 and 287 selected SNPs respectively as described in the Method section.

[d] Combined PRSs were generated using the formula $\alpha_0 + \alpha_1 PRS_1 + \alpha_2 PRS_2$ where α_0, α_1 and α_2 are the weights obtained by fitting a logistic regression model with breast cancer as outcome, PRS_1 and PRS_2 as explanatory variables using the validation dataset. The weights for the considered combination of PRSs can be found in Table S5.

[†]Adjusted for first the 10 principal components and study, and standardised to SDs in controls of each PRS.

[*]Fixed effect meta-analysis of three prospective cohorts, SCHS, CKB and KCPS-II. HR per SD and AUC of individual studies can be found in Figure S5.

Table 2. Mean, standard deviation, and the association of polygenic risk scores (PRS) with breast cancer risk in women of South Asian ancestry

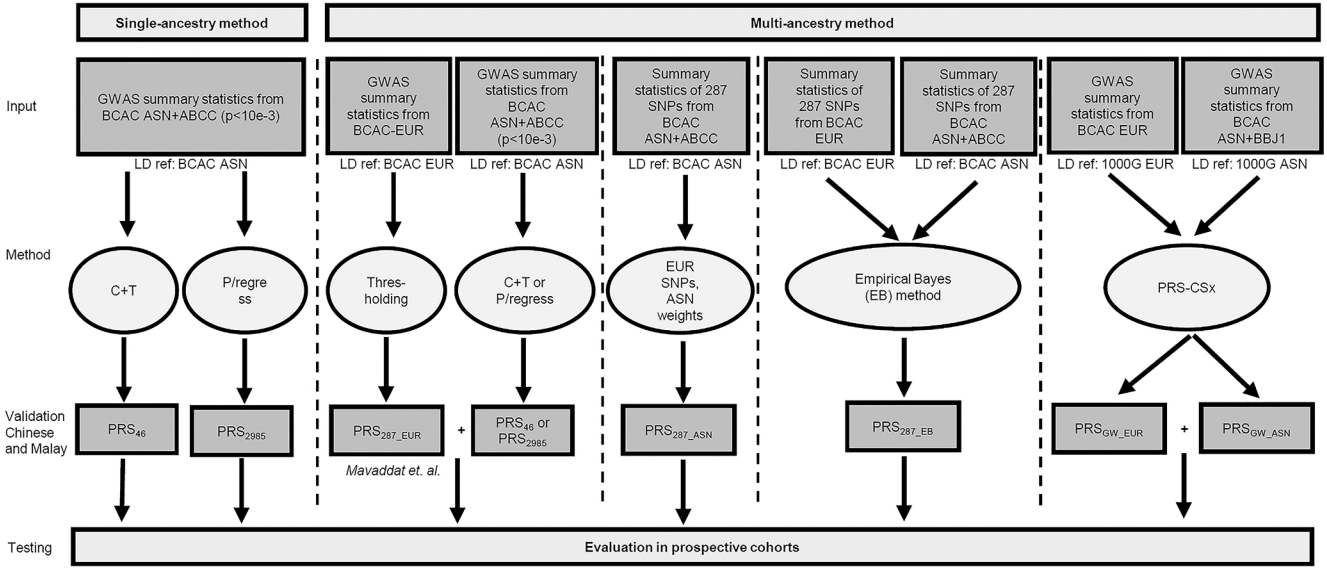
Method	PRS developed based on East Asians[a]	Validation set[b]			
		Cases Mean (SD)	Control Mean (SD)	OR per SD[†] (95% CI)	AUC
[1] Clumping and Thresholding	[a]PRS_{46}	-0.490 (0.388)	-0.548 (0.387)	1.18 (1.06-1.31)	0.546
[2] Penalised regression	[a]PRS_{2985}	0.059 (0.381)	-0.048 (0.376)	1.32 (1.19-1.46)	0.581
[3] EUR SNPs+ EUR weights	[a]PRS_{287_EUR}	0.482 (0.570)	0.251 (0.608)	1.49 (1.34-1.67)	0.614
[4] EUR SNPs +ASN weights	[a]PRS_{287_ASN}	-0.552 (0.493)	-0.720 (0.479)	1.43 (1.28-1.58)	0.592
[5] EUR SNPs+ EB weights	[a]PRS_{287_EB}	0.084 (0.521)	-0.127 (0.545)	1.50 (1.35-1.67)	0.613
Combine [1] + [3]	[c]PRS_{46} + PRS_{287_EUR}	-0.212 (0.420)	-0.376 (0.444)	1.48 (1.33-1.65)	0.611
Combine [2] + [3]	[c]PRS_{2985} + PRS_{287_EUR}	-0.166 (0.419)	-0.347 (0.441)	1.53 (1.37-1.71)	0.620
Combine [1] + [4]	[c]PRS_{46} + PRS_{287_ASN}	0.008 (0.431)	-0.135 (0.420)	1.42 (1.28-1.57)	0.591
Combine [2] + [4]	[c]PRS_{2985} + PRS_{287_ASN}	0.036 (0.425)	-0.121 (0.413)	1.46 (1.32-1.62)	0.602
Combine [1] + [5]	[c]PRS_{46} + PRS_{287_EB}	-0.157 (0.438)	-0.328 (0.455)	1.49 (1.33-1.66)	0.610
Combine [2] + [5]	[c]PRS_{2985} + PRS_{287_EB}	-0.119 (0.434)	-0.304 (0.449)	1.52 (1.37-1.70)	0.618
[6] PRS-CSx	[c]PRS_{GW_EUR}+ PRS_{GW_ASN}	-0.308 (0.501)	-0.546 (0.502)	1.62 (1.46-1.81)	0.633

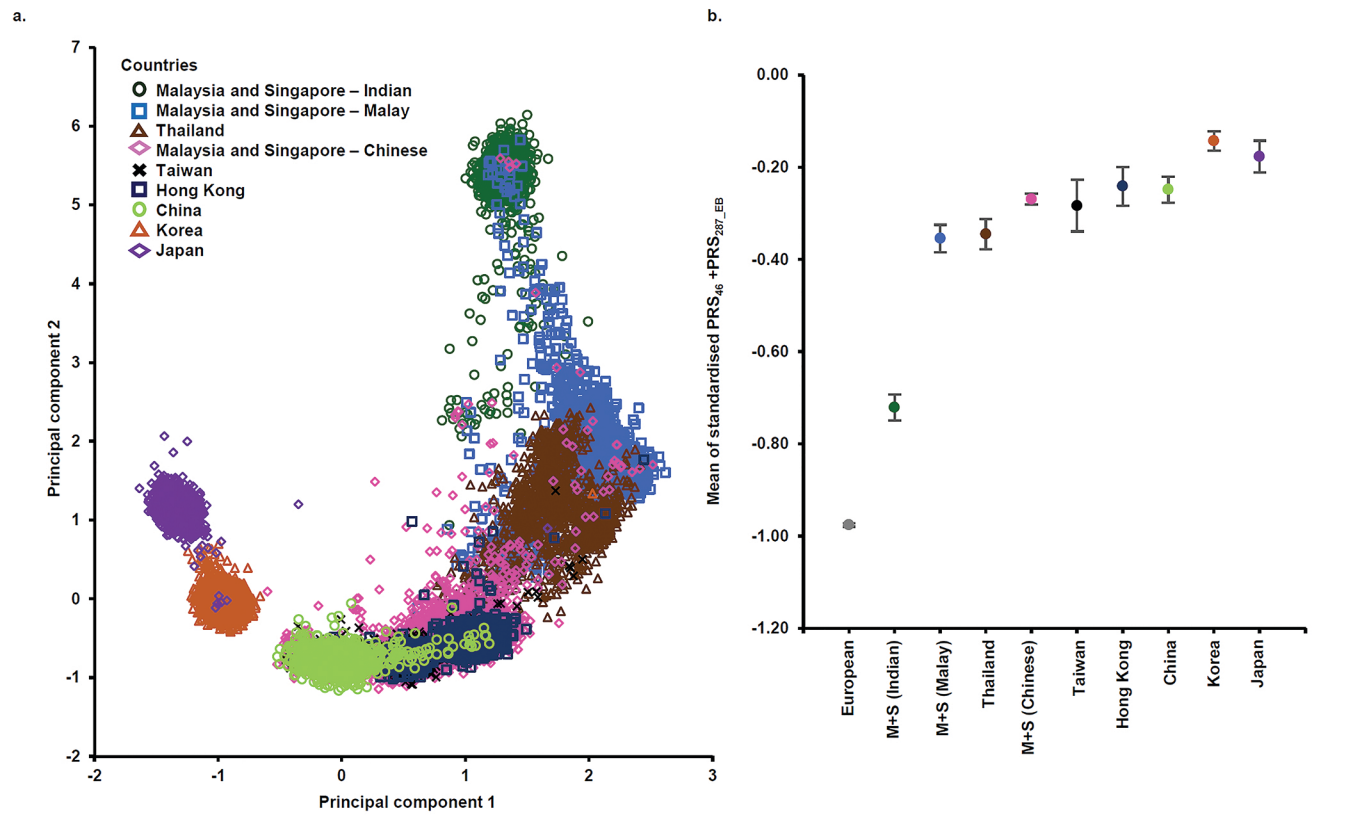
[a] PRSs developed based on Chinese and Malay-ancestry women in the validation dataset as described in Table 1. cohort from Chinese- and Malay- ancestry of MyBrCa and SGBCC as in Table 1.

