







# The role of polygenic risk and susceptibility genes in breast cancer over the course of life

Nina Mars <sup>1</sup>, Elisabeth Widén<sup>1</sup>, Sini Kerminen <sup>1</sup>, Tuomo Meretoja<sup>2,3</sup>, Matti Pirinen <sup>1,4,5</sup>, Pietro della Briotta Parolo<sup>1</sup>, Priit Palta <sup>1,6</sup> & FinnGen\*, Aarno Palotie<sup>1,7,8</sup>, Jaakko Kaprio <sup>1,5</sup>, Heikki Joensuu<sup>3,9</sup>, Mark Daly<sup>1,8</sup> & Samuli Ripatti <sup>1,5,8</sup>✉

Polygenic risk scores (PRS) for breast cancer have potential to improve risk prediction, but there is limited information on their utility in various clinical situations. Here we show that among 122,978 women in the FinnGen study with 8401 breast cancer cases, the PRS modifies the breast cancer risk of two high-impact frameshift risk variants. Similarly, we show that after the breast cancer diagnosis, individuals with elevated PRS have an elevated risk of developing contralateral breast cancer, and that the PRS can considerably improve risk assessment among their female first-degree relatives. In more detail, women with the c.1592delT variant in *PALB2* (242-fold enrichment in Finland, 336 carriers) and an average PRS (10–90<sup>th</sup> percentile) have a lifetime risk of breast cancer at 55% (95% CI 49–61%), which increases to 84% (71–97%) with a high PRS (> 90<sup>th</sup> percentile), and decreases to 49% (30–68%) with a low PRS (< 10<sup>th</sup> percentile). Similarly, for c.1100delC in *CHEK2* (3.7-fold enrichment; 1648 carriers), the respective lifetime risks are 29% (27–32%), 59% (52–66%), and 9% (5–14%). The PRS also refines the risk assessment of women with first-degree relatives diagnosed with breast cancer, particularly among women with positive family history of early-onset breast cancer. Here we demonstrate the opportunities for a comprehensive way of assessing genetic risk in the general population, in breast cancer patients, and in unaffected family members.

<sup>1</sup>Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland. <sup>2</sup>Breast Surgery Unit, Comprehensive Cancer Center, Helsinki University Hospital, Helsinki, Finland. <sup>3</sup>University of Helsinki, Helsinki, Finland. <sup>4</sup>Helsinki Institute for Information Technology HIIT and Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland. <sup>5</sup>Department of Public Health, University of Helsinki, Helsinki, Finland. <sup>6</sup>Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia. <sup>7</sup>Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Analytic and Translational Genetics Unit, Department of Medicine, and the Department of Neurology, Massachusetts General Hospital, Boston, MA, USA. <sup>8</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>9</sup>Comprehensive Cancer Center, Helsinki University Hospital, Helsinki, Finland. \*A list of authors and their affiliations appears at the end of the paper. ✉email: [samuli.ripatti@helsinki.fi](mailto:samuli.ripatti@helsinki.fi)

In women, breast cancer is the most commonly diagnosed cancer and the leading cause of cancer-related deaths<sup>1</sup>. Approximately 5–10% of all breast cancers are estimated to develop due to high-impact germline mutations in breast cancer susceptibility genes, with up to 30% due to pathogenic mutations in *BRCA1* and *BRCA2* and with a smaller proportion carrying mutations in other susceptibility genes, such as *PTEN*, *TP53*, *CHEK2*, *PALB2* and *STK11*<sup>2</sup>. While pathogenic mutations in *BRCA1* and *BRCA2* are less common in Finns<sup>3</sup>, two frameshift mutations, c.1592delT (rs180177102) in *PALB2* and c.1100delC (rs555607708) in *CHEK2* have an unusually high allele frequency in Finland, which provides a unique opportunity to explore the impact of these mutations in the population. *PALB2* (Partner and Localizer of *BRCA2*) encodes a key tumour suppressor protein that functions through affecting *BRCA2* nuclear localisation and DNA damage response functions, and through interacting with *BRCA1*<sup>4</sup>. The second gene, *CHEK2* (Checkpoint kinase 2), is a tumour suppressor gene encoding a serine/threonine-protein kinase involved in DNA repair, cell cycle arrest and apoptosis<sup>5</sup>.

Beyond genetic predisposition caused by high-risk mutations in breast cancer susceptibility genes, breast cancer has a highly polygenic mode of inheritance. Large-scale genetic screens have to date identified over a hundred loci associated with risk of breast cancer<sup>6</sup>. These variants, and many more yet to be discovered, represent common genetic variation acting through a wide range of molecular pathways, in contrast to the rare, high-risk pathogenic variants in high-risk breast cancer susceptibility genes that often disrupt a specific pathway involved in maintaining integrity of DNA repair processes. Individually, the common variants have very small effect sizes with odds ratios usually ranging from 0.85 to 1.20, but their cumulative impact in breast cancer risk has been shown to be considerably larger<sup>7</sup>. This cumulative effect can be captured in a single measure by a polygenic risk score (PRS), the summed contribution of many common risk variants, which is able to identify women at over 3-fold risk of breast cancer, compared to women with an average risk<sup>7,8</sup>. By improving identification of these women at high risk of breast cancer, it could serve as a new tool for personalised, risk-based breast cancer screening<sup>7,9,10</sup>.

Here, we comprehensively assess the impact of germline genetic variation on risk of breast cancer and show (1) how a high breast cancer PRS compares to high-risk mutations in breast cancer susceptibility genes, (2) how the PRS modifies the risk of breast cancer in women carrying pathogenic mutations in the *PALB2* and *CHEK2* genes and (3) that the PRS has utility for informing about risk of contralateral breast cancer, and about the risk in first-degree relatives. We use data from the FinnGen study, which combines nationwide health registries with genomic information for 122,978 women from across the country, representing 5% of the Finnish adult female population.

## Results

We studied 122,978 women in FinnGen, with the mean age at the end of follow-up 58.5 (inter-quartile range, IQR 45.1–72.2, range 16.0–106.0). In FinnGen, 8401 (6.8%) women have been diagnosed with breast cancer, with mean age at disease onset of 58.6 (IQR 50.4–66.3, range 21.3–98.3 years). We first tested the association of three polygenic risk scores on breast cancer risk: a 313 SNP score<sup>7</sup>, a genome-wide score by Mars et al.<sup>10</sup> derived using LDpred software and a new genome-wide score derived using PRS-CS<sup>11</sup>. In our data, the genome-wide scores outperformed the 313 SNP score with hazard ratio (HR) estimates per standard deviation at 1.55 (95% confidence interval, CI 1.52–1.58), 1.63 (CI 1.60–1.67) and 1.71 (CI 1.68–1.75) of the PRS for 313 SNP score, LDpred score and PRS-CS score, each scaled

**Table 1 Comparison of effect sizes for three polygenic risk scores (PRS) with and without excluding the *PALB2* and *CHEK2* loci and regions around them.**

	HR (95% CI) <i>PALB2</i> and <i>CHEK2</i> included	HR (95% CI) <i>PALB2</i> and <i>CHEK2</i> excluded
PRS <sub>313</sub>	1.55 (1.52–1.58)	1.54 (1.51–1.57)
PRS <sub>LDpred</sub>	1.63 (1.60–1.67)	1.59 (1.55–1.62)
PRS <sub>CS</sub>	1.71 (1.68–1.75)	1.71 (1.68–1.75)

Out of the three PRSs tested, the PRS built with the software PRS-CS had the strongest association with breast cancer, and was therefore chosen for our analyses. Each PRS was scaled separately to mean zero and unit variance to obtain hazard ratios (HR) per standard deviation. To have a PRS independent of the *PALB2* and *CHEK2* variants, we excluded the variants within the *CHEK2* gene ±3 Mb, and variants within the *PALB2* gene ±2 Mb. CI confidence interval, PRS<sub>313</sub> a PRS with 313 genetic variants, with 260 variants polymorphic in FinnGen, PRS<sub>LDpred</sub> a PRS built with the software LDpred, PRS<sub>CS</sub> a PRS built with the software PRS-CS.

