## **Supplementary Methods and Figures**

Loveday C, Garrett A, Law P. et al. Analysis of rare disruptive germline mutations in 2,135 enriched BRCA-negative breast cancers excludes additional high-impact susceptibility genes.

# **Supplementary Methods**

# **Supplementary Figure S1 Ancestry analysis via principal components**

# **Supplementary Figure S2 Quality by depth correlations**

# **Supplementary Figure S3 Variant calibration quantile-quantile plots**

# **Supplementary Figure S4 Phenotypic enrichment in cases**

# **Supplementary Figure S5 Frequency of oncogenic cancer genevariants in cases vs controls**

# **Supplementary Figure S6 *PPM1D* variant allele frequencies**

**Supplementary References**

**Supplementary Methods**

## **Case Subjects**

Probands and their family members were recruited to the ICR Breast and Ovarian Cancer Susceptibility (BOCS) study from24 UK genetics centres between 2007 and 2015 on the basis of being affected with breast or ovarian cancer and having a significant family history and/or bilateral disease and/or both breast and ovarian cancer and/or disease diagnosed age < 30 years and/or triple-negative breast cancer. All cases included in this analysis had been diagnosed with pathologically confirmed invasive breast cancer and had either (i) bilateral disease, (ii) early-onset breast cancer (< 30 years), (iii) a significant family history of breast cancer (defined as breast cancer in at least one first degree-relative or two second degree relatives) or (iv) concomitant ovarian cancer. All cases included in this analysis were negative for germline *BRCA1* and *BRCA2* mutations/exon level deletions on routine screening undertaken at the recruiting genetics centre and/or performed on enrolment to the study. Histology and receptor status were not routinely collected and thus were available for only a limited number of the 2,135 eligible cases that underwent exome sequencing. Histological subtype information was available for 1,323 cases, comprising N = 1,104 ductal, N = 180 lobular, and N = 39 tubular. ER status was known for N = 1,030 (of which 185 were ER negative), PR status was known for N = 739 (of which 237 were PR negative) and HER2 status was known for N = 700 (of which 568 were HER2 negative). A total of N = 536 samples had a result for all three receptors (of which 75 were triple negative). Informed written consent was obtained from all participants and the research was approved by the London Multicentre Research Ethics Committee (MREC/01/2/18).

## **Whole exome sequencing**

Whole exome sequencing of lymphocyte DNA was performed using either i) 50 ng genomic DNA and the Nextera DNA sample preparation kit (Illumina, San Diego, CA, USA) or ii) 1.5 µg genomic DNA and the TruSeq exome enrichment kit (Illumina). The captured libraries were amplified by PCR with the supplied paired end PCR primers. Sequencing was performed using anlluminaHiSeq 2000 or HiSeq 2500 (high output mode) using v3 chemistry and generating 2 × 101 bp reads. CASAVA (v1.8) was used to generate per read FASTQ files from per cycle BCL basecall files. Paired end FASTQ files were generated for an initial total of *N* = 2,430 samples. Truncated or malformed FASTQ files were excluded (*N* = 1 sample).

To process FASTQ files we used the nf-core/sarek pipeline (v2.5.2; Garcia et al.1), a workflow built using NextFlow (Tommaso et al.2). This pipeline was used to perform preprocessing according to Genome Analysis Tool Kit (GATK) best practices (McKenna et al.3), variant calling using HaplotypeCaller (Poplin et al.4) in GVCF mode, and quality control using bamQC (Okonechnikov et al.5). Intermediate gVCFs output by the Sarek pipeline were further processed using the following GATK (v4.1.0.0) tools: GenomicsDBImport (to combine), GenotypeGVCF (to produce genotype VCFs), VariantRecalibrator (to build a recalibration model to score variant quality for filtering purposes) and ApplyBQSR (to apply the recalibration). Variant normalization (left alignment and splitting of multi-allelics) was performed on the recalibrated VCFs using bcftools (v.1.9) (Li et al.6). Normalised variants were then annotated using Ensembl Variant Effect Predictor (VEP; McLaren et al.7), restricting annotations to Ensembl canonical transcripts using the ‘–pick’ option. VEP plug ins were used to further annotate protein truncating variants (PTVs) and missense variants with Loss-Of-Function Transcript Effect Estimator (LOFTEE; Karczewski et al.8) predictions and rare exome variant ensemble learner (REVEL; Ioannidis et al.9) scores. Variant calls from Genome Aggregation Database (gnomAD; Karczewski et al.8) were also normalised and annotated as above.