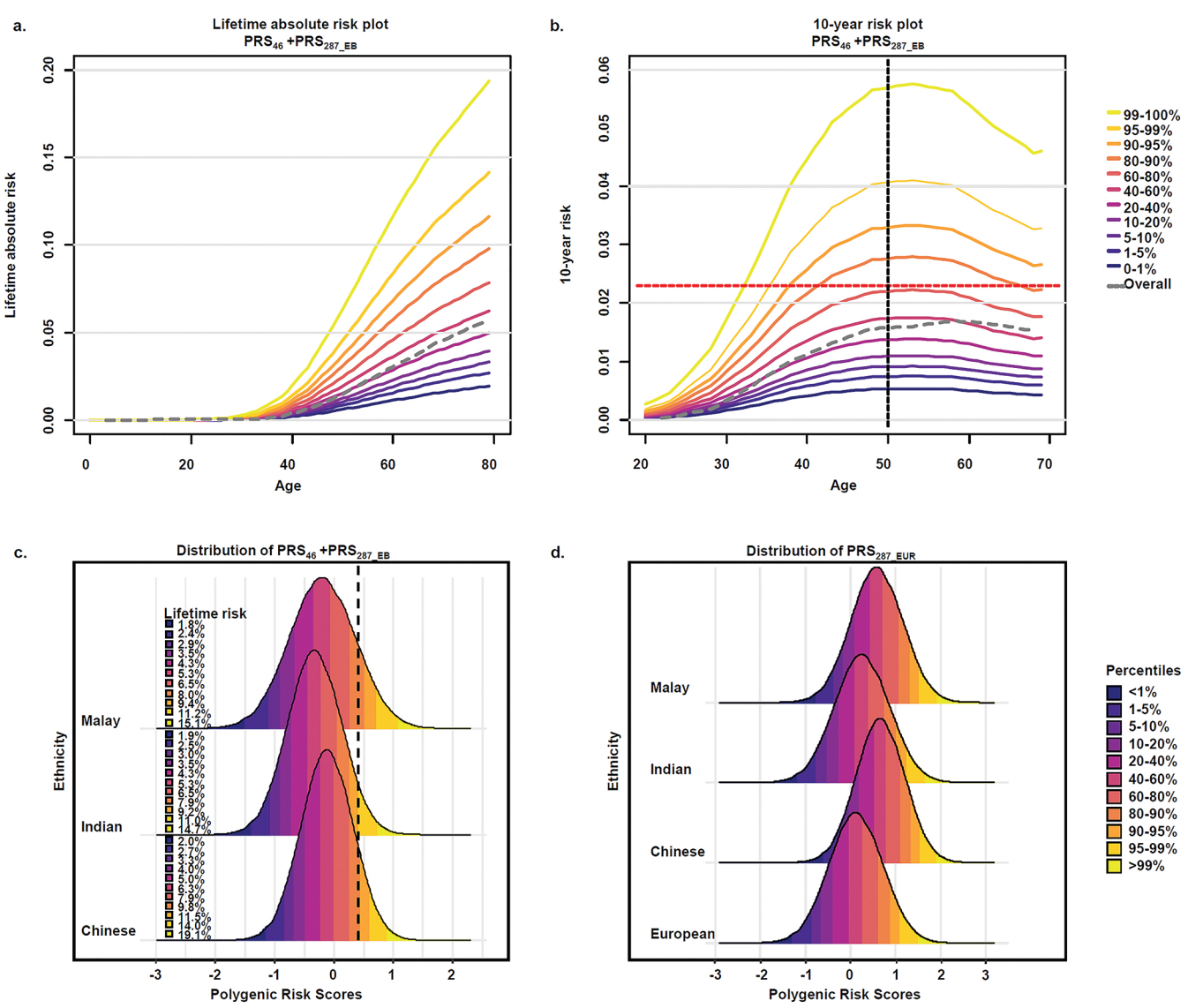
[b]Evaluation of PRSs performance in 585 breast cancer cases and 1,018 controls of Indian-ancestry women in the validation dataset (Table S1).

[c] Combined PRSs were generated using the formula $\alpha_0 + \alpha_1 PRS_1 + \alpha_2 PRS_2$ where α_0, α_1 and α_2 are the weights estimated from East Asian ancestry women as described in Table 1. The weights for the considered combination of PRSs can be found in Table S5.

[†]Adjusted for first the 10 principal components and study, and standardised to SDs in controls of each PRS.


Single-ancestry method
Multi-ancestry method
Input
Method
Validation Chinese and Malay
Testing
GWAS summary statistics from BCAC ASN+ABCC (p<10e-3)
GWAS summary statistics from BCAC-EUR
GWAS summary statistics from BCAC ASN+ABCC (p<10e-3)
Summary statistics of 287 SNPs from BCAC ASN+ABCC
Summary statistics of 287 SNPs from BCAC EUR
Summary statistics of 287 SNPs from BCAC ASN+ABCC
GWAS summary statistics from BCAC EUR
GWAS summary statistics from BCAC ASN+BBJ1
LD ref: BCAC ASN
LD ref: BCAC EUR
LD ref: BCAC ASN
LD ref: BCAC ASN
LD ref: BCAC EUR
LD ref: BCAC ASN
LD ref: 1000G EUR
LD ref: 1000G ASN
C+T
P/regress
Thres-holding
C+T or P/regress
EUR SNPs, ASN weights
Empirical Bayes (EB) method
PRS-CSx
PRS46
PRS2985
PRS287_EUR
+
PRS46 or PRS2985
PRS287_ASN
PRS287_EB
PRSGW_EUR
+
PRSGW_ASN
Mavaddat et. al.
Evaluation in prospective cohorts


a.
Countries
Malaysia and Singapore – Indian
Malaysia and Singapore – Malay
Thailand
Malaysia and Singapore – Chinese
Taiwan
Hong Kong
China
Korea
Japan
Principal component 2
Principal component 1
b.
Mean of standardised PRS46 +PRS287_EB
European
M+S (Indian)
M+S (Malay)
Thailand
M+S (Chinese)
Taiwan
Hong Kong
China
Korea
Japan


a.
Lifetime absolute risk plot
PRS46 +PRS287_EB
Lifetime absolute risk
0.00
0.05
0.10
0.15
0.20
0
20
40
60
80
Age
b.
10-year risk plot
PRS46 +PRS287_EB
10-year risk
0.00
0.01
0.02
0.03
0.04
0.05
0.06
20
30
40
50
60
70
Age
99-100%
95-99%
90-95%
80-90%
60-80%
40-60%
20-40%
10-20%
5-10%
1-5%
0-1%
Overall
c.
Distribution of PRS46 +PRS287_EB
Lifetime risk
1.8%
2.4%
2.9%
3.5%
4.3%
5.3%
6.5%
8.0%
9.4%
11.2%
15.1%
1.9%
2.5%
3.0%
3.5%
4.3%
5.3%
6.5%
7.9%
9.2%
11.0%
14.7%
2.0%
2.7%
3.3%
4.0%
5.0%
6.3%
7.9%
9.8%
11.5%
14.0%
19.1%
Ethnicity
Malay
Indian
Chinese
-3
-2
-1
0
1
2
Polygenic Risk Scores
d.
Distribution of PRS287_EUR
Ethnicity
Malay
Indian
Chinese
European
-3
-2
-1
0
1
2
3
Polygenic Risk Scores
Percentiles
<1%
1-5%
5-10%
10-20%
20-40%
40-60%
60-80%
80-90%
90-95%
95-99%
>99%

Supplementary materials

Polygenic risk scores for prediction of breast cancer risk in Asian populations

Ho et al.

Figure S1. Schema for development and validation of PRS

The development of PRSs were conducted using clumping + thresholding method, lasso penalised regression, integration of Asian weights into European PRS, linear combinations of multiple PRSs and Bayesian polygenic prediction method. All the PRSs were subsequently evaluated in the prospective cohorts.

