BRCA1 and *BRCA2* 5' non-coding region variants identified in breast cancer patients alter promoter activity and protein binding.

Leslie J. Burke[1*], Jan Sevcik[1,2*], Gaetana Gambino[1,3*], Emma Tudini[1,4], Eliseos J. Mucaki[5], Ben C. Shirley[6], Philip Whiley[1,4], Michael T. Parsons[4], Kim De Leeneer[7], Sara Gutiérrez-Enríquez[8], Marta Santamariña[9], Sandrine M. Caputo[10], Elizabeth Santana dos Santos[10,11,12], Jana Soukupova[2], Marketa Janatova[2], Petra Zemankova[2], Klara Lhotova[2], Lenka Stolarova[2], Mariana Borecka[2], Alejandro Moles-Fernández[8], Siranoush Manoukian[13], Bernardo Bonanni[14], ENIGMA consortium, Stacey L. Edwards[4], Marinus J. Blok[15], Thomas van Overeem Hansen[16], Maria Rossing[16], Orland Diez[8,17], Ana Vega[9], Kathleen B.M. Claes[7], David E. Goldgar[18], Etienne Rouleau[19], Paolo Radice[20], Paolo Peterlongo[21], Peter K. Rogan[5,6], Maria Caligo[3], Amanda B. Spurdle[4] and Melissa A. Brown[1^]

1. School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, Australia.

2. Institute of Biochemistry and Experimental Oncology, First Faculty of Medicine. Charles University in Prague, Czech Republic.

3. Section of Molecular Genetics, Department of Laboratory Medicine, University Hospital of Pisa, Pisa, Italy.

4. Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Australia.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/humu.23652.

5. University of Western Ontario, Department of Biochemistry, Schulich School of Medicine and Dentistry, Canada.

6. CytoGnomix Inc., London, Ontario, N5X 3W1, Canada.

7. Center for Medical Genetics, Ghent University Hospital, and Cancer Research Institute Ghent (CRIG), Ghent University, Ghent, Belgium.

8. Oncogenetics Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain.

9. Fundación Pública Galega de Medicina Xenómica-SERGAS, Grupo de Medicina Xenómica-USC, CIBERER, IDIS, Santiago de Compostela, Spain.

10. Service de Génétique, Department de Biologie des Tumeurs, Institut Curie, Paris, France.

11. Department of oncology, Center for Translational Oncology, Cancer Institute of the State of São Paulo - ICESP, São Paulo, Brazil.

12. A.C.Camargo Cancer Center, São Paulo, Brazil.

13. Unit of Medical Genetics, Department of Medical Oncology and Hematology, Fondazione IRCCS (Istituto di Ricovero e Cura a Carattere Scientifico) Istituto Nazionale dei Tumori (INT), Milan, Italy.

14. Division of Cancer Prevention and Genetics, Istituto Europeo di Oncologia, Milan, Italy.

15. Department of Clinical Genetics, Maastricht University Medical Centre, Maastricht, The Netherlands.

16. Center for Genomic Medicine, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark.

17. Area of Clinical and Molecular Genetics, University Hospital Vall d'Hebron (UHVH), Barcelona, Spain.

18. Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah, USA.

19. Gustave Roussy, Villejuif, France.

20. Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy.

21. IFOM, Fondazione Istituto FIRC di Oncologia Molecolare, Milan, Italy.

* These authors contributed equally to this study and are shared first authors

^ Corresponding author: l.burke@uq.edu.au

Key words

Breast cancer; *BRCA1; BRCA2*; promoter; transcription; variants of unknown clinical significance (VUS)

Abstract

The widespread use of next generation sequencing for clinical testing is detecting an escalating number of variants in non-coding regions of the genome. The clinical significance of the majority of these variants is currently unknown, which presents a significant clinical challenge. We have screened over 6000 early-onset and/or familial breast cancer cases collected by the ENIGMA consortium for sequence variants in the 5' non-coding regions of breast cancer susceptibility genes *BRCA1* and *BRCA2,* and identified 141 rare variants with global minor allele frequency <0.01, 76 of which have not been reported previously. Bioinformatic analysis identified a set of 21 variants most likely to impact transcriptional regulation, and luciferase reporter assays detected altered promoter activity for four of these variants. Electrophoretic mobility shift assays demonstrated that three of these altered the binding of proteins to the respective *BRCA1* or *BRCA2* promoter regions, including NFYA binding to *BRCA1*:c.-287C>T and PAX5 binding to *BRCA2*:c.-296C>T. Clinical classification of variants affecting promoter activity, using existing prediction models, found no evidence to suggest that these variants confer a high risk of disease. Further studies are required to determine if such variation may be associated with a moderate or low risk of breast cancer.

Introduction

Genetic susceptibility to breast cancer (BC) is complex. Multiple germline variants have been identified over the past 25 years that are broadly categorized as high, moderate and low risk. High risk variants are generally rare, have a major deleterious effect on gene function, are sufficient to confer a high risk of disease and are highly penetrant within a family. Nonsense, splicing, large deletions and some missense changes in *BRCA1* and *BRCA2* fall into this category (reviewed in (Walsh et al., 2006). There is also evidence that some alleles confer a moderate risk of cancer. These can include hypomorphic variants in known "high-risk" cancer syndrome genes (Shimelis et al., 2017; Spurdle et al., 2012), or clear loss-of-function alleles in other genes such as *CHEK2*, *PALB2* and *ATM* (Couch et al., 2017). Low risk variants, largely identified by genome-wide association studies (GWAS), are usually common and cause subtle functional effects, such as small but significant changes in gene expression due to altered activity of proximal and distal regulatory elements (reviewed in (Bogdanova, Helbig, & Dork, 2013; Ghoussaini, Pharoah, & Easton, 2013; Skol, Sasaki, & Onel, 2016). Evidence suggests that combinations of low, moderate and high-risk variants could confer a clinically significant risk of disease (Ding et al., 2012; Kuchenbaecker et al., 2017; Sawyer et al., 2012). Identification and evaluation of all such variants is therefore crucial for accurately predicting BC risk.

Use of next generation sequence analysis for germline clinical testing of cancer cases is identifying an increasing number of variants in non-coding regions of cancer susceptibility genes, including promoters, untranslated regions (UTRs), and introns. There are currently no firm recommendations for assessing the relevance of non-coding region variants to clinical testing of Mendelian disease genes, and so the vast majority of such variants are deemed of

uncertain clinical significance. This adds to the clinical challenge presented by variants of uncertain significance (VUSs), namely that they complicate test reporting and genetic counseling, limit patient eligibility for intensive surveillance and gene-targeted therapies and prevent gene testing and guided management of relatives (reviewed in (Amendola et al., 2015; Eccles et al., 2013; Plon et al., 2011). It is therefore essential that the functional and clinical significance of variants mapping to non-coding regions of the genome is determined.

Gene expression is controlled at many levels with key regulatory elements being housed in non-coding regions of the genome such as gene promoters, introns, long-range elements and 5' and 3' untranslated regions. The *BRCA1* gene is regulated at the transcriptional and post-transcriptional level, with functional proximal and distal regulatory elements being described in the promoter, introns and UTRs, by us and others (Brewster et al., 2012; Brown et al., 2002; Santana dos Santos et al., 2017; Saunus et al., 2008; Tan-Wong, French, Proudfoot, & Brown, 2008; Wardrop, Brown, & kConFab, 2005; Wiedemeyer, Beach, & Karlan, 2014). Although less studied, the *BRCA2* promoter has also been mapped and characterized (reviewed in (Wiedemeyer et al., 2014).

Common and rare variation in regulatory elements upstream of genes has been shown to alter gene expression and be associated with disease risk (reviewed in (Betts, French, Brown, & Edwards, 2013; Diederichs et al., 2016; Millot et al., 2012). We and others have described germline cancer-associated variants in the regulatory regions, including large deletions in the *BRCA1* promoter (Brown et al., 2002), and single nucleotide variants in the promoter and/or 5'UTR of *BRCA1* and *BRCA2* (Evans et al., 2018; Santana dos Santos et al., 2017), *MLH1* promoter (Hitchins et al., 2011), *POLG* promoter (Popanda et al., 2013), *PTEN* promoter (Heikkinen et al., 2011), *TERT* promoter (Horn et al., 2013), *KLHDC7A* and *PIDD1*

promoters (Michailidou et al., 2017), *BRCA1* 3'UTR (Brewster et al., 2012) and BC associated SNPs in long-range enhancers of *CCND1* (French et al., 2013).

Cancer risk-associated variants within regulatory regions are anticipated to mediate an effect on trans-acting regulatory factors (e.g. transcription factors (TF) and miRNAs), by disrupting binding of regulatory factors and interactions between regulatory elements such as promoter-enhancer interactions. For example, a variant in a *Cyclin D1* transcriptional enhancer has been associated with altered binding of the ELK4 transcription factor (French et al., 2013) and a variant within the *BRCA1* 3'UTR has been shown to introduce a functional mir-103 binding site (Brewster et al., 2012). In addition, a dominantly inherited 5' UTR *BRCA1* variant was recently shown to be associated with *BRCA1* promoter hypermethylation, known to impact TF binding, and associated allelic loss of *BRCA1* expression in two families affected by breast and ovarian cancer (Evans et al., 2018).

In this paper, we describe 141 germline variants in the *BRCA1* and *BRCA2* promoter, identified by members of the ENIGMA consortium in early onset or familial BC patients with no known pathogenic variants in the coding region of these genes. Using a combination of bioinformatic and experimental analyses, we have prioritized and analyzed a subset of variants that are most likely to affect the regulation of *BRCA1* and *BRCA2* and thus have the most potential to contribute to BC risk. Transcription factor binding site affinity changes resulting from these variants were subsequently analyzed by information theory-based analyses. In parallel, we have assessed if these variants exhibited the features expected for a high-risk pathogenic *BRCA1* or *BRCA2* variant, on the basis of available clinical and population data.

Materials and methods

Study design

An overview of the study design is shown in Figure 1. Collection of variants at all sites enabled an initial catalogue of variants from which variants were prioritized for functional analysis. Additional screening was carried out at three sites, Maastricht (M), Santiago (S), Prague (Pr), that included additional patients (M, S, Pr) and controls (Pr) that expanded the list of variants (Pr), the number of patients (M, S, Pr) and included control subjects (Pr).

Clinical and control samples

Clinical and genetic data were collected and analyzed in accordance with local human ethics guidelines of the institutions contributing to this study. All participating individuals provided informed consent for their data to be used for research purposes. An overview of the samples analyzed is shown in Table 1. Clinical samples were collected from nine European sites and were originally selected for *BRCA1* and *BRCA2* testing using ascertainment criteria that included family history and young age of BC diagnosis. Female patients who did not carry a pathogenic variant in *BRCA1* or *BRCA2* coding regions or splice junctions were selected for testing of variation in the *BRCA1* and *BRCA2* 5' regions. The controls were: 661 healthy female individuals recruited through the Immunohematology and Transfusion Medicine Service of INT and Associazione Volontari Italiani Sangue (AVIS) of Milan; 312 healthy females above 60 years of age and with no malignancy in the first filial generation recruited through First Faculty of Medicine, Charles University in Prague (Lhota et al., 2016; Soukupova, Zemankova, Kleiblova, Janatova, & Kleibl, 2016); and 130 healthy females without cancer diagnosis recruited in Santiago de Compostela.