## **Case exclusions and sample/variant level quality control**

Sample and variant level QC metrics were selected to align filtering of case data to that of control data in gnomAD.

### Pre variant calling sample QC

Individual sample level QC was performed to exclude case samples for which raw sequencing metrics obtained via bamQC were outliers, defined as greater than 3 median absolute deviations (i) below the median number of mapped reads; (ii) below the median for proportion of bases at > 30X; (iii) above the median error rate. Summary sequencing metrics are shown in Supplementary Table S3. A total of *N* = 162 samples were excluded at this stage.

### Post variant calling sample QC

Individual sample level QC was performed to exclude case samples for which raw variant calling metrics were outliers, defined as 4 median absolute deviations (MAD) from the median in either directionfor any of the following: (i) number of SNVs; (ii) number of INDELs; (iii) insertion:deletion ratio; (iv) transition:transversion (TiTv) ratio; (v) heterozygous:homozygous ratio. Outliers were identified on a pulldown chemistry aware basis. A total of *N* = 29 samples were excluded based on these metrics (Supplementary Table S3).

### Cryptic relatedness

We checked for cryptic relatedness (up to 3-degree relationships) amongst the case series using the ‘--relatedness2’ option from vcftools (v0.1.16), which calculates kinship coefficients based on the method of Manichaikul et al. 2010, (N = 1 case) (Manichaikul et al.10; Supplementary Table S3).

### Occult BRCA-mutation

Cases were also removed where WES revealed a pathogenic variant in *BRCA1or BRCA2*, defined as being a Pathogenic/Likely Pathogenic (P/LP) variant on ClinVar with one or more stars, classified using ACMG criteria and/or (ii) those predicted as high confidence protein truncating variants (HC-PTVs) on account of a high confidence LOFTEE annotation. A total of *N* = 42 samples were excluded based on these metrics (Supplementary Table S3).

### Ancestry exclusions

We performed ancestry analysis using plinkQC (Meyer11), referencing 1000 Genomes phase 3 data (1000 Genomes Project Consortium et al.12). Principal component analysis (PCA) was performed via PLINK using a set of 20,604 shared autosomal single nucleotide variants (SNPs) of missingness < 5% and a minor allele frequency (MAF) > 1% (Purcell et al.13). The first two principal components were fed into plinkQC, using non-Finnish European samples within the 1000 Genomes dataset as the target reference population and a scaling factor of 1.5. All case samples for whom Euclidean distance from the centre fell outside the radius specified by the maximum Euclidean distance of the 1000 Genomes non-Finnish European data, multiplied by the scaling factor, were considered non-European and excluded from the study, excluding *N* = 60 case samples from the analysis (Supplementary Figure S1, Supplementary Table S3).

### Variant level QC

Variant level QC (*N* = 817,385 starting variants) was performed on the case data, removing variants: (i) with excess heterozygosity, as defined by an inbreeding coefficient < -0.3; (ii) for which no sample had a high-quality genotype (equal to depth >= 10, genotype quality >= 20 and minor allele balance > 0.2 for heterozygous genotypes). *N* = 123,245 variants were removed using these metrics (Supplementary Table S4).

### Sanger validation

Sanger sequencing of known breast cancer susceptibility genes was performed in index cases with primers designed using standard approaches. The QIAGEN Multiplex PCR kit (QIAGEN, Hilden, Germany) was used for DNA amplification, and amplicons were bidirectionally sequenced with BigDye Terminator cycle sequencing kits (Thermo Fisher Scientific, Waltham, MA, USA) and an ABI 3730 sequencer (Life Technologies, Carlsbad, CA, USA). Sequencing traces were analysed with Mutation Surveyor software. On comparison of WES variant calls to calls from Sanger sequencing of samples for a subset of rare variants in known cancer predisposition genes, we found high concordance; of 176 variants detected by Sanger, 167 were detected by WES (concordance = 95%).

## **Control Subjects**

For controls, we used WES data from the Genome Aggregation Database (gnomAD; version 2)8, which comprised 51,377 Non-Finnish European non-cancer individuals (updated 06/03/2019, date of accession 21/05/2020).

