ACCOMPLISHMENTS, COLLABORATIVE PROJECTS AND FUTURE INITIATIVES IN BREAST CANCER GENETIC PREDISPOSITION

EDITED BY: Paolo Peterlongo, Nandita Mitra and Luis G. Carvajal-Carmona

frontiers Research Topics

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

ACCOMPLISHMENTS, COLLABORATIVE PROJECTS AND FUTURE INITIATIVES IN BREAST CANCER GENETIC PREDISPOSITION

Topic Editors:
Paolo Peterlongo, IFOM - The FIRC Institute for Molecular Oncology, Italy
Nandita Mitra, University of Pennsylvania, United States
Luis G. Carvajal-Carmona, University of California Davis, Davis, United States

In this eBook, we described the accomplishments, collaborative projects and future initiatives in the field of breast cancer genetic predisposition. More specifically, the articles included focused on aspects such as mutation screening in unexplored populations, identification and characterization of novel predisposing genes and mutations, and population screening.

Table of Contents

Editorial: Accomplishments, Collaborative Projects and Future Initiatives in Breast Cancer Genetic Predisposition

Paolo Peterlongo[1*] and Luis G. Carvajal-Carmona[2,3,4]

[1] Genome Diagnostics Program, IFOM The FIRC Institute for Molecular Oncology, Milan, Italy, [2] Genome Center, University of California, Davis, Davis, CA, United States, [3] Department of Biochemistry and Molecular Medicine, School of Medicine, University of California, Davis, Sacramento, CA, United States, [4] Population Sciences and Health Disparities Program, University of California Davis Comprehensive Cancer Center, Sacramento, CA, United States

Editorial on the Research Topic

Accomplishments, Collaborative Projects and Future Initiatives in Breast Cancer Genetic Predisposition

Since the discovery of breast cancer genes *BRCA1* and *BRCA2 (BRCA1/2)* over two decades ago, much has been accomplished in the field of breast cancer genetic predisposition. On one hand, novel genes harboring rare pathogenic variants, most of which act in the same *BRCA1/2* pathway, causing increasing disease risk have been identified. In addition, several single nucleotide polymorphisms (SNPs) that modify the breast cancer risk in individuals with *BRCA1/2* mutations are now known. These moderate-to-high penetrant genetic variants now represent key elements for improving risk prediction in familial cases. On the other hand, hundreds of common low-risk SNPs have been discovered and can be incorporated into prediction models to improve the identification of women at risk of breast cancer in the general population. Moreover, multifactorial analyses, family studies and high throughput functional assays have been developed to validate candidate genes, classify the variants of uncertain significance (VUS) detected by gene-panel next generation sequencing in clinical and research settings, and to measure the risk magnitude conferred by known pathogenetic variants. Articles in the present Frontiers in Oncology e-book explore these aspects of breast cancer predisposition further.

Individuals who carry *BRCA1/2* pathogenic variants have an average cumulative risk of developing breast cancer, by age 80 years, of ∼70% (1). Thanks to the efforts of the collaborators of the *PALB2* Interest Group (http://www.palb2.org/), *PALB2* is now considered the third high-risk gene with pathogenic variants associated with 44% lifetime risk of developing breast cancer (2). The moderate-penetrance genes *ATM* and *CHEK2* are also associated with breast cancer, conferring a 20% average lifetime risk (3, 4). More recently, *BARD1, RAD51D, BRIP1,* and *RAD51C* have been proposed as risk factors for triple-negative breast cancer [TNBC; (5)], indicating that the risk associated with pathogenic variants in each gene may vary by tumor subtype. Support for this hypothesis is the latest emerging breast cancer gene *FANCM* which has also shown to confer a higher risk for TNBC (6–8). All these genetic factors explain only about half of the familial cases, hence novel breast cancer genes or alleles are yet to be detected (9). In this e-book, the impact of *BRCA1/2* mutations and of novel genes was investigated in unexplored populations and in breast cancer progression. Solano et al. studied the *BRCA1/2* mutation spectra in high-risk Ashkenazi Jewish population from Argentina. They reported that, in addition to carriers of known Ashkenazi