Identification of variants

Regions containing the *BRCA1* and *BRCA2* promoter and 5'UTR were sequenced using a range of standard DNA sequencing technologies, and bioinformatic filtering pipelines. Variants mapping to the 2400bp region (hg19; chr17:41,278,514 – 41,276,114) of *BRCA1* and the 2000bp region (hg19; chr13: 32,888,597-32,890,597) of *BRCA2* were considered for further analysis. The identified variants in *BRCA1* and *BRCA2* 5' non-coding regions are numbered whereby the first translated nucleotide of the translation initiation codon is +1 (http://varnomen.hgvs.org/) using the Mutalyzer website (https://mutalyzer.nl/). *BRCA1* is described using NC_000017.10 (hg19 genomic sequence) and NM_007294.3 (transcript). *BRCA2* is described using NC_000013.10 (hg19 genomic sequence) and NM_000059.3 (transcript).

Bioinformatic analysis of variants

As an initial screen, each variant submitted for study was assessed for population frequency using intersection of the variants with dbSNP (version 138 or 150, as the study progressed) within the UCSC Genome browser and Variant Effect Predictor at ENSEMBL (http://www.ensembl.org/info/docs/tools/vep/index.html). Variants with a global minor allele frequency (MAF) of <0.01 were included in subsequent bioinformatic analyses. Further details of bioinformatics analyses to map active regulatory elements and prioritize variants for functional assays are contained in *Supplementary Methods*. Variants were considered to be high priority for experimental analysis if they contained all of the following features: 1) resided in DNaseI or FAIRE peaks, 2) coincided with high scores for DNaseI (Base Overlap Signal >40) or FAIRE (Base Overlap Signal >10) in a breast cell line, 3) resided in a region of breast cell specific TF binding, 4) overlapped with a TF consensus motif and 5) were

within an evolutionarily conserved element with a high Phastcons score (>0.75). Medium priority variants lacked one or two of these features, whereas low priority variants had only one or none of these features.

In silico transcription factor binding analysis

All rare variants were analyzed *in silico* using an information-theory (IT) based method (Caminsky et al., 2016; Mucaki et al., 2016) and a modified version of the Shannon pipeline utilizing TF information models built from ENCODE ChIP-seq data sets (Lu, Mucaki, & Rogan, 2017) to assess potential effects of variants on TF binding. Details of analyses are contained in *Supplementary Methods*.

Experimental analysis of variants

Promoter reporter assays

The 499bp *BRCA1* (chr17:41,277,787-41,277,289) and 750bp *BRCA2* (chr13:32,889,230-32,889,979) promoter regions were cloned into pCR-Blunt vector (Life Technologies). Site-directed mutagenesis was used to introduce variants using the primers listed in *Supplementary Table 1*. Plasmids were purified using the QIAprep miniprep kit (Qiagen) as per the manufacturer's instructions. Plasmid preparations were validated using restriction digest and DNA sequencing and inserts were shuttled into pGL3-Basic luciferase reporter vector (Promega). All plasmids for transfection were analyzed for DNA conformation on a 1% w/v agarose gel and only plasmids possessing a supercoiled conformation were used for transfections. Transfection details are described in *Supplementary methods*.

The luciferase based reporter assay was performed as described previously (Brewster et al., 2012). Positive controls were B1-Ets, *BRCA1*:c.-330_-329delinsTT, that decreases *BRCA1* promoter activity in MCF7 cells (Atlas, Stramwasser, Whiskin, & Mueller, 2000) and B2-Ets

(E2Fmut1: *BRCA2*:c.-282_-281delinsAA), that has been shown to decrease *BRCA2* promoter activity in MCF7 cells (Davis, Miron, Andersen, Iglehart, & Marks, 1999). Statistical analyses were performed in GraphPad Prism using one -way ANOVA followed by Tukey's post hoc test and values $p<0.05$ were deemed statistically significant.

Electrophoretic mobility shift assay (EMSA)

Nuclear proteins were extracted as described in *Supplementary methods* and EMSAs were carried out using a Pierce LightShift Chemiluminescent EMSA Kit (Thermofisher, USA) with modifications described in *Supplementary methods*. For competition and supershift studies, nuclear extracts were initially incubated with unlabeled double stranded (ds) competitor probes or antibodies in binding buffer before addition of the biotinylated probe and incubation at room temperature. Positive controls for *BRCA1* and *BRCA2* DNA binding were sequences surrounding the B1-Ets and B2-Ets mutations described above.

Qualitative and quantitative classification of variants

Variants were classified according to the ENIGMA classification criteria for variation in *BRCA1* and *BRCA2* (https://enigmaconsortium.org/) to determine whether any of the prioritized variants were associated with a high risk of disease. See *Supplementary methods* for further details.

Results

Identification and prioritization of sequence variants in *BRCA1* and *BRCA2* 5′ non-coding regions.

The 5' non-coding region of *BRCA1* and *BRCA2* in early onset or familial BC patients with no known *BRCA1* or *BRCA2* germline pathogenic variant were sequenced at nine different sites as part of an approved ENIGMA (https://enigmaconsortium.org/) project. For the

BRCA1 5' region, 6475 patients were sequenced at 8 different sites along with 1103 controls. For the *BRCA2* 5' region, 6603 patients were sequenced at 8 different sites as well as 442 controls.

After excluding variants with global MAF >0.01 at time of variant identification, a total of 141 unique single nucleotide variants and short insertions/deletions were identified, 81 in *BRCA1* and 60 in *BRCA2* (Supp. Tables 2 and 3). To evaluate the potential of these rare variants to impact gene regulation, we initially undertook a comprehensive bioinformatic analysis. Promoter regions of *BRCA1* and *BRCA2* were defined by bioinformatic predictors including chromatin marks (Figure 2). These regions show the characteristic histone H3 epigenetic marks, including H3K4me3, H3K27ac and H3K9ac, as well as occupancy by multiple TFs. Of the variants identified in cases only, 22 *BRCA1* and 23 *BRCA2* variants resided within the minimal promoter regions.

To predict the potential impact of variants on promoter activity, we prioritized variants using breast cell specific data for chromatin accessibility and TF occupancy along with evolutionary conservation. Due to the limited breast cell specific TF ChIP-seq data, we also included ENCODE TF ChIP-seq and TF consensus motif data from all cell lines. A total of 9 *BRCA1* and 12 *BRCA2* variants were selected for further functional analysis (Figure 2, Table 2, 3).

BRCA1 and *BRCA2* promoter activity is altered by 5' non-coding sequence variants

To examine the potential effect of the 21 prioritized *BRCA1* and *BRCA2* 5' non-coding variants on regulatory activity, promoter activity was measured using luciferase assays in MCF7 and MDA-MB-468 BC cell-lines. Two of the nine prioritized *BRCA1* variants decreased *BRCA1* promoter activity relative to the WT construct (Figures 3A, B). *BRCA1*:c.-

315del significantly decreased the *BRCA1* promoter luciferase activity in both cell lines, while *BRCA1*:c.-192C decreased luciferase activity in the MCF7 cell line. Furthermore, one variant, *BRCA1*:c.-287T, displayed increased activity relative to the WT construct in the MCF7 cell line. For *BRCA2*, one of the 12 variants, *BRCA2*:c.-296T, decreased *BRCA2* promoter activity relative to the WT construct in the MCF7 cell line (Figures 3C, D).

In silico analyses of *BRCA1* and *BRCA2* 5' variants predict alterations in TF binding

BRCA1 and *BRCA2* promoters are regulated by a complex array of DNA-binding proteins and transcriptional coactivators and co-repressors (reviewed in (McCoy, Mueller, & Roskelley, 2003; Mueller & Roskelley, 2003; Wiedemeyer et al., 2014). *In silico* analysis was carried out to examine whether the *BRCA1* and *BRCA2* promoter variants shown to alter luciferase activity (see above) are likely to affect binding of trans-acting protein factors in breast cells.

Interrogation of ENCODE ChIP-seq datasets derived from breast cell lines show that, although the number of datasets is limited, TFs bind to regions encompassing the prioritized variants (Figure 2 and Supp. Figure 1). ENCODE ChIP-seq data from other cell lines indicate that some variants are located within consensus motifs for specific TFs associated with these regions (Table 2, 3; Supp. Figure 1). *BRCA1:c.*-287C>T overlaps with the consensus binding motif for CCAAT Box binding factors and *BRCA2:c.*-296C>T is located within the consensus motif for PAX5.

IT analysis of the prioritized variants showed that the binding strengths of several TFs are predicted to be altered by the *BRCA1* and *BRCA2* variants (Table 4. and Supp. Table 4). All of the variants that altered promoter activity were predicted to have consequences on TF binding. *BRCA1*:c.-287C>T and *BRCA2:*c.-296C>T are predicted to disrupt binding of

CCAAT Box binding factors and PAX5, respectively. *BRCA1*:c.-315del is predicted to disrupt the binding of TCF7L2 but creates a POU2F2 (also known as Oct-2) binding site. *BRCA1*:c.-192T>C is predicted to strengthen a RFX5 site and creates an ETS1 site.

5′ variants in *BRCA1* and *BRCA2* alter protein-DNA interactions in EMSA analyses

To examine potential alterations in the binding of nuclear proteins from breast cells by the *BRCA1* and *BRCA2* promoter variants that altered luciferase activity, we carried out EMSA analysis. For *BRCA1*, two of three analyzed variants; c.-315del and c.-287C>T displayed allele-specific protein binding (Figure 4). For probes containing the region surrounding the *BRCA1*:c.-315del variant, changing the WT sequence to the variant sequence, resulted in the enhanced binding of a slower migrating band (Figure 4A,B). For probes containing the region surrounding the *BRCA1*:c.-287C>T variant, introduction of the variant sequence resulted in almost complete loss of protein binding to the probe (Figure 4A).

To determine if the DNA-protein interactions were specific, competition experiments were performed. In the case of *BRCA1*:c.-315del all bands were competed by both the WT and the variant containing probes in two cell lines (Figure 5A,B). For *BRCA1*:c.-287C>T, only the WT probe was able to compete for binding (Figure 5C). The non-specific probe from an unrelated region of the *BRCA1* promoter did not compete any bands showing that the bands seen in the EMSA were specific.

Analysis of the regions of the *BRCA2* promoter using EMSA revealed that region containing the *BRCA2*:c.-296C>T variant bound nuclear proteins from MCF7 nuclear extracts and that this interaction was dramatically reduced by introduction of the variant sequence (Figure 6A). Competition experiments showed that these interactions were specific and not competed by a non-specific probe from an unrelated region of the *BRCA1* promoter (Figure 6A).