## **Quality Control, calibration and coverage normalization for case/control variant sets**

### Region filtering and hard filtering

We removed variants residing in genomic regions of high sequence repetitiveness, as defined by the GA4GH Benchmarking team (<https://data.nist.gov/od/id/mds2-2190>), as these regions are associated with an excess of false positive variant calls (variants removed: *N* = 248,565 in cases and *N* = 5,576,575 in controls). Additionally, we applied the following GATK metric hard filters to both case and control variant call sets to further reduce the burden of likely sequencing artifacts: QualByDepth< 2; FisherStrand> 60; StrandOddsRatio> 3; RMSMappingQuality< 40; MappingQualityRankSumTest< -2.5; (variants removed: *N* = 10,747 in cases and *N* = 371,307 in controls). These filters were applied such that if a shared variant was removed from one dataset it was also removed from the other, irrespective of its value in the second dataset. We applied a MAF filter to remove variants from both datasets with a MAF > 0.5% in either dataset (*N* = 71,122 in cases and *N* = 13,646 in controls). Finally, we removed variants that did not have an entry of ‘PASS’ for variant quality score recalibration (VSQR) (*N* = 7,024 in cases and *N* = 73,148 in controls) (Supplementary Table S4).

### Calibration using QualByDepth

We used the GATK QualByDepth (QD) metric to calibrate the inclusion of rare SNVs and INDELs in burden testing using the approach described by Guo et al. 2018. (Guo et al.14). We first assessed whether QD was sufficiently correlated between the two datasets at the variant MAF cut off of interest (MAF < 0.5%) (Supplementary Figure 2). The correlations were r2 = 0.31 for SNVs (*N* = 294,116) and r2 = 0.39 for INDELs (*N* = 9,490). These values are low but expected, due to the high levels of variability in coverage and quality scores at individual base positions between the two datasets. Next, to calibrate burden testing for SNVs, QD filters were calculated at 25th, 20th, 15th, 10th, 5th percentiles for cases and controls separately (Supplementary Table S5). A synonymous variant burden test applied to every pairwise combination, measuring inflation of the test statistic using a modification of the genomic control metric used in genome-wide association studies (GWAS), as described by Guo et al.14 (the lambda delta 95 approach). We selected the 5th centile in cases and controls, as these values allowed us to maximise the number of variants to be used in the subsequent burden tests without significant inflation of the test statistic (i.,e. demonstrated a normal distribution upon visual inspection and a lambda delta 95 close to 1) (Supplementary Figure S3). The same approach was used to calibrate indels, calibrated using all nonsynonymous SNVs (i.e. all protein altering indels and SNVs) rather than synonymous SNVs (Supplementary Figure S3).

### Coverage Normalisation

We calculated per sample per base depth of coverage in cases using GATK (v3.8-0) DepthOfCoverage requiring a minimum mapping quality of 20, including reference bases. We retrieved the equivalent coverage data generated for gnomAD exomes v2.1.1 (<https://gnomad.broadinstitute.org/downloads>).

To normalize coverage between datasets prior to burden testing, we utilized the underlying principlesof the intersect and the binomial methods14, testing a range of parameters and assessing for evidence of inflation of the modified genomic control test statistic.Ultimately, we used a three-step approach to coverage normalization that did not significantly lead to inflation of the test statistic. Firstly, using the intersect method, we includedsites with a read depth of > 10X in ≥ 70% of cases and >10X in ≥ 70% of controls. For sites not qualifying for inclusion via the intersect test, we performed a binomial test as follows: binomtest = k = x, n = x+y, p = s/(s+t), where x = number of cases with coverage 10X, y = number of controls covered at 10X, s = total number of cases, and t = total number of controls. We retained those positions whereby the proportion of cases vs controls with coverage ≥ 10X was*P*> 0.000001. We only considered overlapping sites with coverage at proportion covered at 10X of ≥ 10% for the biniomial test (positions below 10% were automatically disqualified). Finally, if ≥80% of basepositions in a canonical protein codingEnsembl transcript were qualified by the previous steps, then all positions within the transcript were ultimately included.The combined intersect and binomial method, with whole transcriptinclusion at ≥ 80% of passed bases, resulted in *N* = 31,341,042 sites, median and IQR for all transcripts was 100% (70-100%) and known breast cancer gene transcripts was 100% (100-100%), respectively.