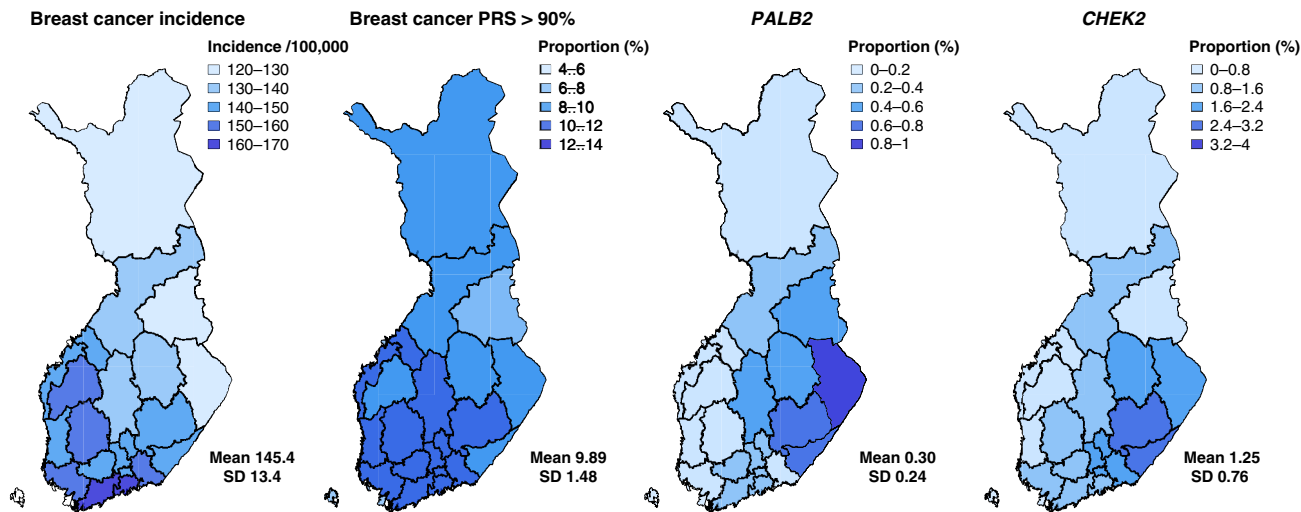
separately to mean zero and unit variance. We therefore chose the PRS built with PRS-CS for subsequent analyses (Table 1 and Supplementary Table 1).

We then studied the allele frequencies, geographic variation and risk estimates for the two Finnish-enriched, high-impact breast cancer mutations. The allele frequency for rs180177102 (*PALB2*) was 0.0014 (242-fold enrichment compared to non-Finnish non-Estonian Europeans, NFEE<sup>12</sup>), with 336 heterozygote mutation carriers included in the analyses. The allele frequency for rs555607708 (*CHEK2*) was 0.0064 (3.7 times enriched in Finns compared to NFEE), with 1641 heterozygotes and 7 homozygote individuals.

**Geographic variation of genetic risk.** Considering Finns have passed internal genetic bottlenecks, we first aimed to characterise any geographic distribution for both the *PALB2* and *CHEK2* mutations, and for the PRS. Both the *PALB2* and *CHEK2* mutations had more carriers in Eastern Finland, with the proportion of carriers ranging from close to 0 in Western Finland, to 2.8% for *CHEK2* in South Karelia and to 0.8% for *PALB2* in North Karelia (Fig. 1). In contrast, we observed slightly higher proportions of individuals with high PRS in Western and Southern Finland, in line with breast cancer incidence.

**The effect of frameshift mutations in *PALB2* and *CHEK2*.** Both *PALB2* and *CHEK2* conferred considerably elevated risk for breast cancer (Table 2). The *PALB2* variant conferred a risk increase for breast cancer with a HR of 4.99 (95% CI 4.02–6.20,  $p = 6.76 \times 10^{-48}$ ), corresponding to a lifetime risk by age 80 of 56.1% (95% CI 50.8–61.4%). The *CHEK2* variant conferred a risk increase for breast cancer with HR 2.19 (95% CI 1.91–2.51),  $p = 3.90 \times 10^{-29}$ , corresponding to a lifetime risk of 31.7% (95% CI 29.5–33.9%). Comparing to women with a PRS between the 10th and 90th percentiles (lifetime risk 15.5%, 95% CI 15.3–15.7%), women with PRS above the 90th percentile had a similar effect size as *CHEK2* mutation carriers (HR 2.38, 95% CI 2.26–2.50,  $p = 1.98 \times 10^{-230}$ ) and their similar lifetime risk was similar (32.5%, 95% CI 31.6–33.4%). However, a high PRS affected a nearly 7-fold larger group of women (Table 1; results excluding first-degree relatives in Supplementary Table 2). Estimating these while accounting for competing risks (non-breast cancer related death) yielded 4.6%, 4.9% and 3.2% lower estimates for lifetime risks in carriers of the *PALB2* and *CHEK2* mutations, and women with high PRS, respectively (Supplementary Figs. 1–3 and Supplementary Table 3).

**PRS modifies the risk in *PALB2* and *CHEK2* mutation carriers.** Next, we estimated how the PRS modifies breast cancer risk in the



**Fig. 1 Geographic variation in genetic risk.** The risk is compared to age-standardised breast cancer incidence. The proportion of women with the breast cancer polygenic risk score (PRS) above the 90th percentile in each region is estimated with respect to the PRS distribution of the whole country. The *PALB2* and *CHEK2* maps show across different regions the proportion of women carrying at least one risk allele for the variants. The areas represent region of birth obtained from Statistics Finland. The national breast cancer incidence in women was obtained from the Finnish Cancer Registry (publicly available at <https://cancerregistry.fi/statistics/>) with diagnosis C50 (International Classification of Diseases for Oncology, 3rd edn, ICD-O-3). The incidence represents the mean of 5-year age-standardised incidences (based on the 2014 Finnish population, calculated for each hospital district over 1998–2007). The mean and standard deviation were calculated over the different regions. Variants: rs180177102 (c.1592delT) for *PALB2* and rs555607708 (c.1100delC) for *CHEK2*. *CHEK2* and polygenic risk score plots are based on 122,978 women, and *PALB2* on 109,371 women. Colour contrasts were chosen approximately based on the standard deviation for each map.

**Table 2 Risk for breast cancer events in the population in carriers of the *PALB2* and *CHEK2* frameshift mutations, and in the top decile of the polygenic risk score (PRS).**

	<i>PALB2</i>	<i>CHEK2</i>	PRS > 90%
Number of individuals	336	1648	12,298
Number of cases	84	214	1821
Lifetime risk of breast cancer, % (95% CI)	56.1 (50.8–61.4)	31.7 (29.5–33.9)	32.5 (31.6–33.4)
Mean age at disease onset in cases (SD)	53.1 (10.4)	56.5 (12.0)	57.8 (11.3)

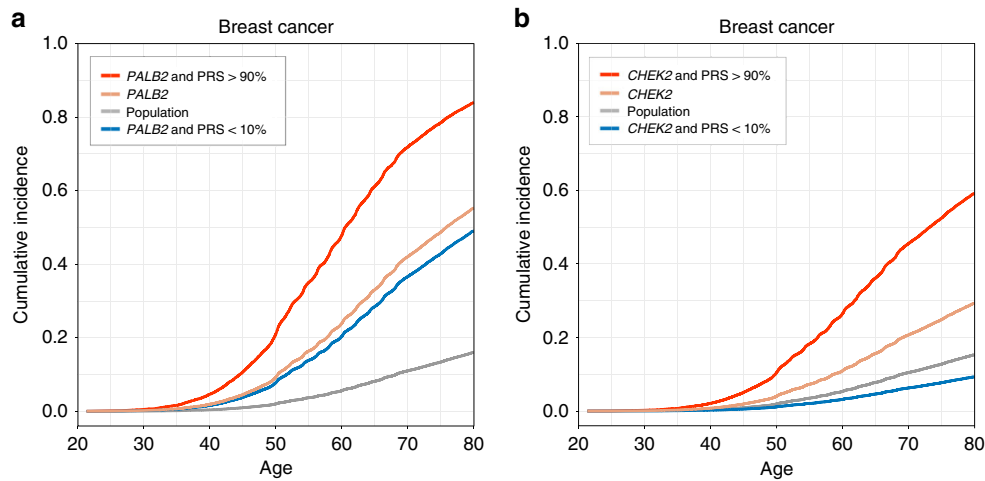
Lifetime risk was estimated by age 80. The variants were rs180177102 (c.1592delT) for *PALB2* and rs555607708 (c.1100delC) for *CHEK2*. The *PALB2* analysis was done in 109,371 women, and the *CHEK2* and PRS analyses in 122,978 women. CI confidence interval, SD standard deviation.

mutation carriers. For both *PALB2* and *CHEK2*, a high PRS further increased the breast cancer risk. In terms of lifetime risk for breast cancer by age 80, women with the *PALB2* mutation and average PRS (10–90th percentile) had a lifetime risk of 55.3% (95% CI 49.4–61.2%), which increased to 83.9% (71.2–96.6%) among women with a high PRS (>90th percentile), and decreased to 49.1% (30.6–67.6%) in women with a low PRS (<10th percentile; Fig. 2 and Tables 3 and 4). Women with *CHEK2* and an average PRS had a lifetime risk of 29.3% (95% CI 26.8–31.8%) which doubled to 59.2% (52.1–66.3%) in women with a high PRS and decreased to 9.3% (4.5–14.1%) in women with low PRS.

