

2-22-2016

# Prioritizing Variants in Complete Hereditary Breast and Ovarian Cancer (HBOC) Genes in Patients Lacking known BRCA Mutations.

Natasha G Caminsky

Eliseos J Mucaki

Ami M Perri

Ruipeng Lu

Joan H M Knoll

*See next page for additional authors*

Follow this and additional works at: <https://ir.lib.uwo.ca/biochempub>

 Part of the [Biochemistry Commons](#)

---

## Citation of this paper:

Caminsky, Natasha G; Mucaki, Eliseos J; Perri, Ami M; Lu, Ruipeng; Knoll, Joan H M; and Rogan, Peter K, "Prioritizing Variants in Complete Hereditary Breast and Ovarian Cancer (HBOC) Genes in Patients Lacking known BRCA Mutations." (2016). *Biochemistry Publications*. 182.

<https://ir.lib.uwo.ca/biochempub/182>

---

**Authors**

Natasha G Caminsky, Eliseos J Mucaki, Ami M Perri, Ruipeng Lu, Joan H M Knoll, and Peter K Rogan

# Prioritizing Variants in Complete Hereditary Breast and Ovarian Cancer Genes in Patients Lacking Known *BRCA* Mutations

Natasha G. Caminsky,<sup>1</sup> Eliseos J. Mucaki,<sup>1</sup> Ami M. Perri,<sup>1</sup> Ruipeng Lu,<sup>2</sup> Joan H. M. Knoll,<sup>3,4</sup> and Peter K. Rogan<sup>1,2,4,5\*</sup>

<sup>1</sup>Department of Biochemistry, Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada; <sup>2</sup>Department of Computer Science, Faculty of Science, Western University, London, Ontario, Canada; <sup>3</sup>Department of Pathology and Laboratory Medicine, Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada; <sup>4</sup>Cytogenomix Inc., London, Ontario, Canada; <sup>5</sup>Department of Oncology, Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada

Communicated by Michael Dean

Received 8 September 2015; revised 22 January 2016; accepted revised manuscript 16 February 2016.

Published online 22 February 2016 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22972

**ABSTRACT:** *BRCA1* and *BRCA2* testing for hereditary breast and ovarian cancer (HBOC) does not identify all pathogenic variants. Sequencing of 20 complete genes in HBOC patients with uninformative test results ( $N = 287$ ), including noncoding and flanking sequences of *ATM*, *BARD1*, *BRCA1*, *BRCA2*, *CDH1*, *CHEK2*, *EPCAM*, *MLH1*, *MRE11A*, *MSH2*, *MSH6*, *MUTYH*, *NBN*, *PALB2*, *PMS2*, *PTEN*, *RAD51B*, *STK11*, *TP53*, and *XRCC2*, identified 38,372 unique variants. We apply information theory (IT) to predict and prioritize noncoding variants of uncertain significance in regulatory, coding, and intronic regions based on changes in binding sites in these genes. Besides mRNA splicing, IT provides a common framework to evaluate potential affinity changes in transcription factor (TFBSs), splicing regulatory (SRBSs), and RNA-binding protein (RBBSs) binding sites following mutation. We prioritized variants affecting the strengths of 10 splice sites (four natural, six cryptic), 148 SRBS, 36 TFBS, and 31 RBBS. Three variants were also prioritized based on their predicted effects on mRNA secondary ( $2^\circ$ ) structure and 17 for pseudoexon activation. Additionally, four frameshift, two in-frame deletions, and five stop-gain mutations were identified. When combined with pedigree information, complete gene sequence analysis can focus attention on a limited set of variants in a wide spectrum of functional mutation types for downstream functional and co-segregation analysis.

Hum Mutat 00:1–13, 2016. © 2016 Wiley Periodicals, Inc.

**KEY WORDS:** 3' untranslated region; breast neoplasm; ovarian; tumor suppressor; genetic testing; information theory; RNA stability

## Introduction

Currently, the lifetime risk for a woman to develop breast cancer (BC) is 12.3% and 1.3% in the case of ovarian cancer (OC)

[Howlander et al., 2014]). Approximately 5–10% of all BC cases are hereditary in nature, versus 25% for OC, where relative risk (RR) of BC or OC with one affected first degree family member is estimated at 2.1 and 3.1, respectively [Stratton et al., 1998; Walsh et al., 2011]. Two highly penetrant genes, *BRCA1* and *BRCA2*, are associated with a large proportion of hereditary breast and ovarian cancer (HBOC) cases. However, the estimated rate of linkage to these genes is significantly higher than the proportion of pathogenic mutations identified in HBOC families [Ford et al., 1998], suggesting unrecognized or unidentified variants in *BRCA1/2*.

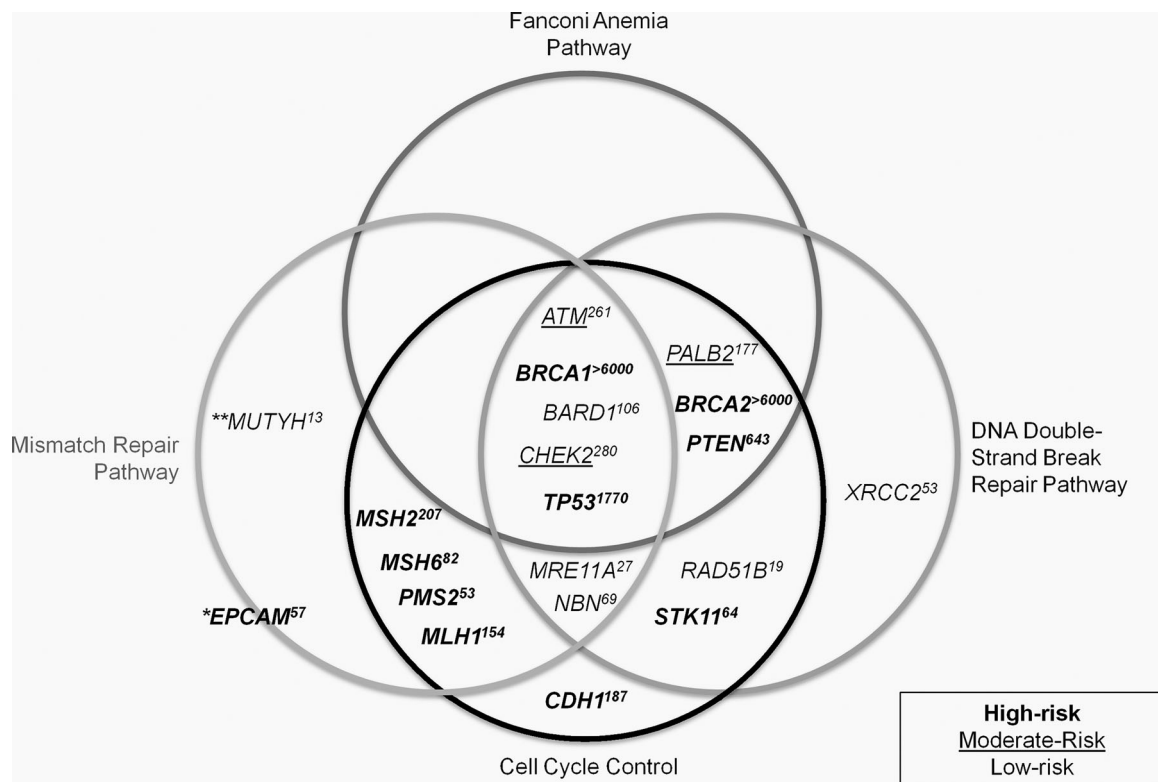
Clinical *BRCA1/2* testing is restricted primarily to coding regions. Limitations on how variants can be interpreted, lack of functional validation, and mutations in other genes contribute to uninformative results. The heritability that is not associated with *BRCA* genes is likely due to other genetic factors rather than environmental causes, specifically moderate- and low-risk susceptibility genes [Antoniu and Easton, 2006]. Hollestelle et al. [2010] point out the challenges in estimating increased risks associated with mutations in these genes, as the disease patterns are often incompletely penetrant, and require large pedigree studies to confidently assess pathogenicity.

Next-generation sequencing (NGS) of gene panels for large cohorts of affected and unaffected individuals has become an increasingly popular approach to confront these challenges. Numerous HBOC gene variants have been catalogued, including cases in which RR has been determined; however, the literature is also flooded with variants lacking a clinical interpretation [Cassa et al., 2012]. It is not feasible to functionally evaluate the effects of all the variant(s) of uncertain significance (VUS) identified by NGS. Further, *in silico* assessment of protein coding variants has not been entirely reliable [Rogan and Zou, 2013; Vihinen, 2013]. Several approaches have been developed to better assess variants from exome and genome-wide NGS data [Duzkale et al., 2013; Kircher et al., 2014]. Nevertheless, there is an unmet need for other methods that quickly and accurately bridge variant identification and classification.

To begin to address this problem, we sought to provide potentially novel interpretations of noncoding sequence changes, based on disruption or acquisition of interactions with proteins that recognize nucleic acid binding sites. Information theory (IT) based analysis predicts changes in sequence binding affinity, and it has been applied and validated for use in the analysis of splice sites (SSs), SRBSs [Rogan et al., 1998, 2003; Mucaki et al., 2013; Caminsky et al., 2015] and TFBSs [Gadiraju et al., 2003]. A unified framework based on IT requires binding genome-scale site data devoid of consensus sequence bias [Schneider, 1997], for example, photoactivatable-ribonucleoside-enhanced cross-linking and

Additional Supporting Information may be found in the online version of this article.

\*Correspondence to: Peter K. Rogan, Department of Biochemistry, Schulich School of Medicine and Dentistry, Western University, London, Ontario N6A 2C1, Canada. E-mail: progan@uwo.ca



**Figure 1.** Common genomic pathways among 20 HBOC genes, including risk and relevant literature. The left, top, and right circles indicate sequenced genes that play important roles in the mismatch repair (MMR), Fanconi anemia, and DNA double-strand break repair pathways, respectively. The bottom circle contains genes involved in cell cycle control. Genes considered to present a high risk of breast and/or ovarian cancer when mutated are bolded, moderate-risk genes are underlined, and low-risk genes are in normal font. The estimated number of articles listing a gene's association with breast or ovarian cancer (based on a systematic search in PubMed [performed June 2015]) is indicated in superscript. **\*\*MUTYH** is only high risk in the case of bi-allelic mutations. **\*EPCAM** is not involved in any pathways, but is associated with hereditary nonpolyposis colorectal cancer (HNPCC) by virtue of the fact that 3' deletions of *EPCAM* can cause epigenetic silencing of *MSH2*, causing Lynch syndrome protein. See Supp. Table S1 for citations and further evidence supporting this gene set.

immunoprecipitation (PAR-CLIP), ChIP-Seq, and a comprehensive, validated set of SSs. Although these data sources are heterogeneous, the IT models and binding site affinities derived from them are uniformly scaled (in units of bits). Thus, binding interactions involving disparate proteins or other recognition molecules can be measured and directly compared.