To determine the effect of these variants on the binding of specific TFs, competition and supershift analyses were performed. *BRCA1*:c.-287C>T overlaps with the consensus binding motif for CCAAT Box binding factors, NFYA and NFYB (Table 2, Supp. Figure. 1A) and IT analysis predicts that the variant disrupts binding of these TFs (Table 4). Consistent with these predictions, supershift experiments show that *BRCA1*:c.-287C>T disrupts binding of NFYA to this region (Figure. 5D). In addition, we analysed *BRCA2*:c.-296C>T, which maps within the consensus binding motif for PAX5 (Table 2, Supp. Figure 1B) and is predicted by IT analysis to disrupt binding of PAX5 (Table 4), by cross-competition experiments using known PAX5 binding sites from *hCD19* (Kozmik, Wang, Dorfler, Adams, & Busslinger, 1992) and *hDAO* (Tran et al., 2015) genes. These experiments show that known PAX5 binding sites compete efficiently for binding of nuclear proteins to the *BRCA2* promoter region, indicating that PAX5 binding is reduced as a consequence of the nucleotide sequence change (Figure. 6B). In contrast, supershift experiments for POU2F2 (Oct-2) showed no evidence for *BRCA1*:c.-315del causing a change in binding of POU2F2 in the cell-line used (data not shown).

Clinical classification of *BRCA1* and *BRCA2* 5' non-coding sequence variants

Variants were classified according to the ENIGMA guidelines, which are calibrated for classification of variants as high risk, using available population frequency and/or clinical data (Supp. Tables 5 and 6). In this context the term pathogenicity refers to a variant that confers a high risk of disease. Importantly, these classification guidelines do not identify those variants that confer a moderate or low risk of disease.

Of those variants identified in cases only, 26/70 (37%) of BRCA1 variants had been reported in dbSNP at study initiation (max global frequency 0.006) (Supp. Table 2), and 22/54 (41%)

of *BRCA2* variants observed in cases only were identified in dbSNP (max global frequency 0.006) (Supp. Table 3). Review of variant frequency in public reference groups identified 21 variants that were classifiable, as Not Pathogenic, based on frequency in control groups (Supp. Table 5): six *BRCA1* and five *BRCA2* variants were observed at >1% frequency in population subgroups (stand-alone evidence against pathogenicity, when detected in a non-founder outbred population group); six *BRCA1* and four *BRCA2* variants occurred at frequency 0.001-0.01 (range 0.0014 – 0.0076) in at least five individuals in the reference set, which combined with a low assumed prior is considered sufficient as evidence against pathogenicity (Supp. Table 5). Frequency data from controls screened for this study also supported the frequency-based classifications for 8 of these 21 variants (Supp. Table 5).

Segregation analysis for 7 informative families aided classification for 6 variants, while histopathology LRs derived for 24 tumors altered classification for 10 variants (Supp. Table 6). Combining findings from qualitative and quantitative methods, most variants (113/141; 80%) remained Class 3 Uncertain, largely due to a lack of data.

A total of 27/141 (19%) variants were classified as Not Pathogenic or Likely Not Pathogenic. Of the 21 variants prioritized for functional analysis, eight variants (38%) were classified as Not Pathogenic or Likely Not Pathogenic based on frequency information and/or multifactorial analysis (Table 5), including two variants (*BRCA1:c.*-192T>C and *BRCA2:c.*-296 C>T) that were shown to decrease promoter activity and in the case of *BRCA2:c.*-296 C>T also resulted in perturbed TF binding. Taken together this analysis indicates that none of the variants shown to affect function in this study are associated with a high risk of disease. This analysis is silent, however, on whether these variants may confer a moderate or low risk of disease.

Discussion

Next generation sequencing and gene panel testing enable rapid analysis of gene regions that have previously not been included in standard screening procedures, including promoters, UTRs, introns and extragenic regions. It is hypothesized that variants in these regions have potential to modulate gene expression (Stranger et al., 2005; Stranger et al., 2007) and impact on relative disease risk, possibly in collaboration with multiple other low, moderate and high-risk variants (Manolio et al., 2009). This extends and validates our previous study (Santana dos Santos et al., 2017) by using a larger number patients analysed over nine geographical locations, identifying additional breast cancer associated variants, and showing that a subset of these variants modulate binding of specific transcription factors. Further, we have compared results from our bioinformatics and functional analysis to variant classifications based on ENIGMA *BRCA1/2* guidelines for high-risk variation in these genes.

Through targeted sequencing of over 6000 early onset/familial BC patients, we identified 141 single nucleotide variants and small indels mapping to the 5' non-coding regions of BRCA1 and BRCA2. Of these, four: *BRCA1*:c.-315del, *BRCA1*:c.-287C>T, *BRCA1*:c.-192T>C and *BRCA2*:c.-296C>T caused a significant change in promoter activity. The observed alterations in *BRCA1* and *BRCA2* promoter activity are of a similar magnitude to that seen with other germline variants associated with BC risk (Michailidou et al., 2017), including a variant in the *TERT* promoter, which creates a new binding site for Ets factors and results in a 1.2-1.5 fold increase in luciferase activity in a promoter reporter assay (Horn et al., 2013), and variants in the promoters of *KLHDC7A* and *PIDD1* (Michailidou et al., 2017). Whilst this supports the hypothesis that moderate change in promoter activity can be associated with disease risk, further work is needed to confirm this.

One of the four variants significantly altered luciferase activity in both tested cell lines, whereas the remaining three variants only affected luciferase activity in MCF7 cells. This may reflect the differential availability of crucial TFs in MDA-MB-468 cells (Kao et al., 2009) and highlights the importance of undertaking that assays for functional activity of variants in more than one cell line. Three variants, *BRCA1*:c.-380G>A, *BRCA2*:c.-296C>T and *BRCA2*:c.-218G>A, were also analyzed in our earlier paper (Santana dos Santos et al., 2017). Although the cell-lines used in the two studies were different (MDA-MB-231 in (Santana dos Santos et al., 2017) and MCF7 and MDA-MB-468 here), the trends are the same in five out of six analyses. The difference for *BRCA2*:c.-296C>T, which causes a significant decrease in MDA-MB-231 and MCF7 cells, but not MDA-MB-468 cells, may again be indicative of differential gene expression in breast cancer cell lines (Kao et al., 2009). Overall however, the consistency of results performed in two separate laboratories underscore the robustness of the assay system.

Some variants were associated with a decrease in promoter activity, whilst others were associated with an increase. As transcription factors can function as activators or repressors, a variant associated change in transcription factor binding can result in either a decrease or an increase in promoter (or other regulatory element) activity. Differences in the quanta and direction of promoter activity have been reported previously eg (Fraile-Bethencourt et al., 2018; Santana dos Santos et al., 2017) and have also been shown to differ between cell-lines potentially reflecting the availability of transcription factors or co-factors (eg. Zn).

Three of the variants, *BRCA1*:c.-315del, *BRCA1*: c.-287C>T and *BRCA2*:c.-296C>T, altered protein binding. ENCODE ChIP-seq data from breast cancer cell-lines indicate candidate proteins that are bound to the genomic regions containing these variants (Figure 2 and Supp. Figure 1). These include E2F1, CEBPB, GATA3, Max, ELF1, GABP and FOXA1 for

BRCA1 and E2F1, MYC, ELF1, GABP, Max and PML for *BRCA2*. Interestingly, a number of these factors have previously been implicated in breast cancer.

In addition, ENCODE ChIP-seq data from cell lines derived from tissues other than breast indicate that the variants that affect protein binding are located within consensus motifs for specific TFs associated with these regions (Tables 2 and 3; Supp. Figure. 1). *BRCA1*:c.-287C>T overlaps with the consensus binding motif for CCAAT Box binding factors, *BRCA1*:c.-315del is located in a consensus motif for CREB/ATF proteins although the deletion does not modify this motif and *BRCA2*:c.-296C>T is located within the consensus motif for PAX5. IT analysis also predicts that all these variants alter TF binding (Table 4. and Supp. Table 4). We show that *BRCA1*:c.-287C>T disrupts the binding of NFYA to the *BRCA1* promoter region. Furthermore, we present evidence that *BRCA2*:c.-296C>T disrupts the binding of PAX5. *BRCA1*:c.-315del lies in the so called positive regulator region that has been shown to bind GABPα, CREB and AP-1 proteins (Atlas, Stramwasser, & Mueller, 2001; Atlas et al., 2000; Graves, Zhou, MacDonald, Mueller, & Roskelley, 2007; Suen & Goss, 1999; Thakur & Croce, 1999). While these proteins are generally considered activators of transcription, repression of promoter activity by *BRCA1*:c.-315del suggests the recruitment of an additional transcriptional repressor or co-repressor to this region. IT analysis predicts creation of a binding site for POU2F2, a known repressor, however we found no evidence to suggest that this variant increased POU2F2 binding in the cell-line used, although it is possible that changes may be observable in other cell-lines. Biochemical studies, including mass spectrometry, will be required to validate and discover other alterations in TF binding.

One variant, *BRCA1*:c.-287C>T, increased promoter activity and decreased protein:DNA interactions. This increase in promoter activity was unanticipated since this variant is within a consensus motif for the CCAAT box binding proteins, NFYA and NFYB, and mutation of

this CCAAT box has previously been shown to reduce *BRCA1* promoter activity in MCF7 cells (Bindra et al., 2005; Xu, Chambers, & Solomon, 1997). This variant also decreases promoter activity in MDA-MB-231 cells (Santana dos Santos et al., 2017). Here we show that the *BRCA1*:c.-287C>T variant reduces NFYA binding. Importantly, NFY proteins can function as transcriptional activators or repressors depending on recruitment of co-repressors or co-activators (Peng & Jahroudi, 2002; Peng et al., 2007) and recruitment of TFs to neighbouring sequences (Zhu et al., 2012) indicating possible mechanisms for divergent activities of NFY proteins at this site.

BRCA1:c.-192T>C, which lies in the 5'UTR, decreased reporter activity but did not bind any proteins from MCF7 nuclear extracts in EMSA analysis. Possibly, EMSA binding conditions are not optimal for binding of factors to this sequence or alternatively, this reduction in promoter activity could be by post-transcriptional mechanisms as seen for *BRCA2*:c.-26G>A (Gochhait et al., 2007).

Using existing prediction models developed for high risk variants, population frequency and clinical information classified 27 variants as 'Not Pathogenic' or 'likely Not Pathogenic'. This included two *BRCA1* and six *BRCA2* variants with functional assay data available, six with no statistically significant effect on promoter activity, and two which decreased promoter activity *in vitro*. These two variants, *BRCA1:c.*-192T>C and *BRCA2:c.*-296C>T, were observed in population subgroup controls; notably *BRCA1:c.*-192T>C was observed at a frequency of >1%, which is considered *stand-alone* evidence against pathogenicity (defined as high risk of cancer) for *BRCA1/2* variation. This suggests that promoter region variants, irrespective of bioinformatic prediction or functional assay results, are unlikely to be associated with a high risk of cancer. This is consistent with current evidence from ENIGMA studies (de la Hoya et al., 2016) suggest that an allele resulting in only ~20-30% expression

of *BRCA1* transcript/s encoding functional transcripts is not associated with high risk of breast cancer. The low impact of these variants on risk is likely to reflect the complex interplay of TFs and DNA elements, and possible redundancy in the system. For example, a variant in one TF binding site within a cluster may be buffered by other binding sites and thus insufficient on its own to reduce gene expression markedly (Lu and Rogan, biorxiv preprint https://doi.org/10.1101/283267).