To test for possible interaction between mutation carriers and the PRS, we first compared the PRS effect size in pooled mutation carriers (*PALB2* and *CHEK2*) and in non-carriers. In both carriers and non-carriers, hazard ratios for the top and bottom decile of the PRS were very similar (reference group PRS 10–90%; Table 5). This was observed also in *PALB2* and *CHEK2* mutation carriers separately. For *PALB2*, the HR per SD in carriers was 1.81 (95% CI 1.34–2.44,  $p = 1.05 \times 10^{-4}$ ), in carriers of *CHEK2*, 1.86 (1.60–2.16,  $p = 6.58 \times 10^{-16}$ ), and in carriers of neither the *PALB2* nor the *CHEK2* mutation, the HR was 1.71 (1.67–1.74,  $p < 1.00 \times 10^{-300}$ ). Similarly, in a formal test for interaction by introducing an interaction term in the regression model, we found no evidence of an interaction between the PRS and

mutations for neither the *PALB2* variant ( $p = 0.18$ ), nor the *CHEK2* variant ( $p = 0.45$ ).

**PRS refines risk assessment in first-degree relatives.** Next, we evaluated how the PRS modifies the risk conferred by a positive first-degree family history. Family history was assessed in 7715 mother–daughter pairs and 12,086 pairs of sisters, separately for family history of early-onset (age < 45) and late-onset (age ≥ 45) breast cancer. For both, PRS stratified women for breast cancer risk, but the stratification was more pronounced in family history of early-onset disease (Fig. 3 and Supplementary Table 4). Women with an average PRS (between the 10th and 90th percentiles) and positive family history of early-onset breast cancer had a lifetime risk at 32.5% (95% CI 24.0–41.0%) – a risk similar to women with a high PRS (>90th percentile) in the full dataset (32.5%, 31.6–33.4%). A combination of family history of early-onset breast cancer and a high PRS further increased the risk to 49.0% (30.1–67.9%), but with only one breast cancer case in the bottom decile we were unable to estimate the impact of a low PRS. We then tested whether family history adds to risk assessment if we know the woman’s PRS. When adjusting with a continuous PRS, the effect size for family history of early-onset breast cancer was attenuated, from HR 2.80 (95% CI 1.81–4.33,



**Fig. 2 The impact of polygenic risk in PALB2 and CHEK2 mutation carriers.** Adjusted survival curves showing how the polygenic risk score (PRS) affects the breast cancer risk conferred by the PALB2 (panel A) and CHEK2 (panel B) frameshift mutations. Population level was defined as women with PRS between the 10th and 90th percentiles. The PALB2 analysis was done in 109,371 women and CHEK2 analysis in 122,978 women. Adjusted survival curves Cox proportional hazards model.

**Table 3 Impact of polygenic risk score (PRS) on the breast cancer risk conferred by the PALB2 frameshift mutation.**

	HR (95% CI)	p value	Lifetime risk, % (95% CI)	Cases	Controls
PALB2 & PRS >90%	11.8 (6.95–20.0)	$5.14 \times 10^{-20}$	83.9 (71.2–96.6)	14	18
PALB2	4.82 (3.77–6.16)	$5.66 \times 10^{-36}$	55.3 (49.4–61.2)	65	211
PRS 10–90%	1.00 (reference)	–	16.0 (15.8–16.2)	5605	81,645
PALB2 & PRS <10%	4.00 (1.66–9.63)	0.002	49.1 (30.6–67.6)	5	23

Population level was defined as women with PRS between the 10th and 90th percentiles. The estimates were obtained from a Cox proportional hazards model on 109,371 women, without adjusting for multiple comparisons. All tests were two-tailed.

**Table 4 Impact of polygenic risk score (PRS) on the breast cancer risk conferred by the CHEK2 frameshift mutation.**

	HR (95% CI)	p value	Lifetime risk, % (95% CI)	Cases	Controls
CHEK2 & PRS >90%	5.71 (4.33–7.52)	$5.31 \times 10^{-35}$	59.2 (52.1–66.3)	51	133
CHEK2	2.12 (1.81–2.48)	$2.35 \times 10^{-20}$	29.3 (26.8–31.8)	158	1163
PRS 10–90%	1.00 (reference)	–	15.3 (15.1–15.5)	6116	90,945
CHEK2 & PRS <10%	0.58 (0.24–1.41)	0.23	9.3 (4.5–14.1)	5	138

Population level was defined as women with PRS between the 10th and 90th percentiles. The estimates were obtained from a Cox proportional hazards model in 122,978 women, without adjusting for multiple comparisons. All tests were two-tailed.

**Table 5 To test for interaction in all 122,978 women, we compared the polygenic risk score (PRS) effect size in pooled mutation carriers (pooling PALB2 and CHEK2) and in non-carriers.**

	PRS < 10%	PRS 10–90%	PRS > 90%
Mutation	0.42 (0.23–0.79)	1.00 (reference)	2.44 (1.82–3.28)
No mutation	0.38 (0.34–0.43)	1.00 (reference)	2.37 (2.25–2.50)

The table shows the hazard ratios and 95% confidence intervals for the bottom and top deciles, comparing them to women with an average risk (PRS between the 10th and 90th percentiles).

$p = 4.08 \times 10^{-6}$ ), to HR 2.32 (1.50–3.60,  $p = 1.72 \times 10^{-4}$ ). Also for late-onset, the association was attenuated, from HR 1.30 (1.07–1.57,  $p = 0.01$ ), to HR 1.09 (0.90–1.33,  $p = 0.37$ ).

**High PRS increases risk for contralateral breast cancer.** Lastly, we tested the association between the PRS contralateral breast

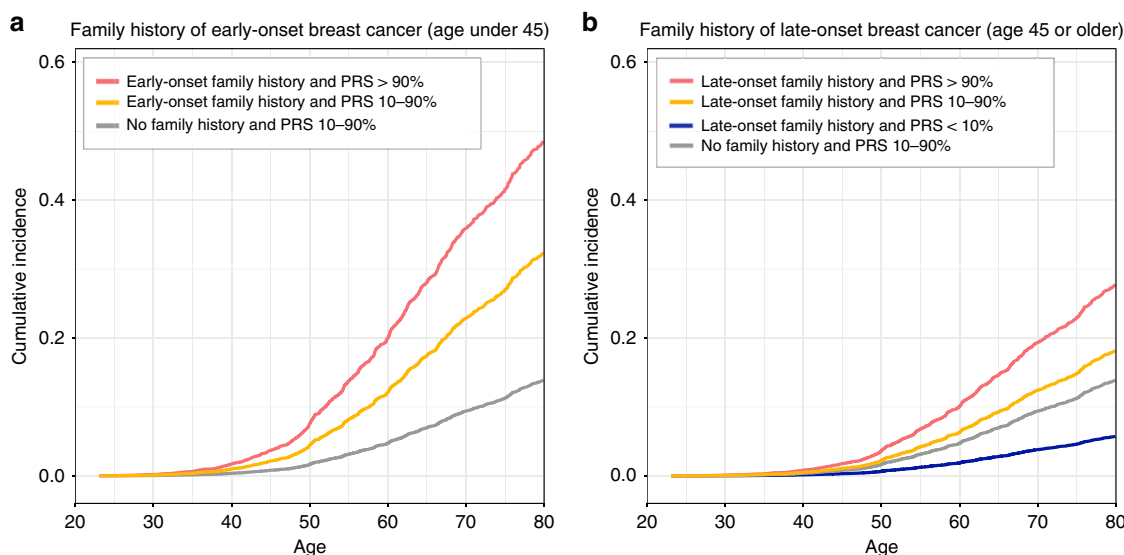
cancer among breast cancer patients. With PRS between the 10th and 90th percentile as reference, a high PRS (>90th percentile) was associated with risk of contralateral breast cancer with HR 1.60 (95% CI 1.25–2.04,  $p = 0.0002$ ), with 97 individuals out of 1604 cases with a high PRS being diagnosed with contralateral breast cancer.