We have described a unified IT framework for the identification and prioritization of variants in coding and noncoding regions of *BRCA1*, *BRCA2*, and five other HBOC genes (*ATM*, *CDH1*, *CHEK2*, *PALB2*, and *TP53* [Mucaki et al., in press; biorxiv preprint: <http://dx.doi.org/10.1101/031419>]). This approach was applied to a cohort of 102 individuals lacking *BRCA* mutations with a history of HBOC. This distinguished prioritized variants from flagged alleles conferring small changes to regulatory protein binding site sequences in 70.6% of cases (Mucaki et al., submitted).

In the present study, we have sequenced 13 additional genes that have been deemed HBOC susceptibility loci (*BARD1*, *EPCAM*, *MLH1*, *MRE11A*, *MSH2*, *MSH6*, *MUTYH*, *NBN*, *PMS2*, *PTEN*, *RAD51B*, *STK11*, and *XRCC2* [Minion et al., 2015]). These genes encode proteins with roles in DNA repair, surveillance, and cell cycle regulation (Fig. 1; for further evidence supporting this gene set see Supp. Table S1 [Apostolou and Fostira, 2013; Al Bakir and Gabra, 2014]), and are associated with specific disease syndromes that confer an increased risk of BC and OC, as well as many other types of cancer (Supp. Table S2). High-risk genes confer >4 times increased risk of BC compared to the general population. *BRCA1* and

*BRCA2* are estimated to increase risk 20-fold [Antoniou et al., 2003]. Pathogenic variants in other high-risk genes, *CDH1*, *PTEN*, *STK11*, and *TP53*, are rarely seen outside of their associated syndromes and account for <1% of hereditary BC cases [Maxwell and Domchek, 2013]. *EPCAM*, *MLH1*, *MSH2*, *MSH6*, and *PMS2* have also been proposed to harbor high-risk BC alleles, but the RR is still controversial [Maxwell and Domchek, 2013]. Genes with moderate-risk alleles, *ATM*, *CHEK2*, and *PALB2*, cause between a two- and four-fold increased risk of BC [Apostolou and Fostira, 2013; Maxwell and Domchek, 2013]. The remaining genes (*BARD1*, *MRE11A*, *MUTYH*, *NBN*, *RAD51B*, and *XRCC2*) are newly identified and currently associated with unknown risks for HBOC (Fig. 1).

We report NGS of hybridization-enriched, complete genic, and surrounding regions of these genes, followed by variant analysis in 287 consented patients from Southwestern Ontario, Canada, with previously uninformative HBOC test results (this set of patients is different from our submitted study, except for six previously anonymous individuals who subsequently consented to participate). We then reduced the set of potentially pathogenic gene variants in each individual by prioritizing the results of coding and IT analyses. After applying a frequency-based filter, the IT-based framework prioritizes variants based on their predicted effect on the recognition of sequence elements involved in mRNA splicing, transcription and untranslated region (UTR) binding, combined with UTR secondary structure and coding variant analysis. Our approach integrates disparate sources of information,

including bioinformatic analyses, likelihood ratios based on familial segregation, allele frequencies, and published findings to prioritize disease-associated mutation candidates.

## Methods

### Ethics and Patient Recruitment

Recruitment and consent of human participants was approved by the University of Western Ontario Research Ethics Board (Protocol 103746). Patients were enrolled from January, 2014 through March, 2015 at London Health Sciences Centre (LHSC). Patients met the following criteria: male or female, aged between 25 and 75 years, >10% risk of having an inherited mutation in a breast/OC gene, diagnosed with BC and/or OC, and previously receiving uninformative results for a known, pathogenic *BRCA1* or *BRCA2* variant in either the patient or other relatives (by protein truncation test, denaturing high-performance liquid chromatography, and/or multiplex ligation-dependent probe amplification).

The median age of onset for patients ( $N = 287$ ; Supp. Fig. S1) with BC was 48 ( $N = 277$ ) and 46 for OC ( $N = 17$ ), and seven were diagnosed with both BC and OC. Furthermore, 31 patients had bilateral BC (98 patients at diagnosis; 23 developed tumors on the opposite side after the initial occurrence), one had bilateral OC, and 13 have had recurrent BC in the same breast. There was a single case of male BC (Supp. Table S3).

### Probe Design, Sample Preparation, and Sequencing

Probes for sequence capture were designed by *ab initio* single copy analysis, as described in Mucaki et al. (submitted) and Dorman et al. [2013]. The probes covered 1,103,029 nt across the 21 sequenced genes, including the negative control gene *ATP8B1* (see Supp. Methods for gene names, GenBank accession numbers, and OMIM reference numbers). This set of genes was proposed for evaluation at the evidence-based network for the Interpretation of Germline Mutant Alleles (ENIGMA) Consortium Meeting (2013). Other genes that have been found to be mutated in HBOC could not be included (e.g., *BRIP1*, *RAD50*, *RAD51C*, and *RAD51D*; [Heikkinen et al., 2003; Seal et al., 2006; Janatova et al., 2015]).

Patient DNA extracted from peripheral blood was either obtained from the initial genetic testing at LHSC Molecular Genetics Laboratory or isolated from recent samples. NGS libraries were prepared using modifications to a published protocol [Gnirke et al., 2009] described in Mucaki et al. (submitted), and all postcapture pull-down steps were automated (Supp. Methods). An Illumina Genome Analyzer Iix instrument in our laboratory was used for sequencing.

Library preparation and resequencing were repeated for samples with initial average coverage below our minimum threshold ( $<30\times$ ). To ensure that the proper sample was resequenced, the variant call format (VCF) files from each run were compared to all others in the run using VCF compare (<http://vcftools.sourceforge.net/>). VCF files from separate runs for the resequenced patients were concordant, except for minor differences in variant call rates due to differences in coverage. The aligned reads from both runs were then merged (with BAMtools; <http://sourceforge.net/projects/bamtools/>).

Samples were demultiplexed and aligned using Consensus Assessment of Sequencing and Variation [CASAVA; v1.8.2; DePristo et al., 2011] and Complex Reads Analysis & Classification (CRAC; v1.3.0; <http://crac.gforge.inria.fr/>). Aligned BAM files were then preprocessed for variant calling with Picard (v.1.109;

<http://broadinstitute.github.io/picard/>) (MarkDuplicates, AddOrReplaceReadGroups, FixMateInformation). The Genome Analysis Toolkit (GATK; v3.1; <http://www.broadinstitute.org/gatk/>) was then used for variant calling using the modules “Indel realigner” and the “Unified Genotyper.” Variants flagged by bioinformatic analysis (see the section *Variant Analysis*) were also assessed by manual inspection with the Integrative Genome Viewer v2.3 (IGV; <http://www.broadinstitute.org/igv/>). Variants in this study are written in HGVS notation, are based on cDNA sequence, and comply with journal guidelines.

### Information Models

Models for natural SSs and splicing regulatory factors (SRFs) are described in Mucaki et al. [2013]. These models were used to predict deleterious effects on natural splicing, the activation of cryptic SSs, and changes in binding of splicing enhancers and silencers. In addition, using a combination of cryptic site activation and hnRNPA1 site prediction, pseudoexon formation was also assessed.

We previously built models for TFBSs ( $N = 83$ ) using ENCODE ChIP-seq data [ENCODE Project Consortium, 2012; Mucaki et al., submitted]. Due to the inclusion of the additional genes, eight additional transcription factors (TFs) were identified from the literature and ENCODE ChIP-seq data from BC cell lines with evidence of binding and potentially regulating these genes. However, models for three of these TFs passed our quality control criteria (TFIIB150 [*BDPI*], PBX3 and ZNF274; described in Lu et al., submitted). Supp. Table S4 contains the full list of TFs ( $N = 86$ ) and indicates which genes exhibit evidence of promoter or other binding events. Noise models ( $N = 5$ ), reflecting motifs of interacting cofactors or sequence-specific histone modifying events, were excluded (Supp. Methods).

Information weight matrices,  $R_i(b,l)$ , for sequences bound by RNA-binding proteins (RBPs) were derived from frequency matrices published in the Catalog of Inferred Sequence Binding Preferences of RNA binding protein (CISBP-RNA; <http://cisbp-rna.cabr.utoronto.ca/>) and RNA-Binding Protein Database (BPDB; <http://rbpdb.cabr.utoronto.ca/>). These  $R_i(b,l)$ s were used to compute changes in binding affinity due to SNVs, using conservative minimum information thresholds described in Mucaki et al. (submitted). Finally, predicted changes in UTR structure resulting from variants were determined using SNPfold [<http://ribosnitch.bio.unc.edu/snpfold/>; Halvorsen et al., 2010]. Significant changes in UTR structure and stability were represented using mfold (<http://unafold.rna.albany.edu/?q=mfold>).

### Variant Analysis

Information analysis has been used in the interpretation of variant effects on binding sites containing these changes, whether this involves the creation or strengthening, or the abolition or weakening of a site [Rogan et al., 1998]. This analysis was applied to all variants identified by NGS. Changes in information are directly related to changes in thermodynamic entropy and thus binding affinity [Rogan et al., 1998]. For example, a 1.0 bit change in information corresponds to at least a twofold change in binding affinity. Information theoretical analysis of SSs and SRF binding sites has been extensively used and proven to be reliable and robust (85.2% accuracy when compared to variants validated by expression studies) [Caminsky et al., 2015].



Information analysis was automated and thresholds for changes were applied programmatically based on our previously validated criteria [Rogan et al., 1998, 2003; von Kodolitsch et al., 2006; Dorman et al., 2014]. This reduced manual review of prioritized variants, databases, and the literature. A minimum 1.0 bit threshold was set for variants predicted to affect natural SSs or that activate a cryptic SS by exceeding the strength of cognate natural sites. Variants affecting splicing regulatory, transcription, and RBP binding sites were assessed more stringently and had a minimum threshold of 4.0 bits, i.e.  $\geq 16$ -fold, in order to be flagged for further assessment. A population frequency filter was also applied to variants with allele frequencies  $>1\%$  (in the NCBI Short Genetic Variations database (dbSNP)) or  $>5\%$  of our patient cohort. Such variants were eliminated from further consideration.

To assess coding changes affecting predicted protein chain length or amino acid(s) composition, we used SNPnexus (<http://hsnp-nexus.org/>). Insertion/deletions (indels) and nonsense mutations were noted, and missense variants were further assessed with *in silico* tools (Mutation Assessor, <http://mutationassessor.org/>; PolyPhen2, <http://genetics.bwh.harvard.edu/pph2/>; PROVEAN/SIFT, <http://provean.jcvi.org/>) by referencing the published literature and consulting mutation databases (listed in Supp. Table S5; see Mucaki et al. [submitted] for more details on variant analysis). Variants remained prioritized unless there was clear evidence (co-segregation analysis or functional assays) supporting the nonpathogenicity of the variant.

*EPCAM* mutations in familial cancer are limited to 3' deletions causing epigenetic silencing of *MSH2*, and there is currently no evidence of other types of variants that alter its mRNA transcript or protein product [Ligtenberg et al., 2009]. Therefore, with the exception of indels, none of the variants flagged in *EPCAM* were prioritized. We chose to prioritize variants in *MUTYH* using the same framework as all other genes, despite *MUTYH* pathogenicity resulting from biallelic variants [Jones et al., 2002], because it is possible that a second *MUTYH* mutation remains unrecognized.