Given that moderate and low risk variants often occur in >1% of the population, and that the remaining 13 variants had insufficient evidence available to assess clinical significance, we cannot exclude the possibility that *BRCA1/2* promoter region variants, in particular those with proven functional effect, may be associated with a moderate or low risk of cancer. This indicates an urgent need to further develop prediction models to accommodate criteria for moderate or low risk variants by extending the *BRCA1/2*-specific criteria developed by ENIGMA (http://www.enigmaconsortium.org/), or even the generic variant classification criteria developed by the American College of Medical Genetics for Mendelian disorders (Richards et al., 2015).

This study has evaluated the significance of single nucleotide variants and small indels mapping to the 5' non-coding region of *BRCA1* and *BRCA2* using bioinformatic, biological and biochemical analyses in combination with consideration of clinical data that informs qualitative and quantitative variant classification. We present data to suggest that a subset of these variants have functional effects on gene regulation. We also present evidence that variants mapping to and affecting the function of *BRCA* promoters, are not likely to be associated with a high risk of cancer. We propose that studies of differing design, such as very large-scale case-control sequencing studies able to detect rare variation, will be required to address if a low to moderate risk of cancer may be associated with *BRCA1/2* regulatory

region variation that has not been captured to date by genome-wide association genotyping platforms. We believe that the bioinformatic and functional analysis presented will be important to define the design and interpretation of such future sequencing studies. We also believe that this study highlights the challenges associated with classifying variants with respect to low or moderate disease risk, and the need to be cautious in the clinical use of information on individual variants that is likely to be one of many factors contributing to disease risk.

Acknowledgements

The authors would like to acknowledge all the patients that were involved in this study. This work was supported by grants from the National Health and Medical Research Council (ID1104808) and Cancer Council Queensland (ID1044008 and ID1026095) to M.A.B. A.B.S. is supported by an NHMRC Senior Research Fellowship (ID1061779). S.L.E. is supported by an NHMRC Senior Research Fellowship (ID1135932). This work was supported by the grants from the National Cancer Institute (INCa: INCA-DGOS_8706 to S.M.C.), the Ministry of Health of the Czech Republic (AZV 16-33444A: J.Sevcik, J.S., M.J., P.Z., K.L., L.S., M.B), Fondazione Pisa (G. G. and M. C. #2016), the Spanish Instituto de Salud Carlos III (ISCIII) funding (to O.D and S.O.G), an initiative of the Spanish Ministry of Economy and Innovation and partially supported by European Regional Development FEDER Funds: FIS PI12/02585 and PI15/00355 (to O.D.) and PI13/01711 and PI16/01218 (to S.G-E.). S.G-E. is supported by the Miguel Servet Progam (CP10/00617). Partial funding also came from a CIBERER grant (ER17P1AC7112/2017) and Fundación Mutua Madrileña to A.V. and a Ghent University Special Research Fund (BOF15/GOA/011) to K.B.M.C. P.K.R. is supported by the Canadian Breast Cancer Foundation, Canadian Foundation for Innovation,

Canada Research Chairs Secretariat, and the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant RGPIN-2015-06290). T.vO.H and M.R. were supported by The Research Council of The Capital Region of Denmark (Grant E-22283-02). P.P. and P.R.were supported by Investigator Grants (#4017 to P.P and #15547 to P.R.) from the Italian Association for Cancer Research (AIRC).

Disclosure statement

B.C.S is an employee of and P.K.R is co-founder of CytoGnomix, which has developed algorithms and software for interpretation of variants within transcription factor binding sites.

References

Amendola, L. M., Dorschner, M. O., Robertson, P. D., Salama, J. S., Hart, R., Shirts, B. H., . . . Jarvik, G. P. (2015). Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome Res, 25*(3), 305-315. doi:10.1101/gr.183483.114

Atlas, E., Stramwasser, M., & Mueller, C. R. (2001). A CREB site in the BRCA1 proximal promoter acts as a constitutive transcriptional element. *Oncogene, 20*(48), 7110-7114. doi:10.1038/sj.onc.1204890

Atlas, E., Stramwasser, M., Whiskin, K., & Mueller, C. R. (2000). GA-binding protein alpha/beta is a critical regulator of the BRCA1 promoter. *Oncogene, 19*(15), 1933-1940. doi:10.1038/sj.onc.1203516

Betts, J. A., French, J. D., Brown, M. A., & Edwards, S. L. (2013). Long-range transcriptional regulation of breast cancer genes. *Genes Chromosomes Cancer, 52*(2), 113-125. doi:10.1002/gcc.22020

Bindra, R. S., Gibson, S. L., Meng, A., Westermark, U., Jasin, M., Pierce, A. J., . . . Glazer, P. M. (2005). Hypoxia-induced down-regulation of BRCA1 expression by E2Fs. *Cancer Res, 65*(24), 11597-11604. doi:10.1158/0008-5472.CAN-05-2119

Bogdanova, N., Helbig, S., & Dork, T. (2013). Hereditary breast cancer: ever more pieces to the polygenic puzzle. *Hered Cancer Clin Pract, 11*(1), 12. doi:10.1186/1897-4287-11-12

Brewster, B. L., Rossiello, F., French, J. D., Edwards, S. L., Wong, M., Wronski, A., . . . Peterlongo, P. (2012). Identification of fifteen novel germline variants in the BRCA1 3'UTR reveals a variant in a breast cancer case that introduces a functional miR-103 target site. *Hum Mutat, 33*(12), 1665-1675. doi:10.1002/humu.22159

Brown, M. A., Lo, L. J., Catteau, A., Xu, C. F., Lindeman, G. J., Hodgson, S., & Solomon, E. (2002). Germline BRCA1 promoter deletions in UK and Australian familial breast cancer patients:

Identification of a novel deletion consistent with BRCA1:psiBRCA1 recombination. *Hum Mutat, 19*(4), 435-442. doi:10.1002/humu.10055

Caminsky, N. G., Mucaki, E. J., Perri, A. M., Lu, R., Knoll, J. H., & Rogan, P. K. (2016). Prioritizing Variants in Complete Hereditary Breast and Ovarian Cancer Genes in Patients Lacking Known BRCA Mutations. *Hum Mutat, 37*(7), 640-652. doi:10.1002/humu.22972

Couch, F. J., Shimelis, H., Hu, C., Hart, S. N., Polley, E. C., Na, J., . . . Dolinsky, J. S. (2017). Associations Between Cancer Predisposition Testing Panel Genes and Breast Cancer. *JAMA Oncol, 3*(9), 1190-1196. doi:10.1001/jamaoncol.2017.0424

Davis, P. L., Miron, A., Andersen, L. M., Iglehart, J. D., & Marks, J. R. (1999). Isolation and initial characterization of the BRCA2 promoter. *Oncogene, 18*(44), 6000-6012. doi:10.1038/sj.onc.1202990

de la Hoya, M., Soukarieh, O., Lopez-Perolio, I., Vega, A., Walker, L. C., van Ierland, Y., . . . Spurdle, A. B. (2016). Combined genetic and splicing analysis of BRCA1 c.[594-2A>C; 641A>G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms. *Hum Mol Genet, 25*(11), 2256-2268. doi:10.1093/hmg/ddw094

Diederichs, S., Bartsch, L., Berkmann, J. C., Frose, K., Heitmann, J., Hoppe, C., . . . Wullenkord, R. (2016). The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. *EMBO Mol Med, 8*(5), 442-457. doi:10.15252/emmm.201506055

Ding, Y. C., McGuffog, L., Healey, S., Friedman, E., Laitman, Y., Paluch-Shimon, S., . . . Consortium of Investigators of Modifiers of, B. (2012). A nonsynonymous polymorphism in IRS1 modifies risk of developing breast and ovarian cancers in BRCA1 and ovarian cancer in BRCA2 mutation carriers. *Cancer Epidemiol Biomarkers Prev, 21*(8), 1362-1370. doi:10.1158/1055-9965.EPI-12-0229

Eccles, S. A., Aboagye, E. O., Ali, S., Anderson, A. S., Armes, J., Berditchevski, F., . . . Thompson, A. M. (2013). Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer. *Breast Cancer Res, 15*(5), R92. doi:10.1186/bcr3493

Evans, D. G. R., van Veen, E. M., Byers, H. J., Wallace, A. J., Ellingford, J. M., Beaman, G., . . . Newman, W. G. (2018). A Dominantly Inherited 5' UTR Variant Causing Methylation-Associated Silencing of BRCA1 as a Cause of Breast and Ovarian Cancer. *Am J Hum Genet, 103*(2), 213-220. doi:10.1016/j.ajhg.2018.07.002

Fraile-Bethencourt, E., Valenzuela-Palomo, A., Diez-Gomez, B., Infante, M., Duran, M., Marcos, G., . . . Velasco, E. A. (2018). Genetic dissection of the BRCA2 promoter and transcriptional impact of DNA variants. *Breast Cancer Res Treat*. doi:10.1007/s10549-018-4826-7

French, J. D., Ghoussaini, M., Edwards, S. L., Meyer, K. B., Michailidou, K., Ahmed, S., . . . Dunning, A. M. (2013). Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *Am J Hum Genet, 92*(4), 489-503. doi:10.1016/j.ajhg.2013.01.002

Ghoussaini, M., Pharoah, P. D. P., & Easton, D. F. (2013). Inherited genetic susceptibility to breast cancer: the beginning of the end or the end of the beginning? *Am J Pathol, 183*(4), 1038-1051. doi:10.1016/j.ajpath.2013.07.003

Gochhait, S., Bukhari, S. I., Bairwa, N., Vadhera, S., Darvishi, K., Raish, M., . . . Bamezai, R. N. (2007). Implication of BRCA2 -26G>A 5' untranslated region polymorphism in susceptibility to sporadic breast cancer and its modulation by p53 codon 72 Arg>Pro polymorphism. *Breast Cancer Res, 9*(5), R71. doi:10.1186/bcr1780

Graves, M. L., Zhou, L., MacDonald, G., Mueller, C. R., & Roskelley, C. D. (2007). Regulation of the BRCA1 promoter in ovarian surface epithelial cells and ovarian carcinoma cells. *FEBS Lett, 581*(9), 1825-1833. doi:10.1016/j.febslet.2007.03.072

Heikkinen, T., Greco, D., Pelttari, L. M., Tommiska, J., Vahteristo, P., Heikkila, P., . . . Nevanlinna, H. (2011). Variants on the promoter region of PTEN affect breast cancer progression and patient survival. *Breast Cancer Res, 13*(6), R130. doi:10.1186/bcr3076

Hitchins, M. P., Rapkins, R. W., Kwok, C. T., Srivastava, S., Wong, J. J., Khachigian, L. M., . . . Ward, R. L. (2011). Dominantly inherited constitutional epigenetic silencing of MLH1 in a cancer-affected family is linked to a single nucleotide variant within the 5'UTR. *Cancer Cell, 20*(2), 200-213. doi:10.1016/j.ccr.2011.07.003

Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., . . . Kumar, R. (2013). TERT promoter mutations in familial and sporadic melanoma. *Science, 339*(6122), 959-961. doi:10.1126/science.1230062

Kao, J., Salari, K., Bocanegra, M., Choi, Y. L., Girard, L., Gandhi, J., . . . Pollack, J. R. (2009). Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PLoS One, 4*(7), e6146. doi:10.1371/journal.pone.0006146

Kozmik, Z., Wang, S., Dorfler, P., Adams, B., & Busslinger, M. (1992). The promoter of the CD19 gene is a target for the B-cell-specific transcription factor BSAP. *Mol Cell Biol, 12*(6), 2662-2672.