**Discussion**

Using large-scale biobank data combining longitudinal nationwide health registry data with genomic information, we show that over the life course, the breast cancer PRS strongly alters the breast cancer incidence in high-impact mutation carriers. After breast cancer diagnosis, individuals with an elevated PRS have an increased likelihood of developing contralateral breast cancer, and the PRS can considerably improve risk assessment among their female first-degree relatives.

The breast cancer PRS strongly altered the risk of breast cancer in PALB2 and CHEK2 mutation carriers, substantially increasing the risk of breast cancer in women with a high PRS, and lowering the risk in women with a low PRS. Deciding on appropriate





**Fig. 3** Impact of the polygenic risk score (PRS) in estimating the breast cancer risk of women with a first-degree relative diagnosed with breast cancer.

**a** Shows the impact of family history of early-onset breast cancer, and **b** the impact of family history of late-onset breast cancer. Adjusted survival curves based on Cox proportional hazards models. Risk estimated in 7715 mother–daughter pairs and 12,086 full sibling-pairs (sisters). The pairs of first-degree relatives were inferred with KING by a kinship coefficient ranging between 0.177 and 0.354 (inference based on 57 K unlinked variants). Due to the sample size, we were unable to assess impact of a low PRS (<10th percentile) with early-onset family history.

surveillance and risk-reduction strategies is a clinical challenge particularly for moderate-risk mutations such as those in *CHEK2*<sup>13</sup>, and our results show that additional information provided by the PRS could guide in these decisions. A combination of breast cancer PRS in the top decile and a mutation in the *CHEK2* variant increased the lifetime risk to 59% – a risk comparable to that seen in *PALB2* mutation carriers – whereas those with a PRS in the bottom decile had a risk similar to the population level.

That PRS modifies the risk in *PALB2* and *CHEK2* mutation carriers supports previous findings suggesting that common genetic variation at least partly explains the widely observed incomplete penetrance of mutations in breast cancer susceptibility genes<sup>14–17</sup>. This variation is now measurable on an individual level with the breast cancer PRS, which captures a wide range of molecular pathways. Our results are in line with previous studies on *BRCA1*, *BRCA2*, *PALB2* and *CHEK2* mutation carriers, but these studies have used a case–control setting or PRSs consisting of <100 variants<sup>16–19</sup>. We conducted the study in a large longitudinal dataset with 120,000 women, using a more predictive, genome-wide PRS and leveraging the considerable enrichment of the *PALB2* and *CHEK2* variants in an isolated population. With the longitudinal setting, we were also able to quantify the lifetime risk in *PALB2* and *CHEK2* mutation carriers based on observed events over the life course, instead of calculating them using baseline risks from published studies<sup>17,18</sup>.

Harbouring pathogenic mutations in high-risk breast cancer susceptibility genes often prompt intensified medical surveillance and consideration of preventative procedures such as risk-reducing surgery. The lifetime risk estimates for individuals in the top decile of the PRS was comparable to *CHEK2* mutation carriers – both had a risk of 32% by age 80. Considering this, our results also argue for the need of studies on the impact of targeted actions in women with a high PRS only, who currently go undetected. After the diagnosis, patients with elevated PRS had a 1.6-fold elevated risk for contralateral cancer, providing additional evidence of increased breast cancer susceptibility, a finding that might warrant intensified or prolonged surveillance in breast cancer cases with elevated PRS. This finding is in line with earlier

studies showing that familial factors contribute to the risk of contralateral breast cancer<sup>20–22</sup>.

The proportion of mutation carriers and the elevated PRS showed differing geographical distributions. While the elevated PRS distribution followed the breast cancer incidence distribution with highest rates in the early-settlement region in South-Western Finland, the allele frequencies for the *PALB2* and *CHEK2* mutations were highest in the late-settlement region in Eastern Finland. It is likely that the *PALB2* and *CHEK2* mutations have survived both the founder bottleneck in Finland, and the internal bottleneck in the Eastern Finland, therefore being heavily enriched in the Finnish population. These regional differences in both PRS and mutation frequency distributions may have an impact on regional screening strategies.

Finally, the PRS improved risk assessment of first-degree relatives of women with breast cancer, with pronounced stratification particularly for family history of early-onset disease. Family history is an essential factor guiding screening strategies of family members of breast cancer patients<sup>23</sup>, and our results show that PRS could improve the precision of this assessment.

Our study has several limitations. Our findings are limited to individuals of European ancestry and it is important to study the applicability of the results in individuals of admixed and non-European ancestry<sup>24</sup>. The FinnGen study is a mixture of population-based cohorts and samples from hospital biobanks. It is possible that the sampling may introduce biases in some of the estimates. We observed a slightly higher baseline risk compared to the NORDCAN database<sup>25</sup>. However, our key PRS estimates were similar when estimated in a FinnGen subset of population-based cohorts only. Moreover, accounting for the competing risk of mortality from other causes yielded slightly lower estimates for lifetime risks.

In conclusion, we show that a high breast cancer PRS comes with a comparable risk profile to frameshift mutations in breast cancer susceptibility genes *PALB2* and *CHEK2*, and that the PRS strongly modifies breast cancer risk in the mutation carriers. Even after the breast cancer diagnosis, the PRS was associated with breast cancer susceptibility by increasing the risk of contralateral breast cancer, and it considerably improved risk assessment

among the patient's first-degree relatives. These results demonstrate opportunities for a more comprehensive way of assessing genetic risk in the general population, in breast cancer patients and in unaffected family members of breast cancer patients. Optimisation of these strategies in the clinical setting warrant further study.

## Methods

**Participants and endpoints.** The data comprised of 122,978 Finnish women in the FinnGen, Data Freeze 5. FinnGen comprises prospective epidemiological cohorts (initiated as far back as 1992), disease-based cohorts, and hospital biobank samples (Supplementary Table 5). The unique national personal identification number links the genotypes to the Finnish Cancer Registry (available from 1953, with nationwide completeness of solid tumours at 96%<sup>26</sup>), as well as to the national hospital discharge registry (1968-), the national death registry (1969-) and the medication reimbursement registry (1964-). These registries cover the whole population.

Breast cancer cases were identified through the Finnish Cancer Registry with diagnosis C50 (International Classification of Diseases for Oncology, 3rd Edition; ICD-O-3), from the drug reimbursement registry by selecting individuals with a reimbursement code for breast cancer, and from the death registry with ICD-10 C50. Contralateral breast cancer was defined as breast cancer in the opposite breast diagnosed over 6 months after the date of the primary breast cancer diagnosis, obtained from the Cancer Registry.

**Genotyping and imputation.** FinnGen samples were genotyped with Illumina and Affymetrix arrays (Illumina Inc., San Diego, and Thermo Fisher Scientific, Santa Clara, CA, USA), and genotype calls were made with the GenCall or zCall (for Illumina) and the AxiomGT1 algorithm for Affymetrix data. Individuals with ambiguous gender, high genotype missingness (>5%), excess heterozygosity (+4SD) and non-Finnish ancestry were excluded, as well as all variants with high missingness (>2%), low Hardy-Weinberg equilibrium *p*-value (<1e-6) and minor allele count (MAC < 3). Array data pre-phasing was carried out with Eagle 2.3.5<sup>27</sup> with the number of conditioning haplotypes set to 20,000. Genotype imputation was done with Beagle 4.1<sup>28</sup> (as described in <https://doi.org/10.17504/protocols.io.xbgfjw>) by using the SISu v3 population-specific reference panel developed from high-quality data for 3,775 high-coverage (25-30x) whole-genome sequences in Finns.