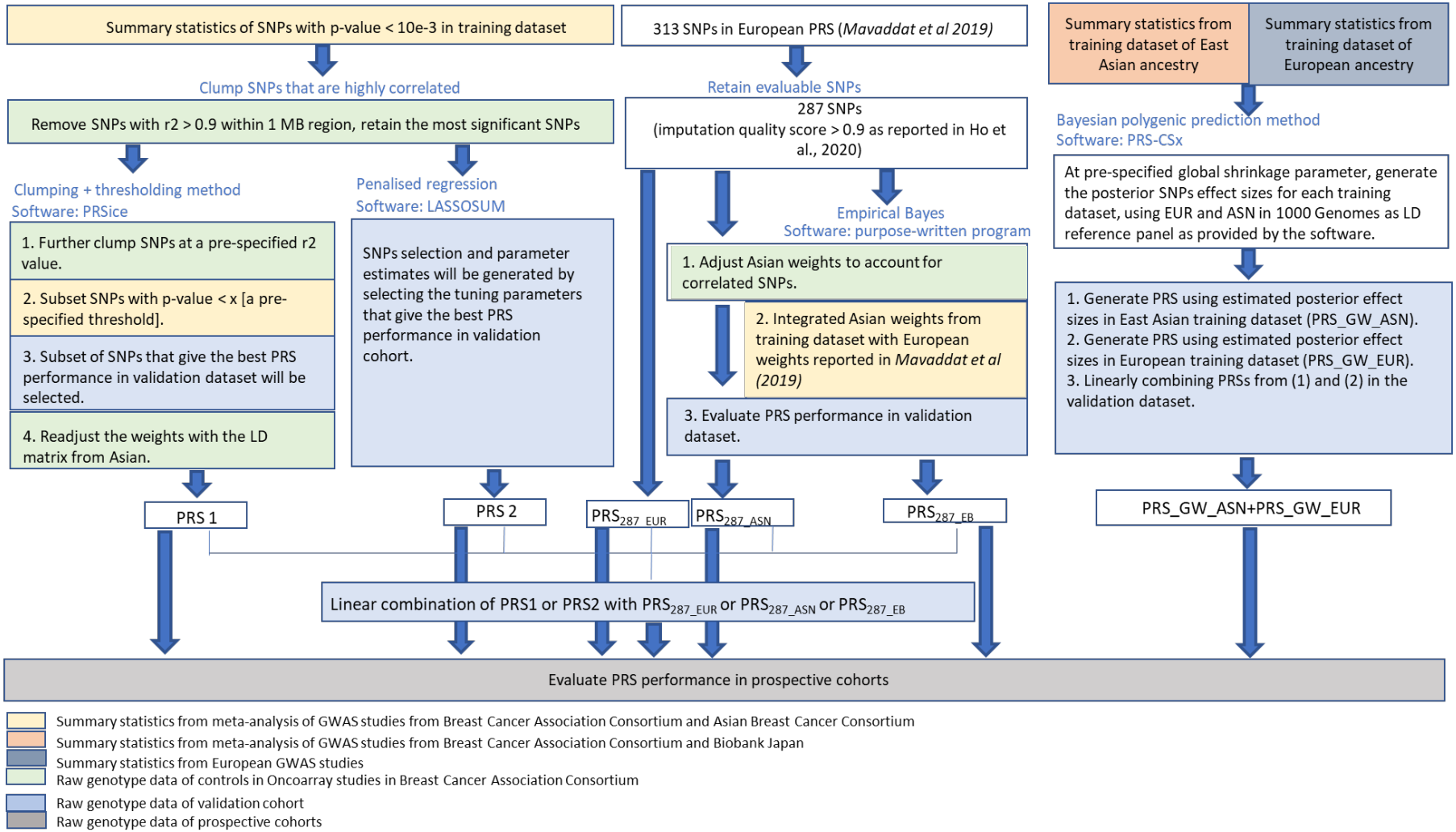


Figure S2. AUCs of PRSs generated using clumping and thresholding method for East Asian ancestry women

PRS performance (AUC) in validation dataset generated using clumping and thresholding method at different p-value threshold. Each point represents a p-value threshold. The best-fit PRS consisted of 46 SNPs at p-value threshold of 5.74×10^{-7}.

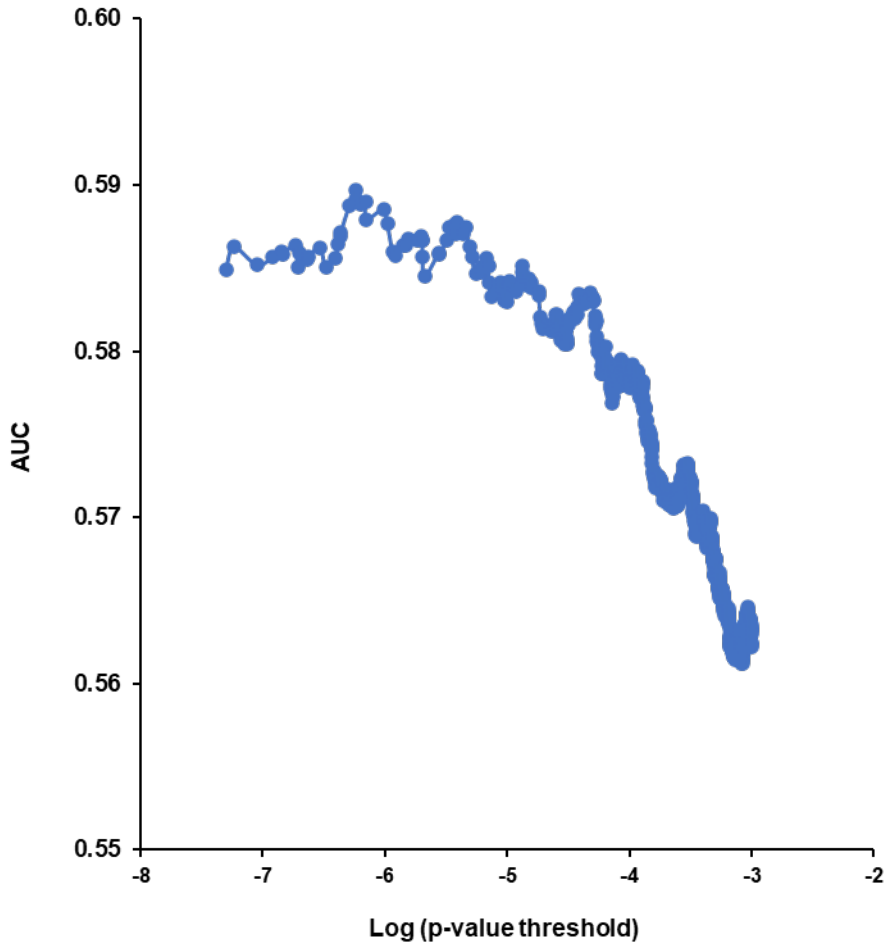


Figure S3. AUCs of best PRSs generated using clumping and thresholding method with different clumping option for East Asian ancestry women

Each point represents the AUC of the best PRS generate at the combination of the given clumping r^2 and clumping size.

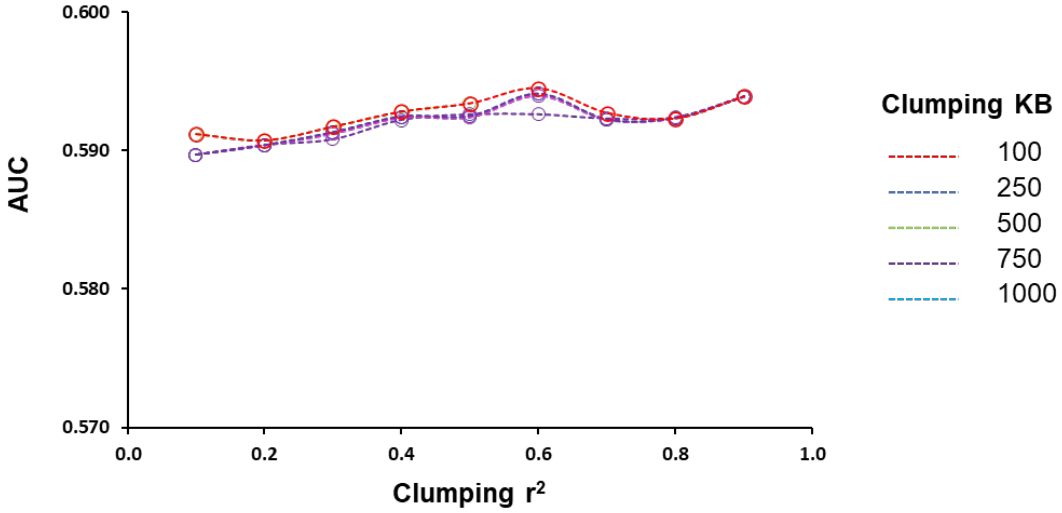


Figure S4. PRSs generating using different values of parameters in penalized regression for East Asian ancestry women

Each point represents correlation between breast cancer status in validation dataset and PRS generated at the given combination of penalty parameter (lambda) and shrinkage parameter (s). The best PRS occurred at penalty parameter (λ) equal to 0.014 and shrinkage parameter (s) equals to 0.9, where 2,985 SNPs were selected.