All protein truncating (nonsense and indels) as well as potentially pathogenic splicing and missense mutations were Sanger sequenced for confirmation (details in Supp. Table S6).

## Negative Control

Variants present in the *ATP8B1* gene were used as negative controls for our variant analysis framework. Initially, it was included in the list of prioritized HBOC genes provided by ENIGMA, but evidence for its association with HBOC is lacking in the published literature. Furthermore, it is not a known susceptibility gene for any type of cancer (mutations in *ATP8B1* cause progressive familial intrahepatic cholestasis [Gonzales et al., 2014]) and is infrequently mutated in breast tumors in several studies (e.g., see Cancer Genome Atlas Network [2012]).

## Likelihood Ratios

Patients with prioritized coding and/or splicing variants, which we consider the most likely to be pathogenic, were selected for co-segregation analysis ( $N = 24$ ) using an online tool that calculates the likelihood of a variant being deleterious based on pedigree information (<https://www.msbi.nl/cosegregation/>; Mohammadi et al. [2009]). Genotypes were assigned based on phenotype such that family members with breast or OC at any age were assigned the same genotype as the patient in our study ("carrier") and family members affected by other cancers, other diseases, or who are

disease free were assigned the "noncarrier" genotype. Because the penetrance parameters cannot be altered from the settings given for *BRCA1* or *BRCA2*, the *BRCA2* option was selected for patients with prioritized variants in non-*BRCA* genes. Penetrance in *BRCA2* is known to be lower than *BRCA1* values [Mohammadi et al., 2009]. Current evidence suggests that mutations in non-*BRCA* genes may be less penetrant than those in the *BRCA* genes [Apostolou and Fostira, 2013]; however, the penetrance of many of these variants remains unknown (Supp. Methods).

## Results

### Variant Analysis

We identified 38,372 unique variants among 287 patients (26,636 intronic, 7,287 intergenic, and 714 coding), on average 1,975 variants per patient, before any filtering criteria were applied. The extensive span of sequences captured in this study, that is, complete genes and flanking regions, constrained the genomic density and sequence coverage that could be achieved; this precluded accurate copy number estimation based solely on read counts.

### Natural site variants

The Shannon Human Splicing Mutation Pipeline (<http://www.mutationforecaster.com>; Shirley et al. [2013]) was used to predict the effect of the 14,458 variants that could potentially affect splicing, of which 244 reduced natural SS strength. Further stringent filtering of the natural SS based on information content changes and allele frequency resulted in seven flagged variants (Supp. Table S7). Henceforth, allele frequency of known variants can be found in their associated supplemental table (where available).

Four of these variants were prioritized (Table 1). A novel synonymous variant in exon 2 of *RAD51B*, c.84G>A (p.Gln28 = ), is predicted to increase exon skipping by weakening the natural splice donor ( $R_{i,final} = 5.2$  bits,  $\Delta R_i = -3.0$  bits). A known *ATM* variant, c.6198+1G>A (8-1D.9-1B [Stankovic et al., 1998]), abolishes the natural donor SS of constitutively spliced exon 42 ( $R_{i,final} = -13.7$  bits,  $\Delta R_i = -18.6$  bits). There is no evidence in public databases for appreciable alternative splicing of this exon in normal breast tissues. The variant will either lead to exon skipping or activation of a pre-existing cryptic site (Fig. 2). An ataxia-telangiectasia patient with this variant exhibited low expression, protein truncation, and abolished kinase activity of ATM [Reiman et al., 2011]. *MLH1* c.306+4A>G causes increased exon skipping (and a decrease in wild-type exon relative expression) due to the weakening ( $R_{i,final} = 6.0$  bits,  $\Delta R_i = -2.6$  bits) of the exon 3 natural donor. Tournier et al. [2008] assessed this variant using an *ex vivo* splicing assay and observed cryptic site activation and exon 3 skipping. *MRE11A* c.2070+2A>T is indicated in ClinVar as likely pathogenic and abolishes the natural donor site of exon 19 ( $R_{i,final} = -11.0$  bits,  $\Delta R_i = -18.6$  bits), while strengthening a cryptic site 5 nt upstream of the splice junction ( $R_{i,final} = 8.1$  bits,  $\Delta R_i = 0.6$  bits). Either cryptic SS activation or complete exon skipping are predicted.

The *BRCA2* variant c.68-7T>A was not prioritized, as its pathogenicity has not been proven. While there is evidence that this variant induces (in-frame) exon skipping [Théry et al., 2011], it did not segregate with disease in HBOC pedigrees, where abnormal splicing was not seen [Santos et al., 2014]. The *ATM* variant c.1066-6T>G, previously reported in Mucaki et al. (submitted), was also not prioritized as the variant does not correlate with BC risk [Ding et al., 2011].

**Table 1. Prioritized Variants Predicted by IT to Affect Natural and Cryptic Splicing**

Gene	Variant	rsID (dbSNP142) allele frequency (%) <sup>c</sup>	Information change			Consequence
			$R_{i,initial}$ (bits)	$R_{i,final}$ (bits)	$\Delta R_i$ (bits)	
<i>ATM</i>	NM_000051.3:c.6198+1G>A [Stankovic et al., 1998; Reiman et al., 2011]	–	4.9	–13.7	–18.6	Abolished natural <sup>d,§</sup>
<i>MRE11A</i>	NM_005591.3:c.2070+2A>T <sup>a</sup>	–	7.6	–11	–18.6	Abolished natural <sup>d,§</sup>
<i>MLH1</i>	NM_000249.2:c.306+4A>G <sup>a</sup> [Tournier et al., 2008]	rs267607733	8.6	6	–2.6	Weakened natural <sup>c</sup>
<i>RAD51B</i>	N M_002877.4:c.84G>A <sup>a</sup> p.Gln28=	Novel	8.2	5.2	–3	Weakened natural <sup>d</sup>
<i>BARD1</i>	NM_000465.2:c.1454C>T <sup>a</sup> p.Ala485Val	Novel	–2.7	4.4	7.1	Created cryptic <sup>e</sup>
<i>BRCA1</i>	NM_007294.2:c.5074+107C>T	rs373676607	–1.3	5.7	7	Created cryptic <sup>f,h</sup>
<i>CDH1</i>	NM_004360.3:c.1223C>G <sup>a</sup> p.Ala408Gly [Schrader et al. 2011]	Novel	–0.6	4.3	4.9	Created cryptic <sup>e</sup>
<i>RAD51B</i>	NM_002877.4:c.958-29A>T <sup>b</sup>	rs34436700 0.78	2.2	4.4	2.2	Strengthened cryptic <sup>f</sup>
<i>STK11</i>	NM_000455.4:c.375-194GT>AC	rs35113943 17.61 rs117211142 0.80	7.5	8.8	1.3	Strengthened cryptic <sup>f</sup>
<i>XRCC2</i>	NM_005431.1:c.122-154G>T	Novel	8.1	10	1.9	Strengthened cryptic <sup>e</sup>

<sup>a</sup>Confirmed by Sanger sequencing.

<sup>b</sup>Ambiguous Sanger sequencing results.

<sup>c</sup>If available

<sup>d</sup>Exon skipping

<sup>e</sup>Exon truncation

<sup>f</sup>Intron retention

<sup>§</sup>Use of alternate isoform

<sup>h</sup>Reduced expression of natural isoform.

### Activation of cryptic splicing

The Shannon Pipeline identified 9,480 variants that increased the strength of at least one cryptic site, of which nine met or exceeded the defined thresholds for information change. Six of these were prioritized (Table 1). A novel *BARD1* variant in exon 6 (c.1454C>T; p.Ala485Val) creates a donor SS ( $R_{i,final} = 4.4$  bits,  $\Delta R_i = 7.1$  bits), which would produce a 58 nt frameshifted exon if activated. The natural donor SS of exon 6, 116 nt downstream of the variant, is stronger (5.5 bits), but the Automated Splice Site and Exon Definition Analysis (ASSEDA, <http://mutationforecaster.com>) server predicts equal levels of expression of both natural and cryptic exons. A *BRCA1* mutation 5074+107C>T downstream of exon 16 is predicted to extend the exon by 105 nt and be slightly more abundant than the natural exon ( $R_{i,total}$  of 8.6 and 8.1 bits, respectively). *CDH1* c.1223C>G (p.Ala408Gly), previously reported in a *BRCA*-negative lobular BC patient with no family history of gastric cancer [Schrader et al., 2011], creates a cryptic donor site ( $R_{i,final} = 4.3$  bits,  $\Delta R_i = 4.9$  bits) in exon 9, 97 nt downstream of the natural acceptor. While residual splicing of the normal exon is still expected, the cryptic is predicted to become the predominant splice form (~twice as abundant).

*STK11* c.375-194GT>AC (rs35113943 and rs117211142) and the novel *XRCC2* c.122-154G>T both strengthen strong preexisting cryptic sites exceeding the  $R_{i,total}$  values of their respective natural exons. Finally, a known *RAD51B* variant 29 nt upstream of exon 10: c.958-29A>T strengthens a cryptic acceptor site ( $R_{i,final} = 4.4$  bits,  $\Delta R_i = 2.2$  bits) that, if activated, would produce a transcript retaining 21 intronic nucleotides.

The remaining cryptic site variants (Supp. Table S7) were not prioritized. The novel *BRCA2* c.7618-269..7618-260del10 variant is predicted to create a cryptic site with an exon having a lower  $R_{i,total}$  value (5.2 bits) than the natural exon (6.6 bits). *PMS2* c.1688G>T (p.Arg563Leu; rs63750668; three patients) does not segregate with disease. Drost et al. [2013] demonstrated that this variant does not impair DNA repair activity. Finally, *RAD51B* c.728A>G (p.Lys243Arg; rs34594234; 7 patients) predicts an increase in the