**Variants.** We chose two previously reported Finnish-enriched frameshift variants for our main analyses, rs180177102 (c.1592delT) in *PALB2* and rs555607708 (c.1100delC) in *CHEK2*. Genotype data batches with an imputation INFO score < 0.8 were excluded. This excluded 13,607 women from analyses involving the *PALB2* variant (mainly older disease-based cohorts), but no exclusions were needed for *CHEK2*. *PALB2* mutation carrier status was ignored in analyses involving the *CHEK2* variant, and vice versa. Women homozygous for the *CHEK2* variant were analysed jointly with the heterozygotes.

**Polygenic risk score.** To choose our breast cancer PRS, we compared three scores: (1) a previously published PRS with 313 SNPs<sup>7</sup>, (2) another previously published, genome-wide PRS<sup>10</sup> built with the software LDpred<sup>29</sup> and (3) a genome-wide PRS we built with the software PRS-CS (PRS-CS-auto, with 1000 Genomes Project European sample, *N* = 503, as the external LD reference panel) using HapMap3 variants<sup>11</sup>. For the LDpred and PRS-CS PRSs, the input weights came from a large independent genome-wide association study (GWAS)<sup>6</sup>. To have a PRS independent of the *PALB2* and *CHEK2* variants, we excluded the variants within the *CHEK2* gene ±3 Mb, and variants within the *PALB2* gene ±2 Mb (Supplementary Fig. 4). Out of these three, the PRS-CS score showed the strongest association for breast cancer and was therefore chosen for subsequent analyses (Table 1 and Supplementary Table 1). All three PRSs showed acceptable goodness-of-fit (Supplementary Fig. 5). The final variant count for the PRS-CS PRS with *PALB2* and *CHEK2* excluded was 1,074,667.

A high PRS was defined as a PRS above the 90th percentile, as it corresponds to a lifetime risk of ≥30%, which guidelines consider as the threshold for high risk<sup>23</sup>. Correspondingly, we defined a PRS below the 10th percentile as a low PRS. Individuals between the 10th to 90th percentiles served as the reference category.

**Geographic variation.** Geographic variation is reported by region of birth (obtained from Statistics Finland) as the proportion of individuals with (1) the frameshift mutations in the *PALB2* or *CHEK2* variants, and (2) high PRS (>90th percentile). The benchmark for these analyses was age-standardised (age in 2014) breast cancer incidence for the whole Finnish population, calculated as the mean of 5-year incidences for each hospital district over 1998–2007. The incidence data was obtained from the Finnish Cancer Registry (publicly available at <https://cancerregistry.fi/statistics/>). Polygen data for the Finnish map were obtained from GADM (<https://gadm.org/data.html>).

**Population structure-related bias analysis.** A population structure-related bias analysis was performed by following the approach described in detail in Kerminen et al.<sup>30</sup>. In brief, the method measures the accumulation of PRS differences between the Western and Eastern subpopulations of Finland using a “random PRS”, made from a randomly chosen set of independent ( $r^2 < 0.1$ ) variants with minor allele frequency >0.05 that are not associated with breast cancer (breast cancer GWAS<sup>6</sup> *p*-value >0.5). If such random PRS accumulated differences between the subpopulations, that could indicate a population genetic bias in effect estimates of the GWAS, rather than a real difference in genetic susceptibility of breast cancer between the subpopulations. We found no evidence of such bias (Supplementary Fig. 6), which indicates that any detected geographic variation in the PRS is unlikely to result from a population genetic bias.

**Risk assessment in first-degree relatives.** The pairs of first-degree relatives were inferred with KING v2.2.4<sup>31</sup> by a kinship coefficient ranging between 0.177 and 0.354 (inference based on 57 K unlinked variants). To analyse the impact of family history in first-degree relatives, we randomly sampled one female relative for each woman who had at least one first-degree relative in the dataset. For mother–daughter pairs, the mother was assigned as the index relative. For sisters, we randomly assigned one to be the index relative, irrespective of age. If both women in the pair were breast cancer cases, we used the year of diagnosis to assign the woman diagnosed earlier as the index. Some individuals appeared several times as non-index individuals, which may occur when, for instance, a woman is the daughter of one index individual and the sister of another – we therefore randomly sampled the data to contain each non-index individual only once. We then inferred the risk of breast cancer in these unique non-index individuals. We analysed separately family history of early-onset (age < 45) and late-onset (age ≥ 45) breast cancer.

**Statistical analysis.** We estimated HRs and 95% CIs with the Cox proportional hazards model, and used Schoenfeld residuals and log-log inspection for assessing the proportional hazards assumption. Start of follow-up was set at birth, and follow-up ended at the first record of the endpoint of interest, death or at the end of follow-up on 31 December 2018, whichever came first. All tests were two-tailed. In all survival analyses, we used age as the time scale, with 63 batches and the first 10 principal components as covariates. The only exception was the analysis on contralateral breast cancer, where follow-up started from the diagnosis, and age was included as a covariate.

Goodness-of-fit for the PRS was assessed with a method proposed by May & Hosmer for a Cox proportional hazards model<sup>32</sup>. In line with previous studies on breast cancer susceptibility genes, we assessed lifetime risk (cumulative incidence without competing risks) by age 80<sup>14,33</sup>. Lifetime risk was estimated from the adjusted survival curves, with 95% CIs obtained by normal approximation. The adjusted survival curves were plotted with the R package *survminer*. This presents the expected survival curves separately for subgroups, based on the Cox model. To estimate the covariate-adjusted cumulative incidence functions in the presence of competing risks, we used the Stata module *stcompadj*<sup>34</sup>. The competing event was non-breast cancer causes of death and covariates were assumed to have similar effects the main and competing event.

Interactions between the PRS and the pathogenic mutations were assessed (1) by comparing the PRS effect sizes in pooled and non-pooled mutation carriers and non-carriers (with the PRS scaled to zero mean and unit variance within the whole dataset), and (2) formally by introducing an interaction term for the mutation and the continuous PRS. For data and variant handling and PRS calculation, we used BCFtools versions 1.7 and 1.9, and PLINK 2.0. For statistical analyses, we used R 3.6.3 and Stata 16.0 (College Station, TX, USA). Cromwell and WOMtool were used for workflow handling.

**Ethics statement.** The FinnGen project is approved by the Finnish Institute for Health and Welfare (THL), approval number THL/2031/6.02.00/2017, amendments THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019), Digital and population data service agency VRK43431/2017-3, VRK/6909/2018-3, the Social Insurance Institution (KELA) KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019 and Statistics Finland TK-53-1041-17.

Patients and control subjects in FinnGen provided informed consent for biobank research, based on the Finnish Biobank Act. Alternatively, older research cohorts, collected prior the start of FinnGen (in August 2017), were collected based on study-specific consents and later transferred to the Finnish biobanks after approval by Valvira, the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Valvira. The Ethics Review Board of the Hospital District of Helsinki and Uusimaa approved the FinnGen study protocol Nr HUS/990/2017.

The Biobank Access Decisions for FinnGen samples and data utilised in FinnGen Data Freeze 5 include: THL Biobank BB2017\_55, BB2017\_111, BB2018\_19, BB\_2018\_34, BB\_2018\_67, BB2018\_71, BB2019\_7 Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, Auria Biobank AB17-5154, Biobank Borealis of Northern Finland\_2017\_1013, Biobank of Eastern Finland 1186/2018, Finnish Clinical Biobank Tampere MH0004, Central Finland

Biobank 1-2017 and Terveystalo Biobank STB 2018001. Analyses of potential geographic bias of PRS were done with THL biobank permission BB2019\_44.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

### Data availability

The FinnGen data may be accessed through Finnish Biobanks' FinBB portal (web link: [www.finbb.fi](http://www.finbb.fi), email: [info.fingenious@finbb.fi](mailto:info.fingenious@finbb.fi)). The GWAS summary statistics used for constructing our main PRS are available at <http://bcac.ccg.medschl.cam.ac.uk/bcacdata/oncoarray/oncoarray-and-combined-summary-result/gwas-summary-results-breast-cancer-risk-2017/>, with contact information at <http://bcac.ccg.medschl.cam.ac.uk/contact/>. The weights for our main PRS are available at PGS Catalog ([pgs-info@ebi.ac.uk](https://www.PGSCatalog.org/score/PGS000335/)) at <https://www.PGSCatalog.org/score/PGS000335/>, and the previously published PRSs at <https://www.PGSCatalog.org/score/PGS000004/> and <https://www.PGSCatalog.org/score/PGS000332/>. The remaining data are available within the Article, Supplementary Information or available from the authors upon request.