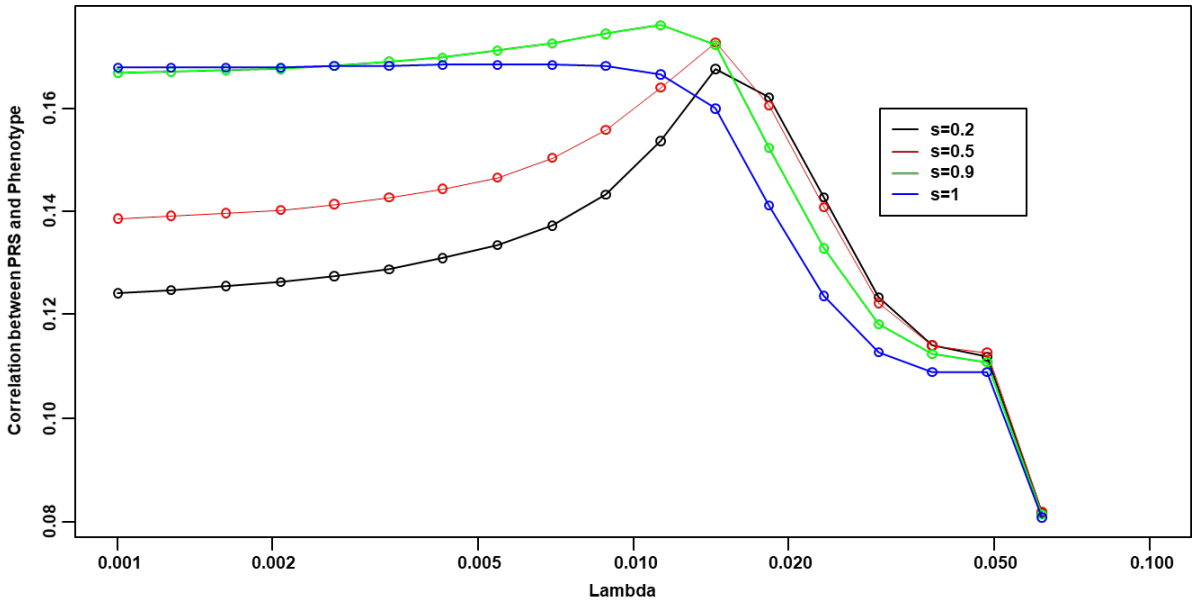
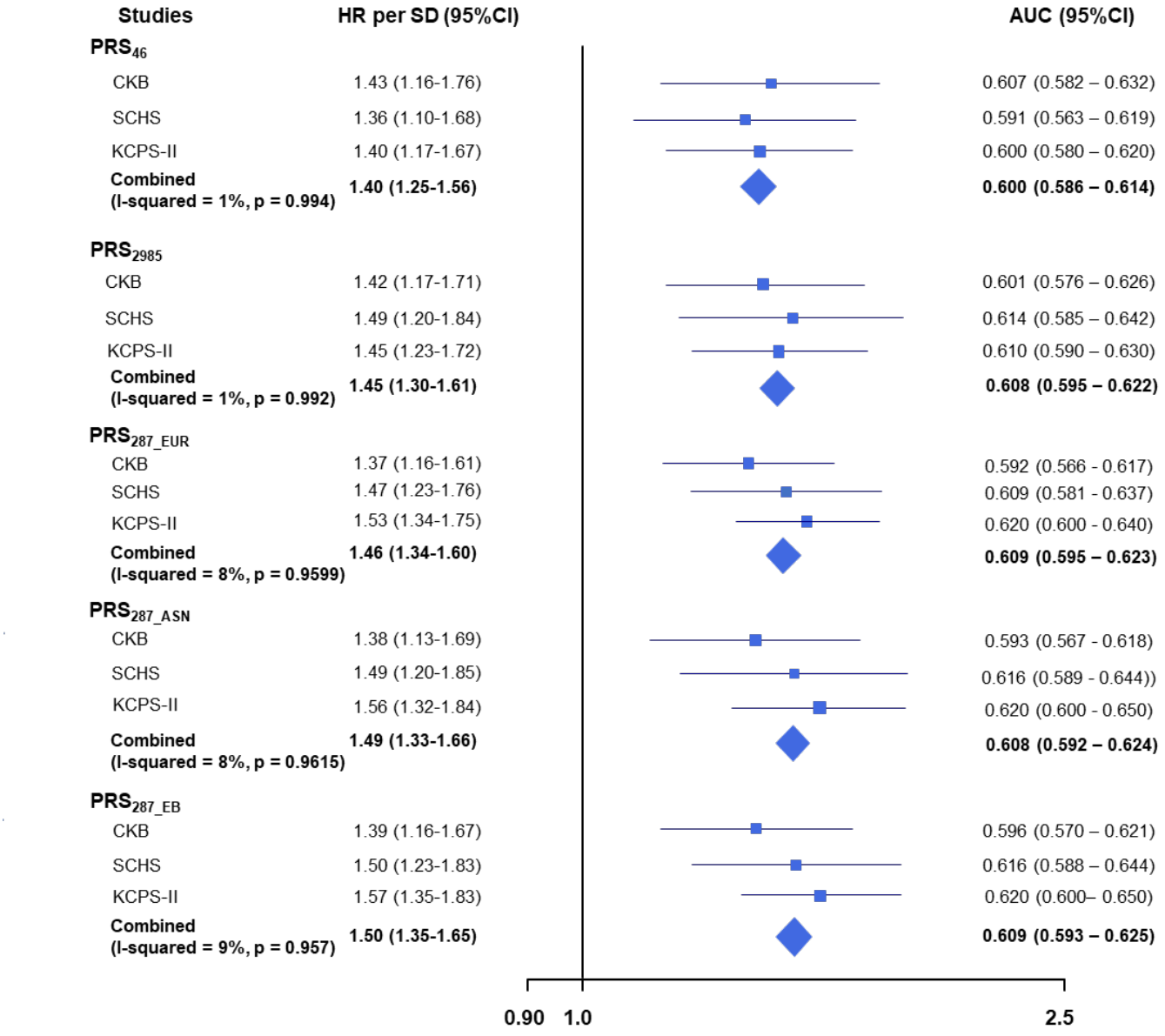


Figure S5. Performance of the PRSs in test dataset.

Forest plot showed the association between standardised PRSs and breast cancer risk in three prospective cohorts, China Kadoorie Biobank (CKB), Singapore Chinese Health Study (SCHS) and Korean Cancer Prevention Study II. The squares represent the hazard ratios (HRs), the horizontal lines represent the corresponding 95% confidence intervals and the diamond shapes represent the overall estimates. I-squared and p-value (two-sided) for heterogeneity were obtained by fitting a random-effects model and using generalized Q-statistic estimator (the rma() command in R).

(a) Performance of Asian-based and European-based PRSs



(b) Performance of PRSs generated by linear combinations of multiple PRSs.

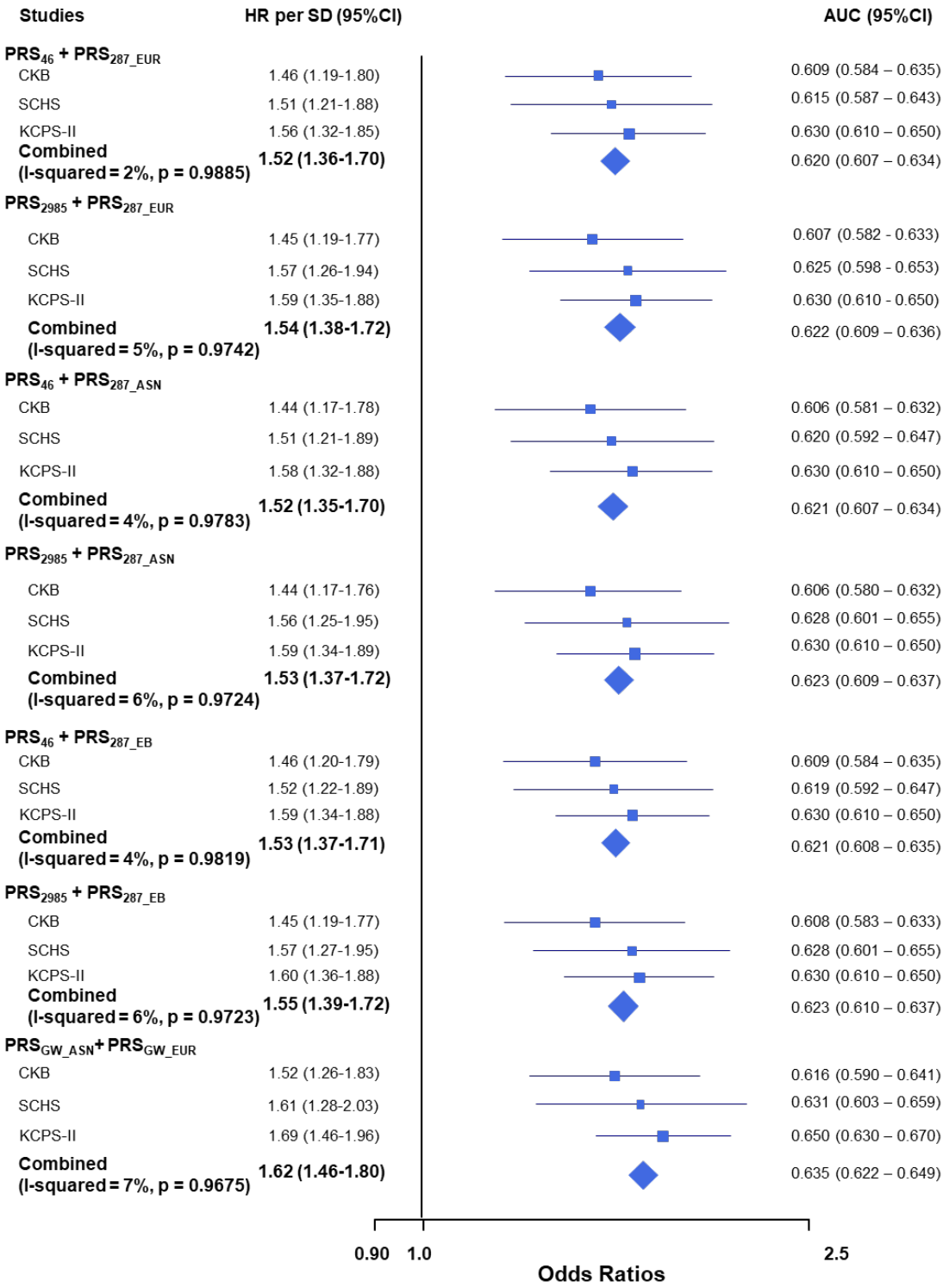


Figure S6. Distribution of 10-years absolute breast cancer risk at age 40 by birth cohort.

Dashed vertical line equals to 2.3% (average 10-year absolute risk of a 50-years old European women). Area under the curve represents the proportion of women who would have absolute risk at age 40 greater than a specific risk threshold. For example, area to the right of the vertical line for the blue curve represent proportion of women who were born after 1979 who would have absolute risk at age 40 greater than 2.3%. The breast cancer incidence for birth cohort 1960-1969 and 1970-1979 were observed and determined using breast cancer incidence in Singapore from 1968 to 2017. For birth cohort 1980-1989, breast cancer incidences were projected by assuming an annual increase in breast cancer incidence of 3.9%[1].