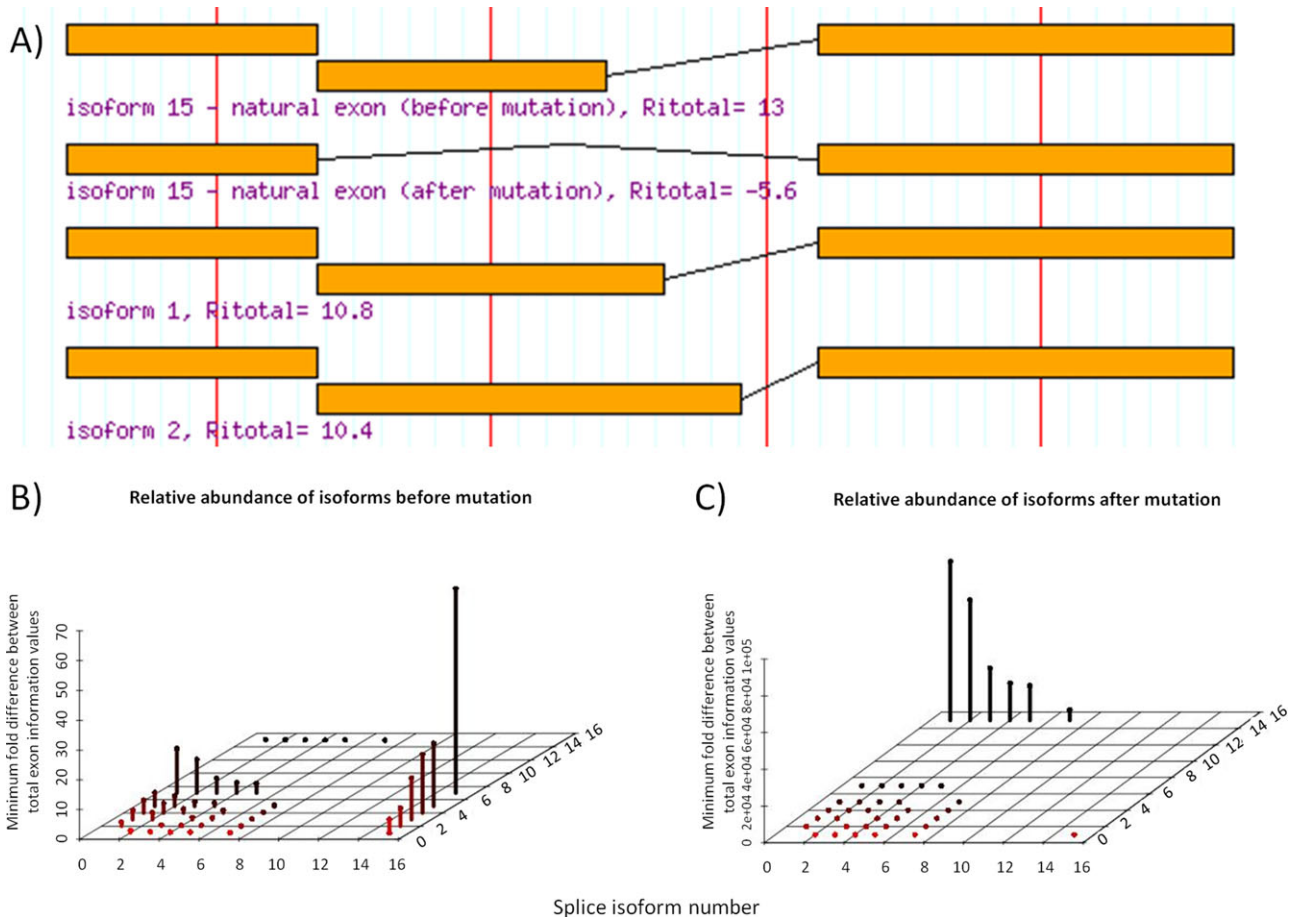
abundance of the cryptic exon; however, the natural exon remains the predominant isoform.

### Pseudoexon activation

Pseudoexons arise from creation or strengthening of an intronic cryptic SS in close proximity to another intron site of opposite polarity. Our analysis detected 623 variants with such intronic cryptic sites, of which 17 were prioritized (among nine genes), occurring within 250 nt of a preexisting site of opposite polarity, with an hnRNPA1 site within 5 nt of the acceptor of the predicted pseudoexon (Supp. Table S8). Three are novel (*BRCA2* c.7007+824C>T, *BRCA2* c.8332-1130G>T, and *PTEN* c.802-796C>A) and the remainder were present in dbSNP. Seven of these variants (*BARD1* c.1315-168C>T, *BRCA2* c.631+271A>G, *MLH1* c.1559-1732A>T, *MRE11A* c.1783+2259A>G, *MSH6* c.260+1758G>A, *PTEN* c.79+4780C>T, and *RAD51B* c.1037-1012C>A), although rare, occur in multiple patients, and one patient had predicted pseudoexons in both *BARD1* and *RAD51B*.

### SRF binding

Variants within exons or within 500 nt of a natural SS ( $N = 9,998$ ) were assessed for their potential effect on SRF binding sites (SRF-BSs). Initially 216 unique variants were flagged (Supp. Table S9), but after considering each in the context of the SRF function and location within the gene [Caminsky et al., 2015], we prioritized 148, of which 57 are novel. Some prioritized variants affect distant SRFs that may activate cryptic sites, but were not predicted to affect natural splicing. Of the 88 suitable prioritized variants for which exon definition analysis was performed (where initial  $R_{i,total}$  of the exon > SRF gap surprisal value), 55 were predicted to induce or contribute to increased exon skipping. For example, an uncommon *ATM* missense variant within exon 41, c.6067G>A (p.Gly2023Arg; rs11212587), strengthens an hnRNPA1 site ( $R_{i,final} = 5.2$  bits,



**Figure 2.** Predicted isoforms and relative abundance as a consequence of *ATM* natural splice variant c.6198+1G>A. **(A)** Intronic *ATM* variant c.6198+1G>A abolishes the natural donor of exon 42 ( $R_{i, initial} = 4.9$  bits,  $\Delta R_i = -18.6$  bits), and would either result in exon skipping (causing a frame-shift; isoform 15 after mutation), or possibly activate a downstream cryptic site (isoform 1 maintains reading frame, isoform 2 would not). **(B)** The peaks in plot display the predicted abundance (Y-axis) of a splice isoform (X-axis) relative to another predicted isoform (Z-axis). In the wild-type mRNA, the natural exon (isoform 15) has the highest predicted relative abundance. Before mutation, it is predicted to be approximately fivefold stronger than isoform 1 and 2. **(C)** After mutation, isoform 1 and 2 is now > 100,000-fold stronger than isoform 15 (abolished wild-type exon). Isoform 2 to be slightly less abundant than 1.

$\Delta R_i = 4.7$  bits) 30 nt from the natural donor, and is predicted to induce exon 41 skipping ( $\Delta R_{i, total} = -9.5$  bits).

### TF binding

To assess potential changes to TFBSs, variants occurring from 10 kb upstream of the start of transcription through the end of the first intron were analyzed by IT, flagging 88 (of 4,530 identified; Supp. Table S10). Considering the gene context of each TFBS and extent of information change, we prioritized 36 variants. The following example illustrates the rationale for highlighting these variants: *BRCA1* c.-19-433A>G abolishes a binding site for HSF 1 ( $R_{i, initial} = 5.5$  bits,  $\Delta R_i = -7.8$  bits). While HSF 1 is known to be a transcriptional activator associated with poor BC prognosis [Santagata et al., 2011], the specific effect of reduced HSF 1 binding to *BRCA1* has not been established. Similarly, *MLH1* c.-4285T>C (rs115211110; five patients) significantly weakens a C/EBP $\beta$  site ( $R_{i, initial} = 10.1$  bits,  $\Delta R_i = -6.3$  bits), a TF that has been shown to play a role in BC development and progression [Zahnaw, 2009]. Another *MLH1* variant, c.-6585T>C (novel), greatly decreases the binding strength ( $R_{i, initial} = 12.5$  bits,  $\Delta R_i = -10.8$  bits) of the NF- $\kappa$ B p65 subunit, which is activated

in ER-negative breast tumors [Biswas et al., 2004]. Two prioritized variants (*PMS2* c.-9059G>C and *XRCC2* c.-163C>A) weaken PAX5 binding sites, a TF which when overexpressed can result in mammary carcinoma cells regaining epithelial cell characteristics [Vidal et al., 2010].

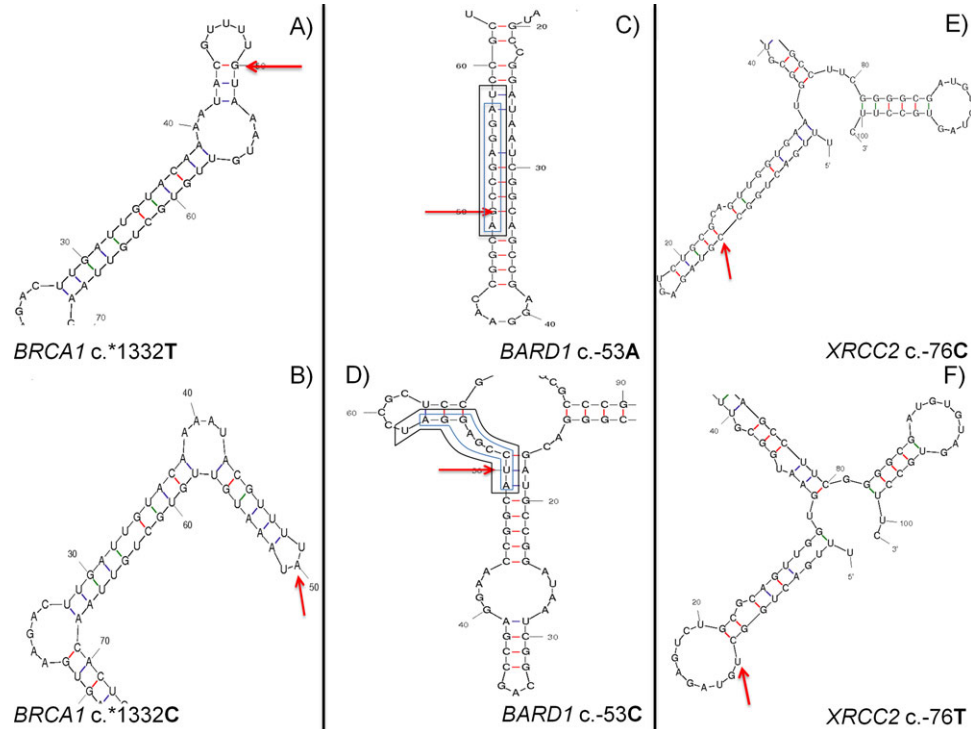
### Alterations to mRNA structure

A total of 1,355 variants were identified in the 5' and 3' UTRs of the patients. Analysis of these variants with SNPfold flagged three unique variants ( $P < 0.05$ ) in *BRCA1*, *BARD1*, and *XRCC2* (Table 2). The predicted mRNA 2 $^{\circ}$  structures of the reference and variant sequences are shown in Figure 3 (generated with mfold). The *BRCA1* variant occurs in the 3' UTR of all known transcript isoforms (NM\_007294.3:c.\*1332T>C; rs8176320; 3 patients). The most likely inferred structure consisting of a short arm and a larger stem loop is destabilized when the variant nucleotide is present (Fig. 3A and B). The *BARD1* variant falls within the 5' UTR of a rare isoform (XM\_005246728.1:c.-53G>T; rs143914387; five patients) and is within the coding region of a more common transcript (NM\_000465.2:c.33G>T; p.Gln11His). While the top ranked



**Table 2. Variants Predicted by SNPfold to Significantly Affect UTR Structure**

Gene	Variant	UTR position	rsID (dbSNP142) allele frequency (%) <sup>a</sup>	Rank	P-value
<i>BARD1</i>	XM_005246728.1:c.-53G>T (c.33G>T p.Gln11His)	5'UTR	rs1439143870.04	6/600	0.01
<i>BRCA1</i>	NM_007294.3:c.*1332T>C NM_007299.3:c.*1438T>C	3'UTR	rs81763200.42	13/450	0.03
<i>XRCC2</i>	NM_005431.1:c.-76C>T	5'UTR	rs5475387310.08	3/300	0.01

<sup>a</sup>If known.