### Code availability

The full genotyping and imputation protocol for FinnGen is described at <https://doi.org/10.17504/protocols.io.xbgfjw>.

Received: 15 July 2020; Accepted: 6 November 2020;

Published online: 14 December 2020

### References

- Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
- Economopoulou, P., Dimitriadis, G. & Psyrri, A. Beyond BRCA: new hereditary breast cancer susceptibility genes. *Cancer Treat. Rev.* **41**, 1–8 (2015).
- Vehmanen, P. et al. Low proportion of *BRCA1* and *BRCA2* mutations in Finnish breast cancer families: evidence for additional susceptibility genes. *Hum. Mol. Genet.* **6**, 2309–2315 (1997).
- Ducy, M. et al. The tumor suppressor *PALB2*: Inside out. *Trends Biochem. Sci.* **44**, 226–240 (2019).
- Nevanlinna, H. & Bartek, J. The *CHEK2* gene and inherited breast cancer susceptibility. *Oncogene* **25**, 5912–5919 (2006).
- Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
- Mavaddat, N. et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).
- Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
- Lee, A. et al. BOADICEA: A comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet. Med.* **21**, 1708–1718 (2019).
- Mars, N. et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat. Med.* **26**, 549–557 (2020).
- Ge, T., Chen, C. Y., Ni, Y., Feng, Y. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
- Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
- Tung, N. et al. Counselling framework for moderate-penetrance cancer-susceptibility mutations. *Nat. Rev. Clin. Oncol.* **13**, 581–588 (2016).
- Antoniou, A. C. et al. Breast-cancer risk in families with mutations in *PALB2*. *New Engl. J. Med.* **371**, 497–506 (2014).
- Antoniou, A. C. et al. A comprehensive model for familial breast cancer incorporating *BRCA1*, *BRCA2* and other genes. *Br. J. Cancer* **86**, 76–83 (2002).
- Kuchenbaecker, K. B. et al. Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in *BRCA1* and *BRCA2* mutation carriers. *J. Natl. Cancer Inst.* **109**, djw302 (2017).
- Muranen, T. A. et al. Genetic modifiers of *CHEK2*\*1100delC-associated breast cancer risk. *Genet. Med.* **19**, 599–603 (2017).
- Gallagher, S. et al. Association of a polygenic risk score with breast cancer among women carriers of high- and moderate-risk breast cancer genes. *JAMA Netw. Open* **3**, e208501 (2020).
- Fahed, A. C. et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.* **11**, 3635 (2020).

- Reiner, A. S. et al. Breast cancer family history and contralateral breast cancer risk in young women: An update from the women's environmental cancer and radiation epidemiology study. *J. Clin. Oncol.* **36**, 1513–1520 (2018).
- Narod, S. A., Kharazmi, E., Fallah, M., Sundquist, K. & Hemminki, K. The risk of contralateral breast cancer in daughters of women with and without breast cancer. *Clin. Genet.* **89**, 332–335 (2016).
- Robson, M. E. et al. Association of common genetic variants with contralateral breast cancer risk in the WECARE study. *J. Natl. Cancer Inst.* **109**, djx051 (2017).
- National Collaborating Centre for Cancer. NICE clinical guidelines, no. 164. Familial breast cancer: classification and care of people at risk of familial breast cancer and management of breast cancer and related risks in people with a family history of breast cancer (2013).
- Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
- Danckert, B. et al. NordCAN: cancer incidence, mortality, prevalence and survival in the Nordic countries, version 8.2 (26.03.2019). <http://www.dep.iaarc.fr/nordcan/> Accessed on 13 July 2020.
- Leinonen, M. K., Miettinen, J., Heikkinen, S., Pitkaniemi, J. & Malila, N. Quality measures of the population-based Finnish cancer registry indicate sound data quality for solid malignant tumours. *Eur. J. Cancer* **77**, 31–39 (2017).
- Loh, P. R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
- Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
- Vilhjalmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
- Kerminen, S. et al. Geographic variation and bias in the polygenic scores of complex diseases and traits in Finland. *Am. J. Hum. Genet.* **104**, 1169–1181 (2019).
- Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
- May, S. & Hosmer, D. W. A simplified method of calculating an overall goodness-of-fit test for the Cox proportional hazards model. *Lifetime Data Anal.* **4**, 109–120 (1998).
- Yang, X. et al. Ovarian and breast cancer risks associated with pathogenic variants in *RAD51C* and *RAD51D*. *J. Natl. Cancer Inst.* **112**, djao30 (2020).
- Coviello E. stcompadj: Stata Module To Estimate The Covariate-adjusted Cumulative Incidence Function In The Presence Of Competing Risks. Statistical Software Components S457063 (Department of Economics, Boston College, 2009).

### Acknowledgements

We would like to thank Sari Kivikko, Huei-Yi Shen and Ulla Tuomainen for management assistance. Following biobanks are acknowledged for collecting the FinnGen project samples: Auria Biobank (<https://www.auria.fi/biopankki>), THL Biobank (<https://thl.fi/web/thl-biopankki>), Helsinki Biobank (<https://www.terveyskyla.fi/helsinginbiopankki>), Biobank Borealis of Northern Finland (<https://www.oulu.fi/university/node/38474>), Finnish Clinical Biobank Tampere ([https://www.tays.fi/en-US/Research\\_and\\_development/verenluovutus/biopankkitoiminta](https://www.tays.fi/en-US/Research_and_development/verenluovutus/biopankkitoiminta)), Biobank of Eastern Finland (<https://ita-suomenbiopankki.fi>), Central Finland Biobank (<https://www.ksshp.fi/fi-FI/Potilaalle/Biopankki>), Finnish Red Cross Blood Service Biobank (<https://www.veripalvelu.fi/verenluovutus/biopankkitoiminta>) and Terveystalo Biobank (<https://www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/>). All Finnish Biobanks are members of BBMRI.fi infrastructure ([www.bbmrif.org](http://www.bbmrif.org)). We also thank study participants for their generous participation at THL Biobank and the National FINRISK study. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and by twelve industry partners (AbbVie Inc, AstraZeneca UK Ltd, Biogen MA Inc, Celgene Corporation, Celgene International II Sàrl, Genentech Inc, Merck Sharp & Dohme Corp, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc and Novartis AG). This work was supported by the Sigrid Jusélius Foundation (to S.R., A.P., M.P., and H.J.); University of Helsinki HiLIFE Fellow grants 2017–2020 (to S.R.); Academy of Finland Center of Excellence in Complex Disease Genetics (grant number 312062 to S.R., 312074 to A.P., 312075 to M.D.; 312073 to J.K.; 312076 to M.P.); Academy of Finland (grant number 331671 to N.M., grant number 285380 to S.R., 128650 to A.P., 308248 to J.K., 288509 to M.P., 218068 and 131449 to H.J.); The Finnish Innovation Fund Tekes (grant number 2273/31/2017 to E.W.); Foundation and the Horizon 2020 Research and Innovation Programme (grant number 667301 (COSYN) to A.P.); Cancer Foundation Finland sr (to T.M.); Cancer Society of Finland (to H.J.); and Jane and Aatos Erkkö Foundation (to H.J.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

### Author contributions

S.R. and N.M. conceived and designed the study. N.M., S.K. and P.D.B.P. carried out the statistical and computational analyses with advice from S.R., E.W., T.M., M.P., A.P., J.K.,



H.J. and M.D. Quality control of the data was carried out by N.M., S.K., P.D.B.P. and P.P. All authors provided critical inputs to interpretation of the data. The manuscript was written and revised by N.M. and S.R. with comments from all of the co-authors. All co-authors have approved the final version of the manuscript.

### Competing interests

A.P. is a member of the Pfizer Genetics Scientific Advisory Panel. H.J. has a co-appointment at Orion Pharma, has received fees from Neutron Therapeutics, and owns stocks of Orion Pharma and Sartar Therapeutics. The remaining authors declare no competing interests.

### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-020-19966-5>.