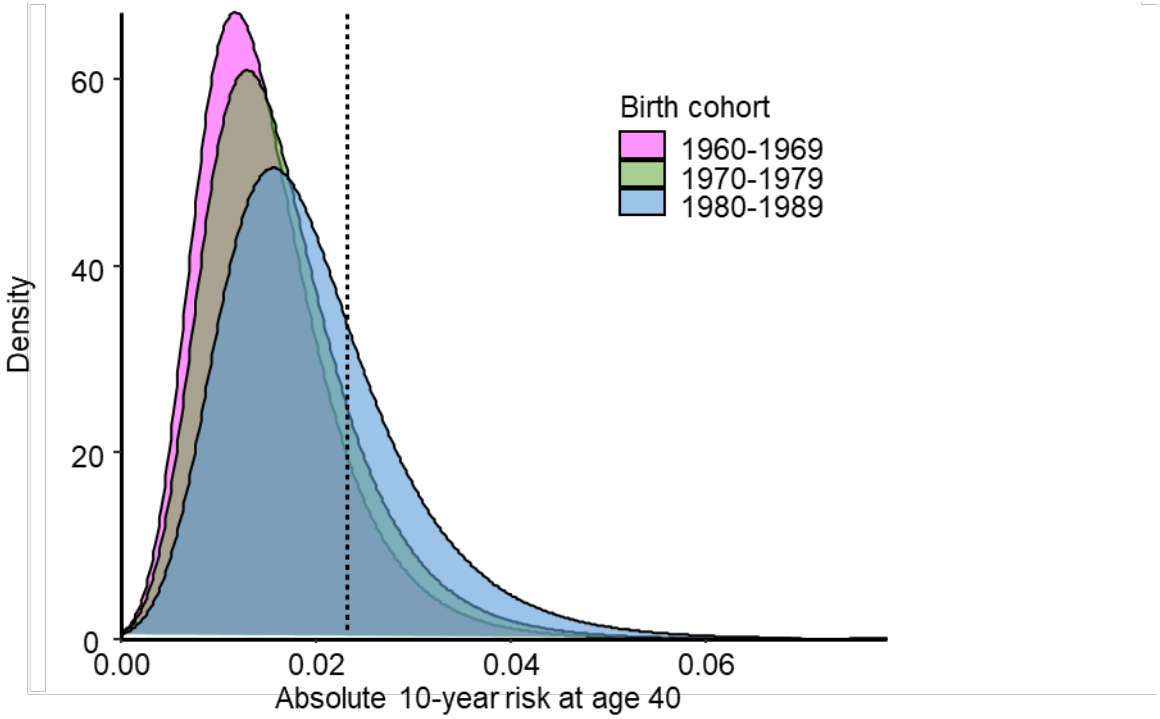
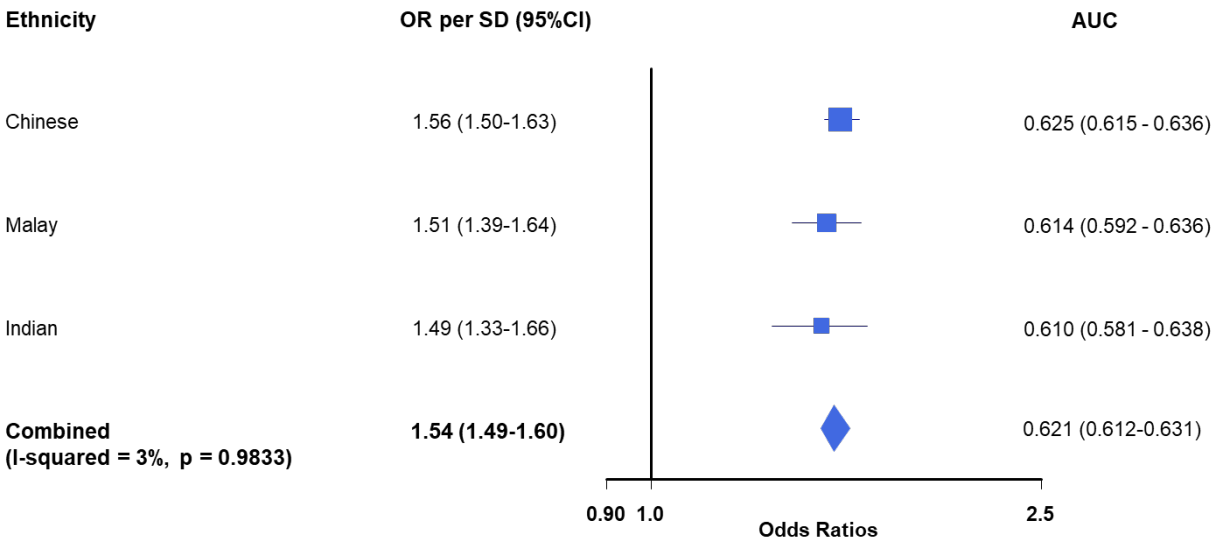


Figure S7. Performance of the PRS_{46} + PRS_{287_EB} in Chinese, Malay and Indian women from Malaysia and Singapore

Forest plot showed the association between standardised PRSs and breast cancer risk in Chinese, Malay and Indian women from Malaysia and Singapore (validation cohort). Odds ratios (ORs) and AUCs were generated using data from Malaysia Breast Cancer Genetics (MyBrCa) and Singapore Breast Cancer Cohort (SGBCC) studies, stratified by ethnicity. The squares represent the odds ratios (ORs), the horizontal lines represent the corresponding 95% confidence intervals and the diamond shapes represent the overall estimates. I-squared and p-value (two-sided) for heterogeneity were obtained by fitting a random-effects model and using generalized Q-statistic estimator (the rma() command in R). The number of cases and controls for each ethnicity, ORs and corresponding 95% confidence intervals are tabulated in Table S8.



Please refer to attached excel for Table S1, S2, S3, and S4.

Table S1. Participating studies and the number of individuals used in polygenic risk scores evaluation analyses.

Table S2. SNPs and beta coefficient SNPs used in the construction of. PRS_{287}, PRS_{46}, PRS_{235} and PRS_{2985}.

Table S3. SNPs and beta coefficient SNPs used in the construction of. PRS_{GW_ASN}.

Table S4. SNPs and beta coefficient SNPs used in the construction of. PRS_{GW_EUR}.

Table S5. Weights used in the linear combinations of multiple PRSs

PRS combination	Weight, α_1 †	Weights, α_2 †	α_0 †	*w**
East Asian ancestry				
$\alpha_1 PRS_{46} + \alpha_2 PRS_{287\_EUR} + \alpha_0$	0.17389	0.33479	-0.31324	0.66
$\alpha_1 PRS_{2985} + \alpha_2 PRS_{287\_EUR} + \alpha_0$	0.19846	0.31648	-0.47000	0.61
$\alpha_1 PRS_{46} + \alpha_2 PRS_{287\_ASN} + \alpha_0$	0.14457	0.32984	0.57515	0.70
$\alpha_1 PRS_{2985} + \alpha_2 PRS_{287\_ASN} + \alpha_0$	0.17360	0.30808	0.39362	0.64
$\alpha_1 PRS_{46} + \alpha_2 PRS_{287\_EB} + \alpha_0$	0.14893	0.35354	-0.05224	0.70
$\alpha_1 PRS_{2985} + \alpha_2 PRS_{287\_EB} + \alpha_0$	0.17373	0.33482	-0.19889	0.66
$\alpha_1 PRS_{GW\_ASN} + \alpha_2 PRS_{GW\_EUR} + \alpha_0$	0.16856	0.38484	0.54881	0.70
South Asian ancestry				
$\alpha_1 PRS_{46} + \alpha_2 PRS_{287\_EUR} + \alpha_0$	0.03168	0.38924	-0.74574	0.92
$\alpha_1 PRS_{2985} + \alpha_2 PRS_{287\_EUR} + \alpha_0$	0.16089	0.33732	-0.75920	0.68
$\alpha_1 PRS_{46} + \alpha_2 PRS_{287\_ASN} + \alpha_0$	0.01452	0.33817	-0.08530	0.96
$\alpha_1 PRS_{2985} + \alpha_2 PRS_{287\_ASN} + \alpha_0$	0.16498	0.27290	-0.19377	0.62
$\alpha_1 PRS_{46} + \alpha_2 PRS_{287\_EB} + \alpha_0$	0.01339	0.39594	-0.52023	0.97
$\alpha_1 PRS_{2985} + \alpha_2 PRS_{287\_EB} + \alpha_0$	0.14950	0.33877	-0.54253	0.69
$\alpha_1 PRS_{GW\_ASN} + \alpha_2 PRS_{GW\_EUR} + \alpha_0$	0.07095	0.44277	0.39168	0.86

†Combined PRSs were generated using the formula $\alpha_0 + \alpha_1 PRS_1 + \alpha_2 PRS_2$ where α_0, α_1 and α_2 are the weights obtained by fitting a logistic regression model with breast cancer as outcome, and PRS_1 and PRS_2 are explanatory variables using East Asian ancestry women (top panel) or South Asian ancestry women (bottom panel) in the validation dataset. Here PRS_1 represents the Asian-based PRS and PRS_2 represent the European-based PRS. The PRSs were standardized to respective standard deviation (SD) of the controls in the validation dataset.

*Contribution of the European based PRS to the linear combination, where $w = \alpha_2 / (\alpha_1 + \alpha_2)$ and (1-w) represents the contribution of Asian based PRS to the linear combination.

Table S6. PRSs generated using different values of global shrinkage parameter in PRS-CSx for East Asian ancestry women

Global shrinkage parameter	PRS	East Asian Ancestries			
		Cases Mean (SD)	Control Mean (SD)	OR per SD† (95% CI)	AUC
ϕ=10^{-6}	PRS_{GW_ASN}	0.296 (0.109)	0.249 (0.108)	1.53 (1.48-1.59)	0.620
	PRS_{GW_EUR}	0.139 (0.169)	0.065 (0.169)	1.55 (1.50-1.61)	0.623
	*PRS_{GW_ASN} + PRS_{GW_EUR}	0.068 (0.460)	-0.143 (0.459)	1.58 (1.52-1.64)	0.628
ϕ=10^{-4}	PRS_{GW_ASN}	-0.142 (0.184)	-0.208 (0.183)	1.44 (1.39-1.49)	0.601
	PRS_{GW_EUR}	-0.183 (0.209)	-0.281 (0.211)	1.59 (1.53-1.65)	0.629
	*PRS_{GW_ASN} + PRS_{GW_EUR}	0.082 (0.492)	-0.158 (0.488)	1.62 (1.56-1.68)	0.636
ϕ=10^{-2}	PRS_{GW_ASN}	1.446 (0.513)	1.339 (0.516)	1.23 (1.19-1.28)	0.558
	PRS_{GW_EUR}	0.017 (0.432)	-0.135 (0.436)	1.41 (1.36-1.46)	0.597
	*PRS_{GW_ASN} + PRS_{GW_EUR}	0.346 (0.382)	-0.111 (0.382)	1.46 (1.41-1.51)	0.605
ϕ=1	PRS_{GW_ASN}	5.873 (0.995)	5.713 (1.005)	1.17 (1.13-1.21)	0.545
	PRS_{GW_EUR}	4.878 (0.848)	4.668 (0.846)	1.27 (1.22-1.31)	0.567
	*PRS_{GW_ASN} + PRS_{GW_EUR}	0.002 (0.283)	-0.078 (0.284)	1.31 (1.27-1.36)	0.579

†Adjusted for first the 10 principal components and study, and standardised to SDs in controls of each PRS.

