An investigation of the association of genetic susceptibility risk with somatic mutation burden in breast cancer

Bin Zhu*,1, Anwesha Mukherjee2, Mitchell J Machiela1, Lei Song1, Xing Hua1, Jianxin Shi1,Montserrat Garcia-Closas1, Stephen J Chanock1 and Nilanjan Chatterjee*,1,3

1Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institute of Health, Bethesda, MD 20892, USA; 2Department of Statistics, Florida State University, Tallahassee, FL 32306, USA and 3Department of Biostatistics and Department of Oncology, Johns Hopkins University, Baltimore, MD 21205, USA

Background: Genome-wide association studies have reported nearly 100 common germline susceptibility loci associated with the risk for breast cancer. Tumour sequencing studies have characterised somatic mutation profiles in breast cancer patients. The relationship between breast cancer susceptibility loci and somatic mutation patterns in breast cancer remains largely unexplored.

Methods: We used single-nucleotide polymorphism (SNP) genotyping array data and tumour exome sequencing data available from 638 breast cancer patients of European ancestry from The Cancer Genome Atlas (TCGA) project. We analysed both genotype data and, when necessary, imputed genotypes for 90 known breast cancer susceptibility loci. We performed linear regression models to investigate possible associations between germline risk variants with total somatic mutation count (TSMC), as well as specific mutation types. We examined individual SNP genotypes, as well as a multi-SNP polygenic risk score (PRS). Models were statistically adjusted for age at diagnosis, stage, oestrogen-receptor (ER) and progesterone-receptor (PR) status of breast cancer. We also performed stratified analyses by ER and PR status.

Results: We observed a significant inverse association (P = 8.75 × 10⁻⁶; FDR = 0.001) between the risk allele in rs2588809 of the gene RAD51B and TSMC across all breast cancer patients, for both ER⁺ and ER⁻ tumours. This association was also evident for different types of mutations. The PRS analysis for all patients, with or without rs2588809, showed a significant inverse association (P = 0.01 and 0.04, respectively) with TSMC. This inverse association was significant in ER⁺ patients with the ER⁺-specific PRS (P = 0.02), but not among ER⁻ patients for the ER⁻-specific PRS (P = 0.39).

Conclusions: We observed an inverse association between common germline risk variants and TSMC, which, if confirmed, could provide new insights into how germline variation informs our understanding of somatic mutation patterns in breast cancer.

Breast cancer is a complex disease, in which the characteristics of germline susceptibility loci as well as the spectrum of somatic alterations have begun to emerge, largely due to the capacity to conduct large-scale genome-wide studies. Genetic susceptibility loci implicated in breast cancer include highly penetrant rare variants in genes such as BRCA1 and BRCA2, moderately penetrant low-frequency variants in genes such as ATM, PALB2 and CHEK2, and multiple low-penetrant common variants identified more recently through genome-wide association studies (GWAS; Mavaddat et al, 2010). Specifically, GWAS have reported nearly 100 common susceptibility loci for breast cancer, marked by common single-nucleotide polymorphisms (SNPs; Long et al, 2012; Siddiq et al, 2012; Bojesen et al, 2013; Garcia-Closas et al, 2013; Michailidou et al, 2013; Cai et al, 2014; Milne et al, 2014; Michailidou et al, 2015). Although each SNP individually exhibits small effect size, in combination they can explain a substantial

*Correspondence: Dr B Zhu; E-mail: bin.zhu@nih.go or Dr N Chatterjee; E-mail: nilanjan@jhu.edu

Received 4 February 2016; revised 17 June 2016; accepted 30 June 2016; published online 28 July 2016

© 2016 Cancer Research UK. All rights reserved 0007 – 0920/16
Association of germline-somatic variants in BRCA

proportion of the variation in the familial risk for breast cancer, as well as risk in the general population (Easton et al, 2015; Mavaddat et al, 2015; Michailidou et al, 2015).

In parallel, whole-exome and genome-tumour sequencing studies have been conducted to define the landscape of somatic mutations of breast cancer (Banerji et al, 2012; Curtis et al, 2012; Ellis et al, 2012; Koboldt et al, 2012; Nik-Zainal et al, 2012; Shah et al, 2012; Stephens et al, 2012; Nik-Zainal et al, 2016), particularly focusing on identifying candidate driver genes, namely the genes harbouring mutations that confer selective growth advantage. In addition, studies of somatic mutation patterns for breast cancer have identified distinct mutational signatures across the genome. The number of somatic mutations (or mutational burden) in coding exons varies widely between breast tumours, and it has been related to age at diagnosis and to tumour grade (Stephens et al, 2012). Although genetic susceptibility variants show distinct associations with different pathological subtypes of breast cancer, particularly those defined by ER status (Mavaddat et al, 2010; Siddiq et al, 2012; Garcia-Closas et al, 2013), little is known of the relationship between the inherited and somatic genetic components. A recent survey of cancer predisposition genes, defined as genes harbouring high- or moderate-penetrant risk variants, has suggested that a large fraction of these genes could be oncogenic when mutated somatically (Rahman, 2014). In contrast, a study has reported no evidence that cancer susceptibility regions that harbour common low-penetrant susceptibility are preferentially selected for altered somatic mutation frequencies in cancer patients (Machiela et al, 2015).