## **Statistical analyses**

### Gene set definitions

We considered breast cancer susceptibility genes as being those predominantly associated with breast cancer (*BRCA1*, *BRCA2*, *PALB2*, *ATM*, *CHEK2*), those for whom only association with ER-negative breast cancer has been demonstrated (*BARD1*, *RAD51C*, *RAD51D*) and those associated with breast cancer as part of a broad syndrome of pleomorphic tumour susceptibility (*TP53*, *CDH1*, *NF1*, *PTEN*, *STK11*). We considered very high penetrance as OR ≥ 10, high penetrance as OR ≥ 5 and intermediate/moderate penetrance as OR ≥ 2.

We considered a set of known cancer susceptibility genes (CSGs) with an autosomal dominant pattern of inheritance, as defined in Rahman, 201415 (*N* = 86 genes, Supplementary Table S6).

We considered a set of genes involved in DNA repair using the core set of genes from Knijnenburg et al.16(*N* = 276 genes, Supplementary Table S6).

We considered a set of oncogenes that are recurrently targeted by single residue mutations in a somatic context, as defined by the OncoKB database17, an expert curated database of genomic alterations in cancer (*N* = 686 genes, Supplementary Table S6).

### Comparison of Pathogenic/Likely Pathogenic variants in CSGs

Variants were defined as Pathogenic/Likely Pathogenic (P/LP) on the basis of having either (i) a classification of P/LP classification on ClinVar, with one or more stars and classified using ACMG criteria and/or (ii) MAF ≤ 0.5% and predicted to truncate the protein and with a ‘high confidence’ LOFTEE annotation15.

### T0.5 burden testing

We performed a collapsed burden test of the following classes of rare variants (defining rare as an individual variant MAF ≤ 0.5%): Class I, predicted truncating SNVs and indels with a ‘high confidence’ LOFTEE annotation; Class II, predicted truncating SNVs and indels with a ‘high confidence’ LOFTEE annotation + nonsynonymous SNVs predicted damaging by REVEL (> 0.7); Class III, all nonsynonymous SNVs. The selection of REVEL and threshold of 0.7 were selected on the basis of superior performance in cancer susceptibility genes 9, 18. Sequencing filters were relaxed to assess for additional (more common) rare variants in established breast/ovarian cancer susceptibility genes with known “founder” mutations, using ClinVar annotations to ensure identification of all pathogenic variants present. A Bonferroni-corrected threshold was imposed.

Differences between cases with and without a qualifying variant in an established breast cancer susceptibility gene were assessed using Kolmogorov-Smironov test (two-sample, comparing CDFs for a) age at breast cancer diagnosis and b) genetic enrichment score) or Chi-squared test (two-sided, differences in proportions of cases with and without bilateral disease).

### Follow-up of association signals in UK Biobank data

We followed putative association signals in UK Biobank, using the data quarriable through the AstraZeneca PheWAS Portal19. A threshold of p<0.05 was applied as nominal evidence of replication.

### Gene set enrichment analysis

g:Profiler was used to perform gene set enrichment analysis of genes with a *P* value in any class of burden test ≤ 0.01 and odds ratio > 1, using gene ontology (GO) terms and KEGG pathways for gene grouping20, 21.

### 

### Clonal hematopoiesis analysis for PTVs in PPM1D

In order to look for evidence that protein truncating variants in *PPM1D* (restricted to LOFTEE high confidence variants) were somatic mosaic in origin, we checked whether such variants had a lower variant allele frequency (VAF) compared to synonymous variants in the same gene, utilising the approach of Karczewski et al. (2020)8. Cumulative distribution functions (CDFs) were calculated for PTV and synonymous VAFs separately in increments of 0.1 (ranging from 0 to 1). Counts were summed for these distributions, with the floor of each bin used as its representative value. To compare the resulting distributions, the Kolmogorov-Smironov test (two-sample) was applied. Moods median test (two-sided) was also applied to look for a difference in median values between the two variant types.

