Germline copy number variants (CNVs) are pervasive in the human genome but potential disease associations with rare CNVs have not been comprehensively assessed in large datasets. We analysed rare CNVs in genes and non-coding regions for 86,788 breast cancer cases and 76,122 controls of European ancestry with genome-wide array data. Gene burden tests detected the strongest association for deletions in BRCA1 ($P = 3.7E^{-18}$). Nine other genes were associated with a $p$-value < 0.01 including known susceptibility genes CHEK2 ($P = 0.0008$), ATM ($P = 0.002$) and BRCA2 ($P = 0.008$). Outside the known genes we detected associations with $p$-values < 0.001 for either overall or subtype-specific breast cancer at nine deletion regions and four duplication regions. Three of the deletion regions were in established common susceptibility loci. To the best of our knowledge, this is the first genome-wide analysis of rare CNVs in a large breast cancer case-control dataset. We detected associations with exonic deletions in established breast cancer susceptibility genes. We also detected suggestive associations with non-coding CNVs in known and novel loci with large effects sizes. Larger sample sizes will be required to reach robust levels of statistical significance.
Copy number variants (CNVs) are pervasive in the human genome but are more challenging to detect with current technologies than single nucleotide variants (SNVs). Recent comprehensive sequencing projects\textsuperscript{1,2} have characterised CNVs in large sample sets. The gnomAD project identified a median of 3,505 deletions and 723 duplications covering more than 50 base pairs per genome. Most deletions and duplications tend to be rare with longer variants tending to be rarer, suggesting negative selection against these variants. At the population level the 1000 Genomes project has mapped a large proportion of inherited CNVs\textsuperscript{3} and observed that 65% had a frequency below 2%.

While somatic copy number alterations play a major role in the pathogenesis of breast tumours\textsuperscript{4,5}, some germline CNVs are known to be associated with inherited risk of breast cancer. Rare loss of function variants in susceptibility genes such as \textit{BRCA1} and \textit{CHEK2} are associated with a large increase in risk\textsuperscript{6}. While the majority of these variants are single nucleotide mutations and short indels, they also include longer deletions and duplications. It has been reported that up to a third of loss of function variants in susceptibility genes such as \textit{CHEK2} and \textit{APOBEC3B} gene-coding region increases breast cancer risk\textsuperscript{12} and analysis of the tumours of the germline deletion carriers showed an increase in APOBEC-mediated somatic mutations\textsuperscript{13}. A deletion in a regulatory region was identified as a likely causal variant at the 2q35 locus\textsuperscript{14,15}.

Large-scale genome-wide association studies (GWAS) have established breast cancer associations with common variants at more than 150 loci, mostly in non-coding regions\textsuperscript{8-11}. At two of the loci, deletions imputed from the 1000 Genomes reference panel have been identified as likely causal variants. A deletion of the \textit{APOBEC3B} gene-coding region increases breast cancer risk\textsuperscript{12} and analysis of the tumours of the germline deletion carriers showed an increase in APOBEC-mediated somatic mutations\textsuperscript{13}.

Results

Summary of CNVs detected. After quality control we detected a mean of 2.9 deletions (standard deviation 1.6) and 2.5 duplications (SD 2.0) per sample. Supplementary Data 5 shows the mean length, probe coverage, and segment z-scores of called CNVs. Duplications tended to be longer than deletions: for example, deletions called on OncoArray covered a mean of 45 Kilobases (Kb) (SD 106 Kb) over 9.8 probes (SD 17.2), while duplications covered a mean of 109 Kb (SD 202 Kb) over 18.9 probes (SD 36.5). CNV calls observed in multiple samples were concentrated in a small proportion of probes (Supplementary Data 6), with <11% of probes having frequency >0.01% and <2% of probes having frequency >0.5%.

We identified called CNVs which overlapped for at least 90% of their length with rare deletions and duplications (frequency <1% identified by the 1000 Genomes Project (Supplementary Data 5 and Supplementary Fig. 1)). Forty-nine percent of OncoArray deletions and 47% of iCOGs deletions matched a 1000 Genomes Project variant while 29% of OncoArray duplications and 20% of iCOGs duplications matched. In total, we identified CNVs closely matching 3273 of the deletion variants published by the 1000 Genomes Project (~9% of total) and 1255 of their duplication variants (~24% of the total).

CNVs associated with overall risk. Association results were derived for 1301 regions containing deletions and 992 regions containing duplications. QQ plots are shown in Fig. 2a for deletions and 2b for duplications. There was no evidence for inflation in the test statistics for duplications (\textit{lambda} = 0.98; \textit{lambda}$_{1000}$ = 1.00) and minimal evidence for deletions (\textit{lambda} = 1.11; \textit{lambda}$_{1000}$ = 1.00).

Seven deletion and two duplication regions were associated with breast cancer risk at \textit{p} < 0.001 (Table 1); of these, deletions within the \textit{BRCA1} region achieved \textit{p} < 10$^{-6}$. The results for all regions are shown in Supplementary Data 7 and 8 and include statistics on the number of probes covered by the calls. The results for individual probes covered by the regions analysed are in Supplementary Data 9–12.

The \textit{BRCA1} locus contains multiple deletions across the gene. The \textit{CHEK2} region (OR 1.94, \textit{p} = 0.0003) covers the whole gene but nearly all the calls correspond to a deletion of exons nine and ten, which was previously observed in 1% of breast cancer cases and 0.4% of controls in Poland\textsuperscript{17}. We observed the deletion in 0.9% of Polish cases and 0.5% of controls.

The most significant association (OR = 0.69 \textit{p} = 0.00001) for duplications covers a large region on 17p13.3 with multiple long variants overlapping shorter duplications. The OncoArray results by probe show the strongest associations at a series of probes (17: 815429–850542) in the first intron of \textit{NXN}, with the lowest \textit{P}-value at 17: 819141 (OR = 0.45, \textit{p} = 0.002). The most significant probe position on iCOGs was also in this region (17:836631, OR = 0.58, \textit{p} = 0.09) (Fig. 3).

Associations with risk of breast cancer subtypes. We repeated the analyses restricting cases to those with ER-positive, ER-negative, and triple negative disease. Deletions and duplications with \textit{p}-values below 0.001 are shown in Tables 2 and 3 and the results for all regions are in Supplementary Data 13 and 14. An association was observed for \textit{BRCA1} for all subtypes, with the exception of duplications for ER-positive disease. The odds ratio for \textit{BRCA1} deletions was higher for ER-negative disease (OR = 27.03; 95% CI, 15.66–46.67) than ER-positive (OR = 2.81; CI, 1.56 to 5.08; \textit{P} = 8.46E–28 for the difference), while for \textit{CHEK}2 the odds ratio was higher for ER-positive disease (OR = 2.32; CI,1.56–3.44) than ER-negative (OR = 1.36; CI,0.66 to 2.82; \textit{P} = 0.11 for the difference), consistent with the known subtype-specific associations for deleterious variants in these genes\textsuperscript{18}.