In this report, using the breast cancer data from The Cancer Genome Atlas (TCGA) study (Koboldt et al, 2012), we examined associations between established breast cancer susceptibility loci and exome-wide single-nucleotide substitution counts observed in tumour tissues. As mutations carrying the aetiological signature of tumours are expected to be present across the whole genome, and not only in the specific genes, examination of association of cancer risk factors and total somatic mutation count (TSMC) of substitutions in the whole exome, might be a powerful approach to explore relationships between germline variants and somatic substitutions. For example, biologic age, which is the strongest risk factor for many cancers and is often a surrogate for cumulative carcinogenic events, has been shown to be directly associated with TSMC across many cancers (DePinho, 2000). For other major risk factors, such as smoking for lung cancer and sun exposure for melanoma, specific signature mutations, as hallmarks of exposure, are observed across the whole genome (Pfeifer et al, 2002; 2005).

METHODS

Data were extracted from germline genotypes generated using the Affymetrix Genome-Wide Human SNP Array 6.0 on circulating leucocyte DNA drawn from 638 breast cancer cases of European ancestry. To remove subjects who may not be Caucasians, cases were selected on the basis of principal component analysis, which combined common SNP genotypes with ones from HapMap (Altshuler et al, 2010b) reference samples (Supplementary Figure 1). We performed genotype imputation for restricted European ancestry samples using IMPUTE2 (Howie et al, 2009) with haplotypes generated by the 1000 Genome Project (Phase3; Altshuler et al, 2010a) as the reference.

Genotyped or imputed dosage data were available for 90 established SNPs representing common susceptibility loci with minor-allele frequencies (MAFs) > 0.01. All loci had reported breast cancer risk associations below the threshold for genome-wide significance (P < 5 × 10⁻⁸). We initially selected 94 SNPs from the study by Michailidou et al (2015) and removed two SNPs (rs7726159 and rs2380205) that were not genome-wide significant in the study by Michailidou et al (2013), and one SNP that is rare (rs17879961, MAF = 0.0049). Except for one SNP not present in the 1000 Genomes Project reference panel, the remaining 90 breast cancer susceptibility SNPs passed quality filter with IMPUTE2 info score > 0.8 (Supplementary Table 1).

Somatic mutation data were obtained from whole-exome sequencing of TCGA breast cancer tumour samples. Mutation counts were extracted from the Mutation Annotation File (version 2.1.1.0 curated) generated by the Washington University Genome Institute. Details about sample preparation, sequencing protocol, and mutation calling pipeline are described elsewhere (Koboldt et al, 2012). For clinical information, we retrieved age at diagnosis, oestrogen-receptor (ER) status, progesterone-receptor (PR) status and tumour stage.

We used the somatic mutation burden, overall as TSMC or by mutation-specific types, as the outcome variable to perform linear regression analysis of association with SNP genotypes, individually or collectively as a polygenic risk score (PRS). Analyses were adjusted for subject and tumour characteristics, including age at diagnosis, ER status, PR status and tumour stage. Subjects with extremely low or high numbers of TSMC (bottom 1% and top 3% of subjects, respectively) were excluded as outliers. The mutation counts were log₁₀ transformed and results were presented after standardising each type of log-transformed mutation count to have unit standard deviation so that effect-sizes are comparable across TSMC of different mutation types. At each locus, the genotype was coded based on the number of risk alleles (0, 1 or 2). For the current sample size, it was estimated that the study has 80% power to detect effect size of 0.5-s.d. unit change in mutation count (in log₁₀ scale) per copy of a SNP allele with a population frequency of 0.33 at 5% type I error. Power curves for additional effect size and risk allele frequency combinations are illustrated in Supplementary Figure 2.