founder mutations, up to 7% of tested individuals were positive for other *BRCA1/2* pathogenic variants. In a second study, Torrezan et al. performed whole exome sequencing in Brazilian breast cancer probands that were negative for causal variants in most known predisposition genes. Beside a very rare and novel pathogenic variants in *ATM* and *BARD1*, respectively, the authors found rare and possibly damaging variants in several candidate genes. Tang et al. aimed at the identification of genes associated with the progression of breast cancer. The authors developed a free-scale gene co-expression networks to explore associations between gene sets and clinical features, and to identify candidate biomarkers. Breast cancer is not exclusively a female disease and about 1% of all cases arise in males. As reviewed by Rizzolo et al., 13% of male breast cancer (MBC) are due to pathogenic variants in *BRCA2* while *CHEK2* and *PALB2* account for a smaller proportion of cases. In their study these authors suggest that monoallelic mutation of *MUTYH* gene, which cause the recessive MUTYH-associated polyposis (MAP) syndrome may also cause MBC.

The identification of common SNPs associated with breast cancer risk and of those that modify the breast cancer risk in individuals with *BRCA1/2* pathogenic variants is the most significant success of the Breast Cancer Association Consortium (BCAC, http://bcac.ccge.medschl.cam.ac.uk/) and of the Consortium of Investigators of Modifiers of *BRCA1/2* (CIMBA, http://cimba.ccge.medschl.cam.ac.uk/). A meta-analysis of genome wide association studies (GWAS), which combined data from 122,977 breast cancer cases and 105,974 controls, resulted in the identification of a total of 172 risk-associated SNPs explaining about half of the familial relative risk [the risk of first-degree relatives of breast cancer patients of developing the disease; (10)]. Ten additional SNPs were found by the analyses of breast cancer cases with estrogen receptor (ER) negative tumors (11), bringing the total number of known breast cancer SNPs to 182 SNPs. While individually these risk factors are not clinically relevant, they can be combined into polygenic risk scores (PRS) that can be predictive of cancer risk in *BRCA1/2* mutation carriers and in the general population (1, 12). As discussed in this e-book, in a review by Rivandi et al., many of these GWAS-identified SNPs are located outside coding regions and are tags for mostly unknown, causal or functional variants. Hence, their identification would provide better estimates of the explained familial relative risk, thereby improving polygenetic PRSs and increase our understanding of the biological mechanisms involved in breast cancer susceptibility. The success of BCAC and CIMBA in identifying several low risk alleles, resides in the capability of coordinating the efforts of over 180 worldwide groups or studies contributing DNA samples and data from breast cancer cases and control and from *BRCA1/2* mutation carriers. The French Genetic Modifiers of *BRCA1* and *BRCA2* (GEMO) Group, described in this e-book by Lesueur et al., is one of the larger studies within CIMBA. GEMO was initiated in 2006 and today involves 32 clinics and 17 diagnostics laboratories that, as of April 2018, collected 5,303 participants.

As discussed above, many variants in *BRCA1/2* and in other established or candidate breast cancer genes have uncertain clinical significance. These VUSs, which are typically rare missense variants, represent a serious clinical problem as carrier risk estimates are often unclear. The Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA; https://enigmaconsortium.org) was formed to determine the clinical significance of variants in *BRCA1/2* and other known or suspected breast cancer genes (13). To this aim, ENIGMA gathers pathologists, epidemiologists, geneticists, bio-informaticians, genetic counselors, and molecular biologists into working groups to assess the clinical relevance of variants by applying statistical approaches and multifactorial likelihood models, studying tumor markers, or performing functional assays. Two articles in this e-book provide insights into VUSs classification. The first study, by Zuntini et al., investigated whether co-segregation analyses, integrated with functional data and *in silico* predictions, could improve VUSs interpretation and counseling in carrier families. In the second study, Caleca et al., used an *in vitro* assay specifically designed to test the *BRCA2* and *PALB2* gene products interaction and showed initial pathogenicity evidence for two very rare missense variants in these genes. A special class of VUS are those suspected to cause mRNA splicing defects. One of the ENIGMA working groups was established to improve the clinical classification of likely spliceogenic variants. Members of this working group contributed articles exploring some of the aspects of this variant class. For example, Fraile-Bethencourt et al. identified eight spliceogenic variants in exon 16 of *BRCA2* by minigene assay, highlighting the efficiency of this approach for clinical classification. *In silico* tools for splicing defect prediction may play a key role in VUS analysis. Moles-Fernández et al. used 99 *in vitro*-validated variants to evaluate the performance of six commonly used splicing *in silico* tools. Finally, Farber-Katz et al. conducted analyses of *BRCA1/2* variants using a novel RNA-massively parallel sequencing assay capable to perform quantitative and qualitative analysis of transcripts. Similarly, Lattimore et al. utilized targeted RNA-seq to re-assess *BRCA1/2* mRNA isoform expression patterns in lymphoblastoid cell lines. Recommendations from these two studies will facilitate the application of targeted RNA-seq approaches for the quantitative characterization of *BRCA1* and *BRCA2* germline splicing alterations.