Kuchenbaecker, K. B., McGuffog, L., Barrowdale, D., Lee, A., Soucy, P., Dennis, J., . . . Antoniou, A. C. (2017). Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. *J Natl Cancer Inst, 109*(7). doi:10.1093/jnci/djw302

Lhota, F., Zemankova, P., Kleiblova, P., Soukupova, J., Vocka, M., Stranecky, V., . . . Kleibl, Z. (2016). Hereditary truncating mutations of DNA repair and other genes in BRCA1/BRCA2/PALB2-negatively tested breast cancer patients. *Clin Genet, 90*(4), 324-333. doi:10.1111/cge.12748

Lu, R., Mucaki, E. J., & Rogan, P. K. (2017). Discovery and validation of information theory-based transcription factor and cofactor binding site motifs. *Nucleic Acids Res, 45*(5), e27. doi:10.1093/nar/gkw1036

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753. doi:10.1038/nature08494

McCoy, M. L., Mueller, C. R., & Roskelley, C. D. (2003). The role of the breast cancer susceptibility gene 1 (BRCA1) in sporadic epithelial ovarian cancer. *Reprod Biol Endocrinol, 1*, 72. doi:10.1186/1477-7827-1-72

Michailidou, K., Lindstrom, S., Dennis, J., Beesley, J., Hui, S., Kar, S., . . . Easton, D. F. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature, 551*(7678), 92-94. doi:10.1038/nature24284

Millot, G. A., Carvalho, M. A., Caputo, S. M., Vreeswijk, M. P., Brown, M. A., Webb, M., . . . Group, E. C. F. A. W. (2012). A guide for functional analysis of BRCA1 variants of uncertain significance. *Hum Mutat, 33*(11), 1526-1537. doi:10.1002/humu.22150

Mucaki, E. J., Caminsky, N. G., Perri, A. M., Lu, R., Laederach, A., Halvorsen, M., . . . Rogan, P. K. (2016). A unified analytic framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer. *BMC Med Genomics, 9*, 19. doi:10.1186/s12920-016-0178-5

Mueller, C. R., & Roskelley, C. D. (2003). Regulation of BRCA1 expression and its relationship to sporadic breast cancer. *Breast Cancer Res, 5*(1), 45-52. doi:10.1186/bcr557

Nair, B., Johar, K., Priya, A., & Wong-Riley, M. T. (2016). Specificity protein 4 (Sp4) transcriptionally regulates inhibitory GABAergic receptors in neurons. *Biochim Biophys Acta, 1863*(1), 1-9. doi:10.1016/j.bbamcr.2015.10.005

Peng, Y., & Jahroudi, N. (2002). The NFY transcription factor functions as a repressor and activator of the von Willebrand factor promoter. *Blood, 99*(7), 2408-2417.

Peng, Y., Stewart, D., Li, W., Hawkins, M., Kulak, S., Ballermann, B., & Jahroudi, N. (2007). Irradiation modulates association of NF-Y with histone-modifying cofactors PCAF and HDAC. *Oncogene, 26*(54), 7576-7583. doi:10.1038/sj.onc.1210565

Phillips, R. J., Tyson-Capper Nee Pollard, A. J., Bailey, J., Robson, S. C., & Europe-Finner, G. N. (2005). Regulation of expression of the chorionic gonadotropin/luteinizing hormone receptor gene in the human myometrium: involvement of specificity protein-1 (Sp1), Sp3, Sp4, Sp-like proteins, and histone deacetylases. *J Clin Endocrinol Metab, 90*(6), 3479-3490. doi:10.1210/jc.2004-1962

Plon, S. E., Cooper, H. P., Parks, B., Dhar, S. U., Kelly, P. A., Weinberg, A. D., . . . Hilsenbeck, S. (2011). Genetic testing and cancer risk management recommendations by physicians for at-risk relatives. *Genet Med, 13*(2), 148-154. doi:10.1097/GIM.0b013e318207f564

Popanda, O., Seibold, P., Nikolov, I., Oakes, C. C., Burwinkel, B., Hausmann, S., . . . Schmezer, P. (2013). Germline variants of base excision repair genes and breast cancer: A polymorphism in DNA polymerase gamma modifies gene expression and breast cancer risk. *Int J Cancer, 132*(1), 55-62. doi:10.1002/ijc.27665

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., . . . Committee, A. L. Q. A. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med, 17*(5), 405-424. doi:10.1038/gim.2015.30

Santana dos Santos, E., Caputo, S. M., Castera, L., Gendrot, M., Briaux, A., Breault, M., . . . Rouleau, E. (2017). Assessment of the functional impact of germline BRCA1/2 variants located in non-coding regions in families with breast and/or ovarian cancer predisposition. *Breast Cancer Res Treat*. doi:10.1007/s10549-017-4602-0

Saunus, J. M., French, J. D., Edwards, S. L., Beveridge, D. J., Hatchell, E. C., Wagner, S. A., . . . Brown, M. A. (2008). Posttranscriptional regulation of the breast cancer susceptibility gene BRCA1 by the RNA binding protein HuR. *Cancer Res, 68*(22), 9469-9478. doi:10.1158/0008-5472.CAN-08-1159

Sawyer, S., Mitchell, G., McKinley, J., Chenevix-Trench, G., Beesley, J., Chen, X. Q., . . . James, P. A. (2012). A role for common genomic variants in the assessment of familial breast cancer. *J Clin Oncol, 30*(35), 4330-4336. doi:10.1200/JCO.2012.41.7469

Shimelis, H., Mesman, R. L. S., Von Nicolai, C., Ehlen, A., Guidugli, L., Martin, C., . . . for, N. C. (2017). BRCA2 Hypomorphic Missense Variants Confer Moderate Risks of Breast Cancer. *Cancer Res, 77*(11), 2789-2799. doi:10.1158/0008-5472.CAN-16-2568

Skol, A. D., Sasaki, M. M., & Onel, K. (2016). The genetics of breast cancer risk in the post-genome era: thoughts on study design to move past BRCA and towards clinical relevance. *Breast Cancer Res, 18*(1), 99. doi:10.1186/s13058-016-0759-4

Soukupova, J., Zemankova, P., Kleiblova, P., Janatova, M., & Kleibl, Z. (2016). [CZECANCA: CZEch CAncer paNel for Clinical Application-- Design and Optimization of the Targeted Sequencing Panel for the Identification of Cancer Susceptibility in High-risk Individuals from the Czech Republic]. *Klin Onkol, 29 Suppl 1*, S46-54. doi:10.14735/amko2016S46

Spurdle, A. B., Whiley, P. J., Thompson, B., Feng, B., Healey, S., Brown, M. A., . . . Consortium, E. (2012). BRCA1 R1699Q variant displaying ambiguous functional abrogation confers intermediate breast and ovarian cancer risk. *J Med Genet, 49*(8), 525-532. doi:10.1136/jmedgenet-2012-101037

Stranger, B. E., Forrest, M. S., Clark, A. G., Minichiello, M. J., Deutsch, S., Lyle, R., . . . Dermitzakis, E. T. (2005). Genome-wide associations of gene expression variation in humans. *PLoS Genet, 1*(6), e78. doi:10.1371/journal.pgen.0010078

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., . . . Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nat Genet, 39*(10), 1217-1224. doi:10.1038/ng2142

Suen, T. C., & Goss, P. E. (1999). Transcription of BRCA1 is dependent on the formation of a specific protein-DNA complex on the minimal BRCA1 Bi-directional promoter. *J Biol Chem, 274*(44), 31297-31304.

Tan-Wong, S. M., French, J. D., Proudfoot, N. J., & Brown, M. A. (2008). Dynamic interactions between the promoter and terminator regions of the mammalian BRCA1 gene. *Proc Natl Acad Sci U S A, 105*(13), 5160-5165. doi:10.1073/pnas.0801048105

Thakur, S., & Croce, C. M. (1999). Positive regulation of the BRCA1 promoter. *J Biol Chem, 274*(13), 8837-8843.

Tran, D. H., Shishido, Y., Chung, S. P., Trinh, H. T., Yorita, K., Sakai, T., & Fukui, K. (2015). Identification of two promoters for human D-amino acid oxidase gene: implication for the differential promoter regulation mediated by PAX5/PAX2. *J Biochem, 157*(5), 377-387. doi:10.1093/jb/mvu084

Walsh, T., Casadei, S., Coats, K. H., Swisher, E., Stray, S. M., Higgins, J., . . . King, M. C. (2006). Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. *JAMA, 295*(12), 1379-1388. doi:10.1001/jama.295.12.1379

Wardrop, S. L., Brown, M. A., & kConFab, I. (2005). Identification of two evolutionarily conserved and functional regulatory elements in intron 2 of the human BRCA1 gene. *Genomics, 86*(3), 316-328. doi:10.1016/j.ygeno.2005.05.006

Wiedemeyer, W. R., Beach, J. A., & Karlan, B. Y. (2014). Reversing Platinum Resistance in High-Grade Serous Ovarian Carcinoma: Targeting BRCA and the Homologous Recombination System. *Front Oncol, 4*, 34. doi:10.3389/fonc.2014.00034

Xu, C. F., Chambers, J. A., & Solomon, E. (1997). Complex regulation of the BRCA1 gene. *J Biol Chem, 272*(34), 20994-20997. doi:10.1074/jbc.272.34.20994

Zhu, X., Wang, Y., Pi, W., Liu, H., Wickrema, A., & Tuan, D. (2012). NF-Y recruits both transcription activator and repressor to modulate tissue- and developmental stage-specific expression of human gamma-globin gene. *PLoS One, 7*(10), e47175. doi:10.1371/journal.pone.0047175

Figure Legends

Figure 1: Overview of Study Design

Outline of the workflow of variant collection, prioritization and analysis.