*Combined PRSs were generated using the formula $\alpha_0 + \alpha_1 PRS_1 + \alpha_2 PRS_2$ where α_0, α_1 and α_2 are the weights obtained by fitting a logistic regression model with breast cancer as outcome, and PRS_1 and PRS_2 are explanatory variables using either East Asian ancestry women in the validation dataset. The PRSs were standardized to standard deviation (SD) of the controls in the validation dataset.

Table S7. Mean, standard deviation, and the association of polygenic risk scores (PRS) with breast cancer risk in women of South Asian ancestry

PRS developed based on South Asians	Validation cohort[a]			
	Cases Mean (SD)	Control Mean (SD)	OR per SD[†] (95% CI)	AUC
[b]PRS_{46} + PRS_{287_EUR}	-0.478 (0.376)	-0.631 (0.400)	1.50 (1.34-1.67)	0.614
[b]PRS_{2985} + PRS_{287_EUR}	-0.467 (0.404)	-0.641 (0.425)	1.53 (1.37-1.71)	0.620
[b]PRS_{46} + PRS_{287_ASN}	-0.493 (0.354)	-0.614 (0.344)	1.43 (1.28-1.59)	0.592
[b]PRS_{2985} + PRS_{287_ASN}	-0.482 (0.385)	-0.625 (0.374)	1.46 (1.32-1.63)	0.603
[b]PRS_{46} + PRS_{287_EB}	-0.477 (0.384)	-0.632 (0.401)	1.50 (1.35-1.67)	0.613
[b]PRS_{2985} + PRS_{287_EB}	-0.467 (0.409)	-0.641 (0.422)	1.53 (1.37-1.70)	0.618
[b]PRS_{GW_EUR} + PRS_{GW_ASN}	-0.438 (0.482)	-0.699 (0.480)	1.63 (1.47-1.83)	0.636

[a]Evaluation of PRSs performance in 585 breast cancer cases and 1,018 controls of Indian ancestry women in the validation dataset (Table S1).

[b]Combined PRSs were generated using the formula $\alpha_0 + \alpha_1 PRS_1 + \alpha_2 PRS_2$ where α_0, α_1 and α_2 are the weights obtained by fitting a logistic regression model with breast cancer as outcome, and PRS_1 and PRS_2 are explanatory variables using Indian-ancestry women in the validation dataset. The weights for the considered combinations of PRSs can be found in bottom panel of Table S5.

[†]Adjusted for first the 10 principal components and study, and standardised to SDs in controls of each PRS.

Table S8. Breast cancer odds ratio by percentiles of PRS_{46}+PRS_{287_EB} in East Asian ancestry women

Percentiles	Control	Cases	Estimated OR (95% CI)	Predicted OR
<1	67	29	0.53(0.33-0.82)	0.31
1-5	265	97	0.43(0.33-0.55)	0.35
5-10	232	137	0.49(0.39-0.60)	0.53
10-20	664	381	0.67(0.58-0.78)	0.63
20-40	1327	897	0.79(0.71-0.89)	0.79
40-60	1328	1134	1	1
60-80	1327	1505	1.33(1.19-1.48)	1.26
80-90	664	915	1.61(1.41-1.83)	1.58
90-95	232	544	1.91(1.63-2.24)	1.88
95-99	265	579	2.55(2.16-3.02)	1.85
>99	67	174	3.01(2.25-4.06)	3.24

PRS_{46}+PRS_{287_EB} was categorised into quantiles based on the PRS distribution in controls of validation dataset of East Asian ancestry women. The middle quintile was used as the reference category. Observed odds ratios (ORs) were compared with those predicted under a theoretical polygenic model in which the log OR depends-linearly on the PRS (see Supplementary Methods).

Table S9. Association between PRS_{46} + PRS_{287_EB} and breast cancer risk in validation cohort by ethnicity

Ethnicity*	Cases	Control	Cases Mean (SD)		Control Mean (SD)		OR per SD (95% CI)†	AUC (95% CI)
PRS_{46}+PRS_{287_EB}								
Chinese	5230	5153	0.076	(0.439)	-0.118	(0.439)	1.56 (1.50 - 1.63)	0.625 (0.615 - 0.636)
Malay	1086	1335	-0.007	(0.466)	-0.197	(0.556)	1.51 (1.39 - 1.64)	0.614 (0.592 - 0.636)
Indian	585	1018	-0.157	(0.438)	-0.328	(0.455)	1.49 (1.33 - 1.66)	0.610 (0.581 - 0.638)

†Adjusted for first 10 principal components and study, and standardised to SDs in controls of each PRSs.

*Self-declared ethnicity was used.

Supplementary Methods

Study populations and genotyping

The study population was divided into three datasets. PRSs were developed using training and validation datasets and evaluated in the testing dataset. The training dataset included women of East Asian ancestry from three sources: (a) 20,198 women (10,020 invasive cases and 10,179 controls) participating in 11 studies in Breast Cancer Association Consortium (BCAC); (b) 23,928 women (11,993 invasive cases and 11,935 controls) participating in 5 studies in the Asia Breast Cancer Consortium (ABCC); and (c) 79,550 (6,325 invasive cases and 73,225 controls) participating in Biobank Japan (BBJ). Except for studies in BCAC where raw genotype data was available, only summary statistics were available for the remaining studies. Training set 1 included summary statistics of variants with p-value < 10^{-3} from a meta-analysis of BCAC and ABCC studies as described in Shu et al, (2020)[30], training set 2 included summary statistics of variants from a meta-analyses of BCAC studies and BBJ. Given that summary statistics from ABCC studies were not available (only summary statistics of meta-analysed BCAC and ABCC studies were available) and a portion of samples in BBJ had been included in ABCC studies, we could not include ABCC studies in the meta-analysis of BCAC studies and BBJ. Finally, publicly available summary statistics from European GWAS were included as training set 3[23].

The validation set included 14,633 women of Easy Asian (Chinese-ancestry and Malay ancestry) or South Asian ancestry participating in two multi-ethnic case-control studies: (a) 6,993 women (3,384 invasive cases and 3,609 controls) participating in Malaysian Breast Cancer Genetics (MyBrCa) study; and (b) 7,640 women (3,593 invasive cases and 4,047

controls) participating in the Singapore Breast Cancer Cohort (SGBCC) study. Given that East Asians and South Asians are genetically distinguishable, we further divided the validation dataset into – set 1 included 13,030 (6,392 cases invasive cases and 6,638 controls) women of Chinese or Malay ancestry, and set 2 included 1,603 (585 invasive cases and 1,018 controls) women of Indian ancestry. Samples in the development dataset were genotyped using one of the five arrays: iCOGS[31], OncoArray[32], Affymetrix Genome-Wide Human SNP Array 6.0, Illumina Multi-Ethnic Genotyping Array, Illumina HumanOmiExpress[33] and Illumina HumanExome-12v1_A Beadchip[34-36] (Table S1).

The best PRSs were evaluated in the testing set comprising 89,898 women from three prospective cohorts of East Asian ancestry: (a) 10,021 women who had not had any cancer diagnosis prior to recruitment into Singapore Chinese Health Study (SCHS)[25,26] , of which 413 registry-confirmed breast cancer developed over 195,317 person years of prospective follow-up; (b) 38,864 women without any cancer diagnosis prior to recruitment into China Kadoorie Biobank (CKB)[28], of which 476 developed breast cancers over 423,396 person years of prospective follow-up; and (c) 41,031 without any cancer diagnosis prior to recruitment into Korean Cancer Prevention Study Biobank(KCPS-II)[29] of which 705 developed breast cancers over 406,556 person years of prospective follow-up. For all studies, follow-up started six months after recruitment and was censored at age of breast cancer diagnosis, age at risk-reducing mastectomy, age of diagnosis of any cancer, age of death, or age on 31 December 2015 (for SCHS), age on 31 December 2017 (for CKB) and age on 31 December 2017 (for KCPS-II) whichever came first. Samples in SCHS and KCPS-II were genotyped using the Illumina Global Screening Array[25,29], while samples in CKB were genotyped using custom-designed Affymetrix Axiom arrays[28]. Analyses using CKB data were conducted under research approval

2020-0047. Supplementary Table 1 summarises the study design and the number of breast cancer cases and controls in each study.

Genotype calling, quality control procedure and imputation have been described previously[23,30,31,33,37,38]. All data were imputed using the 1000 Genomes Project (Phase 3) data as the reference panel[39], except BioBank Japan, for which the HapMap Phase II (release 22) [40] was used. SNPs with overall minor allele frequency in controls > 0.01 and imputation $r^2 > 0.9$ for OncoArray studies, imputation $r^2 > 0.7$ for Biobank Japan and imputation $r^2 > 0.3$ for other studies in the training and validation dataset were included in this analysis. Since all samples in the validation sets were genotyped using OncoArray, a higher threshold was imposed for OncoArray to ensure accurate determination of PRS in the validation datasets.