In summary, articles in the present e-book move the field of breast cancer genetics in several aspects, ranging from characterizing genetic variation in new populations to developing and applying tools for variant re-classification. Since the discovery of the role of *BRCA1/2* on breast cancer risk, much has been learned and through the tremendous international, multi- and inter-disciplinary efforts of consortia such as BCAC, CIMBA, and ENIGMA. The next decade promises to illuminate many new aspects of breast cancer risk prevention on families and in the general population.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

REFERENCES

1. Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips KA, Mooij TM, Roos-Blom MJ, et al. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA*. (2017) 317:2402–16. doi: 10.1001/jama.2017.7112

2. Antoniou AC, Casadei S, Heikkinen T, Barrowdale D, Pylkäs K, Roberts J, et al. Breast-cancer risk in families with mutations in PALB2. *N Engl J Med*. (2014) 371:497–506. doi: 10.1056/NEJMoa1400382

3. Easton DF, Pharoah PD, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, et al. Gene-panel sequencing and the prediction of breast-cancer risk. *N Engl J Med*. (2015) 372:2243–57. doi: 10.1056/NEJMsr1501341

4. Schmidt MK, Hogervorst F, van Hien R, Cornelissen S, Broeks A, Adank MA, et al. Age- and tumor subtype-specific breast cancer risk estimates for CHEK2*1100delC carriers. *J Clin Oncol*. (2016) 34:2750–60. doi: 10.1200/JCO.2016.66.5844

5. Shimelis H, LaDuca H, Hu C, Hart SN, Na J, Thomas A, et al. Triple-Negative breast cancer risk genes identified by multigene hereditary cancer panel testing. *J Natl Cancer Inst*. (2018) 110:855–62. doi: 10.1093/jnci/djy106

6. Kiiski JI, Pelttari LM, Khan S, Freysteinsdottir ES, Reynisdottir I, Hart SN, et al. Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. *Proc Natl Acad Sci USA*. (2014) 111:15172–7. doi: 10.1073/pnas.1407909111

7. Peterlongo P, Catucci I, Colombo M, Caleca L, Mucaki E, Bogliolo M, et al. FANCM c.5791C>T nonsense mutation (rs144567652) induces exon skipping, affects DNA repair activity and is a familial breast cancer risk factor. *Hum Mol Genet*. (2015) 24:5345–55. doi: 10.1093/hmg/ddv251

8. Neidhardt G, Hauke J, Ramser J, Groß E, Gehrig A, Müller CR, et al. Association between loss-of-function mutations within the FANCM gene and early-onset familial breast cancer. *JAMA Oncol*. (2017) 3:1245–8. doi: 10.1001/jamaoncol.2016.5592

9. Couch FJ, Nathanson KL, Offit K. Two decades after BRCA: setting paradigms in personalized cancer care and prevention. *Science*. (2014) 343:1466–70. doi: 10.1126/science.1251827

10. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*. (2017) 551:92–4. doi: 10.1038/nature24284

11. Milne RL, Kuchenbaecker KB, Michailidou K, Beesley J, Kar S, Lindström S, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat Genet*. (2017) 12:1767–78. doi: 10.1038/ng.3785

12. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am J Hum Genet*. (2019) 1:21–34. doi: 10.1016/j.ajhg.2018.11.002

13. Spurdle AB, Healey S, Devereau A, Hogervorst FB, Monteiro AN, Nathanson KL, et al. ENIGMA–evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Hum Mutat*. (2012) 33:2–7. doi: 10.1002/humu.21628

BRCA1 and *BRCA2* Mutations Other Than the Founder Alleles Among Ashkenazi Jewish in the Population of Argentina

Angela R. Solano[1,2*], Natalia C. Liria[1], Fernanda S. Jalil[1], Daniela M. Faggionato[1], Pablo G. Mele[2], Alejandra Mampel[3], Florencia C. Cardoso[1] and Ernesto J. Podesta[2]

[1] Genotipificación y Cáncer Hereditario, Centro de Educación Médica e Investigaciones Clínicas "Norberto Quirno" (CEMIC), Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina, [2] Facultad de Medicina, Instituto de Investigaciones Biomédicas (INBIOMED), Universidad de Buenos Aires-CONICET, Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina, [3] Hospital Universitario, Instituto de Genética, Universidad Nacional de Cuyo, Mendoza, Argentina

In Ashkenazi Jewish (AJ) high risk families 3 mutations [2 in *BRCA1* (c. 68_69del and c.5266dup) and 1 in *BRCA2* (c.5946del)] account for the majority of high risk breast and ovarian cancer cases in that ethnic group. Few studies with limited number of genotyped individuals have expanded the spectrum of mutations in both *BRCA* genes beyond the 3 mutation panel. In this study, 279 high risk individual AJ were counseled at CEMIC (Centro de Educación Médica e Investigaciones Clínicas), and were genotyped first for the 3 recurrent mutation panel followed by Next Generation Sequencing (NGS) of *BRCA1 BRCA2* in 76 individuals who tested negative for the first genotyping step. Of 279 probands (259 women), 55 (50 women) harbored one of the 3 mutations (19.7%); Of 76 fully sequenced cases (73 women), 6 (5 women) (7.9%) carried a pathogenic mutation: in *BRCA1*, c.2728C>T - p.(Gln910*); c.5407-?_(*1_?)del and c.5445G>A - p.(Trp1815*); in *BRCA2*, c.5351dup - p.(Asn1784Lysfs*3); c.7308del - p.(Asn2436Lysfs*33) and c.9026_9030del - p.(Tyr3009Serfs*7). Of 61 mutation carriers the distribution was as follows: 11 cancer free at the time of genotyping, 34 female breast cancer cases with age range 28–72 years (41.6 ± 9.3), 3 male breast cancer cases with age range 59–75 years (65 ± 7.3), 6 breast and ovarian cancer cases with age range 35–60 years (breast 40.4 ± 5.2; ovary 47.8 ± 7.2) and 7 ovarian cancer cases with age range 41–77 years (60.6 ± 13.3). This information proved highly useful for counseling, treatment, and prevention for the patient and the family. In conclusion comprehensive *BRCA1/2* testing in AJ high risk breast ovarian cancer cases adds valuable clinically relevant information in a subset of cases estimated up to 7% and is therefore recommended.

Keywords: non-founder Ashkenazi *BRCA1/2* mutations, Ashkenazi Jewish, hereditary breast and ovary cancer, *BRCA1*, *BRCA2*

INTRODUCTION

Inherited pathogenic mutations in *BRCA1* (OMIM*113705) or *BRCA2* (OMIM*600185) substantially increase lifetime risk for breast, ovarian and to a lesser extent other cancer types defining individuals who carry *BRCA1* or *BRCA2* cancer predisposing mutations is valuable for both cancer cases and unaffected family members: targeted treatment in the form of PARP inhibitors is available for mutation carrying patients (1, 2), and early detection schemes and risk-reducing strategies are offered to asymptomatic mutation carriers (3, 4).

Near 300 *BRCA1* and *BRCA2* pathogenic mutations in the Argentine population have been reported by our group (5, 6) routinely deposited in Leiden Open Variation Database 3.0 (7) and Leiden Open Variation Database–Chapter for Argentina (8), most previously described in other world populations with ~10% of novel pathogenic variants.