**Figure 3.** Predicted RNA structure change due to variants flagged by SNPfold using mfold. Wild-type (**A**, **C**, and **E**) and variant (**B**, **D**, and **F**) structures are displayed. The variant nucleotide is marked with an arrow. (**A**) Predicted wild-type structure of *BRCA1* 3' UTR surrounding c.\*1332G>A. (**B**) *BRCA1* 3' UTR structure due to c.\*1332G>A variant, extending arm length while reducing hairpin size. (**C**) *BARD1* 5' UTR structure of rare isoform (XM\_005246728.1:c.-53G>T). Two overlapping preexisting RBP sites (SRSF7 [outer box] and SRSF2 [inner box]) are predicted and either could occupy this location if accessible. (**D**) 2° *BARD1* 5' UTR structure of the region predicted only with sequence containing the c.-53T mutation. The primary predicted c.-53T structure is identical to wild type (with one disrupted C-G bond leading to a 4.1 kcal/mol lower  $\Delta G$ ). The variant both weakens and abolishes the preexisting SRSF7 and SRSF2 sites, respectively. (**E**) *XRCC2* structure within common 5' UTR surrounding c.-76C>T variant. (**F**) *XRCC2* 5' UTR structure predicted from c.-76T sequence, containing a hairpin not found in wild type. This hairpin may allow for the binding of previously inaccessible nucleotides including the altered nucleotide.

isoform following mutation is similar to the wild-type structure, the second-ranked isoform ( $\Delta G = +1.88$  kcal/mol) is distinctly different, creating a loop in a long double-stranded structure (Fig. 3C and D). The *XRCC2* variant is within its common 5' UTR (NM\_005431.1:c.-76C>T) and is located 11 nt downstream from the 5' end of the mRNA. The variant nucleotide disrupts a potential GC base pair, leading to a large stem-loop that could allow access for binding of several RBPs (Fig. 3E and F). The variant simultaneously strengthens PUM2 ( $R_{i,initial} = 2.8$  bits,  $\Delta R_i = 4.4$  bits, positions 11–17) and RBM28 sites ( $R_{i,initial} = 4.0$ ,  $\Delta R_i = 3.6$  bits, positions 10–13); however, there is a stronger NCL site (8.3 bits, positions 20–31) in the area that is not affected and may compete for binding.

### RBP binding

Using IT models of 76 RBBSs, 33 UTR variants were prioritized (Supp. Table S11) from the initial list of 1,367 UTR variants. Interestingly, one of the three variants that destabilized the mRNA

was also flagged using our RBP scan. The *BARD1* c.-53A>C variant weakens a predicted 8.3 bit SRSF7 site ( $\Delta R_i = -3.0$  bits) while simultaneously abolishing a predicted 9.7 bit SRSF2 site ( $\Delta R_i = -29.7$  bits) (Fig. 3C and D).

## Exonic Protein-Altering Variants

### Protein truncating

Of the 714 identified coding variants, six were indels, each of which were found in a single patient, and two preserved the reading frame. Two indels were novel (*BRCA1*:c.3550\_3551insA [p.Gly1184Glufs] and *CDH1*:c.30\_32delGCT [p.Leu11del]). Previously reported indels were detected in *CHEK2* and *PALB2*. In addition, five nonsense mutations, which have been previously reported by others, were found in six different patients (Table 3; details in Supp. Table S12).

**Table 3. Variants Resulting in Premature Protein Truncation**

Gene	Exon	Variant	rsID (dbSNP142) allele frequency (%) <sup>a</sup>	Details
Frameshift insertions/deletions				
<i>BRCA1</i>	10 of 23	NM_007294.2:c.3550_3551insA <sup>b</sup> p.Gly1184Glufs	Novel	STOP at p.1187 676 AA short
<i>PALB2</i>	4 of 13	NM_024675.3:c.757_758delCT <sup>c</sup> p.Leu253Ilefs	rs180177092	STOP at p.255 932 AA short
<i>PALB2</i>	9 of 13	NM_024675.3:c.2920_2921delAA <sup>c</sup> p.Lys974Glufs	rs180177126	STOP at p.979 208AA short
Insertions/deletions with conserved reading frame				
<i>CDH1</i>	1 of 16	NM_004360.3:c.30_32delGCT <sup>d</sup> p.Leu11del	Novel	Loss of one AA frame and AA sequence conserved
<i>CHEK2</i>	4 of 14	NM_007194.3:c.483_485delAGA <sup>c</sup> p.Glu161del	–	Loss of one AA frame and AA sequence conserved
Stop codons				
<i>ATM</i>	13 of 63	NM_000051.3:c.1924G>T <sup>c</sup> p.Glu642Ter	–	2415 AA short
<i>ATM</i>	62 of 63	NM_000051.3:c.8977C>T <sup>c</sup> p.Arg2993Ter	–	64 AA short
<i>BRCA1</i>	23 of 23	NM_007294.2:c.5503C>T <sup>b</sup> p.Arg1835Ter	rs41293465	28 AA short
<i>PALB2</i>	13 of 13	NM_024675.3:c.3549C>G <sup>c</sup> p.Tyr1183Ter	rs118203998	4 AA short

<sup>a</sup>If known.<sup>b</sup>Not confirmed through Sanger sequencing.<sup>c</sup>Confirmed by Sanger sequencing.<sup>d</sup>Ambiguous Sanger sequencing results.

AA: amino acid.

### Missense variants

Of the 155 unique missense variants (Supp. Table S13), 119 were prioritized by consulting published literature, disease- and gene-specific databases. All are of unknown clinical significance and 21 have not been previously reported.

Missense variants that have been previously described as detrimental include the *ATM* variant c.7271T>G (p.Val2424Gly; rs28904921; two patients), which replaces a hydrophobic residue by glycine in the conserved FAT domain and confers a ninefold increase (95% CI) in BC risk [Goldgar et al., 2011]. Functional studies, assessing ATM kinase activity in vitro with TP53 as a substrate, showed that cell lines heterozygous for the mutation had less than 10% of wild-type kinase activity, such that this variant is expected to act in a dominant-negative manner [Chenevix-Trench et al., 2002]. The *CHEK2* variant c.433C>T (p.Arg145Trp; rs137853007; one patient) results in rapid degradation of the mutant protein [Lee et al., 2001]. Finally, the *PMS2* variant c.2T>C (p.Met1Thr) is listed in ClinVar as pathogenic and would be expected to abrogate correct initiation of translation. This variant has not been reported in BC families, but is associated with colorectal cancer (CRC) [Senter et al., 2008].

### Variant Prioritization

We prioritized an average of 18.2 variants in each gene, ranging from seven (*XRCC2*) to 61 (*ATM*), an average of 0.41 variants/kb, and an average of 0.65 variants/patient (Table 4). *ATM* had the second greatest gene probe coverage (103,511 nt captured), the highest number of unique prioritized variants, and was among the top genes for number of prioritized variants per kilobase (0.59).

In total, our framework allowed for the prioritization of 346 unique variants in 246 patients, such that 85.7% of tested patients ( $N = 287$ ) had at least one prioritized variant. Most patients (84.7%) harbored fewer than four prioritized variants. The distribution of patients with prioritized variants was similar across eligibility groups (Table 5). Although Class 5 (91.1% of patients with prioritized variants) and Class 8 (100% with prioritized variants, with a single patient in this category) deviated to a greater extent from the mean variants/category, these differences were not significant,  $\chi^2(4, N = 246) = 0.98, p > 0.90$ . The distribution of prioritized variants among mutation types is nine protein truncating, 28 mRNA splicing, 34

**Table 4. Comparing Counts of Prioritized Variants**

Gene	Unique prioritized variants	Unique patients	Gene probe coverage (nt)	Prioritized variants/patient	Prioritized variants/kb
<i>ATM</i>	61	102	103,511	0.60	0.59
<i>ATP8B1</i>	21	37	94,793	0.57	0.22
<i>BARD1</i>	17	46	73,735	0.37	0.23
<i>BRCA1</i>	19	24	52,075	0.79	0.36
<i>BRCA2</i>	24	28	73,332	0.86	0.33
<i>CDH1</i>	21	32	61,312	0.66	0.34
<i>CHEK2</i>	12	13	28,372	0.92	0.42
<i>MLH1</i>	18	25	50,553	0.72	0.36
<i>MRE11A</i>	17	31	64,713	0.55	0.26
<i>MSH2</i>	18	17	112,437	1.06	0.16
<i>MSH6</i>	19	23	25,216	0.83	0.75
<i>MUTYH</i>	8	16	21,439	0.50	0.37
<i>NBN</i>	11	21	57,067	0.52	0.19
<i>PALB2</i>	26	46	25,319	0.57	1.03
<i>PMS2</i> <sup>a</sup>	8	15	11,726	0.53	0.68
<i>PTEN</i> <sup>b</sup>	15	23	86,059	0.65	0.17
<i>RAD51B</i> <sup>c</sup>	22	47	62,465	0.47	0.35
<i>STK11</i>	12	20	28,373	0.60	0.42
<i>TP53</i>	11	30	23,544	0.37	0.47
<i>XRCC2</i>	7	10	19,942	0.70	0.35

<sup>a</sup>Homologous to other genomic regions, thus fewer probes designed within gene.<sup>b</sup>*PTEN* has pseudogene *PTENP1*, thus fewer probes covering exonic regions.<sup>c</sup>Probes limited to 1,000 nt surrounding all exons, and 10,000 nt up- and downstream of gene.

affecting RBBS and/or UTR structure, 36 affecting TFBS, 119 missense, and 149 affecting SRFBS, of which 29 were prioritized into multiple categories (Supp. Tables S14 and S15 show this information by gene and patient, respectively).

All prioritized protein-truncating ( $N = 10$ ), and selected splicing ( $N = 7$ ) and missense ( $N = 5$ ) variants were verified by bidirectional Sanger sequencing as they were more likely to be pathogenic (taking into account available published studies). Of the protein-truncating variants, four nonsense, one indel with a conserved reading frame, and two frameshifts were confirmed (Table 3). Six splicing variants and all missense were confirmed (Table 1 and Supp. Table S13). An additional 145 prioritized variants, including 88 noncoding variants, were confirmed upon resequencing of patient gDNA. Of the 57 resequenced coding variants, 13 were prioritized for their

**Table 5. Distribution of Recruited Patients among Eligibility Groups**

Eligibility group <sup>a</sup>	Number of patients within eligibility group	Number of patients with prioritized variants
Breast cancer <60 years, and a first- or second-degree relative with ovarian cancer or male breast cancer (5)	68	62
Breast and ovarian cancer in the same individual, or bilateral breast cancer with the first case <50 years (6)	37	32
Two cases of ovarian cancer, both <50 years, in first- or second-degree relatives (7)	72	59
Two cases of ovarian cancer, any age, in first- or second-degree relatives (8)	1	1
Three or more cases of breast or ovarian cancer at any age (10)	109	92
	287	246

The risk categories for individuals eligible for screening for a genetic susceptibility to breast or ovarian cancers are determined by the Ontario Ministry of Health and Long Term-Care Referral Criteria for Genetic Counseling.

<sup>a</sup>Numbers in parentheses correspond to eligibility group designation.

noncoding effects (12 SRFBS, two cryptic site strengthening; one variant prioritized for both). These variants can be found in Supp. Table S15 (where “coverage” column contains two or more coverage values).