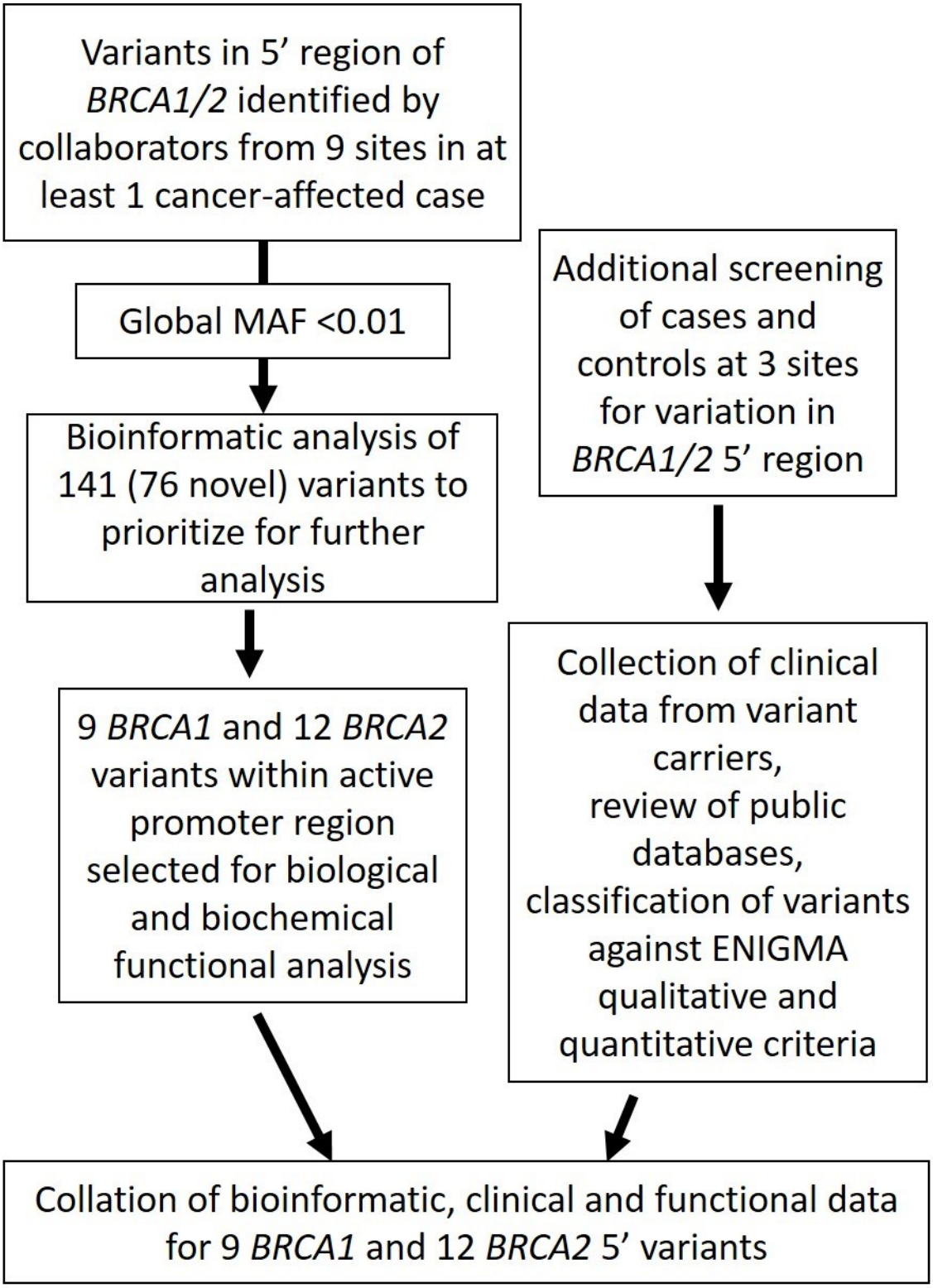

Figure 2. Variants identified in the 5' regions of *BRCA1* and *BRCA2* map to predicted regulatory elements.

Snapshots of the UCSC genome browser showing regions of *BRCA1* (A) and *BRCA2* (B) analyzed by targeted sequencing with available ENCODE regulatory marks derived from MCF7 cells. Chromatin segregation states from regulatory region annotation are shown (MCF7 States). The *BRCA1* and *BRCA2* genomic regions used for functional analyses are highlighted in grey. Prioritized variants within these regions are indicated.

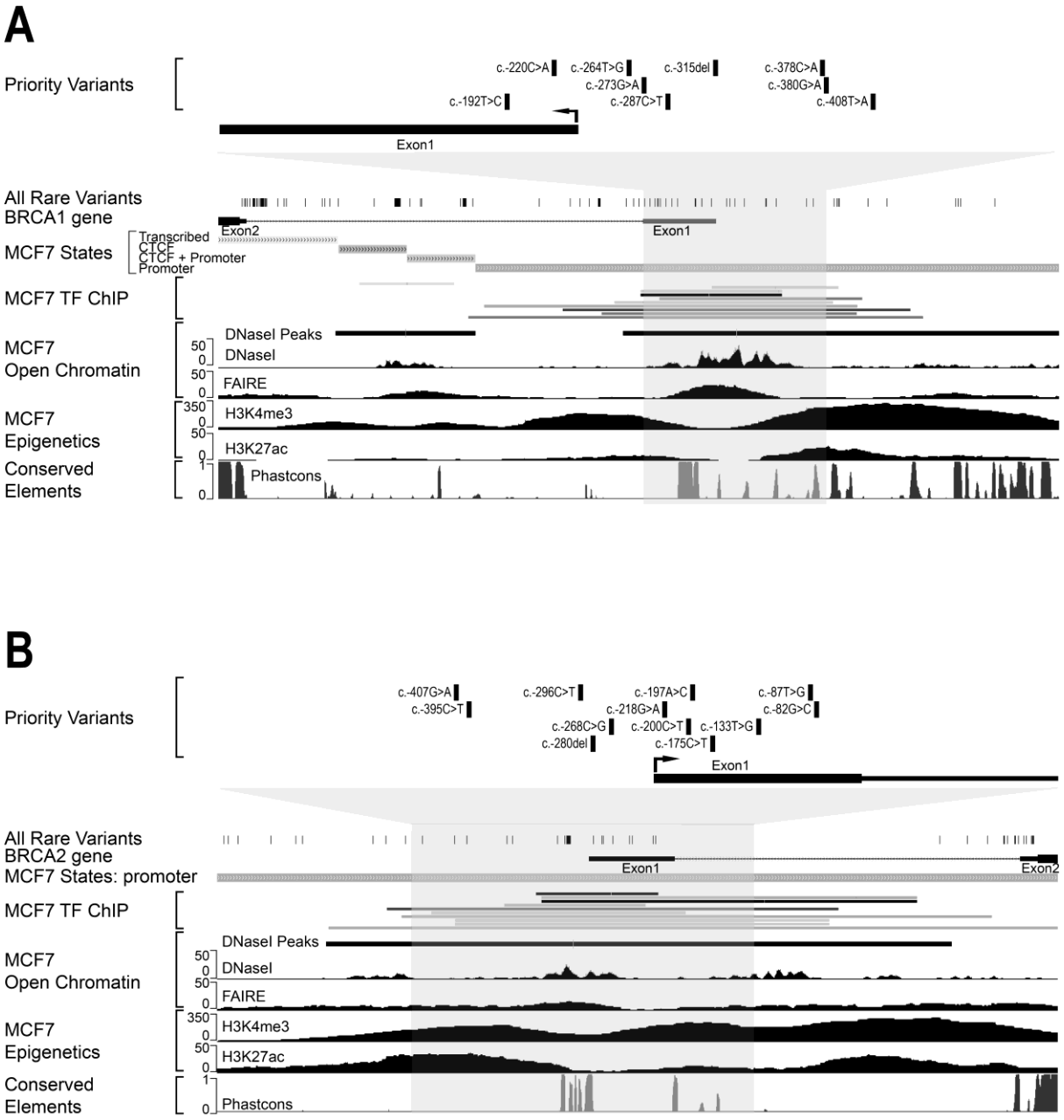

Figure 3. Variants mapping to the 5' regions of *BRCA1* and *BRCA2* alter promoter activity in MCF7 and MDA-MB-468 breast cancer cells.

MCF7 (A,C) and MDA-MB-468 cells (B,D) cells were transfected with pGL3 vectors where luciferase expression is controlled by a portion of the *BRCA1* (B1) (A, B) or *BRCA2* (B2) (C,D) promoter. Cells were transfected with plasmids containing the wild-type promoter sequence (WT; grey bars), positive control (B1-Ets or B2-Ets; striped bars) or the indicated variants (black bars). Luciferase expression was normalized to a co-transfected pRL-TK plasmid. Data represent the average of three independent biological replicates ± SD. The horizontal dotted line represents WT promoter activity set at 1.0 fold. The vertical dotted lines demarcate individual experiments that include WT, positive control and variant containing plasmids. (* $p<0.05$; ** $p<0.01$, *** $p<0.005$, **** $p<0.0001$).

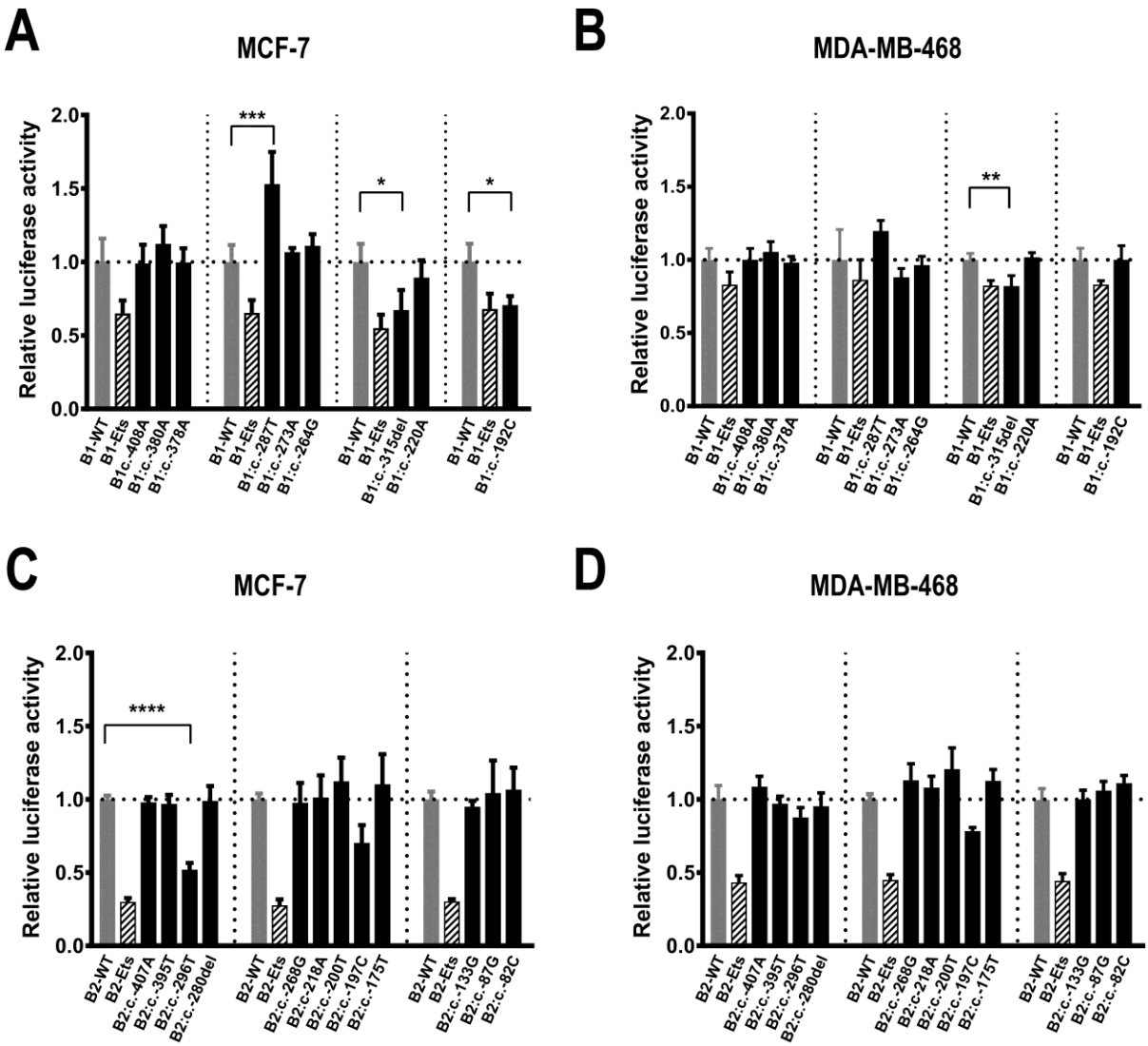

Figure 4. Variants in the 5' regions of *BRCA1* alter DNA:protein complex formation.