For each subject, the PRS reflected the total genetic susceptibility burden based on the 90 independent SNPs, and was defined as the weighted combination of each SNP genotype with the weights defined by previously reported log-odds-ratio of association of the SNPs with breast cancer (Supplementary Table 1). We used log-odds-ratio estimates for overall breast cancer and subtypes defined by ER status reported in the study by Michailidou et al (2015) for all non-correlated SNPs, and estimates in the study by Mavaddat et al (2015) for three correlated SNPs with conditional independent signals in 11q13 (rs554219, rs75915166 and rs78540526). In addition, we examined mutation burden with the following specific types: mutations from thymine (or adenine on the other strand) to other nucleotides, from cytosine (or guanine on the other strand) to other nucleotides, transition mutations, transversion mutations and APOBEC-mediated mutations, defined as cytosine to thymine and cytosine to guanine substitutions in the TCW motifs (W is either adenine or thymine; Roberts et al, 2013).

We further considered the somatic copy-number burden in the analysis. The processed segments of copy-number variation (CNV) were downloaded (in November 2015). Following previous work (Laddha et al, 2014), we used the magnitude 0.2 as the threshold to identify amplifications and deletions, and required at least 10 markers included in the CNV segment. The total number of CNV segments across the genome was calculated and treated as a covariate in the linear regression model.

RESULTS

The age at diagnosis of 638 breast cancer patients ranged from 26 to 90 years old with a median of 59.5 years. We first examined each characteristic without adjustment for other characteristics.
(Table 1). Specifically, TSMC was higher for patients with older age at diagnosis \((P = 4.02 \times 10^{-4}\) for age groups, Table 1; \(P = 5.04 \times 10^{-4}\) for the trend of age, Supplementary Figure 3), low PRS \((P = 0.04\) for PRS groups, Table 1; \(P = 0.01\) for the trend of PRS, Figure 1), negative vs positive ER status \((P = 1.74 \times 10^{-10}\), negative vs positive PR status \((P = 2.87 \times 10^{-15}\), and late vs early stages \((P = 4.08 \times 10^{-3}\). In addition, TSMC were significantly associated with patient group defined by both ER and PR status \((P = 2.95 \times 10^{-14}\). In an analysis of all characteristics (PRS, age at diagnosis, ER status, PR status and stage) simultaneously fitted in a linear regression model (Supplementary Table 4), we observed that the TSMC was associated with age at diagnosis \((P = 2.3 \times 10^{-6}\), tumour stage \((P = 3.05 \times 10^{-3}\) and \(1.68 \times 10^{-3}\) for stage II vs stage I and for stage III/IV vs stage I, respectively) and PR status \((P = 3.96 \times 10^{-7}\), but not with ER status \((P = 0.18\). Further, in stratified analyses by mutation type, we observed that both ER \(^+\) and PR \(^+\) tumours were significantly associated with a lower mutation count of thymine to other nucleotides, particularly ER \(^+\) tumours; whereas PR \(^+\), but not ER \(^+\), tumours were associated with lower mutation count of cytosine to other nucleotides (comparisons of positive vs negative ER and PR status; Supplementary Tables 5 and 6).

In an analysis of association of the mutation count and individual breast cancer susceptibility SNPs, rs2588809 in RAD51B was inversely associated with TSMC \((P = 8.75 \times 10^{-6}\), Table 2) with \(P\)-value of 0.001 adjusted for multiple comparisons using the Benjamini–Hochberg false-discovery rate (FDR; Benjamini and Hochberg, 1995). Statistical significance of the association was evident across all types of mutations and breast cancer subtypes (Table 2). Two other SNPs, rs11814448 in DNAJC1 and rs13387042, which localises to a gene-poor region of chromosome 2q35, also showed possible inverse associations (FDR = 0.25) with TSMC (Supplementary Table 1 for all subjects, Supplementary Table 2 for ER \(^+\) subjects and Supplementary Table 3 for ER \(^-\) subjects).

PRS for overall breast cancer was inversely associated with TSMC \((P = 1.34 \times 10^{-2}\), Figure 1) as well as for all different types of mutations (Table 3). We observed a significant trend \((P = 2.28 \times 10^{-2}\), Figure 1 right panel) in that the strength of association of the individual SNPs with TSMC (measured by the regression coefficient) tended to be larger for those with larger reported odds ratio of association with breast cancer risk, but this trend was