### Negative control

*ATP8B1* was sequenced and analyzed in all patients as a negative control (Supp. Table S16). We prioritized 21 *ATP8B1* variants with an average of 0.22 variants/kb and 0.57 variants/patient. This is lower than the prioritization rate for many of the documented HBOC genes. This result illustrates that the proposed method represents a screening rather than a diagnostic approach, as some variants may be incorrectly prioritized.

### Pedigree Analysis

Pathogenic *BRCA2* variants within a region of exon 11 have been associated with a high incidence of OC. We therefore verified whether there were a high number of OC cases in the families of patients prioritized with exon 11 *BRCA2* variants ( $N = 3$ ). The family of the patient with *BRCA2* variant c.4828G>A (p.Val1610Met; diagnosed with BC at 65) has three reported cases of BC/OC, one of which is OC (diagnosed at 74), two degrees of separation from the proband. The patient with *BRCA2*:c.6317T>C (p.Leu2106Pro; diagnosed with BC at 52) has three other affected family members, two with OC and one with BC. Finally, four patients found to have the *BRCA2* variant c.5199C>T (p.Ser1733 = ) do not have any family members with reported cases of OC.

We also selected patients with prioritized mismatch repair variants ( $N = 8$  in 10 patients) to assess the incidence of reported CRC cases in these families. Notably, the patient with mutation *MSH2*:c.1748A>G (p.Asn583Ser) had five relatives with CRC. A similar analysis of prioritized *CDH1* variants did not reveal any patients with a family history of gastric cancer.

### LR Analyses

We carried out co-segregation analysis of 25 patients with prioritized pathogenic variants (four nonsense, four frameshift, two in-frame deletions, six missense, four natural splicing, and six cryptic splicing; including a patient who exhibited prioritized natural and cryptic SS variants). We compared these findings with those from patients ( $N = 25$ ) harboring moderate-priority variants (variants prioritized through IT analysis only) and those in whom no variants were flagged or prioritized ( $N = 14$ ). In instances where disease alleles could be transmitted through either founder parent, the lineage with the highest likelihood ratio (LR) was reported. For patients with likely pathogenic variants, the LRs ranged from

0.00 to 70.96 (Table 6 and Supp. Table S17). Disease co-segregation was supported ( $LR > 1.0$ ) in 18 patients, and the remainder were either neutral ( $LR < 1.0$  [Mohammadi et al., 2009]) or could not be analyzed either due to missing pedigree information or limited numbers of affected individuals in a family. Patient 10–6F (*PALB2*: c.757\_758delCT) exhibited the highest likelihood ( $LR = 70.96$ ). Prioritized variants with neutral evidence include a variant that abolishes a natural SS in *MRE11A*, c.2070+2T>A ( $LR = 0.03$ ), and an in-frame deletion c.483\_485delAGA in *CHEK2* ( $LR = 0.00$ ).

## Discussion

Rare noncoding and/or non-truncating mutations can confer an increased risk of disease in BC [Tavtigian et al., 2009]. This study determined both coding and noncoding sequences of 20 HBOC-related genes, with the goal of discovering and prioritizing rare variants with potential effects on gene expression. This work emphasizes results from the analysis of noncoding variants, which are abundant in these genes, yet have been underrepresented in previous HBOC mutation analyses. Nevertheless, alterations to mRNA binding sites in *BRCA*, and lower risk or rare HBOC genes, have been shown to contribute to HBOC (exonic splicing enhancers (ESEs) in *ATM* [Heikkinen et al., 2005], *BARD1* [Ratajska et al., 2011], and *BRCA* genes [Gochhait et al., 2007; Sanz et al., 2010]). We prioritized 346 unique variants that were predicted to result in four nonsense, three frameshift, two indels with preserved reading frame, 119 missense, four natural splicing, six cryptic splicing, 17 pseudoexon activating, 148 SRFBS, 36 TFBS, three UTR structure, and 31 RBBS mutations (Supp. Table S14). Among these variants, 101 were novel (see Supp. Table S18 for references to previously identified variants). Compared to our initial seven-gene panel (Mucaki et al., submitted), the inclusion of the additional genes in this study prioritized at least one variant in 15% additional patients (increased from 70.6 to 85.7%).

The *BRCA* genes harbor the majority of known germline pathogenic variants for HBOC families [Chong et al., 2014]. However, a large proportion of the potentially pathogenic variants identified in our study were detected in *ATM*, *PALB2*, and *CHEK2*, which, although of lower penetrance, were enriched because the eligibility criteria excluded known *BRCA1* and *BRCA2* carriers. *BRCA1* and *BRCA2* variants were nevertheless prioritized in some individuals. We also had expected intragenic clustering of some *BRCA* coding variants [Mucaki et al., 2011]. For example, pathogenic variants occurring within exon 11 of *BRCA2* are known to be associated with higher rates of OC in their families [Lubinski et al., 2004]. We identified three variants in exon 11; however, there was no evidence of OC in these families. Overall, *ATM* and *PALB2* had the highest number of prioritized variants (61 and 26, respectively). However, only 12 variants were prioritized in *CHEK2*; potentially pathogenic variants

**Table 6. LR Values for Patients with Prioritized Truncating, Splicing, and Selected Missense Variants**

Genes <sup>a</sup>	Variant		Category	UWO ID	LR
	mRNA	Protein			
<i>ATM</i>	c.1924G>T	p.Glu642Ter	Nonsense	10-2F	7.46 <sup>MGM</sup> , 9.61 <sup>MGF</sup>
	c.6198+1G>A	–	Natural splicing	8-1D.9-1B	1.00
	c.7271T>G	p.Val2424Gly	Missense	10-1F	1.44
<i>BARD1</i>	c.8977C>T	p.Arg2993Ter	Nonsense	12-1D	1.96 <sup>P</sup>
	c.1454C>T	p.Arg2993Ter	Nonsense	12-4G.13-5D	5.30 <sup>P</sup>
<i>BRCA1</i>	c.3550_3551insA	p.Ala485Val	Cryptic splicing	8-1D.9-1B	1.00
	c.5503C>T	p.Gly1184Glufs	Frameshift indel	11-6H	3.36 <sup>P</sup>
<i>BRCA2</i>	c.10095delCins11	p.Arg1835Ter	Nonsense	8-5D.9-5D	41.99
	c.30_32delGCT	p.Ser3366Asnfs	Frameshift indel	15-4E	3.71
<i>CDH1</i>	c.1223C>G	p.Leu11del	Inframe deletion	10-4A	1.00
	c.470T>C	p.Ala408Gly	Cryptic splicing	15-3G	2.14
<i>CHEK2</i>	c.433C>T	p.Ile157Thr	Missense	12-2G	2.86
	c.3549C>G	p.Arg145Trp	Missense	15-5G	19.44 <sup>P</sup>
<i>PALB2</i>	c.757_758delCT	p.Tyr1183Ter	Nonsense	4-3C.5-4G.14-4A	3.48
	c.2920_2921delAA	p.Leu253Ilefs	Frameshift indel	15-1E	1.78
	c.2T>C	p.Lys974Glufs	Frameshift indel	10-6F	70.96
<i>PMS2</i>	c.2T>C	p.Lys974Glufs	Frameshift indel	8-3A.9-3A	5.03
<i>RAD51B</i>	c.84G>A	p.Met1Thr	Missense	11-4H	16.53 <sup>P</sup>
	c.958-29A>T	p.Gln28=	Leaky splicing	8-1H.9-1E	3.51 <sup>P</sup>
<i>STK11</i>	c.375-194GT>AC	–	Cryptic splicing	10-4B	7.44 <sup>P</sup>
			Cryptic splicing	10-5A	2.67 <sup>M</sup>

LR values in favor of neutrality are not shown.

<sup>a</sup>RefSeq accessions: *ATM* (NM\_000051.3), *BARD1* (NM\_000465.2), *BRCA1* (NM\_007294.2), *BRCA2* (NM\_000059.3), *CDH1* (NM\_004360.3), *CHEK2* (NM\_007194.3), *PALB2* (NM\_024675.3), *PMS2* (NM\_000535.5), *RAD51B* (NM\_002877.4), *STK11* (NM\_000455.4).

P, paternal; M, maternal; MGF, maternal grandfather; MGM, maternal grandmother.

may have been underrepresented during sequence alignment as a consequence of the known paralogy with *CHEK2P2*.

Fewer *TP53*, *STK11*, and *PTEN* variants were prioritized, as pathogenic variants in these genes tend to be infrequent in patients who do not fulfill the clinical criteria for their associated syndromes (Li-Fraumeni syndrome, Peutz-Jeghers syndrome, and Cowden syndrome, respectively [Hollestelle et al., 2010]), although they have been indicated as near moderate to high-risk genes in BC [Easton et al., 2015]. This underrepresentation of prioritized variants may be supported by the negative Residual Variation Intolerance Scores (RVIS) for these genes [Petrovski et al., 2013], which are likely indicative of purifying selection. Although the density of prioritized variants in these genes is below average (18.2 per gene), the total number was nonetheless important (*TP53* = 11, *STK11* = 12, *PTEN* = 15).

The fundamental difference between IT and other approaches such as combined annotation-dependent depletion [CADD; Kircher et al., 2014] is that IT depends only on positive experimental data from the same or closely related species. CADD does not appear to account for unobserved reversions or other hidden mutations [e.g. perform a Jukes–Cantor correction; Jukes and Cantor, 1969], nor are the effects of these simulated. Furthermore, the CADD scoring system is *ad hoc*, which contrasts with strong theoretical basis on the IT approach (Rogan and Schneider, 1995) in which information changes in bits represent a formally proven relationship to thermodynamic stability, and therefore can be used to measure binding affinity. This makes it different from other unitless methods with unknown distributions, in which differences in binding affinity cannot be accurately extrapolated from derived scores.

We compared the frequency of all prioritized variants in our patient cohort to the population allele frequencies [1000 Genomes Project, Phase 3; <http://www.1000genomes.org>; 1000 Genomes Project Consortium, 2012] to determine if variants more common in our cohort might be suggestive of HBOC association. Three variants in at least five HBOC patients are present at a much lower

frequency in the general population than in our HBOC population. *NBN* c.\*2129G>T, present in 4.18% of study cohort, is considerably rarer globally (0.38% in 1000 Genomes; <0.1% in other populations). Similarly, the *RAD51B* c.-3077G>T variant (2.09%) is rare in the general population (0.08%). Interestingly, *BARD1* c.33G>T (1.74% of study cohort) has only been reported in the American and European populations in 1000 Genomes (0.29% and 0.20%, respectively) and only Europeans in the Exome Variant Server (0.24%; <http://evs.gs.washington.edu/EVS/>). In Southwestern Ontario, individuals are often of American or European ancestry. The variant was found to be more common in the Exome Aggregation Consortium (ExAC; <http://exac.broadinstitute.org/>) in 1.17% tested Finnish population (0.41% in their non-Finnish European cohort), though no alleles were found in the Finnish populations in 1000 Genomes (*N* = 99). Therefore, the allele frequency of this *BARD1* variant in our HBOC population may simply be enriched in a founder subset of general populations. While we cannot rule out skewing of these allele frequencies due to population stratification, our findings suggest that gene expression levels could be impacted by these variants.