Ancestry informative principal components were available for Asian ancestry samples in the training dataset and the validation dataset, generated using methods as previously described[30]. Briefly, for the BCAC data, continental ancestry was derived by combining the data with the 1000 Genomes Project reference data. Individuals with >40% estimated East Asian ancestry were retained. In the second stage, principal components were generated on the Asian ancestry individuals using a subset of uncorrelated SNPs. Similar ancestry informative principal components were generated for the other dataset.

All studies were approved by the relevant institutional ethics committees and review boards, and all participants provided written informed consent.

Single-SNP association analysis in the training set

Single-SNP association tests were conducted in the BCAC studies separately for the iCOGS and OncoArray datasets, adjusted for age, the first two principal components and country/study

to obtain the per-allele OR for each SNP using Plink 2.0[41]. Single-SNP association analyses in ABCC studies were previously conducted by Shu and colleagues (2020)[30] – only summary statistics from meta-analyses of BCAC and ABCC studies were available for this project. Briefly, the analyses were conducted separately for each study in ABCC, adjusted for age and first two principal components[30]. Combined weights and p-values were derived using fixed-effect meta-analysis with the software METAL[42]. A total of 20,768 SNPs significantly associated with breast cancer risk at p-values < 0.001 were selected. SNPs clumping (within 1Mb windows) was subsequently conducted using the software PRSice v2.11[43] to remove highly correlated SNPs (pairwise correlation $r^2 > 0.9$); the SNP with the lowest p-value for association in the correlated pairs was retained, resulting in 3,050 SNPs for subsequent analyses. Since the raw genotype data were not available for ABCC studies, the correlation r^2 was computed using 4,921 control samples in BCAC OncoArray studies only. Single-SNP association analyses in BBJ1 were previously conducted by *Ishigaki et al (2020)*[33]. Briefly, GWAS was conducted using generalised linear mixed model, adjusted for age and first five principal components. The GWAS summary statistics from BBJ were combined with GWAS from BCAC Asian studies using fixed-effect meta-analysis with the software METAL.

Clumping and thresholding (C+T) method

To account for the joint effect of SNPs used in derivation of the best PRS determined by the C+T method, the SNP weights should ideally be estimated jointly in a single logistic regression model. Raw genotype data of the training set were not available for the joint estimation. However, these weights can be computed from the marginal effect sizes – if $\underline{\gamma}$ are the (conditionally independent) effect sizes and $\gamma'_j = \gamma_j\sqrt{2p_j(1-p_j)}$ are the corresponding normalized effect sizes, where p_j is the effect allele frequency of SNP j, then the predicted

normalised marginal effect sizes of the SNPs β' are given by $\underline{\beta'} = R\underline{\gamma'}$, where R is the matrix of correlations between the SNP genotypes. Thus $\underline{\gamma'} = R^{-1}\underline{\beta'}$, and the optimal weight. The optimal weight for SNP *j* is then given by:

$$\gamma_j = {\gamma'}_j \Big/ \sqrt{2p_j(1-p_j)} \qquad (1)$$

The correlation matrix R was estimated using 4,921 Asian control samples in BCAC OncoArray studies. The concept of this method has been previously described in Section 3.3 of *Prive et al (2020)*[44].

Re-weighting of European-based PRS

For these analyses, we considered PRS based on the 313 SNPs developed in European women[45]. Of the 313 SNPs, only 287 SNPs with imputation info score > 0.9 in OncoArray studies were retained for subsequent analyses. We considered two sets of weights for these SNPs: (i) Asian weights estimated from training set 1 alone; and (ii) weights based on a combination of the Asian (from training set 1) and European weights, allowing for these weights to differ but be correlated.

For (i), the optimal weights taking into account the correlation between SNPs were derived using Equation (1). For (ii), we combined Asian and European weights using an Empirical Bayes approach. In brief, we assume that the true population-specific effect sizes vary from a "global" effect size, β_j, by a normally distributed amount, with variance ξ^2, i.e.

$$\beta'_{jA}, \beta'_{jE} \sim N(\beta_j, \xi^2)$$

where β'_{jA} and β'_{jE} are the unobserved true effect sizes for SNP *j* for Asian and European populations, respectively. Let $\gamma'_{jA} = \gamma_{jA}\sqrt{2p_j(1-p_j)}$ be the normalized weight of SNP *j*

estimated from the training set (using the method described above but p_j is the average effect allele frequency of SNP *j* in Asians and Europeans), and $\gamma'_{jE} = \gamma_{jE}\sqrt{2p_j(1-p_j)}$ be the normalized weight for SNP *j* reported for European populations (γ_{jE} are effect sizes reported in Mavaddat et al[45]). Using Bayes theorem and given that γ'_{jA} and γ'_{jE} are conditional independent given β'_{jA} the posterior distribution of $\beta'_{jA}|\gamma'_{jA}, \gamma'_{jE}$ is given by

$$f(\beta'_{jA}|\gamma'_{jA}, \gamma'_{jE}) \propto f\left(\gamma'_{jA}\middle|\beta'_{jA}\right) f(\beta'_{jA}),$$

where $f\left(\gamma'_{jA}\middle|\beta'_{jA}\right) \sim N(\beta'_{jA}, V_{\gamma'_{jA}})$ and $V_{\gamma'_{jA}}$ is the variance of γ'_{jA} estimated from the training dataset. Here, The estimated posterior effect sizes in Asians, given the data, are therefore:

$$\beta_{jA,EB} = \frac{\tau_A \gamma'_{jA} + \tau_\xi \beta_j}{\tau_A + \tau_\xi} \quad (2)$$

, $\tau_A = 1/V_{\gamma'_{jA}}$ and $\tau_\xi = 1/\xi^2$. The estimated posterior effect sizes for European populations, $\beta_{jE,EB}$, can be obtained using Equation (2) by replacing γ'_{jA} by γ'_{jE} and τ_A by τ_E, where $\gamma'_{jE} = \gamma_{jE}\sqrt{2p_j(1-p_j)}$ τ_E is the corresponding observed inverse variance of γ'_{jE}.

The parameters β_j, ξ^2 were estimated using Expectation-Maximisation (EM) iteration method where in the M-step, β_j was estimated using the formula

$$\beta_j = \frac{1}{2}(\beta_{jA,EB} + \beta_{jE,EB}),$$

since $\frac{1}{2}\left(\beta'_{jA} + \beta'_{jE}\right)$ is an unbiased estimator of β_j, and the variance ξ^2 was derived using the following formula

$$\xi^2 = \frac{1}{2}\, var(\beta_{jA,EB} - \beta_{jE,EB}),$$

since $\beta'_{jA} - \beta'_{jE} \sim N(0, 2\xi^2)$. In the E-step, $\beta_{jA,EB}$ and $\beta_{jE,EB}$ were updated and the algorithm was repeated until β_j and ξ^2 converge. Finally, the optimal weights for each SNP included in PRS derivation were estimated using the following formula

$$\beta_{j,EB} = {\beta_{jA,EB}}\Big/{\sqrt{2p_j(1-p_j)}}.$$

This approach "shrinks" the Asian estimates towards the European estimates, making use of the greater precision in the European estimates but allowing for different Asian weights when the European and Asian estimates differ markedly.

Of the 287 SNPs, the combined Asian weights, $\hat{\beta}_{jA}$, in the training dataset was not available for 48 SNPs with MAF < 0.01, hence for these 48 SNPs, European weights, $\hat{\beta}_{jE}$, were used for PRS construction.

Absolute risk of breast cancer by PRS percentiles

The age-specific absolute risks of developing breast cancer in each PRS percentile, g, were calculated using the following formula:

$$\mathrm{AR_g(t)} = \sum_{\mathrm{u=0}}^{\mathrm{t}} \lambda_{\mathrm{g}}(\mathrm{u}) \cdot \mathrm{S_g(u)} \cdot \mathrm{S_m(u)}$$

where $\lambda_{\mathrm{g}}(u) = \lambda_0(u)\exp(\beta_g)$ is the breast cancer incidence associated with PRS at age u, $\lambda_0(u)$ is the baseline incidence and the corresponding effect size β_g, $\mathrm{S_g(u)}$ is the probability of being breast cancer free at age u, and $\mathrm{S_m(u)}$ is the probability of not dying from a cause other than breast cancer at age *u*. The theoretical effect sizes, $\mathrm{OR_g} = \exp(\beta_g)$, for PRS

interval between two percentiles (*u,v*), using middle quintile as reference (40-60th), were given by

$$\mathrm{OR_g} = \frac{(0.6-0.4)((\Phi(\Phi^{-1}(1-u)+\sigma)-\Phi(\Phi^{-1}(1-v)+\sigma))}{(\mathrm{v}-\mathrm{u})((\Phi(\Phi^{-1}(0.6)+\sigma)-\Phi(\Phi^{-1}(0.4)+\sigma))}$$

where σ is the log OR per unit SD of the continuous PRS[46]. The PRS-specific breast cancer incidences, $\lambda_g(u)$, were calculated iteratively by assuming that the average age-specific breast cancer incidence over all PRS percentiles agreed with the population breast cancer incidence. The details of these methods have been described previously[47]. We calculated lifetime and 10-year absolute risks using Singaporean mortality and breast cancer incidence for Chinese, Malays and Indians in 2017[49,50]. Under polygenic model, the logarithm of risk in the population has been shown to follow a normal distribution[51]. The proportion of population above a specific risk threshold was given by the area under the curve of the distribution of logarithm of 10-year absolute risk at pre-specified age.