**Correspondence** and requests for materials should be addressed to S.R.

**Peer review information** *Nature Communications* thanks Peter Kraft and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

## FinnGen

Aarno Palotie<sup>1,7,8</sup>, Aki Havulinna<sup>1</sup>, Amanda Elliott<sup>1,8</sup>, Anastasia Shcherban<sup>1</sup>, Andrea Ganna<sup>1</sup>, Anu Jalanko<sup>1</sup>, Arto Lehisto<sup>1</sup>, Elina Kilpeläinen<sup>1</sup>, Georg Brein<sup>1</sup>, Ghazal Awaisa<sup>1</sup>, Hannele Laivuori<sup>1</sup>, Henrike Heyne<sup>1</sup>, Jaakko Kaprio<sup>1,5</sup>, Jarmo Harju<sup>1</sup>, Jiwoo Lee<sup>1,8</sup>, Juha Karjalainen<sup>1,8</sup>, Jukka Koskela<sup>1</sup>, Kalle Pärn<sup>1</sup>, Kati Donner<sup>1</sup>, Kristin Tsuo<sup>1,8</sup>, Manuel González Jiménez<sup>1</sup>, Mark Daly<sup>1,8</sup>, Mari Kaunisto<sup>1</sup>, Mari E. Niemi<sup>1</sup>, Mary Pat Reeve<sup>1</sup>, Mervi Aavikko<sup>1</sup>, Mitja Kurki<sup>1,8</sup>, Nina Mars<sup>1</sup>, Oluwaseun A. Dada<sup>1</sup>, Pietro della Briotta Parolo<sup>1</sup>, Priit Palta<sup>1,6</sup>, Risto Kajanne<sup>1</sup>, Samuli Ripatti<sup>1,6,8</sup>, Sina Rüeger<sup>1</sup>, Susanna Lemmelä<sup>1</sup>, Taru Tukiainen<sup>1</sup>, Tiinamaija Tuomi<sup>1</sup>, Timo P. Sipilä<sup>1</sup>, Tuomo Kiiskinen<sup>1</sup>, Vincent Llorens<sup>1</sup>, Adam Ziemann<sup>10</sup>, Anne Lehtonen<sup>10</sup>, Apinya Lertratanakul<sup>10</sup>, Bob Georgantas<sup>10</sup>, Bridget Riley-Gillis<sup>10</sup>, Danjuma Quarless<sup>10</sup>, Fedik Rahimov<sup>10</sup>, Howard Jacob<sup>10</sup>, Jeffrey Waring<sup>10</sup>, Justin Wade Davis<sup>10</sup>, Nizar Smaoui<sup>10</sup>, Relja Popovic<sup>10</sup>, Sahar Esmaeeli<sup>10</sup>, Athena Matakidou<sup>11</sup>, Ben Challis<sup>11</sup>, David Close<sup>11</sup>, Eleonor Wigmore<sup>11</sup>, Slavé Petrovski<sup>11</sup>, Chia-Yen Chen<sup>12</sup>, Ellen Tsai<sup>12</sup>, Heiko Runz<sup>12</sup>, Jimmy Liu<sup>12</sup>, Paola Bronson<sup>12</sup>, Sally John<sup>12</sup>, Sanni Lahdenperä<sup>12</sup>, Stephanie Loomis<sup>12</sup>, Susan Eaton<sup>12</sup>, Yunfeng Huang<sup>12</sup>, Erika Kvikstad<sup>13</sup>, Minal Caliskan<sup>13</sup>, Samir Wadhawan<sup>13</sup>, Elmutaz Shaikho Elhaj Mohammed<sup>14</sup>, Janet van Adelsberg<sup>14</sup>, Joseph Maranville<sup>14</sup>, Marla Hochfeld<sup>14</sup>, Robert Plenge<sup>14</sup>, Shameek Biswas<sup>14</sup>, Steven Greenberg<sup>14</sup>, Andrew Peterson<sup>15</sup>, David Choy<sup>15</sup>, Diana Chang<sup>15</sup>, Edmond Teng<sup>15</sup>, Erich Strauss<sup>15</sup>, Geoff Kerchner<sup>15</sup>, Hao Chen<sup>15</sup>, Hubert Chen<sup>15</sup>, Jennifer Schutzman<sup>15</sup>, John Michon<sup>15</sup>, Julie Hunkapiller<sup>15</sup>, Mark McCarthy<sup>15</sup>, Natalie Bowers<sup>15</sup>, Sarah Pendergrass<sup>15</sup>, Tushar Bhangale<sup>15</sup>, David Pulford<sup>16</sup>, Dawn Waterworth<sup>16</sup>, Diptee Kulkarni<sup>16</sup>, Fanli Xu<sup>16</sup>, Jo Betts<sup>16</sup>, Jorge Esparza Gordillo<sup>16</sup>, Joshua Hoffman<sup>16</sup>, Karen S. King<sup>16</sup>, Meg Ehm<sup>16</sup>, Soumitra Ghosh<sup>16</sup>, Patrick Loerch<sup>17</sup>, Wilco Fleuren<sup>17</sup>, Beryl Cummings<sup>18</sup>, Matt Brauer<sup>18</sup>, Robert Graham<sup>18</sup>, Tim Behrens<sup>18</sup>, Andrey Loboda<sup>19</sup>, Anna Podgornaia<sup>19</sup>, Aparna Chhibber<sup>19</sup>, Audrey Chu<sup>19</sup>, Caroline Fox<sup>19</sup>, Dorothee Diogo<sup>19</sup>, Emily Holzinger<sup>19</sup>, John Eicher<sup>19</sup>, Padhraig Gormley<sup>19</sup>, Vinay Mehta<sup>19</sup>, Anders Malarstig<sup>20</sup>, Åsa Hedman<sup>20</sup>, Catherine Marshall<sup>20</sup>, Christopher Whelan<sup>20</sup>, Heli Lehtonen<sup>20</sup>, Jaakko Parkkinen<sup>20</sup>, Kari Linden<sup>20</sup>, Kirsi Kalpala<sup>20</sup>, Melissa Miller<sup>20</sup>, Nan Bing<sup>20</sup>, Stefan McDonough<sup>20</sup>, Xing Chen<sup>20</sup>, Xinli Hu<sup>20</sup>, Ying Wu<sup>20</sup>, Anthony Muslin<sup>21</sup>, Clarence Wang<sup>21</sup>, Clement Chatelain<sup>21</sup>, Deepak Rajpal<sup>21</sup>, Dongyu Liu<sup>21</sup>, Ethan Xu<sup>21</sup>, Franck Auge<sup>21</sup>, Katherine Call<sup>21</sup>, Kathy Klinger<sup>21</sup>, Marika Crohns<sup>21</sup>, Matthias Gossel<sup>21</sup>, Tai-he Xia<sup>21</sup>, Minna Hendolin<sup>22</sup>, Outi Tuovila<sup>22</sup>, Outi Tuovila<sup>22</sup>,