In total, we observed five deletion and two duplication regions with \textit{p}-values < 0.001 that did not reach \textit{p} < 0.001 in the overall risk analysis. The strongest novel association for ER-positive was for an intronic deletion in \textit{ITGB1} (OR = 3.3, \textit{P} = 0.00007, \textit{P} for difference by ER-status = 0.18). For ER-negative disease the strongest novel association was with an intergenic deletion between \textit{ABC4} (\textit{MRP4}) and \textit{CLDN10} (OR = 2.16, \textit{P} = 0.0002, \textit{P} for difference by ER-status = 0.02). Neither of these associations was significant for the other subtype. For triple negative disease, the strongest evidence of association was found for an intergenic duplication between \textit{TMC3} and \textit{MEX3B} (OR = 2.39, \textit{P} = 0.00009) and for two separate deletions upstream of the \textit{DDX18} gene: 2:118258797–118389164 (OR = 6.56, \textit{P} = 0.00001) and 2:117973154–118107795 (OR = 4.54, \textit{P} = 0.0008). The association at these two deletions was driven by the same samples, with 75% of the carriers of the first deletion observed to have the second deletion and normal copy number at the 62 kb gap between the deletions.

Associations at established common susceptibility loci. Three of the most significant associations were observed within regions harbouring known breast cancer susceptibility loci for breast
Fig. 1 Flow diagram of the calling and analysis pipeline. CNVs were called using the CamCNV pipeline from the intensities from the iCOGs and OncoArray genotyping arrays. The CNVs were assigned to regions, association results generated for each array and then meta-analysed. CNVs covering the coding regions of genes were analysed in gene burden tests.

Fig. 2 QQ plots of association results. Quantile-quantile plots of P-values from association tests of deletion regions (a), duplication regions (b), gene burden analysis for deletions (c), and gene burden for duplications (d).
cancer. The most significant (OR = 1.42; CI 1.21 – 1.67; \(P = 0.0015\)) was upstream of FGF2 and consistent with a 28 kb deletion in the 1000 Genomes Project data (chr10:123433204–123461492). Three independent risk signals have been previously identified at this region.\(^2,20\) The effect size for the CNV was larger than those previously reported for these common SNPs (largest OR = 1.27; CI 1.22 – 1.25). The CNV is in linkage disequilibrium with two of the SNPs identified as likely causally associated variants: rs35054928 (\(D’ = 0.82\) and rs2981578 (\(D’ = 0.88\)). Conditioning on those SNPs reduced somewhat the strength of the association for the deletion (OR = 1.30; CI 1.10 – 1.53; \(P = 0.002\), Supplementary Data 15).

The third strongest signal (OR = 4.9, \(P = 0.00001\)) in the deletion analysis for overall breast cancer was at 8: 132199447–132252439, 144Kb downstream of ADCY8. The strongest GWAS signal in this region lies in an intron of ADCY8 (lead SNP rs73348588, OR = 1.13, \(P = 8.2e^{-7}\)). A 3 kb deletion in intron 4 of KLF12 was associated with ER-negative disease (OR = 2.4, \(P = 0.0007\) for difference by ER-status=0.01). This is 389 kb distant from common SNPs, located between KLF12 and KLF5, associated with ER-negative disease (rs9573140, OR = 0.94, \(P = 3.62e^{-9}\)). The KLF12 and ADCY8 deletions are not in strong linkage disequilibrium with the corresponding GWAS signals and conditioning on these SNPs did not alter the strengths of the association for the CNVs (Supplementary Data 15).

| Type       | Locus          | Chr. (Build37) | Start | End       | Total carriers | Odds ratio | Lower CI | Upper CI | Direction | P-value |
|------------|----------------|---------------|-------|-----------|----------------|------------|----------|----------|-----------|----------|
| Deletion   | BRCA1          | 17            | 41188342 | 41363651  | 195            | 6.27       | 4.02     | 9.79     | ++        | 6.32e–16 |
| Deletion   | Intergenic_FGF2_AT1 | 10         | 123453817 | 123461066 | 630            | 1.42       | 1.21     | 1.67     | ++        | 1.42e–05 |
| Deletion   | Intergenic_ADCY8_EFR3A | 8          | 132199447 | 132250643 | 42             | 4.88       | 2.24     | 10.61    | ++        | 6.36e–05 |
| Deletion   | KHL1           | 13            | 70652321  | 7102996   | 1761           | 0.85       | 0.77     | 0.92     | --        | 2.31e–04 |
| Duplication| CHEK2          | 22            | 29083731  | 29123846  | 141            | 1.94       | 1.35     | 2.79     | ++        | 3.26e–04 |
| Deletion   | Supt3h         | 6             | 44908728  | 45244478  | 32             | 0.23       | 0.14     | 0.52     | --        | 4.25e–05 |
| Duplication| Intergenic_GALNT_C18orf121 | 18        | 33350917  | 33359197  | 233            | 1.92       | 1.32     | 2.78     | +        | 6.24e–04 |

### Gene burden tests

We performed gene burden tests based on CNVs that overlapped exons. Analyses were restricted to genes in which at least 24 CNVs were identified, leaving 645 genes with deletions (Supplementary Data 16) and 1596 genes with duplications (Supplementary Data 17). QQ plots are shown in Fig. 2c for deletions and 2d for duplications. The lambda for inflation was 1.18 (\(\lambda_{0.001} = 1.00\)) for deletions and 1.07 (\(\lambda_{0.001} = 1.00\)) for duplications.

For deletions, we found 10 genes with \(P < 0.01\) (Table 4), the most significant being BRCA1 (OR = 7.66, \(P = 3.72e^{-18}\)). Four of these 10 genes (ATM, BRCA1, BRCA2, and CHEK2) are known breast cancer susceptibility genes.\(^18\) Deletions were also observed in two other known susceptibility genes: PALB2 (23 cases, 9 controls, OR = 2.02, \(P = 0.09\)) and RAD51C (21 cases, 9 controls, OR = 2.04, \(P = 0.08\)). The most significant novel association was for SUPT3H (OR = 0.27, \(P = 0.0004\)).

For duplications we observed 15 genes with \(P < 0.01\) (Table 5). The most significant association was for VPS53 (OR = 0.5, \(P = 0.0009\)). This gene and ABR (OR = 0.61, \(P = 0.008\)) both lie within the region on 17p which had the strongest association in the regional analysis. These associations were also observed in two other known susceptibility genes: PALB2 (23 cases, 9 controls, OR = 2.02, \(P = 0.09\)) and RAD51C (21 cases, 9 controls, OR = 2.04, \(P = 0.08\)). The most significant novel association was for SUPT3H (OR = 0.27, \(P = 0.0004\)).

### Discussion

We used the largest available breast cancer case-control dataset, comprising more than 86,000 cases and 76,000 controls with array genotyping, to test for associations with rare CNVs. Using the intensities from genotyping arrays to detect CNVs is not ideal due to a high level of noise and uncertainty in the calling, particularly for duplications. However, in tests of known CNVs and replication of calls between duplicate samples, the CamCNV method shows reasonable levels of sensitivity and specificity.\(^16\) The main focus of this analysis was low frequency CNVs (<1% frequency) since higher frequency CNVs can generally be studied through imputation to a reference panel. In the 0.05–1% frequency range, we could detect ~20% of the CNVs identified by the 1000 Genomes project. For some loci, we only had evidence from one array because the probes do not exist to detect the variants on the other array. Thus, while this array-based approach provides power to evaluate the CNVs that can be assayed, much denser arrays or direct sequencing would be required to provide a complete evaluation of the contribution of CNVs.