### Table 1. Total in 638 breast tumours stratified by ER and PR status, PRS (by tertile), age at diagnosis and tumour stage

| Status | N   | Mean (± s.d.) | \(P\)-value* |
|--------|-----|---------------|--------------|
| All    | 638 | 1.57 (± 0.31) | 1.74e-10     |
| ER     | –   | 1.71 (± 0.32) | 2.87e-15     |
| +      | 516 | 1.52 (± 0.31) | 2.95e-14     |
| PR     | –   | 1.71 (± 0.33) | 4.17e-02     |
| +      | 452 | 1.51 (± 0.29) | 10.04e-04    |
| ER and PR | – | 1.71 (± 0.31) | 4.02e-04     |
| +      | 440 | 1.51 (± 0.29) | 4.08e-03     |
| PRS \(^a\) |  0.53–0.94 | 1.61 (± 0.31) | 4.17e-02     |
| Age (years) | 0.94 | 1.52 (± 0.34) | 4.02e-04     |
| >55     | 405  | 1.61 (± 0.30) | 4.08e-03     |
| Stage   | I   | 1.49 (± 0.30) | 10.04e-04    |
| II      | 360  | 1.58 (± 0.32) | 10.04e-04    |
| III/IV  | 160  | 1.60 (± 0.30) | 10.04e-04    |

Abbreviations: ER – oestrogen receptor, PR – progesterone receptor; PRS – polygenic risk score; TSMC – total somatic mutation count.
*a by Kruskal–Wallis test without adjustment for other characteristics.
*b Grouped by tertiles of PRS, which is not standardised.
*c PRS is based on 90 SNPs with reported breast cancer overall odds ratios (Supplementary Table 1).

The \(P = 0.01\) under the null hypothesis that the regression line is flat (left). Scatterplot for mutation count coefficient versus breast cancer risk overall odds ratio with \(P = 2.28 \times 10^{-2}\) for the flat line hypothesis (right). The least squares regression lines are shown by the solid lines with 95% confidence intervals in gray shades.

Figure 1. Scatterplot for TSMC vs PRS.
largely influenced by SNP rs11814448 in DNAJC1 (P = 0.37 when excluding SNP rs11814448). The association between PRS and TSMC remained significant (P = 3.81 × 10⁻²) after excluding rs2588809 in RAD51B in the calculation of PRS. Analysis of ER⁺-specific PRS in ER⁺ tumours and ER⁻-specific PRS in ER⁻ tumours, showed a significant inverse association for ER⁺ but not ER⁻ tumours (P = 0.01 for heterogeneity, Supplementary Figure 4). Further, the association observed for ER⁺ tumour appears to be to be present only for ER⁺ PR⁺ tumours (P = 0.01 for heterogeneity, Supplementary Table 7 for PRS and Supplementary Table 8 for RAD51B SNP).

**DISCUSSION**

We reported an inverse association between TSMC in breast tumours and genetic predisposition conferred by common breast cancer susceptibility SNPs. In particular, a highly significant inverse association was observed with respect to the germline risk variant defined by the SNP rs2588809 in the DNA-repair gene RAD51B. Moreover, a significant inverse association was also observed for a PRS for breast cancer that includes genetic predisposition of 90 breast cancer associated loci but this association was only evident among ER⁺ tumours with respect to ER⁺-specific PRS.

The reported inverse association may provide insight into links between germline risk variants and somatic mutations in breast cancer development. There are several possible underlying mechanisms by which the inverse association could arise.

It has been previously shown that genetic susceptibility loci differentially influence distinct subtypes of breast cancer (Stephens et al, 2012), which, in turn, can be related to the number of somatic mutations. For example, many SNPs that have been reported to date from GWAS of breast cancer show differential associations with the risk of ER⁺ and ER⁻ breast cancer. Because the number of ER⁺ tumours included in GWAS to date has been substantially more than the number for ER⁻ tumours (Michaelidou et al, 2015), GWAS discovery has preferentially identified SNPs related to ER⁺ tumours. As the number of total mutations tends to be larger in ER⁻ than ER⁺ tumours (Table 1), an inverse association between the SNPs and mutation counts may be observed if the analysis is not adjusted for ER status (Stephens et al, 2012).

We observed an inverse association between germline risk and TSMC after adjustment for age at diagnosis and tumour characteristics, including ER/PR status and stage. The association with the RAD51B SNP was present for both ER⁺ and ER⁻ tumours, although this SNP is only associated with the risk of ER⁺ tumours (Michaelidou et al, 2013). In contrast, the association of TSMC with PRS was present only for ER⁺-specific PRS in ER⁺ tumours, and this association appeared to be strongest for the ER⁺ PR⁺ subtype. Further tumour characteristics, such as the grade that could not be evaluated in this report due to lack of available data, could explain the reported inverse associations between PRS and TSMC. However, the distinct pattern of