Raimo Pakkanen<sup>22</sup>, Antti Karlsson<sup>23</sup>, Kari Pulkki<sup>23</sup>, Lila Kallio<sup>23</sup>, Petri Virolainen<sup>23</sup>, Arto Mannermaa<sup>24</sup>, Sami Heikkinen<sup>24</sup>, Veli-Matti Kosma<sup>24</sup>, Eija Laakkonen<sup>25</sup>, Jari Laukkanen<sup>25</sup>, Teijo Kuopio<sup>25</sup>, Urho Kujala<sup>25</sup>, Eeva Kangasniemi<sup>26</sup>, Johanna Mäkelä<sup>26</sup>, Jarmo Ritari<sup>27</sup>, Jukka Partanen<sup>27</sup>, Kati Hyvärinen<sup>27</sup>, Mikko Arvas<sup>27</sup>, Anne Pitkäranta<sup>28</sup>, Anu Loukola<sup>28</sup>, Eero Punkka<sup>28</sup>, Miika Koskinen<sup>28</sup>, Olli Carpén<sup>28,29</sup>, Aino Salminen<sup>29</sup>, David Rice<sup>29</sup>, Heikki Joensuu<sup>3,9</sup> & Johanna Mattson<sup>29</sup>, Joni A. Turunen<sup>29</sup>, Juha Sinisalo<sup>29</sup>, Katariina Hannula-Jouppi<sup>29</sup>, Lauri Aaltonen<sup>29</sup>, Marja-Riitta Taskinen<sup>29</sup>, Martti Färkkilä<sup>29</sup>, Paula Kauppi<sup>29</sup>, Pekka Nieminen<sup>29</sup>, Pentti Tienari<sup>29</sup>, Pirkko Pussinen<sup>29</sup>, Sampsa Pikkarainen<sup>29</sup>, Sanna Seitsonen<sup>29</sup>, Terhi Ollila<sup>29</sup>, Tiinamaija Tuomi<sup>29</sup>, Tuomo Meretoja<sup>2,3</sup>, Tuula Salo<sup>29</sup>, Ulla Palotie<sup>29</sup>, Juha Rinne<sup>30</sup>, Kaj Metsärinne<sup>30</sup>, Klaus Elenius<sup>30</sup>, Leena Koulu<sup>30</sup>, Markku Voutilainen<sup>30</sup>, Markus Juonala<sup>30</sup>, Sirkku Peltonen<sup>30</sup>, Ulvi Gursoy<sup>30</sup>, Vesa Aaltonen<sup>30</sup>, Johannes Kettunen<sup>31</sup>, Katri Pylkäs<sup>31</sup>, Marita Kalaoja<sup>31</sup>, Miia Turpeinen<sup>31</sup>, Minna Karjalainen<sup>31</sup>, Raisa Serpi<sup>31</sup>, Reetta Hinttala<sup>31</sup>, Riitta Kaarteenaho<sup>31</sup>, Seppo Vainio<sup>31</sup>, Tuomo Mantere<sup>31</sup>, Anne Remes<sup>32</sup>, Juhani Junttila<sup>32</sup>, Kaisa Tasanen<sup>32</sup>, Kirsi Sipilä<sup>32</sup>, Laura Huilaja<sup>32</sup>, Marja Luodonpää<sup>32</sup>, Nina Hautala<sup>32</sup>, Peeter Karihtala<sup>32</sup>, Saila Kauppila<sup>32</sup>, Terttu Harju<sup>32</sup>, Timo Blomster<sup>32</sup>, Vuokko Anttonen<sup>32</sup>, Hilikka Soinen<sup>33</sup>, Ilkka Harvima<sup>33</sup>, Jussi Pihlajamäki<sup>33</sup>, Kai Kaarniranta<sup>33</sup>, Liisa Suominen<sup>33</sup>, Margit Pelkonen<sup>33</sup>, Maria Siponen<sup>33</sup>, Markku Laakso<sup>33</sup>, Mikko Hiltunen<sup>33</sup>, Mikko Kiviniemi<sup>33</sup>, Päivi Auvinen<sup>33</sup>, Päivi Mäntylä<sup>33</sup>, Reetta Kälviäinen<sup>33</sup>, Valtteri Julkunen<sup>33</sup>, Annika Auranen<sup>34</sup>, Airi Jussila<sup>34</sup>, Hannele Uusitalo-Järvinen<sup>34</sup>, Hannu Kankaanranta<sup>34</sup>, Hannu Uusitalo<sup>34</sup>, Jukka Peltola<sup>34</sup>, Mika Kähönen<sup>34</sup>, Tarja Laitinen<sup>26,34</sup>, Teea Salmi<sup>34</sup>, Elina Järvensivu<sup>35</sup>, Essi Kaiharju<sup>35</sup>, Hannele Mattsson<sup>35</sup>, Kati Kristiansson<sup>35</sup>, Lotta Männikkö<sup>35</sup>, Markku Laukkanen<sup>35</sup>, Markus Perola<sup>35</sup>, Päivi Laiho<sup>35</sup>, Regis Wong<sup>35</sup>, Sini Lähteenmäki<sup>35</sup>, Sirpa Soini<sup>35</sup>, Teemu Niiranen<sup>35</sup>, Teemu Paajanen<sup>35</sup>, Terhi Kilpi<sup>35</sup>, Tero Hiekkalinna<sup>35</sup>, Tuuli Sistonen<sup>35</sup>, Veikko Salomaa<sup>35</sup>, Masahiro Kanai<sup>8</sup>, Wei Zhou<sup>8</sup>, Tomi Mäkelä<sup>36</sup>, Manuel Rivas<sup>37</sup>, Kimmo Palin<sup>3</sup>, Harri Siirtola<sup>38</sup>, Javier Garcia-Tabuenca<sup>38</sup>, Csilla Sipeky<sup>39</sup>, Dhanaprakash Jambulingam<sup>39</sup>, Johanna Schleutker<sup>23,39</sup>, Samuel Heron<sup>39</sup> & Venkat Subramaniam Rathinakannan<sup>39</sup>

<sup>10</sup>Abbvie, North Waukegan Road, North Chicago, Chicago, IL 60064, USA. <sup>11</sup>Astra Zeneca, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge CB2 0AA, UK. <sup>12</sup>Biogen, 225 Binney Street Cambridge, Cambridge, MA 02142, USA. <sup>13</sup>Bristol-Myers-Squibb, 430 East 29th Street, New York, NY 10016, USA. <sup>14</sup>Celgene, 86 Morris Avenue, Summit, NJ 07901, USA. <sup>15</sup>Genentech, 1 DNA Way, South San Francisco, CA 94080, USA. <sup>16</sup>GlaxoSmithKline, 980 Great West Road, Brentford, Middlesex TW8 9GS, UK. <sup>17</sup>Janssen Biotech, 800/850 Ridgeview Drive, Horsham, PA 19044, USA. <sup>18</sup>Maze Therapeutics, 171 Oyster Point Blvd Suite 300, South San Francisco, CA 94080, USA. <sup>19</sup>Merck, 2000 Galloping Hill Road, Kenilworth, NJ 07033, USA. <sup>20</sup>Pfizer, 235 East 42nd Street, New York, NY 10017, USA. <sup>21</sup>Sanofi, The Finnish Institute for Health and Welfare, Paris, France. <sup>22</sup>Business Finland, Porkkalankatu 1, 00180 Helsinki, Finland. <sup>23</sup>Auria Biobank, University of Turku, Hospital District of Southwest Finland, Yliopistonmäki, FI-20014 Turku, Finland. <sup>24</sup>Biobank of Eastern Finland, University of Eastern Finland, Northern Savo Hospital District, P.O. Box 100, FI-70029 KYS Kuopio, Finland. <sup>25</sup>Central Finland Biobank, University of Jyväskylä, Central Finland Health Care District, Keskussairaalanatie 19, FI-40620 Jyväskylä, Finland. <sup>26</sup>Finnish Clinical Biobank Tampere, University of Tampere, Pirkanmaa Hospital District, P.O. Box 2000, FI-33521 Tampere, Finland. <sup>27</sup>Finnish Red Cross Blood Service, Finnish Hematology Registry and Clinical Biobank, Kivihaantie 7, FI-00310 Helsinki, Finland. <sup>28</sup>Helsinki Biobank, Helsinki University and Hospital District of Helsinki and Uusimaa, P.O. Box 400, FI-00029 HUS Helsinki, Finland. <sup>29</sup>Hospital District of Helsinki and Uusimaa, P.O. Box 400, FI-00029 HUS Helsinki, Finland. <sup>30</sup>Hospital District of Southwest Finland, P.O. Box 52, FI-20521 Turku, Finland. <sup>31</sup>Northern Finland Biobank Borealis, University of Oulu, Northern Ostrobothnia Hospital District, P.O. Box 50, FI-90029 Oulu, Finland. <sup>32</sup>Northern Ostrobothnia Hospital District, P.O. Box 10, FI-90029 OYS Oulu, Finland. <sup>33</sup>Northern Savo Hospital District, P.O. Box 100, FI-70029 KYS Kuopio, Finland. <sup>34</sup>Pirkanmaa Hospital District, P.O. Box 2000, FI-33521 Tampere, Finland. <sup>35</sup>THL Biobank/The Finnish Institute for Health and Welfare, P.O. Box 30, 00271 Helsinki, Finland. <sup>36</sup>HiLIFE, University of Helsinki, P.O.Box 3, FI-00014 Helsinki, Finland. <sup>37</sup>University of Stanford, 1265 Welch Road MC5464, MSOB West Wing, Third Floor, Stanford, CA 94305-5464, USA. <sup>38</sup>University of Tampere, P.O. Box 2000, FI-33521 Tampere, Finland. <sup>39</sup>University of Turku, Yliopistonmäki, FI-20014 Turku, Finland.