In support of the reliability of the method, we detected evidence for both deletions and duplications in BRCA1, which was stronger for ER-negative disease, and for deletions in CHEK2, which were stronger for ER-positive disease. The latter appeared
to be driven by a single founder deletion in East European populations. Weaker evidence of association was also observed for deletions in other susceptibility genes (BRCA2, ATM, PALB2, and RAD51C); the ORs were consistent with those seen for deleterious SNVs and indels. In total, around 0.5% of cases in our analysis had a deletion in one of the known susceptibility genes with the proportion rising to ~1% for cases diagnosed under 50 years of age. The majority of coding deletions are expected to affect only part of the gene, with one study observing that a quarter covered only a single exon. To detect all of these using array data would require at least three probes per exon. The OncoArray has this level of coverage for a few genes, including BRCA1 and BRCA2, but the coverage is lower for most genes and many coding CNVs will have been missed.

A key issue is the appropriate level of statistical significance to apply to these analyses. For the gene burden tests, \( P < 2.5 \times 10^{-8} \), as used in exome-sequencing, seems an appropriate level. It is less clear what is appropriate for non-coding variants. A level of \( P < 5 \times 10^{-8} \) has become standard in GWAS and has been shown to lead to acceptable replication, but this seems over-conservative for CNVs, which are more likely to be deleterious. Consistent with this, for at least two of the ~200 common susceptibility loci, the likely causal variant is a CNV, a higher fraction than expected given the relative frequencies of CNVs and SNPs. Based on frequency analysis of whole-genome sequence data Abel et al. estimated that rare CNVs are >800 times more likely to be deleterious than rare SNVs and >300 times more likely than rare indels. On the other hand, the significance level for non-coding CNVs should logically be more stringent than for the gene burden tests. Taken together, a significance level of \( 10^{-6} \) seems appropriate, while associations at \( P < 0.001 \) may be worth following up in future studies. In our analyses, only the association at BRCA1 (both in the overall and gene burden tests) passes the higher threshold. We also calculated Bayesian False Discovery Probabilities (BDFPs) (Supplementary Data 20 and 21) for our associations using prior probabilities of 0.001 for regions and 0.002 for genes. Outside the known genes none of the BDFPs gave a probability below 10%, with the lowest BDFP of 0.11 for the deletion in the FGFR2 locus. For a CNV observed with a frequency of 0.1% (\( n = 91 \) samples in the OncoArray dataset) we had 40% power to detect an association with an odds ratio of 2 but only 1.5% power to detect an association with an odds ratio of 1.5. An OR of 2, comparable to that seen for deleterious variants in ATM and CHEK2, may be plausible for rare coding CNVs or non-coding CNVs that have a significant effect on gene expression. Larger sample sizes will clearly be required to evaluate rare CNVs with more modest associations.

In addition to the BRCA1 and CHEK2 loci, we found associations in three known susceptibility regions identified through GWAS, harbouring FGFR2, ADCY8, and KLF12. In each case, the variants are rarer than the established associated variants, but confer higher risks. The ADCY8 and KLF12 deletions are not in linkage disequilibrium with the associated SNPs. The FGFR2 deletion is in linkage disequilibrium with two of the likely causal common SNPs although there was still evidence of association with the deletion, albeit weaker, after conditional analysis. In-silico and functional analysis clearly demonstrate that FGFR2 is the target of the previously established variants; it will be interesting to establish if the same is true for the CNV.

Excluding loci in known susceptibility regions, the strongest evidence of association was for a 12 kb deletion (13:102121830–102133956) in the second intron of ITGB1 (OR = 3.3, \( P = 0.00007 \) in the ER-positive analysis). This deletes a promoter flanking region (Ensembl ID: ENSR00001563823) and CTCF binding site (Ensembl ID: ENSR00001062398) active in mammary epithelial cells. There is experimental evidence that ITGB1 expression, mediated by the RUNX2 transcription factor, enables breast cancer cells to form bone metastases.

In the gene burden analysis, the strongest novel association was for deletions within SUPT3H, which were associated with a reduced risk. SUPT3H encodes human SPT3, a component of the STAGA complex that acts as a co-activator of the MYC oncprotein. SUPTH is located within the first intron of the RUNX2 transcription factor and the syntenic relationship between the two genes is highly evolutionarily conserved. RUNX2 has a role in mammary gland development and high-grade ductal carcinoma. In our analysis, only the association at SUPT3H (both in the overall and gene burden tests) passes the higher threshold.
Table 2 Subtype associations for deletions.

| Subtype     | Locus                        | Chr. | Start (Build37) | End     | Total carriers | Odds ratio | Lower CI | Upper CI | Direction (OncoArray, iCOGs) | P-value |
|-------------|------------------------------|------|-----------------|---------|----------------|------------|----------|----------|--------------------------------|---------|
| ER Positive | Intergenic_FGFR2_ATE1        | 10   | 123435817       | 123461066 | 478            | 1.52       | 1.27     | 1.81     | ++                             | 5.04E-06|
| ER Positive | CHEK2                        | 22   | 29091788        | 29102967 | 79             | 2.36       | 1.56     | 3.44     | ++                             | 3.03E-05|
| ER Positive | ITGBL1 intronic              | 13   | 102122905       | 102133221 | 58             | 3.29       | 1.83     | 5.92     | +?                             | 7.29E-05|
| ER Positive | Intergenic_ADCY8_EFR3A       | 8    | 132199447       | 132250643 | 31             | 4.95       | 2.20     | 11.30    | ++                             | 1.07E-04|
| ER Positive | BRCA1                        | 17   | 41188342        | 41363651 | 57             | 2.81       | 1.55     | 5.08     | ++                             | 6.41E-04|
| ER Negative | BRCA1                        | 17   | 41188342        | 41363651 | 112            | 2.70       | 1.56     | 4.67     | ++                             | 2.62E-32|
| ER Negative | Intergenic:ABCC4_CLDN10      | 13   | 95991263        | 96004144 | 134            | 2.16       | 1.43     | 3.26     | +?                             | 7.29E-05|
| ER Negative | KLF12 intronic               | 13   | 74356663        | 74357984 | 93             | 2.39       | 1.49     | 3.82     | +?                             | 2.89E-04|
| ER Positive | 17p13.3_VPS53;NXN            | 17   | 13905           | 1559829  | 454            | 0.67       | 0.55     | 0.81     | ---                            | 4.44E-05|
| ER Positive | 21q22.11_HUNK_LINC00159      | 21   | 33410933        | 33863246 | 77             | 2.55       | 1.6      | 4.06     | ++                             | 7.88E-05|
| ER Negative | 15q13                        | 15   | 26440287        | 32797352 | 1250           | 0.83       | 0.75     | 0.93     | ---                            | 6.95E-04|
| ER Negative | BRCA1                        | 17   | 41188342        | 41363651 | 72             | 40.55      | 21.70    | 75.76    | ++                             | 3.64E-31|
| Triple Neg. | BRCA1                        | 17   | 41188342        | 41363651 | 40             | 6.56       | 2.83     | 15.18    | ++                             | 1.13E-05|
| Triple Neg. | Intergenic_DPP10_DDX18       | 2    | 118258797       | 11839864 | 44             | 3.34       | 1.64     | 6.80     | ++                             | 8.84E-04|
| Triple Neg. | Intergenic_DPP10_DDX18       | 2    | 118258797       | 11839864 | 72             | 40.55      | 21.70    | 75.76    | ++                             | 3.64E-31|
| Triple Neg. | Intergenic_DPP10_DDX18       | 2    | 117973154       | 118107795| 48             | 4.54       | 1.88     | 10.97    | ++                             | 7.92E-04|