We applied subpopulation allele frequency analysis for all of our prioritized variants. Supp. Table S19 lists the 49 variants that have allele frequencies >1% in various subpopulations (based on dbSNP). Allele frequencies were as high as 4.2% for the *BRCA2* c.-40+192C>T (8-1G.9-1C), predicted to affect TF binding, in the East Asian subpopulation. Without additional information on patient ethnicities, it is not possible to eliminate prioritized variants that are common in specific subpopulations.

Co-segregation analysis is recommended by the American College of Medical Genetics and Genomics for variant classification [Richards et al., 2015]. Among patients with likely pathogenic, highly penetrant mutations in our cohort (*N* = 24), some variants had LR values consistent with causality, whereas others provided little evidence to support co-segregation among family members (Table 6 and Supp. Table S17). An important caveat, however, was



that the use of *BRCA2* penetrance values in non-*BRCA* genes may have resulted in underestimates of LR values.

In order to evaluate the application of co-segregation analysis in the context of this study, we chose to perform this analysis on patients with moderate priority variants (i.e., variants affecting binding sites) and patients with no flagged or prioritized variants ( $N = 25$  and  $14$ , respectively). LRs ranged from 0.0034 to 78.0 for moderate-priority variants and from 0.0005 to 57.0 for patients with no flagged or prioritized variants (Supp. Fig. S2). The proportion of LR values supporting neutrality and those supporting causation was comparable between patients with prioritized, moderately prioritized, and flagged variants (Supp. Fig. S2). This suggests that co-segregation analysis is only useful in the context of other supporting results for assessing pathogenicity (e.g., likelihood of being pathogenic or benign). Furthermore, the lack of genotype information and at times smaller pedigrees likely also contributed to the lack of concordance between LRs and variant priority.

A small number of patients with a known pathogenic variant carried other prioritized variants. These were likely benign or possibly phenotypic modifiers. One patient possessed five prioritized variants (one missense, one SRFBS, one TFBS, and two RBBs) in addition to a *BRCA1* nonsense mutation (c.5503C>T). While these variants may not directly contribute to causing HBOC, they may act as a risk modifier and alter expression levels [Antoniou and Easton, 2006].

Similarly, genes lacking association with HBOC can be used as a metric for determining a false-positive rate of variant prioritization. In this study, we prioritized 21 *ATP8B1* variants among 37 of our HBOC patients (Supp. Table S16) despite it having not been previously associated with any type of cancer. A variant with a deleterious effect on *ATP8B1* may lead to *ATP8B1*-related diseases, such as progressive familial intrahepatic cholestasis [Gonzales et al., 2014], but should not increase the chances of developing BC. Thus, while our framework may be effective at prioritizing variants, only genes with previous association to a disease should be included in analyses similar to the present study to minimize falsely prioritized variants.

Additional workup of prioritized noncoding and non-*BRCA* variants is particularly important, because with few exceptions [Easton et al., 2015], the pathogenicity of many of the genes and variants has not been firmly established. Furthermore, mutations in several of these genes confer risk to other types of cancer, which alters the management of these patients [Knappskog and Lønning, 2012]. The next step toward understanding the role these prioritized variants play in HBOC is to test family members of the proband and to carry out functional analysis. If this is not possible, then their effects on gene expression could be evaluated using assays for RNA stability and RNA localization. Protein function could be evaluated by binding site assays, protein activity, and quantitative PCR.

A significant challenge associated with VUS analysis, particularly in the case of many of these recent HBOC gene candidates, is the underreporting of variants and thus positive findings tend to be over-represented in the literature [Kraft, 2008]. Hollestelle et al. [2010] argue that a more stringent statistical standard must be applied (i.e.,  $P$ -values of 0.01 should be used as opposed to 0.05) to underreported variants (namely in moderate-risk alleles), because of failure to replicate pathogenic variants, which we have also found [Viner et al., 2014]. In the same way that we use IT-based analysis to justify prioritizing variants for further investigation, variants that are disregarded as lower priority (and that are likely not disease causing) have been subjected to the same thresholds and criteria. Integrating this set of labeled prioritized and flagged, often rare variants from this cohort of *BRCA*-negative HBOC patients, to findings from

exome or gene panel studies of HBOC families should accelerate the classification of some VUS.

Different variant interpretation and reporting guidelines consider the reporting of VUS to be either optional or essential [Wallis et al., 2013; Richards et al., 2015]. In all cases, a reported VUS cannot be the basis for a clinical decision and should be followed up and further investigated. In any case, the number of reported VUS in an individual is frequently too extensive for detailed characterization. Reducing the full set of variants obtained by complete gene sequencing to a prioritized list will be an essential prerequisite for targeting potentially clinically relevant information. Informing patients of prioritized VUS may increase patient accrual and participation [Murphy et al., 2008]. However, it will be critical to explain both the implications and significance of prioritization and the limitations, namely counselling patients to avoid clinical decisions, based on this information [Vos et al., 2012].

## Acknowledgments

We would like to thank the patients who participated; Karen Panabaker for oversight in assessing patient eligibility, directing patient recruitment, and follow-up counseling; and Dr. Peter Ainsworth and Alan Stuart for access to genomic DNA samples and robotic equipment in the MGL at London Health Sciences Centre. Abby Watts-Dickens assisted in patient recruitment. 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 371758–2009). N. G. C. was supported by CIHR Strategic Training Program in Cancer Research and Technology Transfer and the Pamela Greenaway-Kohlmeier Translational Breast Cancer Research Unit. Access to the Shared Hierarchical Academic Research Computing Network (SHARCNET) and Compute/Calcul Canada is gratefully acknowledged.

## Disclosure statement

P. K. R. is the inventor of US Patent 5,867,402 and other patents pending, which underlie the prediction and validation of mutations. He and J. H. M. K. founded Cytogenomix, Inc., which is developing software based on this technology for complete genome or exome mutation analysis.

## References

- 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.
- Al Bakir M, Gabra H. 2014. The molecular genetics of hereditary and sporadic ovarian cancer: implications for the future. *Br Med Bull* 112:57–69.
- Antoniou AC, Easton DF. 2006. Models of genetic susceptibility to breast cancer. *Oncogene* 25:5898–5905.
- Antoniou A, Pharoah PDP, Narod S, Risch HA, Eyfjord JE, Hopper JL, Loman N, Olsson H, Johannsson O, Borg A, Pasini B, Radice P, et al. 2003. Average risks of breast and ovarian cancer associated with *BRCA1* or *BRCA2* mutations detected in case series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet* 72:1117–1130.
- Apostolou P, Fostira F. 2013. Hereditary breast cancer: the era of new susceptibility genes. *BioMed Res Int* 2013:e747318.
- Biswas DK, Shi Q, Baily S, Strickland I, Ghosh S, Pardee AB, Iglehart JD. 2004. NF-kappa B activation in human breast cancer specimens and its role in cell proliferation and apoptosis. *Proc Natl Acad Sci USA* 101:10137–10142.
- Caminsky NG, Mucaki EJ, Rogan PK. 2015. Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. *F1000Res* 3:282.
- Cancer Genome Atlas Network. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* 490:61–70.