## **Power calculations**

Gene level collapsed disease allele frequency in controls was taken as the baseline allele frequency, whilst the frequency in cases was determined by a weighted average of the enrichment found in cases with bilateral disease and/or an affected first degree relative, derivedthrough application of the conditional genotype probability equations detailed in Antoniou and Easton (2003)22. Control MAFcombined values were then sampled between values corresponding to the boundaries imposed by the actual MAFcombined in controls and odds ratios between 2 and 10. Sampling with replacement was performed for the given number of cases and controls, using their respectively calculated MAFcombined as the probability of selecting an individual with a pathogenic variant. A Fisher’s exact test was then performed for each sampling of cases and controls. This process was performed 10,000 times for each frequency/odds ratio combination and for each instance, the frequency of tests that were significant (exome-wide significance threshold of P < 1.3 x 10-06andtargeted gene set significance of P < 6.0 x 10-5) equated to study power. Power to ‘rediscover’ known breast cancer susceptibility genes was computed as above, applying values for the gene-specific OR generated from analysis of 60,466 breast cancer cases and 53,461 ethnicity matched controls by Dorling et al. 202123. Recursion was used to compute MAFcombined values for genes with OR values of 2, 5 and 10 discoverable at 80%, 90% and 99% power.

## **Supplementary Figures**

**Chart, scatter chart

Description automatically generated**

**Supplementary Figure S1 Ancestry analysis via principal components.** Plot showing two principal components for the case samples (blue, N = 2195) versus five superpopulations from 1000 Genomes Project Consortium, with delineation of Finnish and non-Finnish Europeans: NFE, non-Finnish Europeans (purple, N = 565); FIN, Finnish Europeans (black, N = 105); AMR, American (orange, N = 535); SAS, South Asian (yellow, N = 661); EAS, East Asian (grey, N = 621); AFR, African (red, N = 1418). Dashed circle represents the maximum Euclidean distance with a scaling factor of 1.5 for the NFE population. Case samples located outside the dashed circle were excluded from further analysis (N = 60).

Chart

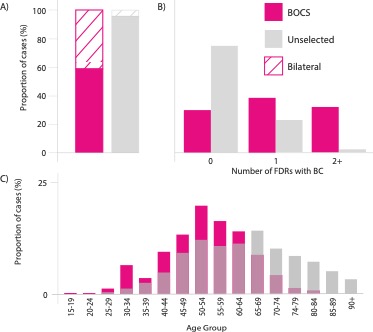
Description automatically generated

**Supplementary Figure S2 Quality by depth correlations.** Density plot with histograms showing the correlation of QualByDepth (QD) scores for variants identified in both cases and controls, with INDELs (N = 9,490 variants) and SNVs (N = 294,116 variants) plotted separately. QD was correlated between the two datasets at the variant MAF cut-off of interest (MAF ≤ 0.005). The correlations were r2 = 0.39 for INDELs and r2 = 0.31 for SNVs. Lighter shading indicates higher density, with tick marks on the shading scale representing increments of N = 100 and N = 8000 for INDELs and SNVs, respectively.

**Shape, polygon

Description automatically generated**

**Supplementary Figure S3 Variant calibration quantile-quantile plots.** QQ plots showing the observed and expected –log10(P values) at the 95th percentile of all genes following filtering on QualByDepth values at the 5th percentile in both cases and control, performed separately for single nucleotide variants (SNVs) and insertions/deletions (INDELs). For SNVs, a burden test was performed on all synonymous SNVs. For indels, a burden test was performed on all nonsynonymous variants (SNVs and and indels). A plot is shown for each coverage normalisation method. We selected the most inclusive combination that was sufficiently calibrated without excessive genomic inflation (i.e. demonstrated a normal distribution upon visual inspection) in order to retain the maximum number of variants, which equated to the intersect binomial plus method.



# 

# **Supplementary Figure S4 Phenotypic enrichment in cases.** Proportional bar plots showing the degree of enrichment of phenotypic features in BOCS cases versus unselected breast cancer cases from the literature A) proportion of bilateral vs unilateral disease; B) proportion of cases with 0, 1 and 2 or more 1st degree relatives with breast cancer; C) proportion of cases by age category. Pink denotes BOCS series and grey denotes unselected series described in the literature24-26.

Application

Description automatically generated with medium confidence

**Supplementary Figure S5 Frequency of oncogenic cancer driver genevariants in cases vs controls.** Frequency of rare oncogenic variants (MAF < 0.5%), collapsed by gene (*N* = 686 somatically mutated cancer driver genes) in cases versus controls. The left panel shows the percentage of cases/controls with a qualifying variant and the right panel shows the associated odds ratio. Genes are ordered by p value. The top 20 associations are shown.