Table 3 Subtype associations for duplications.

| Subtype     | Locus                        | Chr. | Start (Build37) | End     | Total carriers | Odds ratio | Lower CI | Upper CI | Direction (OncoArray, iCOGs) | P-value |
|-------------|------------------------------|------|-----------------|---------|----------------|------------|----------|----------|--------------------------------|---------|
| ER Positive | 17p13.3_VPS53;NXN            | 17   | 13905           | 1559829 | 454            | 0.67       | 0.55     | 0.81     | ---                            | 4.44E-05|
| ER Positive | 21q22.11_HUNK_LINC00159      | 21   | 33410933        | 33863246| 77             | 2.55       | 1.6      | 4.06     | ++                             | 7.88E-05|
| ER Positive | 15q13                        | 15   | 26440287        | 32797352| 1250           | 0.83       | 0.75     | 0.93     | ---                            | 6.95E-04|
| ER Negative | BRCA1                        | 17   | 41188342        | 41363651| 42             | 5.93       | 2.31     | 15.19    | +?                             | 2.09E-04|
| Triple Neg. | BRCA1                        | 17   | 41188342        | 41363651| 35             | 10.80      | 3.33     | 35.02    | +?                             | 7.29E-05|
| Triple Neg. | Intergenic_TMC3_MEX3B        | 15   | 81960409        | 82104822| 231            | 2.39       | 1.55     | 3.71     | ++                             | 9.25E-05|
RUNX2 expression is found in ER-negative tumours. The PCDHB2 association appears to be due to a single variant (5:140739812–140740918) that deletes the first exon but as this gene is part of the protocadherin gamma gene cluster it is also possible that the deletion may be having an effect on one of the five genes that overlap PCDHB2. It also deletes a promoter active in mammary epithelial cells (ENSR00001342785). The next strongest signals were for MEAK7 (OR = 2.19, P = 0.001), a gene implicated in a mTOR signalling pathway, and MAD1L1 (OR = 2.00, P = 0.005), a component of the mitotic spindle-assembly that has been suggested as a possible tumour suppressor.

After BRCA1, the most significant association for ER Negative disease in the gene burden analysis was for CYP2C18 which overlaps CYP2C19 (ER-negative OR = 2.6, P = 0.002; triple-negative OR = 4.4, P = 0.0002). A previous analysis of CNVs and breast cancer in the Finnish population identified a founder mutation reaching an overall frequency of ~ 3% and reported a possible association at this locus for triple negative (OR 2.8, p = 0.02) and ER-negative breast cancer (OR = 2.2, p = 0.048).

The results from duplications are harder to interpret as there are often longer duplications overlapping whole genes with shorter variants covering some part of their length. For the gene burden analysis there was little evidence of strong associations. In the regional analysis, the two strongest associations cover multiple genes. The strongest evidence of association (OR = 0.69, P = 1.1E−05) was for a 1.5 Mb region at the start of chromosome 17 (17:13905–1559829). The probe-specific and gene burden results highlighted some stronger signals within this region, for example within NXN and VPS53, but the direction of effect was consistent across the region with 80% of the OncoArray probes

### Table 4 Gene burden results for deletions.

| Gene      | Cases | Controls | Odds ratio | Lower CI | Upper CI | P-value |
|-----------|-------|----------|------------|----------|----------|---------|
| BRCA1     | 171   | 22       | 7.66       | 4.84     | 12.13    | 3.72E−18|
| CHEK2     | 103   | 48       | 1.83       | 1.29     | 2.61     | 7.66E−04|
| SUPT3H    | 10    | 25       | 0.27       | 0.13     | 0.59     | 9.24E−04|
| PCDHB2    | 25    | 3        | 7.03       | 2.10     | 23.52    | 1.55E−03|
| MEAK7     | 59    | 24       | 2.19       | 1.34     | 3.58     | 1.66E−03|
| ATM       | 35    | 8        | 3.43       | 1.56     | 7.51     | 2.11E−03|
| MAD1L1    | 54    | 25       | 2.00       | 1.23     | 3.26     | 5.53E−03|
| NPHP1     | 477   | 351      | 1.22       | 1.06     | 1.41     | 6.13E−03|
| ZNF320    | 29    | 7        | 3.28       | 1.39     | 7.73     | 6.63E−03|
| BRCA2     | 65    | 33       | 1.81       | 1.17     | 2.81     | 7.82E−03|

### Table 5 Gene burden results for duplications.

| Gene | Cases | Controls | Odds ratio | Lower CI | Upper CI | P-value |
|------|-------|----------|------------|----------|----------|---------|
| VPS53 | 40    | 65       | 0.50       | 0.33     | 0.75     | 9.46E−04|
| ATP12A | 48    | 66       | 0.57       | 0.39     | 0.84     | 3.97E−03|
| USP18 | 12    | 23       | 0.34       | 0.17     | 0.71     | 4.16E−03|
| RPS6KA2 | 10    | 22       | 0.32       | 0.14     | 0.70     | 4.20E−03|
| RSU1   | 21    | 8        | 3.40       | 1.47     | 7.84     | 4.23E−03|
| AC008132.1 | 7    | 17       | 0.26       | 0.10     | 0.66     | 4.49E−03|
| PNPLA4 | 479   | 346      | 1.23       | 1.06     | 1.42     | 5.30E−03|
| NLGN4X | 291   | 320      | 0.79       | 0.67     | 0.93     | 5.55E−03|
| ZNF439 | 31    | 9        | 2.98       | 1.37     | 6.45     | 5.72E−03|
| TRIM6  | 5     | 19       | 0.24       | 0.09     | 0.67     | 6.34E−03|
| RP11-363G10.2 | 13    | 27       | 0.39       | 0.20     | 0.78     | 7.39E−03|
| USP31  | 7     | 18       | 0.30       | 0.12     | 0.73     | 8.02E−03|
| TRDN   | 15    | 29       | 0.42       | 0.22     | 0.80     | 8.26E−03|
| ABR    | 51    | 69       | 0.61       | 0.42     | 0.88     | 8.88E−03|
| DNAJC15 | 109   | 67       | 1.52       | 1.11     | 2.09     | 9.29E−03|

### Table 6 Direction of effect results.

| Category | Analysis ustat | Max. ustat of 50 simulations | Min. ustat of 50 simulations | Simulations with larger/smaller ustat | P-value |
|----------|----------------|-----------------------------|------------------------------|--------------------------------------|---------|
| Deletion regions | 9.48 | 6.96 | −7.81 | 0 | 0.04 |
| Deletion regions minus known | 5.9 | 6.96 | −7.81 | 2 | 0.12 |
| Duplication regions | −9.2 | 6.54 | −6.99 | 0 | 0.04 |
| Gene deletions | 9.18 | 5.64 | −9.67 | 0 | 0.04 |
| Gene deletions minus known | 5.01 | 5.64 | −9.67 | 4 | 0.20 |
| Gene duplications | −4.33 | 5.96 | −11.26 | 33 | 1.29 |

aBRCA1, CHEK2 regions excluded.
bBRCA1, CHEK2, ATM, BRCA2 removed.