EMSA reactions were performed with 3' biotinylated double stranded DNA probes from the *BRCA1* 5' region and nuclear extracts (NE) from (A) MCF7 or (B) MDA-MB-468 cells. DNA probes contained either wild-type (WT) or variant (Var) sequences. Free unbound probe (FP) and probe bound by nuclear proteins (BP) are indicated.

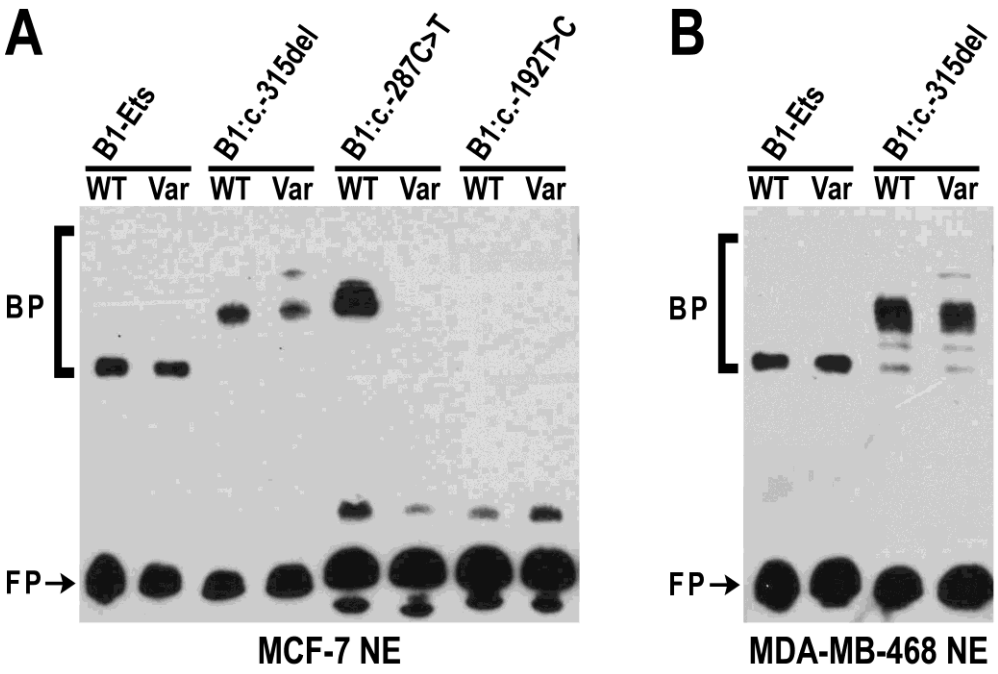

Figure 5. Variant sequences in the *BRCA1* 5' region alter specific DNA:protein complex formation.

Competition EMSAs were performed using 3' biotinylated double stranded DNA probes containing sequences from the *BRCA1* 5' region surrounding the B1:c.-315del (A,B) and B1:c.-287C>T (C) variants. DNA probes containing the wild-type (WT) or variant (Var) sequence were incubated with nuclear extracts from MCF7 cells (MCF7 NE) or MDA-MB 468 cells (468 NE) in the presence (+) or absence (-) of unlabeled WT, Var or non-specific (NS) competitor (Comp) DNA. Free unbound probe (FP) and specific DNA:protein complexes (arrowheads) are indicated. Supershift experiments (D) were performed with the *BRCA1*:c.-287C (WT) probe and antibodies to NFYA, Oct-2 (POU2F2) and PAX5. The supershifted NFYA complex is indicated (*).

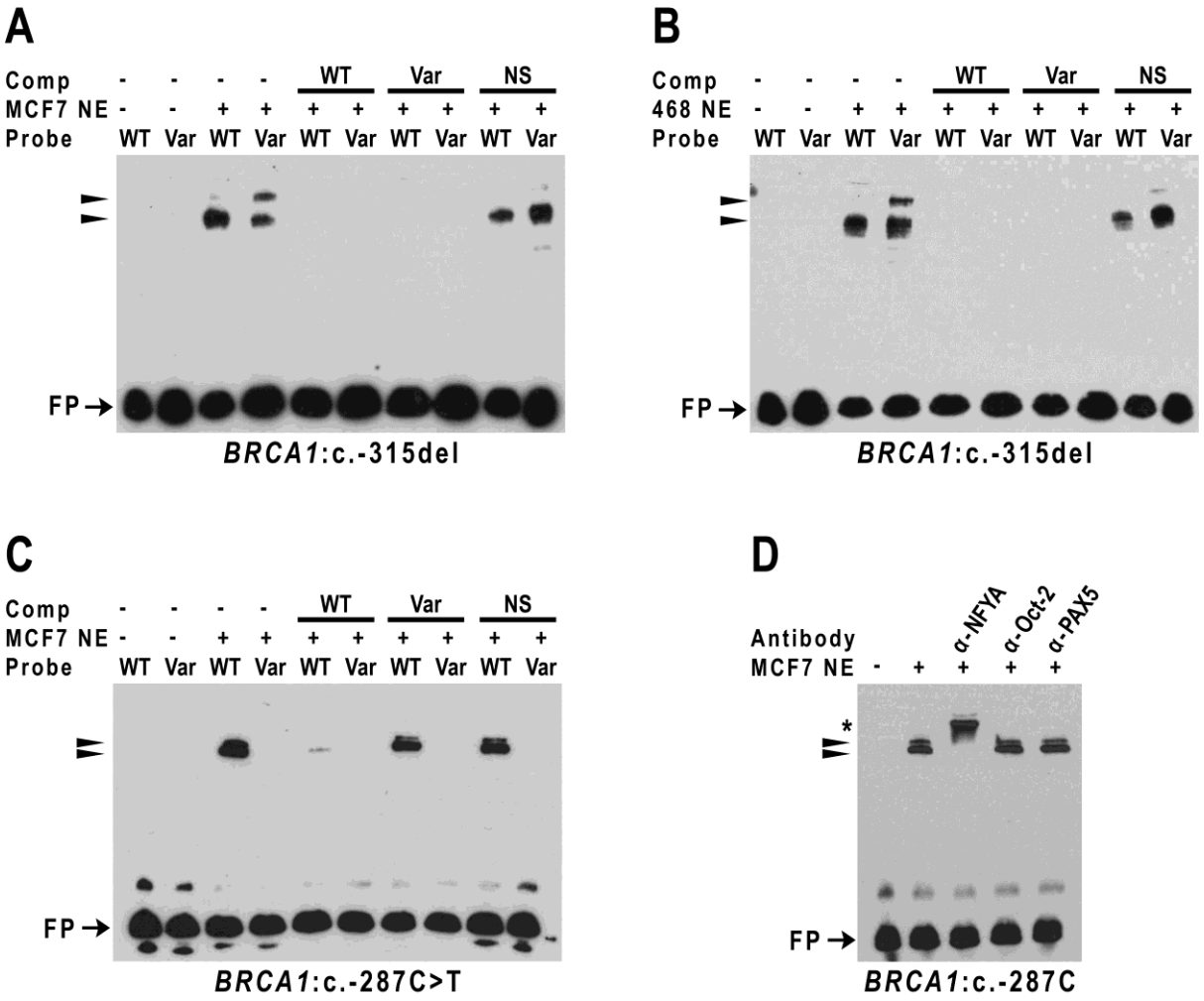

Figure 6. Variants in the 5' region of *BRCA2* alter specific DNA:protein complex formation.

Competition EMSAs (A) were performed using 3' biotinylated double stranded (ds) DNA probes containing sequences from the *BRCA2* 5' region surrounding the BRCA2:c.-296C>T variant. DNA probes containing the wild-type (WT) or variant (Var) sequence were incubated with nuclear extracts from MCF7 cells (MCF7 NE) in the presence (+) or absence (-) of unlabeled WT, Var or non-specific (NS) competitor (Comp) DNA. Cross-competition EMSAs (B) contained *BRCA2* WT sequences and increasing concentrations of ds competitor DNA containing unlabelled WT, Var or PAX5 binding sites from the *hCD19* gene and D-amino acid oxidase gene (*hDAO*). Free unbound probe (FP) and specific DNA:protein complexes (arrowheads) are indicated.