Diagram

Description automatically generated with medium confidence

# **Supplementary Figure S6 *PPM1D* variant allele frequencies.**Violin plot showing the difference in variant allele frequencies (VAF) between high-confidence protein truncating variants (HC-PTVs) and synonymous variants, consistent with a somatic mosaic origin for the former. Boxes demark median and interquartile range values, black dots demark outliers.

## **Supplementary References**

1. Garcia M, Juhos S, Larsson M, et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Res*. 2020;9:63. doi:10.12688/f1000research.16665.2

2. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol*. Apr 11 2017;35(4):316-319. doi:10.1038/nbt.3820

3. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. Sep 2010;20(9):1297-303. doi:10.1101/gr.107524.110

4. Poplin R R-RV, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, Shakir K, Thibault J, Chandran S, Whelan C, Lek M, Gabriel S, Daly MJ, Neale B, MacArthur DG, Banks E. . Scaling accurate genetic variant discovery to tens of thousands of samples. Epub ahead of print. *bioRxiv*. 2018, July 24 2018;doi:10.1101/201178

5. Okonechnikov K, Conesa A, Garcia-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*. Jan 15 2016;32(2):292-4. doi:10.1093/bioinformatics/btv566

6. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. Aug 15 2009;25(16):2078-9. doi:10.1093/bioinformatics/btp352

7. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. Jun 6 2016;17(1):122. doi:10.1186/s13059-016-0974-4

8. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. May 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7

9. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet*. Oct 6 2016;99(4):877-885. doi:10.1016/j.ajhg.2016.08.016

10. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics*. Nov 15 2010;26(22):2867-73. doi:10.1093/bioinformatics/btq559

11. Meyer HV. plinkQC: Genotype quality control in genetic association studies. 2020;doi:10.5281/zenodo.3934294

12. Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. Oct 1 2015;526(7571):68-74. doi:10.1038/nature15393

13. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. Sep 2007;81(3):559-75. doi:10.1086/519795

14. Guo MH, Plummer L, Chan YM, Hirschhorn JN, Lippincott MF. Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. *Am J Hum Genet*. Oct 4 2018;103(4):522-534. doi:10.1016/j.ajhg.2018.08.016

15. Rahman N. Realizing the promise of cancer predisposition genes. *Nature*. Jan 16 2014;505(7483):302-8. doi:10.1038/nature12981

16. Knijnenburg TA, Vockley JG, Chambwe N, et al. Genomic and molecular characterization of preterm birth. *Proc Natl Acad Sci U S A*. Mar 19 2019;116(12):5819-5827. doi:10.1073/pnas.1716314116

17. Chakravarty D, Gao J, Phillips SM, et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol*. Jul 2017;2017doi:10.1200/PO.17.00011

18. Cubuk C, Garrett A, Choi S, et al. Clinical likelihood ratios and balanced accuracy for 44 in silico tools against multiple large-scale functional assays of cancer susceptibility genes. *Genet Med*. Jul 6 2021;doi:10.1038/s41436-021-01265-z

19. Wang Q, Dhindsa RS, Carss K, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature*. Sep 2021;597(7877):527-532. doi:10.1038/s41586-021-03855-y

20. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. May 2000;25(1):25-9. doi:10.1038/75556

21. Gene Ontology C. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res*. Jan 8 2021;49(D1):D325-D334. doi:10.1093/nar/gkaa1113

22. Antoniou AC, Easton DF. Polygenic inheritance of breast cancer: Implications for design of association studies. *Genet Epidemiol*. Nov 2003;25(3):190-202. doi:10.1002/gepi.10261

23. Breast Cancer Association C, Dorling L, Carvalho S, et al. Breast Cancer Risk Genes - Association Analysis in More than 113,000 Women. Feb 4 2021;384(5):428-439. doi:10.1056/NEJMoa1913948

24. Kheirelseid EA, Jumustafa H, Miller N, et al. Bilateral breast cancer: analysis of incidence, outcome, survival and disease characteristics. *Breast Cancer Res Treat*. Feb 2011;126(1):131-40. doi:10.1007/s10549-010-1057-y

25. Brewer HR, Jones ME, Schoemaker MJ, Ashworth A, Swerdlow AJ. Family history and risk of breast cancer: an analysis accounting for family structure. *Breast Cancer Res Treat*. Aug 2017;165(1):193-200. doi:10.1007/s10549-017-4325-2

26. Cancer Research UK Cancer Incidence. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer#heading-Zero