---

COMMUNICATIONS BIOLOGY | https://doi.org/10.1038/s42003-021-02990-6

ARTICLE
having an odds ratio of 0.75 or lower (Fig. 3). This locus has established associations with prostate and colorectal cancer. Interestingly a possible association with ER-positive breast cancer survival was detected for a rare SNP in the first intron of NXN, rs118021774 (HR = 1.83, P = 3.8E−06)35. The detected duplications are not in LD with this SNP. For the 0.4 Mb duplication region on chromosome 21 (OR = 2.23, P = 0.0001) the probe-specific results from OncoArray highlighted that the signal is specific to a shorter intergenic region (21:33421860−33459975) between HUNK and LINC00159.

We observed some evidence of an aggregate directional effect, both for genes and non-genic regions, such that the deletions in aggregate were associated with increased risk. There was also some suggestion that duplications, in aggregate, were associated with a reduced risk. These results suggest that additional associations are present that could be established with a larger dataset. A new GWAS, Confluence (https://dceg.cancer.gov/research/cancer-types/breast-cancer/confluence-project), aims to double the available sample size for breast cancer. This GWAS includes probes specifically designed to assay the most significant CNVs observed in this study (those significant at P > 0.001), and the sample size should be sufficient to confirm or refute these associations.

Methods

Subjects. Data were derived from blood samples from study participants in 66 studies participating in the Breast Cancer Association Consortium (BCAC) and genotyped as part of the OncoArray33, and iCOGS8 collaborations (Supplementary Data 1). Studies included population-based and hospital-based case-control studies, and case-control studies nested within prospective cohorts; we only included data from studies that provided both cases and controls. Phenotype data were based on version 12 of the BCAC database. Cases were diagnosed with either invasive breast cancer or carcinoma-in-situ. Oestrogen receptor (ER) status was determined from medical records or tissue microarray evaluation, where available. Analyses were restricted to participants of European ancestry, as defined by ancestry informative principal components.8,9 Where samples were genotyped on both arrays, we excluded the iCOGS sample as the OncoArray has better genome-wide coverage. After sample quality control (see below), data on 36,980 cases and 34,706 controls with iCOGS genotyping, and 49,808 cases and 41,416 controls with OncoArray genotyping, were available for analysis (Supplementary Data 2).

Arrays. The Illumina iCOGS genotyping array8 includes 211,155 probes for SNVs and short insertions/deletions. Most variants were selected because of previous association in case-control studies for breast prostate and ovarian cancers, or for dense mapping of regions harbouring an association. The OncoArray includes 533,631 probes, of which approximately half were selected from the Illumina HumanCore backbone, a set of SNPs designed to tag the most common variants. The remainder were selected on the basis of evidence of previous association with breast, prostate, ovarian, lung, or colorectal cancer risk. Approximately 32,000 variants on the OncoArray were selected to provide dense coverage of associated loci and known genes. The remainder were mostly selected from lists of common variants ranked by r2-value, with a small number from lists of candidate variants.

CNV calling. CNVs were called using the CamCNV pipeline as previously described36. In brief, the log R (LRR) intensity measurements and B allele frequency (BAF) for each sample at each probe were exported from Illumina’s Genome Studio software. A principal component adjustment (PCA) was applied to the LRR-adjusted data for each array, to remove noise from high intensity probes and those in regions with known common CNVs, the LRRs for each probe were converted to z-scores using the mean and standard deviation from all BCAC samples. Circular binary segmentation was applied to the z-scores ordered by probe position in each sample using the DNACopy package44. This produces a list of segments for each sample by chromosome where the z-score of consecutive probes changes by more than two standard deviations. Segments with a mean probe z-score between −3.7 and −14 were called as deletions and segments with a mean z-score between +2 and +10 as duplications. We restricted the calls to segments covering a minimum of three and a maximum of 200 probes.

As per the CamCNV pipeline, we then excluded deletions with inconsistent B Allele Frequency and CNVs with a shift in LRR at the sample level that was outside the expected range. The additional CNV exclusions are summarised in Supplementary Data 3. To exclude regions with a high level of noise we also excluded CNVs falling within 1 Mb of telomeres and centromeres and a number of immune loci such as the T-cell receptor genes where somatic mutations in the blood are often observed37.

Sample quality control. Standard sample quality control exclusions were performed, as previously described for the SNP genotype analyses8,35. These include exclusions for excess heterozygosity, ancestry outliers, mismatches with other genotyping, and close relatives. A stricker sample call rate of >99% was used for the CNV analysis, compared to >95% used in the genotype analyses. We also excluded any participant for whom a DNA sample was not collected from blood and any that had been whole genome amplified.

In addition, we used two metrics to exclude noisy samples liable to produce an excess of unreliable CNV calls. First, we calculated a derivative log ratio spread (DLRS) figure for each sample as the standard deviation of the differences between LRR for probes ordered by genomic position, divided by the square root of two. This measures the variance in the LRR from each probe to the next averaged over the whole genome and thus is insensitive to large fluctuations such as might be expected between different chromosomes in the same sample. An ideal sample would have a small DLRS as the only variance would come from a small number of genuine CNVs. We calculated the DLRS using dLRRs function in R package ADM3 (https://CRAN.R-project.org/package=ADMM) before and after the PCA. At both stages, we excluded samples with a DLRS more than 3.5 standard deviations above the mean DLRS for that study.

Second, we counted the number of short segments (between three and 200 probes) per sample. Samples with a mean number of 4 or more short segments were excluded. We observed that the distribution of segment counts was skewed to the right with an excess of samples with a large number of segments. We calculated a cut-off for the maximum number of segments using the following formula where x is the segment count for each sample (based on the rationale that the distribution of the true number of segments should be approximately Poisson):

\[ y = x + 3.5 \]

The sample exclusions resulting from these QC steps are summarised in Supplementary Data 4.

Association tests. All analyses were carried out separately for deletions and duplications, since different types of CNV at the same locus do not necessarily have the same effect on risk. As we were only assessing rare CNVs, we treated all carriers as heterozygotes and did not attempt to identify rare homozygotes.

To account for overlapping CNVs and imprecision in the breakpoints, we assigned individual CNVs to regions. To identify the regions, we moved sequentially along each chromosome, identifying the start as an Oncoarray probe position where deletions were observed in at least five samples, and then the end position as the probe position before the first probe where deletions were observed in fewer than five samples. Regions within five probes of each other were then merged together. The process was repeated for duplications. Regions were also merged such that the major susceptibility genes (BRCA1, BRCA2, CHEK2) were within a single region. We then assigned individual CNVs to regions where at least 90% of the CNV’s length fell within the region. For iCOGS, which generally has less dense probe coverage, we first assigned CNVs to the OncoArray regions where they showed >90% overlap. We then assigned any remaining CNVs to regions defined using the iCOGS probes, using the same procedure. Using this approach, 3306 deletion regions were identified from OncoArray data, 812 of which were also observed using iCOGS data, and 541 regions identified using iCOGS alone. For duplications, there were 2203 OncoArray regions, with 854 also observed using iCOGS data, and 483 iCOGS specific regions.