- Cassa CA, Savage SK, Taylor PL, Green RC, McGuire AL, Mandl KD. 2012. Disclosing pathogenic genetic variants to research participants: quantifying an emerging ethical responsibility. *Genome Res* 22:421–428.
- Chenevix-Trench G, Spurdle AB, Gatei M, Kelly H, Marsh A, Chen X, Donn K, Cummings M, Nyholt D, Jenkins MA, Scott C, Pupo GM, et al. 2002. Dominant negative ATM mutations in breast cancer families. *J Natl Cancer Inst* 94:205–215.
- Chong HK, Wang T, Lu H-M, Seidler S, Lu H, Keiles S, Chao EC, Stuenkel AJ, Li X, Elliott AM. 2014. The validation and clinical implementation of BRCAplus: a comprehensive high-risk breast cancer diagnostic assay. *PLoS One* 9:e97408.
- DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Angel G del, Rivas MA, Hanna M, McKenna A, Fennell TJ, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–498.
- Ding H, Mao C, Li S-M, Liu Q, Lin L, Chen Q. 2011. Lack of association between ATM C.1066-6T>G mutation and breast cancer risk: a meta-analysis of 8,831 cases and 4,957 controls. *Breast Cancer Res Treat* 125:473–477.
- Dorman SN, Shirley BC, Knoll JHM, Rogan PK. 2013. Expanding probe repertoire and improving reproducibility in human genomic hybridization. *Nucleic Acids Res* 41:e81.
- Dorman S, Viner C, Rogan P. 2014. Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer. *Sci Rep* 4:7063.
- Drost M, Koppejan H, Wind N de. 2013. Inactivation of DNA mismatch repair by variants of uncertain significance in the PMS2 gene. *Hum Mutat* 34:1477–1480.
- Duzkale H, Shen J, McLaughlin H, Alfares A, Kelly MA, Pugh TJ, Funke BH, Rehm HL, Lebo MS. 2013. A systematic approach to assessing the clinical significance of genetic variants. *Clin Genet* 84:453–463.
- Easton DF, Pharoah PDP, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, Devilee P, Meindl A, Couch FJ, Southey M, Goldgar DE, Evans DGR, et al. 2015. Gene-panel sequencing and the prediction of breast-cancer risk. *N Engl J Med* 372:2243–2257.
- ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.
- Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, Bishop DT, Weber B, Lenoir G, Chang-Claude J, Sobol H, Teare MD, et al. 1998. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am J Hum Genet* 62:676–689.
- Gadiraju S, Vyhldal CA, Leeder JS, Rogan PK. 2003. Genome-wide prediction, display and refinement of binding sites with information theory-based models. *BMC Bioinformatics* 4:38.
- Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 27:182–189.
- Gochhait S, Bukhari SIA, Bairwa N, Vadhera S, Darvishi K, Raish M, Gupta P, Husain SA, Bamezai RNK. 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:R71.
- Goldgar DE, Healey S, Dowty JG, Da Silva L, Chen X, Spurdle AB, Terry MB, Daly MJ, Buys SM, Southey MC, Andrulis I, John EM, et al. 2011. Rare variants in the ATM gene and risk of breast cancer. *Breast Cancer Res* 13:R73.
- Gonzales E, Spraul A, Jacquemin E. 2014. Clinical utility gene card for: progressive familial intrahepatic cholestasis type 1. *Eur J Hum Genet* 22(4). doi:10.1038/ejhg.2013.188.
- Halvorsen M, Martin JS, Broadaway S, Laederach A. 2010. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet* 6:e1001074.
- Heikkinen K, Karppinen S-M, Soini Y, Mäkinen M, Winqvist R. 2003. Mutation screening of Mre11 complex genes: indication of RAD50 involvement in breast and ovarian cancer susceptibility. *J Med Genet* 40:e131.
- Heikkinen K, Rapakko K, Karppinen S-M, Erkko H, Nieminen P, Winqvist R. 2005. Association of common ATM polymorphism with bilateral breast cancer. *Int J Cancer* 116:69–72.
- Hollstelle A, Wasielewski M, Martens JW, Schutte M. 2010. Discovering moderate-risk breast cancer susceptibility genes. *Curr Opin Genet Dev* 20:268–276.
- Howlander N, Noone AM, Krapcho M, Garshell J, Miller D, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). 2014. SEER Cancer Statistics Review, 1975–2011, National Cancer Institute. Bethesda, MD. [http://seer.cancer.gov/csr/1975\\_2011/](http://seer.cancer.gov/csr/1975_2011/), based on November 2013 SEER data submission, posted on the SEER web site April 2014.
- Janatova M, Soukupova J, Stribrna J, Kleiblova P, Vocka M, Boudova P, Kleibl Z, Pohlreich P. 2015. Mutation analysis of the RAD51C and RAD51D genes in high-risk ovarian cancer patients and families from the Czech Republic. *PLoS One* 10:e0127711.
- Jones S, Emmerson P, Maynard J, Best JM, Jordan S, Williams GT, Sampson JR, Cheadle JP. 2002. Biallelic germline mutations in MYH predispose to multiple colorectal adenoma and somatic G:C→T:A mutations. *Hum Mol Genet* 11:2961–2967.
- Jukes T, Cantor C. 1969. Evolution of protein molecules. In: Munro HN, editor. *Mammalian protein metabolism III*. New York: Academic Press. p 21–132.
- Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46:310–315.
- Knappskog S, Lønning PE. 2012. P53 and its molecular basis to chemoresistance in breast cancer. *Expert Opin Ther Targets* 16(Suppl 1):S23–S30.
- von Kodolitsch Y, Berger J, Rogan PK. 2006. Predicting severity of haemophilia A and B splicing mutations by information analysis. *Haemophilia* 12:258–262.
- Kraft P. 2008. Curses—winner’s and otherwise—in genetic epidemiology. *Epidemiology* 19:649–651; discussion 657–658.
- Lee SB, Kim SH, Bell DW, Wahrer DC, Schiripio TA, Jorczak MM, Sgroi DC, Garber JE, Li FP, Nichols KE, Varley JM, Godwin AK, et al. 2001. Destabilization of CHK2 by a missense mutation associated with Li-Fraumeni Syndrome. *Cancer Res* 61:8062–8067.
- Ligtenberg MJL, Kuiper RP, Chan TL, Goossens M, Hebeda KM, Voorendt M, Lee TYH, Bodmer D, Hoenselaar E, Hendriks-Cornelissen SJB, Tsui WY, Kong CK, et al. 2009. Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. *Nat Genet* 41:112–117.
- Lu R, Mucaki EJ, Rogan PK. Preprint. Discovery of primary, cofactor, and novel transcription factor binding site motifs by recursive, thresholded entropy minimization. bioRxiv. doi: <http://dx.doi.org/10.1101/042853>.
- Lubinski J, Phelan CM, Ghadirian P, Lynch HT, Garber J, Weber B, Tung N, Horsman D, Isaacs C, Monteiro ANA, Sun P, Narod SA. 2004. Cancer variation associated with mutation in the BRCA2 gene. *Fam Cancer* 3:1–10.
- Maxwell KN, Domchek SM. 2013. Familial breast cancer risk. *Curr Breast Cancer Rep* 5:170–182.
- Minion LE, Dolinsky JS, Chase DM, Dunlop CL, Chao EC, Monk BJ. 2015. Hereditary predisposition to ovarian cancer, looking beyond BRCA1/BRCA2. *Gynecol Oncol* 137:86–92.
- Mohammadi L, Vreeswijk MP, Oldenburg R, Ouweland A van den, Oosterwijk JC, Hout AH van der, Hoogerbrugge N, Ligtenberg M, Ausems MG, Luijck RB van der, Dommering CJ, Gille JJ, et al. 2009. A simple method for co-segregation analysis to evaluate the pathogenicity of unclassified variants; BRCA1 and BRCA2 as an example. *BMC Cancer* 9:211.
- Mucaki EJ, Ainsworth P, Rogan PK. 2011. Comprehensive prediction of mRNA splicing effects of BRCA1 and BRCA2 variants. *Hum Mutat* 32:735–742.
- Mucaki EJ, Caminsky NG, Perri AM, Lu R, Laederach A, Halvorsen M, Knoll JHM, Rogan PK. *BMC Medical Genomics*, in press. A unified analytic framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer. bioRxiv. doi: <http://dx.doi.org/10.1101/031419>.
- Mucaki EJ, Shirley BC, Rogan PK. 2013. Prediction of mutant mRNA splice isoforms by information theory-based exon definition. *Hum Mutat* 34:557–565.
- Murphy J, Scott J, Kaufman D, Geller G, LeRoy L, Hudson K. 2008. Public expectations for return of results from large-cohort genetic research. *Am J Bioeth* 8:36–43.
- Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 9:e1003709.
- Ratajska M, Antoszewska E, Piskorz A, Brozek I, Borg Å, Kusmirek H, Biernat W, Limon J. 2011. Cancer predisposing BARD1 mutations in breast-ovarian cancer families. *Breast Cancer Res Treat* 131:89–97.
- Reiman A, Srinivasan V, Barone G, Last JJ, Wootton LL, Davies EG, Verhagen MM, Willemsen MA, Weemaes CM, Byrd PJ, Izatt L, Easton DF, et al. 2011. Lymphoid tumours and breast cancer in ataxia telangiectasia; substantial protective effect of residual ATM kinase activity against childhood tumours. *Br J Cancer* 105: 586–591.
- Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, et al. 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* 17:405–424.
- Rogan PK, Faux BM, Schneider TD. 1998. Information analysis of human splice site mutations. *Hum Mutat* 12:153–171.
- Rogan PK, Schneider TD. 1995. Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. *Hum Mutat* 6:74–76.
- Rogan PK, Svojanovsky S, Leeder JS. 2003. Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. *Pharmacogenetics* 13:207–218.
- Rogan PK, Zou GY. 2013. Best practices for evaluating mutation prediction methods. *Hum Mutat* 34:1581–1582.
- Santagata S, Hu R, Lin NU, Mendillo ML, Collins LC, Hankinson SE, Schnitt SJ, Whitesell L, Tamimi RM, Lindquist S, Ince TA. 2011. High levels of nuclear heat-shock factor 1 (HSF1) are associated with poor prognosis in breast cancer. *Proc Natl Acad Sci USA* 108:18378–18383.

- Santos C, Peixoto A, Rocha P, Pinto P, Bizarro S, Pinheiro M, Pinto C, Henrique R, Teixeira MR. 2014. Pathogenicity evaluation of BRCA1 and BRCA2 unclassified variants identified in Portuguese breast/ovarian cancer families. *J Mol Diagn* 16:324–334.
- Sanz DJ, Acedo A, Infante M, Durán M, Pérez-Cabornero L, Esteban-Cardenosa E, Lastra E, Pagani F, Miner C, Velasco EA. 2010. A high proportion of DNA variants of BRCA1 and BRCA2 is associated with aberrant splicing in breast/ovarian cancer patients. *Clin Cancer Res* 16:1957–1967.
- Schneider TD. 1997. Information content of individual genetic sequences. *J Theor Biol* 189:427–441.
- Schrader KA, Masciari S, Boyd N, Salamanca C, Senz J, Saunders DN, Yorida E, Maines-Bandiera S, Kaurah P, Tung N, Robson ME, Ryan PD, et al. 2011. Germline mutations in CDH1 are infrequent in women with early-onset or familial lobular breast cancers. *J Med Genet* 48:64–68.
- Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, Chagtai T, Jayatilake H, Ahmed M, Spanova K, North B, McGuffog L, et al. 2006. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat Genet* 38:1239–1241.
- Senter L, Clendenning M, Sotamaa K, Hampel H, Green J, Potter JD, Lindblom A, Lagerstedt K, Thibodeau SN, Lindor NM, Young J, Winship I, et al. 2008. The clinical phenotype of Lynch syndrome due to germ-line PMS2 mutations. *Gastroenterology* 135:419–428.
- Shirley BC, Mucaki EJ, Whitehead T, Costea PI, Akan P, Rogan PK. 2013. Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences. *Genomics Proteomics Bioinformatics* 11:77–85.
- Stankovic T, Kidd AM, Sutcliffe A, McGuire GM, Robinson P, Weber P, Bedenham T, Bradwell AR, Easton DF, Lennox GG, Haines N, Byrd PJ, et al. 1998. ATM mutations and phenotypes in ataxia-telangiectasia families in the British Isles: expression of mutant ATM and the risk of leukemia, lymphoma, and breast cancer. *Am J Hum Genet* 62:334–345.
- Stratton JF, Pharoah P, Smith SK, Easton D, Ponder BA. 1998. A systematic review and meta-analysis of family history and risk of ovarian cancer. *Br J Obstet Gynaecol* 105:493–499.
- Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F, Byrnes GB, Chuang S-C, Forey N, Feuchtinger C, Gioia L, et al. 2009. Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer. *Am J Hum Genet* 85:427–446.
- Théry JC, Krieger S, Gaildrat P, Révillion F, Buisine M-P, Killian A, Duponchel C, Rousselin A, Vaur D, Peyrat J-P, Berthet P, Frébourg T, et al. 2011. Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. *Eur J Hum Genet* 19:1052–1058.
- Tournier I, Vezain M, Martins A, Charbonnier F, Baert-Desurmont S, Olschwang S, Wang Q, Buisine MP, Soret J, Tazi J, Frébourg T, Tosi M. 2008. A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. *Hum Mutat* 29:1412–1424.
- Vidal LJ-P, Perry JK, Vouyovitch CM, Pandey V, Brunet-Dunand SE, Mertani HC, Liu D-X, Lobie PE. 2010. PAX5alpha enhances the epithelial behavior of human mammary carcinoma cells. *Mol Cancer Res* 8:444–456.
- Vihinen M. 2013. Guidelines for reporting and using prediction tools for genetic variation analysis. *Hum Mutat* 34:275–282.
- Viner C, Dormann SN, Shirley BC, Rogan PK. 2014. Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. *F1000Research* 3:8.
- Vos J, Gómez-García E, Oosterwijk JC, Menko FH, Stoel RD, AsperenCJ van, Jansen AM, Stiggelbout AM, Tibben A. 2012. Opening the psychological black box in genetic counseling. The psychological impact of DNA testing is predicted by the counselees' perception, the medical impact by the pathogenic or uninformative BRCA1/2-result. *Psychooncology* 21:29–42.
- Wallis Y, Payne S, McAnulty C, Bodmer D, Sistermans E, Robertson K, Moore D, Abbs S, Deans Z, Devereau A. 2013. Practice guidelines for the evaluation of pathogenicity and the reporting of sequence variants in clinical molecular genetics. Association for Clinical Genetic Science and Dutch Society of Clinical Laboratory Specialists
- Walsh T, Casadei S, Lee MK, Pennil CC, Nord AS, Thornton AM, Roeb W, Agnew KJ, Stray SM, Wickramanayake A, Norquist B, Pennington KP, et al. 2011. Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. *Proc Natl Acad Sci USA* 108:18032–18037.
- Zahnaw CA. 2009. CCAAT/enhancer-binding protein beta: its role in breast cancer and associations with receptor tyrosine kinases. *Expert Rev Mol Med* 11:e12.