### Table 2. Association between somatic mutation phenotypes and SNP rs2588809 at RAD51B

| Phenotype | Group       | Estimate | S.e. | P-value |
|-----------|-------------|----------|------|---------|
| TSMC      | All subjects| −0.29    | 0.07 | 8.75e–06|
| TSMC      | ER⁺         | −0.25    | 0.07 | 6.24e–04|
| TSMC      | ER⁻         | −0.52    | 0.17 | 2.46e–03|
| TSMC      | ER⁺ PR⁺     | −0.25    | 0.08 | 1.77e–03|
| TSMC      | ER⁺ PR⁻     | −0.27    | 0.22 | 2.13e–01|
| TSMC      | ER⁻ PR⁺     | −0.46    | 0.18 | 1.07e–02|
| T counts | All subjects| −0.27    | 0.07 | 6.59e–05|
| C counts | All subjects| −0.28    | 0.07 | 3.12e–05|
| Transitions | All subjects| −0.29    | 0.07 | 1.02e–05|
| Transversions | All subjects| −0.25    | 0.07 | 1.61e–04|
| APOBEC counts | All subjects| −0.20    | 0.07 | 2.64e–03|

**Abbreviations:** ER = oestrogen receptor; PR = progestrogen receptor; SNP = single-nucleotide polymorphism; TSMC = total somatic mutation count.

### Table 3. Association between somatic mutation phenotypes and PRS

| Phenotype | Group       | Estimate | S.e. | P-value |
|-----------|-------------|----------|------|---------|
| TSMC      | All subjects| −0.19    | 0.08 | 1.34e–02|
| TSMC      | All subjects| −0.16    | 0.08 | 3.81e–02|
| TSMC      | ER⁺         | −0.19    | 0.08 | 1.97e–02|
| TSMC      | ER⁻         | 0.18     | 0.21 | 3.91e–01|
| TSMC      | ER⁺ PR⁺     | −0.24    | 0.09 | 7.24e–03|
| TSMC      | ER⁺ PR⁻     | 0.01     | 0.25 | 9.59e–01|
| TSMC      | ER⁻ PR⁺     | 0.17     | 0.22 | 4.36e–01|
| TSMC      | ER⁻ PR⁻     | 0.21     | 0.08 | 8.18e–03|
| TSMC      | All subjects| −0.18    | 0.08 | 2.21e–02|
| TSMC      | All subjects| −0.19    | 0.08 | 1.54e–02|
| APOBEC counts | All subjects| −0.23    | 0.08 | 4.47e–03|

**Abbreviations:** ER = oestrogen receptor; PR = progestrogen receptor; PRS = polygenic risk score; TSMC = total somatic mutation count.

The reported inverse association may provide insight into links between germline risk variants and somatic mutations in breast cancer development. There are several possible underlying mechanisms by which the inverse association could arise.
association seen for the RAD51B SNP and PRS are unlikely to be both explained by subtype heterogeneity.

We observed an association between higher TSMC with older age at diagnosis, ER+ PR+ and higher stage. Older age at diagnosis was previously reported to be associated with cytosine to thymine substitution in ER+ but not with TSMC, across breast cancer patients, whereas the observation of higher TSMC associated with higher stage is consistent with the previous finding (Stephens et al, 2012). To our best knowledge, we are not aware of other studies reporting an association between TSMC, overall and by mutation type, with respect to joint status ER, PR, stage and age at diagnosis. Our results suggested that although PR status, but not ER, is strongly predictive of overall TSMC, distinct mutation signatures could be associated with ER (cytosine to thymine substitutions) and PR status (cytosine to other nucleotides) when the characteristics were analysed jointly.

It is possible that the inverse association we observe between genetic risk and TSMC is a broader phenomenon that cannot be explained by subtype heterogeneity alone. The best known example of interaction of germline and somatic mutation is the ‘two-hit’ model for carcinogenesis in retinoblastoma (Knudson, 1971). Under this model, the first-hit could be either a germline susceptibility variant or a somatic mutation in an important cancer predisposition gene. Thus subjects with elevated genetic predisposition may require fewer stages to develop a malignancy of the breast than subjects at lower genetic risk. Therefore, it is possible that the observed inverse association is the result of an underlying continuous process of cancer development, in which both germline variants and somatic mutations contribute and perhaps overlap with respect to their relative contributions to development of breast cancer. To further understand the biological basis of our observations, it will be necessary to understand the causal mechanisms that underpin the relationship between common susceptibility alleles and TSMC, as a marker of mutational events critical for development of distinct subtypes of breast cancer. In addition, the current study only recorded presence or absence of ER or PR. By quantifying the magnitudes of ER or PR as quantitative traits in future studies, it may be possible to delineate the relationships between TSMC with ER or PR levels more precisely.
Peltonen L, Abecasis GR, Bustamante CD, Chakravarti A, Gibbs RA, Jorde L, Kaufmann KY, Kent KA, Li T, McGuire AL, McVean GA, Osozou PN, Rotimi CN, Su YY, Tojo LH, Tybulewicz L, Brooks LD, Felsenfeld AL, McEwen JF, Afifi A, Christopher R, Clemen NC, Collins FS, Duncanson A, Green ED, Guyer MS, Peterson JL, Schafer AJ, Abecasis GR, Almasy L, Altshuler DL, Anton A, Brooks LD, Burdick RW, Gibbs RA, Hurles ME, McVean GA. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319): 1061–1073.

Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu FL, Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Cochran G, DeLisi C, Dooling DJ, Dyer TD, Engelhard D, Flicek P, Flannick J, Frankish A, Jhangiani SN, Jones MC, Kefalas P, Kellis M, Kent JS, K想办法 E, Kent A, Li T, McGuire AL, Nekrutenko A, Parkinson M, Pollock S, Shen Y, Sherry ST, Sklar P, Smith KE, Stang P, Stivers D, Voight BF, Waterston RH, Wilson RK, Zhang JF, Zody MC. 1000 Genomes Project Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7317): 52–58.

Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L, Cortes ML, Fernandez-Lopez JC, Peng S, Ardile KG, Aquiles D, Bautista-Pina V, Duke F, Francis C, Jung J, Maffazza Aziz A, Onofrio RC, Parkin M, Pho NH, Quintanar-Jurado V, Ramos AH, Rebollar-Vega R, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Hurles M, Korn JM, Kristiansson K, Lee C, McCarroll SA, Nemesh J, Lewis LR, Ren YR, Wheeler D, Yu W, Muzny DM, Barnes C, Darvishi K, Bonnen PE, de Bakker PIW, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Maier HR, McEwen JE, Manca MC, Marshall PA, Matsud I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng CQ, Brooks LD, McEwen JE. International HapMap 3 Consortium (2010b) Integrating common and rare genetic variation from population-scale sequencing.

Bojesen SE, Pooley KA, Johnatty SE, Beesley J, Michailidou K, Tyrer JP, Anton-Culver H, Bunker CH, Kristensen V, Ness RB, Muir K, Edwards R, Esser D, Krausz T, Lievesley S, Jakes R, Peacey M, Plummer S, Cunningham JM, Eccles M, Fasholtz J, Fearon K, Godwin AK, Green ED, Guyer MS, Peterson JL, Schafer AJ, Abecasis GR, Almasy L, Altshuler DL, Ankerst D, Aslan-Khan M, Ates S, Cui H, Craig M, Foye B, Gao X, Green A, Groden J, Gross S, Hahne F, Hehir K, Houlston R, Hsuung CN, Hutterer R, Imyanitov EN, Kwon YJ, Lappalainen T, Lin Y, Lichter AS, Liberman N, Li X, Liu Q, Lohr H, Lu Y, Kloor MD, Markowetz F, Matias-Guijarro B, Meyer CA, Munroe P, Nannya Y, Narod S, Oravecz T, Poggetti Y, Shibata D, Sigurdsson H, Spurdle AB, van Asperen CJ, Vriezen R, Zheng W, Kachanovsky R, Källberg H, Kato C, Kolonel LN, Lemaître JC, Li J, Liu HH, Longacre TA, Lozano M, Lynch HT, Lysiak AJ, Mandrou L, Manni JJ, Mauk D, McCartney A, McInerney E, McLeod H, Meintjes J, Moreira S, Morley M, Mott R, Murray GI, Nissen H, Oberg R, Olkhovskaya E, Oliner M, Palomeque P, Pekarek A, Personeni N, Price AL, Quail MA, Rackham O, Reddel RR, Ribaud P, Ridolfi S, Rutkoski L, Saltzstein A, Shaw R, Stahelin B, Stamos V, Strudwick C, Sudar D, Talantsev D, Thawley SE, Tischendorf L, Vakkuri V, van der Hout E, Verheijen M, Visser AG, Wang Z, Xu X, Xia J, Yang Q, Zhang Y, Zhao C, Zhu C, Zhu M, Zichner T, Zhang Z, Zelent Z, Zöller M, Zhang R, Zou L. 1000 Genomes Project Consortium (2010a) A map of human genome variation from population-scale sequencing. *Nature* 467(7319): 1000–1018.