Associations were evaluated for each region using logistic regression, with breast cancer status as the outcome, and the presence of a CNV in the region (0 or 1) as a covariate to derive a log odds ratio per deletion/duplication. Statistical significance was evaluated using a likelihood ratio test (based on the above model and one excluding CNV as a covariate). The logistic regression analyses were conducted using m-house software (https://rcge.medschl.cam.ac.uk/software/mlogit). Study and ten ancestry informative principal components, defined separately for each study. Analyses were also performed separately for the iCOGS and OncoArray and then combined in a standard fixed effect meta-analysis using the METAL software (after first deriving the standard error of the log-odds ratio from the likelihood test statistic)36. To avoid spurious findings due to too few observations, we observed that the distribution of sample sizes was skewed to the right with an excess of samples with fewer than 24 deletions or duplications (~0.015% of samples). Associations significant at P < 0.001 were considered noteworthy.

To detect more precisely the location of association signals, we also generated results for each probe. We created a vector of pseudo-genotypes for each probe where we defined that probe was deleted if it was present in fewer than 24 samples and duplications for more than 24 samples. These were then used to test for association with breast cancer status.

To test for association between CNVs affecting the coding sequence of genes, in aggregate, and breast cancer risk, we identified samples with a deletion or duplication overlapping the exons of each gene. Exon positions were downloaded...
from the UCSC Genome Browser hg19 knownGene table. We used logistic regression to generate a log odds ratio (OR) for carriers of coding variants covering each gene, adjusted for study, as above. We generated results for each array and then for carriers combined across both arrays. For the combined analyses we treated studies with samples on both arrays as separate studies.

To calculate BFDPs we assumed a log-normally distributed prior effect size as described by Wakefield24. The prior log(OR) was determined by assuming a 95% probability that the OR was less than some bound \( K \), where \( K = 3 \) for the regional and gene-based analysis, except for BRCA1 and BRCA2 where \( K = 20 \) was assumed. The prior probability of association was assumed to be 0.001 for the regional analysis, 0.099 for BRCA1, BRCA2, ATM, and CHEK2 and 0.002 for other genes. For the gene-based analysis, only positive associations were considered as the prior evidence for all genes was in favour of PTVs being positively associated with risk.

To determine whether there was a tendency for CNVs to be associated with an excess, or deficit, of risk across genes or regions, we computed signed z-scores as the square root of the chi-squared statistic for each gene, multiplied by ±1 depending on whether the effect estimate was positive or negative. These were ranked and normalised summed z-scores, based on the r most significant associations, were derived. The overall test statistic was the maximum summed z-score over all possible values of r:

\[
U = \max_{r \leq n} \sum_{j=1}^{r} z_j
\]

(1)

Where \( n \) is the total number of genes/regions being tested. The significance of \( U \) was then determined by permutation, randomly permuting case-control labels within study 50 times.

Ethical approval. All participating studies were approved by their appropriate ethics review board and all subjects provided written informed consent.

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

Data availability. Full summary statistics for the regions and probes analysed are available in the Supplementary Data. This includes the source data used to produce Figs. 2 and 3. The majority of the OncoArray dataset analysed in this study is available in the dbGap repository, Study ID: phs01265.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gp/epi/). The iCOGS dataset and complete OncoArray code for the CamCNV calling algorithm and a test dataset with OncoArray genotyping is available at the project home page (https://github.com/igd29/CamCNV).

Received: 8 June 2021; Accepted: 1 December 2021; Published online: 18 January 2022

References

1. Abel, H. J. et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* 583, 83–89 (2020).
2. Collins, R. L. et al. A structural variation reference for medical and population genomics. *Nature* 581, 444–451 (2020).
3. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81 (2015).
4. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 534, 47–54 (2016).
5. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* 578, 122–128 (2020).
6. Easton, D. F. et al. Gene-panel sequencing and the prediction of breast-cancer susceptibility. *N. Engl. J. Med.* 372, 2243–2257 (2015).
7. Ewald, I. P. et al. Genomic rearrangements in BRCA1 and BRCA2: A literature review. *Genet. Mol. Biol.* 32, 437–446 (2009).
8. Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* 45, 553–61 (2013).
9. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* 551, 92 (2017).
10. Milne, R. L. et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat. Genet.* 49, 1767 (2017).