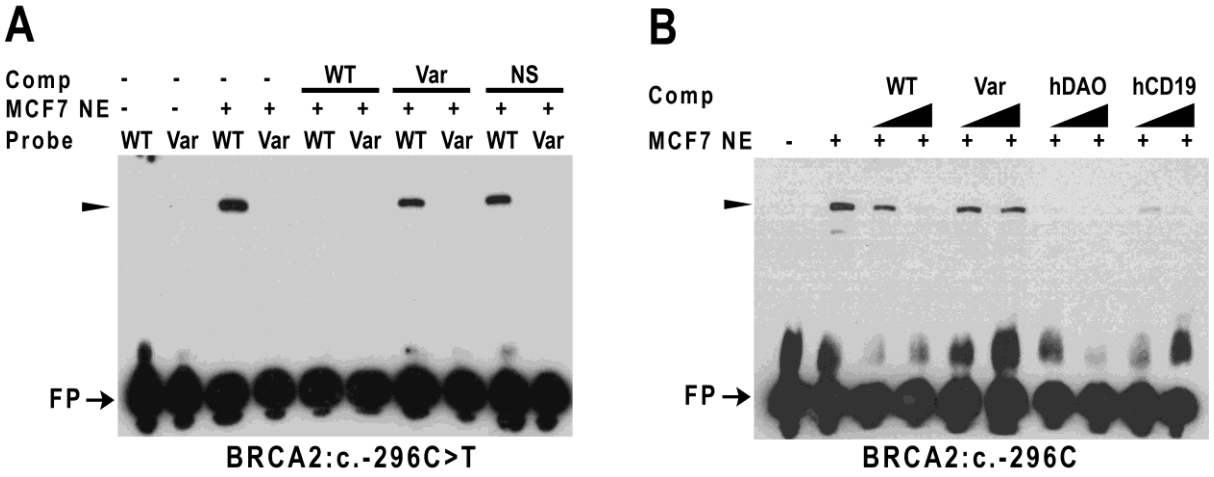

Table 1: Samples used in this study

Location	Institution	Samples	Gene region
Paris	Institut Curie, Saint Cloud	686 cases	*BRCA1* 5'region, *BRCA2* 5'region
Milan	IFOM, Fondazione Instituto FIRC di Oncologia Molecolare	772 cases 661 controls	*BRCA1* 5'region
Pisa	Department of Translational Research and New Technologies in Medicine, University of Pisa	80 cases	*BRCA1* 5'region, *BRCA2* 5'region
Santiago de Compostela	Fundación Pública Galega de Medicina Xenómica-SERGAS, Grupo de Medicina Xenómica-USC, CIBERER, IDIS	270 cases 130 controls	*BRCA1* 5'region, *BRCA2* 5'region
Copenhagen	Center for Genomic Medicine	1157 cases	*BRCA1* 5'region, *BRCA2* 5'region
Ghent	Center for Medical Genetics, Ghent University Hospital	357 cases	*BRCA1* 5'region, *BRCA2* 5'region
Barcelona	Vall d'Hebron Institute of Oncology	192 cases	*BRCA1* 5'region, *BRCA2* 5'region
Prague	CZECANCA – CZEch CAncer panel for Clinical Aplication, Institute of Biochemistry and Experimental Oncology	2961 cases 312 controls	*BRCA1* 5'region, *BRCA2* 5'region
Maastricht	Department of Clinical Genetics, Maastricht University Medical Centre	900 cases	*BRCA2* 5'region

Table 2. *BRCA1* Prioritized variants

Gene	hg19 Position (chr17)	Variant Name†	rsID	Global MAF in dbSNP	TF Motif (ENCODE)‡	Bioinformatic Priority
BRCA1	**g.41277676A>T**	c.-408T>A	Novel		CEBPB	High/Medium
BRCA1	**g.41277648C>T**	c.-380G>A	Novel		RXRA	High/Medium
BRCA1	**g.41277646G>T**	c.-378C>A	rs186775935	0.00040	RXRA	High/Medium
BRCA1	**g.41277583del**	c.-315del	rs901029407	0.00003	ATF1,2,3, CREB1§	Medium
BRCA1	**g.41277555G>A**	c.-287C>T	Novel		NFYA, NFYB	High/Medium
BRCA1	**g.41277541C>T**	c.-273G>A	rs112960339	0.00499		Medium
BRCA1	**g.41277532A>C**	c.-264T>G	rs904148166	0.00003		Medium
BRCA1	**g.41277488G>T**	c.-220C>A	Novel			Medium
BRCA1	**g.41277460A>G**	c.-192T>C	rs113323025	0.00519		Medium

†- Based on NM_007294.3
‡- Overlap with TF motif in ENCODE TF-ChIP datasets from all cells
§- Variant overlaps this motif, but the deletion does not alter the motif sequence

Table 3. *BRCA2* Prioritized variants

Gene	hg19 Position (Chr13)	Variant Name†	rsID	Global MAF in dbSNP	TF Motif (ENCODE) ‡	Bioinformatic Priority
BRCA2	g.32889437G>A	c.-407G>A	rs36221751	0.0018		Medium
BRCA2	g.32889449C>T	c.-395C>T	Novel			Medium
BRCA2	g.32889548C>T	c.-296C>T	rs563971900	0.0004	PAX5	High/Medium
BRCA2	g.32889564delG	c.-280del	Novel		ELF1, GABPA, ELK1,4	High
BRCA2	g.32889576C>G	c.-268C>G	Novel			High/Medium
BRCA2	g.32889626G>A	c.-218G>A	Novel			Medium
BRCA2	g.32889644C>T	c.-200C>T	Novel		MAZ	Medium
BRCA2	g.32889647A>C	c.-197A>C	rs370721506	NA	MAZ	Medium
BRCA2	g.32889669C>T	c.-175C>T	rs55880202	0.0058		Medium
BRCA2	g.32889711T>G	c.-133T>G	Novel			Medium
BRCA2	g.32889757T>G	c.-87T>G	Novel			Medium/Low
BRCA2	g.32889762G>C	c.-82G>C	Novel			Medium/Low

†- Based on NM_000059.3
‡- Overlap with TF motif in ENCODE TF-ChIP datasets from all cells
NA- No data available

Table 4. Information Theory Analysis of Prioritized *BRCA1/2* Variants

Variant Name	TF motif (ENCODE)	Consequences
BRCA1:c.-408T>A	CEBPB	CEBPB site weakened (did not meet stringent filtering thresholds)
BRCA1:c.-380G>A	RXRA	Weak RXRA and IRF3 sites weakened, HNF4G site weakened.
BRCA1:c.-378C>A	RXRA	RXR unchanged, HSF1 site lost and GR site created
BRCA1:c.-315del	ATF1,2,3, CREB1[†]	TCF7L2 site lost and POU2F2 created
BRCA1:c.-287C>T	NFYA, NFYB	NFYA and NFYB sites lost, weak PBX3 site created
BRCA1:c.-273G>A		Altered TF strength did not fulfill stringent filtering thresholds[‡]
BRCA1:c.-264T>G		BHLHE32 and MYC sites created.
BRCA1:c.-220C>A		Altered TF strength did not fulfill stringent filtering thresholds[‡]
BRCA1:c.-192T>C		ETS1 site created, weak RFX5 site strengthened.
BRCA2:c.-407G>A		Weak MEF2A site strengthened, GATA2 site lost.
BRCA2:c.-395C>T		TEAD4 site lost.
BRCA2:c.-296C>T	PAX5	PAX5 site weakened.
BRCA2:c.-280del	ELF1, GABPA, ELK1,4	GABPA site unchanged, MXI1 andTCF3 sites lost.
BRCA2:c.-268C>G		Altered TF strength did not meet filtering thresholds[‡]
BRCA2:c.-218G>A		Altered TF strength did not meet filtering thresholds[‡]
BRCA2:c.-200C>T	MAZ[§]	KLF1 site abolished.
BRCA2:c.-197A>C	MAZ[§]	SP4 weakened, GR site weakened, TCF3 site created
BRCA2:c.-175C>T		Altered TF strength did not fulfill stringent filtering thresholds[‡]
BRCA2:c.-133T>G		Altered TF strength did not fulfill stringent filtering thresholds[‡]
BRCA2:c.-87T>G		Altered TF strength did not fulfill stringent filtering thresholds[‡]

BRCA2:c.-82G>C	Altered TF strength did not fulfill stringent filtering thresholds‡

†- Variant overlaps this motif, but the deletion does not alter the motif sequence
‡- Change in Information did not fulfill stringent filtering criteria, where [A] site $R_i < R_{sequence}$ - 1 Standard Deviation of TF model, or [B] where $\Delta R_i < 4$ bits.
§- No MAZ binding model available

Table 5. Classification of prioritized variants

Gene	Genomic location (hg19)	HGVS c. nomenclature	Luciferase result	Combined interpretation of frequency data & multifactorial analysis	Highest MAF (population, database)	Prior probability of pathogenicity	Segregation Bayes score (# families)	Tumour histopathology likelihood ratio (# tumours)	Combined odds for causality	Posterior probability of pathogenicity§
BRCA1	g.41277676 A>T	c.-408T>A	No effect	Uncertain		0.02				
BRCA1	g.41277648 C>T	c.-380G>A	No effect	Uncertain		0.02		1.67 (1)	1.67	NA
BRCA1	g.41277646 G>T	c.-378C>A	No effect	Uncertain	0.0015 (African, 1000 Genomes)	0.02				
BRCA1	g.41277583 del	c.-315del	Decrease	Uncertain		0.02				
BRCA1	g.41277555 G>A	c.-287C>T	Increase	Uncertain		0.02		0.64 (1)	0.64	NA
BRCA1	g.41277541 C>T	c.-273G>A	No effect	Not pathogenic†	0.0159 (African, 1000 Genomes)	0.02				
BRCA1	g.41277532 A>C	c.-264T>G	No effect	Uncertain		0.02		0.51 (1)	0.51	NA
BRCA1	g.41277488 G>T	c.-220C>A	No effect	Uncertain		0.02				
BRCA1	g.41277460 A>G	c.-192T>C	Decrease	Not pathogenic†	0.0159 (African, 1000 Genomes)	0.02				
BRCA2	g.32889437 G>A	c.-407G>A	No effect	Not pathogenic‡	0.0080 (Prague, this study)	0.02		0.55 (6)	0.55	NA
BRCA2	g.32889449 C>T	c.-395C>T	No effect	Uncertain		0.02				
BRCA2	g.32889548 C>T	c.-296C>T	Decrease	Not pathogenic‡	0.0080 (Prague, this study)	0.02	3.07 (1)	1.91 (8)	5.87	0.1069
BRCA2	g.32889564 delG	c.-280del	No effect	Uncertain		0.02		0.69 (1)	0.69	NA
BRCA2	g.32889576 C>G	c.-268C>G	No effect	Uncertain		0.02				
BRCA2	g.32889626 G>A	c.-218G>A	No effect	Likely not pathogenic		0.02	0.52 (1)	0.72 (1)	0.38	0.0076

BRCA2	g.32889644C>T	c.-200C>T	No effect	Likely not pathogenic		0.02		0.37 (1)	0.37	0.0075
BRCA2	g.32889647A>C	c.-197A>C	No effect	Not pathogenic‡	0.0014 (African, FLOSSIES)	0.02		1.08 (1)	1.08	NA
BRCA2	g.32889669C>T	c.-175C>T	No effect	Not pathogenic†	0.0197 (African, FLOSSIES)	0.02				
BRCA2	g.32889711T>G	c.-133T>G	No effect	Uncertain		0.02				
BRCA2	g.32889757T>G	c.-87T>G	No effect	Uncertain		0.02				
BRCA2	g.32889762G>C	c.-82G>C	No effect	Uncertain		0.02				

†- Not pathogenic based on frequency >1% in an outbred sampleset.
‡- Variant allele assigned a low prior probability of pathogenicity of 0.02 assuming conservatively that 2/100 of such variants might be associated with a high risk of cancer AND allele frequency ≥0.001 and <0.01 (>=5 alleles) in outbred sampleset.
§- Posterior probabilities used to assign IARC 5-tier class as described in Plon et al., (2008).
NA- not applicable: multifactorial classification not assigned as the combined odds of causality were insufficient (≥0.5 and ≤2) to derive a posterior probability of pathogenicity (Vallee et al., 2016).