Bojesen SE, Pooley KA, Johnatty SE, Beesley J, Michailidou K, Tyrer JP, Anton-Culver H, Bunker CH, Kristensen V, Ness RB, Muir K, Edwards R, Esser D, Krausz T, Lievesley S, Jakes R, Peacey M, Plummer S, Cunningham JM, Eccles M, Fasholtz J, Fearon K, Godwin AK, Green ED, Guyer MS, Peterson JL, Schafer AJ, Abecasis GR, Almasy L, Altshuler DL, Ankerst D, Aslan-Khan M, Ates S, Cui H, Craig M, Foye B, Gao X, Green A, Groden J, Gross S, Hahne F, Hehir K, Houlston R, Hsuung CN, Hutterer R, Imyanitov EN, Kwon YJ, Lappalainen T, Lin Y, Lichter AS, Liberman N, Li X, Liu Q, Kloor MD, Markowetz F, Matias-Guijarro B, Meyer CA, Munroe P, Nannya Y, Narod S, Oravecz T, Poggetti Y, Shibata D, Sigurdsson H, Spurdle AB, van Asperen CJ, Vriezen R, Zheng W, Kachanovsky R, Källberg H, Kato C, Kolonel LN, Lemaître JC, Li J, Liu HH, Longacre TA, Lozano M, Lynch HT, Lysiak AJ, Mandrou L, Manni JJ, Mauk D, McCartney A, McInerney E, McLeod H, Meintjes J, Moreira S, Morley M, Mott R, Murray GI, Nissen H, Oberg R, Olkhovskaya E, Oliner M, Palomeque P, Personeni N, Price AL, Quail MA, Rackham O, Reddel RR, Ribaud P, Rutkoski L, Saltzstein A, Shaw R, Stahelin B, Stamos V, Thawley SE, Tischendorf L, Vakkuri V, van der Hout E, Verheijen M, Visser AG, Wang Z, Xu X, Xia J, Yang Q, Zhang Y, Zhao C, Zhu C, Zhu M, Zichner T, Zhang Z, Zou L. 1000 Genomes Project Consortium (2010a) A map of human genome variation from population-scale sequencing. *Nature* 467(7319): 1061–1073.

Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu FL, Bonnen PE, de Bakker PIW, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inoue M, Jiao Y, Kober K, Kolarov N, Kudoh J, Kullberg J, Korhonen LE, Kotelarciuk A, Parkin M, Bastiaens P, Harris A, Lewis LR, Ren YR, Wheeler D, Muzny DM, Barnes J, Darvishi K, Hurles ME, Korn J, Kristiansson K, Lee C, McCarron SA, Nemesh J, Keinan A, Montgomery SB, Pollack S, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng CQ, Brooks LD, McEwen JE. International HapMap 3 Consortium (2010b) Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7317): 52–58.
Machiela MJ, Ho BM, Fisher VA, Hua X, Chanock SJ (2015) Limited evidence that cancer susceptibility regions are preferential targets for somatic mutation. *Genome Biol* 16: 1–11.

Mavaddat N, Antoniou AC, Easton DF, Garcia-Closas M (2010) Genetic susceptibility to breast cancer. *Mol Oncol* 4(3): 174–191.

Maftaz N, Pharoah PP, Michailidou K, Tjonneland A, Menendez J, Rahman S, Canisius S, Dennis J, Lush MJ, Swerdlow A, Ashworth A, Orr N, Schoemaker MJ, Figueroa J, Chanock SJ, Brinton L, Lissowska J, Couch FJ, Olson JE, Vachon C, Pankratz VS, Lambrechts D, Wilders H, Van Ongeval C, Van Limbergen E, Kristensen VN, Anton-Culver H, Neuhausen SL, Arndt V, Schmutzler RK, Sutter C, Yang RX, Schurmann P, Bremer M, Xing W, Chen YQ, Chen YF, Tao L, Chen ZJ, Wang Q, Dong J, Zheng W (2014) Association of germline-somatic variants in BRCA susceptibility loci for breast cancer. *Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer*. *PloS Genet* 8(2): e1002532.

Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, Swerdlow A, Ashworth A, Orr N, Schoemaker MJ, Figueroa J, Chanock SJ, Brinton L, Lissowska J, Couch FJ, Olson JE, Vachon C, Pankratz VS, Lambrechts D, Wilders H, Van Ongeval C, Van Limbergen E, Kristensen VN, Anton-Culver H, Neuhausen SL, Arndt V, Schmutzler RK, Sutter C, Yang RX, Schurmann P, Bremer M, Xing W, Chen YQ, Chen YF, Tao L, Chen ZJ, Wang Q, Dong J, Zheng W (2014) Limited evidence that cancer susceptibility regions are preferential targets for somatic mutation. *Genome Biol* 16: 1–11.
GENICA Network, Andrulis IL, Knight JA, Glendon G, Tchatchou S, iConFab Investigators: Australian Ovarian Cancer Study Group, Matsuo K, Io H, Iwata H, Tajima K, Li J, Brand JS, Brenner H, Dieffenbach AR, Arndt V, Stege Maher, C, Lambrechts D, Peutenman G, Christiaens MR, Smeets A, Jakubowska A, Lubinski J, Jaegers-Bieniek K, Durda K, Hartman M, Hui M, Yen Lim W, Van Chan C, Marme F, Yang R, Bugert P, Lindblom A, Margolin S, Garcia-Closas M, Channock SJ, Lissowska J, Figueroa JD, Bojesen SE, Nordestgaard BG, Flyger H, Hooing MJ, Krüger M, Van den Oweland AM, Koppert J, Zheng W, Deming Halverson S, Shrubsole MJ, Long J, Chang-Claude J, Rudolph A, Seibold P, Flesch-Janys D, Wingquist R, Pylkas K, Jukkola-Vuorinen A, Grip M, Cox A, Cross SS, Reed MW, Schmidt DF, Broeks A, Cornelliens S, Braal L, Kang D, Choi YJ, Park SK, Noh D, Simard J, Dumont M, Goldberg MS, Labreche F, Fasching PA, Hein A, Ekici AB, Beckmann MW, Radice P, Peterlongo P, Azzollini J, Barile M, Sawyer E, Tomlison I, Kerin M, Miller N, Hopper JL, Schmidt DF, Makalic E, Southey MC, Hwang Teo S, Hary P, Sivanandakumaran T, Kay TK, Sheng CH, Ihsiang CN, Yu JC, Hou MF, Guenel P, Truong T, Sanchez M, Mulot C, Blot W, Cai Q, Nevanlinna H, Muralan TA, Aitomaki K, Blomqvist C, Wu AH, Tseng CC, Van Den Berg D, Strand DM, Bogdanova N, Dork T, Muir K, Lophatananon A, Stewart-Brown S, Sirirawanangs P, Mannerama A, Kataja V, Kosma VM, Hartikainen JM, Shu X0, Lu W, Gao YT, Zhang B, Couch FJ, Tolan AE. TNBCYounakakos D, Sangrajrang S, McKay J, Wang X, Olson JE, Vachon C, Purrington K, Severi G, Baglietto L, Haiman CA, Henderson BE, Schumacher F, Le Marchand L, Devilee P, Tollefsen RA, Syneana C, Czene K, Eriksson M, Humphreys K, Darabi H, Ahmed S, Shah M, Pharoah PD, Hall P, Giles GG, Benitez J, Dunning AM, Chevene-Trench G, Easton DF (2014) Common non-synonymous SNPs associated with breast cancer susceptibility: findings from the Breast Cancer Association Consortium. *Hum Mol Genet* 23(22): 6906–6911.

Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau MW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Madden D, Marrison T, Tarpey S, Davies HR, Papaemmanuil E, Stephens PJ, McLaren S, Butler AP, Teague JW, Jonsson G, Garber JE, Silver D, Miron P, Amata F, Boyault A, Langerod A, Butt A, Martens JW, Aparicio SA, Borg A, Salomon AV, Thomas G, Borresen-Dale AL, Richardson AL, Neuberger MS, Futreal PA, Campbell PJ, Stratton MR (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell* 149(5): 979–993.

Nik-Zainal S, Davies H, Staf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov LB, Martin S, Wedge DC, Van Loo P, Ju YS, Smid M, Brinkman AB, Morganella S, Aire M, Lingierde OC, Langerod A, Ringqvist M, Aha S-M, Boyault S, Broeks A, Butler A, Desmedt C, Dirix L, Dronov S, Tlsty T, Costello J, Van de Vijver M, Desmedt C, Sotiriou C, Tutt A, Martens JW, Aparicio SA, Borg A, Salomon AV, Thomas G, Borresen-Dale AL, Richardson AL, Neuberger MS, Futreal PA, Campbell PJ, Stratton MR (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell* 149(5): 979–993.

Pfeifer GP, Ben-David D, Ben-David S, Katz R (2012) Identification of somatic mutations in breast cancer whole-genome sequences. *Nature* 534(7605): 47–54.

Pfeifer GP, Genisken MF, Olivier M, Treytakova N, Hacht SS, Hainaut P (2002) Tobacco smoke carcinomas, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* 21(48): 7435–7451.

Pfeifer GP, You YH, Besaratinia A (2005) Mutations induced by ultraviolet light. *Mutat Res* 571(1-2): 19–31.

Supplementary Information accompanies this paper on British Journal of Cancer website (http://www.nature.com/bjc)