Acknowledgements

Joe Dennis is supported by the CanRisk Cancer Research UK programme grant: PPRPMG-Nov20:100002 and by the Confluence project which is funded with intramural funds from the National Cancer Institute Intramural Research Programme, National Institutes of Health. Logan Walker is supported by a Rutherford Discovery Fellowship (Royal Society of New Zealand). 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 COGS study would not have been possible without the contributions of the following Andrew Berchuck (OCAC), Rosalind A. Eeles, Ali Amin Al Olama, Zsofia Kote-Jarai, Sara Benlloch (PRACTICAL), Antonis Antoniou,
Lesley McGuflg and Ken Offit (CIMBA), Andrew Lee, and Ed Dick, Craig Lucarrini and the staff of the Centre for Genetic Epidemiology Laboratory, the staff of the CNIO genotyping unit, and Daniela Díez, and Yael Neuman, Francioso Basquet Solchon, Sylvie LaBoissière and Frederic Robloux and the staff of the McGill University and Genome Québec Innovation Centre, Sune F. Nielsen, Borge G. Nordestgaard, and the staff of the Copenhagen DNA laboratory, and Julie M. Cunningham, Sharon A. Windebank, Christopher A. Hilker, Jeffrey Meyer and the staff of Mayo Clinic Geno- typing Facility, and thank Magne Annesen. ABCRS thanks the Blood bank Sankün, The Netherlands. ARCTBR Investigators: Christine Clarke, Deborah Marsh, Rodney Scott, Robert Baxter, Desmond Yip, Jane Carpenter, Alison Davis, Nirmala Pathmanathan, Peter Simpson, J. Danny Graham, Mythily Sah- chitananthan. Samples are made available to researchers on a non-exclusive basis. ABCRS thanks Eileen Williams, Elaine Ryder-Mills, Kara Sarges. BCEES thanks Allyson Thompson, Christine Saunders, Terry Steven-Dale, Andrew L. Berghoef, Kimberly Wylie, Rachel Lloyd. The BCNIS study would not have been possible without the contributions of Dr. K. Landsman, Dr. N. Gronich, Dr. A. Flugelmann, Dr. W. Saliba, Dr. F. Lefkowicz, Dr. E. Lianai, Dr. I. Cohen, Dr. S. Kalet, Dr. V. Friedman, Dr. O. Barnert of the NICHCC in Haifa, and all the contributing family medicine, surgery, pathology, and oncoology teams in all medical institutes in Northern Israel. BIOCS thanks Niall McI- nery, Gabrielle Colleran, Andrew Rowan, Angela Jones. The BREOGAN study would not have been possible without the contributions of the following: Angel Carracedo, Víctor Muñoz Garzón, Alejandro Novo Dominguez, Maria Elena Martínez, Sara Miranda, Ponce, Carmen Redondo Maray, Maite Peña Fernández, Manuel Enguix Castelo, Maria Torres, Manuel Calaza (BRIDOGAN), José Antúnez, Máximo Fraña and the staff of the Department of Biostatistics and Biobank of the Norwegian National Cancer Institute – CHUS, Instituto de Investigación Sanitaria de Santiago, IDS, Xerencia de Xestión Integrada de Santiago-SERGAS, Joaquín González-Carreró and the staff of the Depart- ment of Pathology and Biobank of University Hospital Complex of Vigo, Instituto de Investigacion Biomédica Galicia Sur, SERGAS, Vigo, Spain. CRCB thanks study partici- pants, co-investigators, collaborators, and staff of the Canadian Breast Cancer Study, and project coordinators Agnes Lai and Celine Morissette. CCGP thanks Styliani Apostolaki, Anna Margioli, Georgios Nintos, Maria Perraki, Georgia Salonoutrou, Georgiia Sevastaki, Konstantinos Pompadakis. CGPS thanks staff and participants of the Copenhagen General Population Study. For the excellent technical assistance: Dorthe Ulldall Andersen, Maria Birna Arnadottir, Anne Bank, Dorthe Kjeldgård Hansen. The Danish Cancer Biobank is acknowledged for providing infrastructure for the collection of blood samples for the cases. CNIO-BCS thanks Guillermina Pita, Chalo Alonso, Natalia Álvarez, Pilar Zamora, Primitiva Menendez, the Human Genotyping-CEGEN Unit (CNIO). Investigators from the CPS-II cohort thank the participants and Study Man- agement Group for their invaluable contributions to this research. They also acknowl- edge the contribution to this study from central cancer registries supported through the Centres for Disease Control and Prevention National Programme of Cancer Registries, as well as cancer registries supported by the National Cancer Institute Surveillance Epi- demiology and End Results programme. The authors would like to thank the California Teachers Study Steering Committee that is responsible for the formation and main- tenance of the Study within which this research was conducted. A full list of California Teachers Study team members is available at https://www.calteachestudy.org/team. ESTHER thanks Talbot, Sonja Wolf, Dietmar Schütz, and the Copenhagen General Population Study. Danmarks Markedsføringsråd, Jytte Mikkelsen, Niels Holbek, Anna Butterbach. FHRSK thanks NIH for funding and the Manchester NHS Biomedical Research Centre (ES-BRC-1215-20007). GC-HBOC thanks Stefanie Engert, Heide Hel- ver, Uldall Andersen, Maria Birna Arnadottir, Anne Bank, Dorthe Kjeldgård Hansen. The Munich Biocentre (MCB) thankssters thanks Hartwig Ziegler, Sonja Wolf, Volker Hermann, Christa Stegmaier, Katja Neudert, Judith, Neida Ols, Alina Vrienel, Sabine Behrens, Ursula Elblä, Mubhbat Celik, Tíbí Olechs and Stefan Nickles. MBCSG (Milan Breast Cancer Study Group) Paola Radice, Bernardo Peissel, Jacopo Azzollini, Ezica Rosina, Daniela Zaffaroni, Bernardo Bonanni, Irene Ferone, Maria Stavroula Calvello, Arianna Guerrieri Gonzaga, Monica Marabelli, Davide Berti, and Carlotta Cordi. The Medical College of Wisconsin thanks the personnel of the Center for Blood Research. The MCCS was made possible by the contribution of many people, including the original investigators, the teams that recruited the participants and continuing work on follow-up, and the many thousands of Melbourne residents who continue to participate in the study. We thank the coordinators, the research staff, and especially the MMHS parti- cipants for their continued collaboration on research studies in breast cancer. The fol- lowing are NBCS Collaborators: Anne-Lise Børresen-Dale (Prof. Em.), Kristine Kiiski, Carl Blomqvist, Kristiina Aittomäki, Kirsimari Aaltonen, Karl von Smitten, Irja Lotz; and Institute of Occupational Medicine and Maritime Medicine, University Medical Centre Hamburg-Eppendorf, Germany [Volker Harth]. HUBCS thanks Shamil Gantsev. ICICLE thanks Kelly Kohut, Michele Caneppele, Maria Erkkilä. HMBCS thanks Peter Hillemanns, Hans Christiansen and Johann H. Karstens. UKBGS thanks Breast Cancer Now and the Medical Research Council UK for funding of the National Breast Cancer Research Centres (grant numbers 634935 and 633784 for BRIDGES and B-CAST respectively), and the PERSPECTIVE I&I project, 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 Foun- dation. The EU Horizon 2020 Research and Innovation Programme funding source had no role in study design, data collection, data interpretation, or writing of the report. Additional funding for BCAC is provided via the Confluence project which is funded with intramural funds from the National Cancer Institute Intramural Research Programme, National Institutes of Health. Genotyping of the OncoArray was funded by the NIH Grant U19 CA148065, and Cancer Research UK Grant C1287A/16563 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 Quebec and the PSSRIIRI-701 grant, and the Quebec Breast Cancer Foundation. Funding for iCOC came from the European Community’s Seventh Framework Programme under grant agreement no 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (C12168, C1287A/16563), the National Institutes of Health (C1287A/15070, C5047/A10692, C8197/A16565), the National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (U19 CA148537, U19 CA148605, and U19 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 National Cancer Institute for funding of the Post-Cancer GWAS initiative. The Breast Cancer Family and Systematic Breast Cancer Genotyping unit, Jacobs Institute of Science and Technology, Jacobs Institute of Science and Technology, Jacobs Institute of Science and Technology, Jacobs Institute of Science and Technology, Jacobs Institute of Science and Technology, Jacobs Institute of Science and Technology.
Breast Cancer Foundation, the Canadian Institutes of Health Research for the “CHIR Team in Familial Risks of Breast Cancer” programme—grant # CRN-87521 and the Ministry of Economic Development, Innovation and Export Trade—grant # PSR-SIIRT-701. The NBCS has received funding from the R.K. Jepsen Centre for Breast Cancer Research; the Research Council of Norway grant 193387/V50 (to A.-L. Berre森-Dale and V.N. Kristensen) and grant 193387/H10 (to A.-L. Berre森-Dale and V.N. Kristensen), South Eastern Norway Health Authority (grant 39346 to A.-L. Berre森-Dale) and the Norwegian Cancer Society (to A.-L. Berre森-Dale and V.N. Kristensen). The NBHS was supported by NIH grant R01CA100374. Biological sample preparation was conducted by the Survey and Biospecimen Shared Resource, which is supported by P30 CA68485. The Northern California Breast Cancer Family Registry (NC-BCRF) and Ontario Familial Breast Cancer Registry (OFBCR) 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 centres in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organisations imply endorsement by the USA Government or the BCFR. The Carolina Breast Cancer Study (NCBCS) was funded by Komen Foundation, the National Cancer Institute (P50 CA058233, U54 CA156733, U01 CA179715), and the North Carolina University Cancer Research Fund. The NHIS was supported by NIH grants P01 CA87969, UM1 CA186107, and U19 CA148065. The NHIS2 was supported by NIH grants U1M CA176726 and U19 CA148065. The OBCS was supported by research grants from the Finnish Cancer Foundation, the Academy of Finland (grant number 250083, 122715 and Centre of Excellence grant number 251314), the Finnish Cancer Foundation, the Sigfrid Julesius Foundation, the University of Oulu, the University of Oulu Support Foundation and the special Governmental EVO funds for Oulu University Hospital-based research activities. The ORIGO study was supported by the Dutch Cancer Society (RUL 1997-1505) and the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL CP16). The PBCS was funded by Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA. Genotyping for PBCS was supported by the Intramural Research Program of the National Institutes of Health, NCI, Division of Cancer Epidemiology and Genetics. The PLCO is supported by the Intramural Research Programme of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, National Institutes of Health. The RBCS was funded by the Dutch Cancer Society (DDHK 2004-3124, DDHK 2009-4318). The SASBAC study was supported by funding from the Agency for Science, Technology and Research of Singapore (A*STAR), the US National Institute of Health (NIH), and the Susan G. Komen Breast Cancer Foundation. The SCBS was supported by Sheffield Experimental Cancer Medicine Centre and Breast Cancer Now Tissue Bank. SEARCH is funded by Cancer Research UK [C490/ A10124, C490/A16061] and supported by the UK National Institute for Health Research Biomedical Research Centre at the University of Cambridge. The University of Cambridge has received salary support for PDPP from the NHS in the East of England through the Clinical Academic Reserve. The Sister Study (SISTER) is supported by the Intramural Research Programme of the NIH, National Institute of Environmental Health Sciences (Z01-ES044005 and Z01-ES049033). The SMC is funded by the Swedish Cancer Foundation and the Swedish Research Council (VR 2017-00641) grant for the Swedish Infrastructure for Medical Population-based Research (SIMPLER). The SZBCS was supported by Grant PBZ, KBN, 122/05/2004 and the programme of the Minister of Science and Higher Education under the name “Regional Initiative of Excellence” in 2019–2022 project number 002/RD/2018/19 amount of financing 12,000,000 PLN. The TNRCC was supported by: a Specialised Programme of Research Excellence (SFORE) in Breast Cancer (CA116201), a grant from the Breast Cancer Research Foundation, a generous gift from the David F. and Margaret T. Grohne Family Foundation. The UCIBCS component of this research was supported by the NIH [CA58860, CA92040] and the Lon V Smith Foundation [LV394298]. The UKIBCS is funded by Breast Cancer Now and the Institute of Cancer Research (ICR), London. ICR acknowledges NHS funding to the NIHR Biomedical Research Centre. The UKOPS study was funded by The Eve Appeal (The Oak Foundation) and supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. The US3SS study was supported by Massachusetts (K.M.E., R01CA47305), Wisconsin (P.A.N., R01 CA47147) and New Hampshire (L.T.-E., R01CA69664) centres, and Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA. The USRT Study was funded by Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA.