To generate the distribution of birth-cohort specific 10-year absolute risk at age 40, we used birth cohort--specific breast cancer incidences derived using population breast cancer incidence in Singapore from 1968 to 2017[52]. The population incidences were reported in five-year age intervals for calendar year 1968-1972, 1973-1977,…, 2013-2017. By taking the lower-bound of these five-year age interval and the midpoint of calendar-specific interval, the year of birth of the cohort that the reported population incidences were based on were calculated. We took the average of the reported incidences according to three birth cohorts - 1960-1969, 1970-1979 and 1980-1989. Birth cohort-specific incidence were observed for women who were born between 1960-1969 and 1970-1979. For women who were born between 1980-1989, breast cancer incidence was observed up to age 35. The breast cancer incidence for this

birth cohort at age 40 was projected by assuming an annual increase in breast cancer incidence of 3.9%[1].

Supplementary References:

1. Jara-Lazaro, AR, Thilagaratnam S., and Tan PH, Breast cancer in Singapore: some perspectives. *Breast Cancer*, 2010. **17(1)**: p. 23-8.
2. Phuah SY, *et al.* Triple-negative breast cancer and PTEN (phosphatase and tensin homologue) loss are predictors of BRCA1 germline mutations in women with early-onset and familial breast cancer, but not in women with isolated late-onset breast cancer. *Breast cancer research : BCR* **14**, R142, doi:10.1186/bcr3347 (2012).
3. Wu AH, McKean-Cowdin R, Tseng CC. Birth weight and other prenatal factors and risk of breast cancer in Asian-Americans. *Breast Cancer Res Treat.* **14**:917–925 (2011).
4. Kawase T, *et al.* FGFR2 intronic polymorphisms interact with reproductive risk factors of breast cancer: results of a case control study in Japan. *International journal of cancer. Journal international du cancer* **125**, 1946-1952, doi:10.1002/ijc.24505 (2009).
5. Zheng W, *et al.* Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nature genetics* **41**, 324-328, doi:10.1038/ng.318 (2009).
6. Lee KM, *et al.* Genetic polymorphisms of ataxia telangiectasia mutated and breast cancer risk. *Cancer Epidemiol Biomarkers Prev* **14**, 821-825, doi:10.1158/1055-9965.EPI-04-0330 (2005).
7. Hsu HM, *et al.* Breast cancer risk is associated with the genes encoding the DNA double-strand break repair Mre11/Rad50/Nbs1 complex. *Cancer Epidemiol Biomarkers Prev* **16**, 2024-2032, doi:10.1158/1055-9965.EPI-07-0116 (2007).
8. Ding SL, *et al.* Genetic variants of BLM interact with RAD51 to increase breast cancer susceptibility. *Carcinogenesis* **30**, 43-49, doi:10.1093/carcin/bgn233 (2009).
9. Grundy A, *et al.* Shift work, circadian gene variants and risk of breast cancer. *Cancer epidemiology* **37**, 606-612, doi:10.1016/j.canep.2013.04.006 (2013).
10. Kobayashi LC, *et al.* Moderate-to-vigorous intensity physical activity across the life course and risk of pre- and post-menopausal breast cancer. *Breast cancer research and treatment* **139**, 851-861, doi:10.1007/s10549-013-2596-9 (2013).
11. Grundy A, *et al.* Increased risk of breast cancer associated with long-term shift work in Canada. *Occupational and environmental medicine* **70**, 831-838, doi:10.1136/oemed-2013-101482 (2013).
12. Kobayashi LC, *et al.* A case-control study of lifetime light intensity physical activity and breast cancer risk. *Cancer causes & control : CCC* **25**, 133-140, doi:10.1007/s10552-013-0312-z (2014).
13. Kwong A, *et al.* Novel BRCA1 and BRCA2 genomic rearrangements in Southern Chinese breast/ovarian cancer patients. *Breast cancer research and treatment* **136**, 931-933, doi:10.1007/s10549-012-2292-1 (2012).
14. Kwong A, *et al.* Identification of BRCA1/2 founder mutations in Southern Chinese breast cancer patients using gene sequencing and high resolution DNA melting analysis. *PloS one* **7**, e43994, doi:10.1371/journal.pone.0043994 (2012).
15. Shimada N, *et al.* Genetic polymorphisms in estrogen metabolism and breast cancer risk in case-control studies in Japanese, Japanese Brazilians and non-Japanese Brazilians. *J Hum Genet* **54**, 209-215, doi:10.1038/jhg.2009.13 (2009).

16. John EM, *et al.* The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast cancer research : BCR* **6**, R375-389, doi:10.1186/bcr801 (2004).
17. Han SA, *et al.* The Korean Hereditary Breast Cancer (KOHBRA) study: protocols and interim report. *Clin Oncol (R Coll Radiol)* **23**, 434-441, doi:10.1016/j.clon.2010.11.007 (2011).
18. Cho, Y.S. *et al.* A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* **41**, 527-34 (2009).
19. Han, S. *et al.* CASP8 polymorphisms, estrogen and progesterone receptor status, and breast cancer risk. *Breast Cancer Res Treat* **110**, 387-93 (2008).
20. Elgazzar, S. *et al.* A genome-wide association study identifies a genetic variant in the SIAH2 locus associated with hormonal receptor-positive breast cancer in Japanese. *J Hum Genet* **57**, 766-71 (2012).
21. Low, S.K. *et al.* Genome-wide association study of breast cancer in the Japanese population. *PLoS One* **8**, e76463 (2013).
22. Sakaue, S, & Kanai, M., et al. A global atlas of genetic assocaition of 220 deep phenotypes. *medRxiv* (2020)
23. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. Nature 551, 92-94, doi:10.1038/nature24284 (2017)
24. Han S, *et al.* CASP8 polymorphisms, estrogen and progesterone receptor status, and breast cancer risk. *Breast cancer research and treatment* **110**, 387-393, doi:10.1007/s10549-007-9730-5 (2008).
25. Hankin JH, *et al.* Singapore Chinese Health Study: development, validation, and calibration of the quantitative food frequency questionnaire. Nutr Cancer: 39:187–195 (2001).
26. Wu AH, *et al.* Soy intake and breast cancer risk in Singapore Chinese Health Study. Br J Cancer: 99(1):196-200. doi: 10.1038/sj.bjc.6604448 (2008).
27. Pan, R. *et al.* Cancer incidence and mortality: A cohort study in China, 2008-2013. *Int J Cancer*.**1**;141(7):1315-1323 (2017).
28. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol*.**40**(6):1652-66.(2011)
29. Jee, Y.H. *et al.* Cohort Profile: The Korean Cancer Prevention Study-II (KCPS-II) Biobank. *Int J Epidemiol.* **1**;47(2):385-386f (2018)
30. Shu, X. *et al.* Identification of novel breast cancer susceptibility loci in meta-analyses conducted among Asian and European descendants. *Nature communications* **11**, 1217, doi:10.1038/s41467-020-15046-w (2020).
31. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature genetics* **45**, 353-361, 361e351-352, doi:10.1038/ng.2563 (2013).
32. Amos, C. I. *et al.* The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **26**, 126-135, doi:10.1158/1055-9965.EPI-16-0106 (2017).
33. Ishigaki, K. *et al.* Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat Genet*. 2020 Jul;**52(7)**:669-679.
34. Zhang, Y. *et al.* Rare coding variants and breast cancer risk: evaluation of susceptibility Loci identified in genome-wide association studies. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **23**, 622-628, doi:10.1158/1055-9965.EPI-13-1043 (2014).

35. Cai, Q. *et al.* Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. *Nature genetics* **46**, 886-890, doi:10.1038/ng.3041 (2014).
36. Zhang, B. *et al.* Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nature genetics* **46**, 533-542, doi:10.1038/ng.2985 (2014).
37. Dorajoo, R. *et al.* Loci for human leukocyte telomere length in the Singaporean Chinese population and trans-ethnic genetic studies. *Nature communications* **10**, 2491, doi:10.1038/s41467-019-10443-2 (2019).
38. Gan, W. *et al.* Evaluation of type 2 diabetes genetic risk variants in Chinese adults: findings from 93,000 individuals from the China Kadoorie Biobank. *Diabetologia* **59**, 1446-1457, doi:10.1007/s00125-016-3920-9 (2016).
39. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
40. International HapMap, C. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861, doi:10.1038/nature06258 (2007).
41. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).
42. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191, doi:10.1093/bioinformatics/btq340 (2010).
43. Choi, SW, and O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience*. 2019 Jul 1;**8(7)**:giz082.
44. Prive F, Arbel J and Vilhjalmsson BJ. LDpred2: better, faster, stronger. *Bioinformatics*2020 Dec 16;**36(22-23)**:5424-5431.
45. Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *American journal of human genetics* **104**, 21-34, doi:10.1016/j.ajhg.2018.11.002 (2019).
46. Wen, W. *et al.* Prediction of breast cancer risk based on common genetic variants in women of East Asian ancestry. *Breast Cancer Res* **18**, 124, doi:10.1186/s13058-016-0786-1 (2016).
47. Mavaddat, N. *et al.* Prediction of breast cancer risk based on profiling with common genetic variants. *JNCI: Journal of the National Cancer Institute* **107** (2015).
48. Mavaddat, N. *et al*. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet* **104**, 21-34, doi:10.1016/j.ajhg.2018.11.002 (2019).

49. Cancer Incidence in Five Continents. (International Agency for Research on Cancer, Lyon, France, 2014).
50. *Age-Specific Death Rates, Annual* (Government of Singapore, 2017).
51. Pharoah, PDP *et al*. Polygenic susceptibility to breast cancer and implications for prevention. ***Nat. Genet***. **31**(1):33-6 (2002).
52. SINGAPORE CANCER REGISTRY 50TH ANNIVERSARY MONOGRAPH 1968 – 2017.