**Author contributions**

Data management: M.K.R., Q.W., and R.Kc. CNV detection: I.D. and L.W. Statistical/ bioinformatic analysis: J.D., J.T., K.M., and L.Delw. Conceived the OncoXarray and obtained financial support: I.S., P.K., and D.F.E. Provided DNA samples and/or cytogenetic data: T.A., I.A., H.A., N.A., Y.A., K.A., L.B., M.W.B., S.Beh, J.B., M.B., N.B., S.Boj, H.B., J.E.C., J.C.-C., G.C., C.L.C., J.M.C., F.C., A.C., S.C., K.C., P.D., T.D., L.Dons, A.E., M.E., D.G.E., P.F., J.F., O.F., H.F., L.F., M.G., M.G.O., M.G., M.G.D., G.G.A., A.G., P.G., E.H., C.H., P.H., A.Hol, R.H., J.H., A.How, A.Jak, A.Jak, E.I., N.J., M.J., A.Jung, R.Ka, E.K., G.K., Y.K., V.M.K., S.K., P.K., V.N.K., K.K.-S., A.K., J.L., D.L., N.L., M.L., A.L., A.M., S.Man, S.M., S.M., D.M., R.L.M., T.M., R.A.M., H.N., J.O., H.O., T.P., C.P., P.D., P.D.P., D.P.-K., K.P., G.R., E.Sal, D.S., E.I.S., M.S., R.K.S., R.S., J.S., A.S., P.S., M.C.S., A.J.S., R.M.T., I.A.T., L.R.T., M.B.T., I.T., M.A.T., T.T., C.V., C.W., R.W., A.W., X.Y., W.Z., A.Z., D.A., A.D.F.E., ABCTB Investigators, kConFab/AOCS Investigators, NBCS Collaborators, CTS Consortium. Directed the project: D.F.E. All authors read and approved the final version of the manuscript.

**Competing interests**

The authors declare no competing interests.

**Additional information**

**Supplementary information**

The online version contains supplementary material available at https://doi.org/10.1038/s42003-021-02990-6.

**Correspondence**

and requests for materials should be addressed to Joe Dennis.

**Peer review information**

Communications Biology thanks Enrique Hernandez-Lemes and Tao Huang for their contribution to the peer review of this work. Primary Handling Editor: Eve Rogers.

**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.

---

Joe Dennis, Jonathan P. Tyrer, Logan C. Walker, Kyriaki Michailidou, Leila Dorling, Manjeet K. Bolla, Qin Wang, Thomas U. Ahearn, Irene L. Andrusli, Hoda Anton-Culver, Natalia N. Antonenkova, Volker Arndt, Kristian J. Aronson, Laura E. Beane Freeman, Matthias W. Beckmann, Sabine Behrens, Javier Benitez, Laura E. Beane Freeman, Marina Bermisheva, Natalia V. Bogdanova, Stig E. Bojesen, Hermann Brenner, Jenny Chang-Claude, Georgia Chenevix-Trench, Christine L. Clarke, NBCS Collaborators, J. Margriet Collée, CTS
CTS Consortium
James Lacey & Elena Martinez

University of California, San Diego, CA, USA.

ABCTB Investigators
Christine Clarke, Jane Carpenter, Deborah Marsh, Rodney Scott, Robert Baxter, Desmond Yip, Alison Davis, Nirmala Pathmanathan, Peter Simpson, Dinny Graham & Mythily Sachchithananthan

The Westmead Institute for Medical Research, The University of Sydney, Sydney, NSW, Australia.
University of Technology Sydney, Translational Oncology Group, School of Life Sciences, Faculty of Science, Ultimo, NSW, Australia.
School of Biomedical Sciences, University of Newcastle, Newcastle; Hunter Medical Research Institute and NSW Health Pathology North, Newcastle, Australia.
Kolling Institute of Medical Research, University of Sydney, St Leonards, NSW, Australia.
Epigenetics & Transcription Laboratory Melanie Swan Memorial Translational Centre, Sci-Tech, University of Canberra, Canberra, Australia.
Department of Medical Oncology, The Canberra Hospital, Garran, ACT, Australia.
The Canberra Hospital, Garran, ACT, Australia.
The Australian National University, Canberra, ACT, Australia.
Westmead Breast Cancer Institute, Western Sydney Local Health District, Westmead, NSW, Australia.
University of Sydney, Western Clinical School, Westmead, NSW, Australia.
UQ Centre for Clinical Research, Faculty of Medicine, The University of Queensland, Herston, QLD, Australia.

kConFab/AOCS Investigators
Ian Campbell, Georgia Chenevix-Trench, Anna de Fazio, Stephen Fox, Judy Kirk, Geoff Lindeman, Roger Milne, Melissa Southey, Amanda Spurdle & Heather Thorne

Peter MacCallum Cancer Centre, Melbourne, Australia.
Westmead Institute for Cancer Research, Sydney, Australia.
Department of Medicine, Westmead Hospital, Sydney, Australia.
Walter and Eliza Hall Institute, Melbourne, Australia.
Queensland Institute of Medical Research, Brisbane, Australia.