The estimated frequency of pathogenic germline mutations in *BRCA1* and *BRCA2* genes in the general population in several outbred populations vary between 1:300 and 1:800, respectively (4). However, in inbred populations, with AJ as one of the most frequently studied examples, the spectrum of *BRCA* mutations is limited with higher rates in the general population. In this ethnic group 1/40 individuals is a carrier one of 3 recurrent mutations in *BRCA1* [185delAG: c.68_69del (rs386833395) and 5382insC: c.5266dup (rs80357906)], and *BRCA2* [6174delT: c.5946del (rs80359550)]. Such high rates in the general population and consecutive breast and ovarian cancer cases enabled effective use of cancer genetics services for AJ women (9).

For high risk AJ individuals not carrying one of these 3 founder alleles, the probability for other pathogenic mutations in *BRCA1* or *BRCA2* has rarely been reported (10–13). In the current study we report on an extended analysis of *BRCA1/2* by comprehensive next generation sequencing and analysis of point mutations and large rearrangements in Ashkenazi patients with personal and/or family history of cancer qualifying for the panel analysis and testing normal for the 3 founder mutations.

MATERIALS AND METHODS

Study Subjects

The study focused on Ashkenazi Jewish individuals recruited from those referred for counseling and genotyping at CEMIC (Centro de Educación Médica e Investigaciones Clínicas) between 2009 and 2017. Eligibility criteria for patient selection are based on the NCCN guidelines [National Comprehensive Cancer Network. Genetic/Familial High-Risk Assessment: Breast and Ovarian (Version 1.2018) https://www.nccn.org/ Accessed March 30, 2018]. For individuals of AJ origin, with no known familial mutation, were first tested for the 3 AJ specific mutations (see below genotyping methodology). Then, for high risk individuals of AJ who tested negative for the three mutations *BRCA* comprehensive *BRCA* genetic testing was carried out (see below genotyping platform). Study eligibility after genetic counseling required signing an informed consent as part of the routine procedures for genetic analysis (including Ethics Committee approval) at CEMIC, which also complies with the Traditional Pretest Counseling for

FIGURE 1 | Patients of Ashkenazi origin with personal and/or family history (279 probands) were tested for the panel AJ founder mutations; among affected patients 24.3% were detected with a mutation while this rate was 11.2% in the non-affected individuals. Sixty three patients without AJ mutation were subjected to the comprehensive *BRCA1/2* analysis and 6 were diagnosed with a pathogenic mutation (9.5% from the analyzed probands). On the other hand, 98 were healthy individuals with family history of cancers related to *BRCA1/2* and, among them. The non-Ashkenazi mutations were detected always in affected patients.

Susceptibility Testing (purpose of testing) described in the American Society of Clinical Oncology Policy Statement Update (14).

BRCA1/2 Testing Platforms

Genomic DNA of the 279 blood samples was isolated by MagNA Pure® LC instrument with total DNA isolation kit I (Roche Diagnostics). All samples were analyzed by Sanger sequencing of PCR amplified fragments for the 3 founder Ashkenazi mutations c.68_69del and c.5266dup in *BRCA1* and c.5946del in *BRCA2*.

Analysis of comprehensive *BRCA1/2* sequencing and large rearrangements by Multiple Ligation-dependent Probe Amplification assay (MLPA) for eligible individuals was performed by Next Generation Sequencing (NGS) by using the Ion AmpliSeq*BRCA1/2* community panel, as it allows to amplify the entire coding sequences of *BRCA1* and *BRCA2*, including 20–50 bases of adjacent intronic sequence of each exon. The assay is designed to ensure at least 200X total coverage/base. Sequencing of the amplified regions was performed with the next generation platform Personal Genome Machine® System, as previously described (6). Rare coding sequences with low coverage were analyzed by Sanger sequencing to ensure higher coverage rates. The raw signal data and the sequence reads were processed with Ion Torrent Suite software (Thermo Fisher Scientific) on a Torrent server. After data analysis, single nucleotide variants, insertions, deletions, and splice site alterations were registered, and all variants detected were reported. Sanger sequencing was used to confirm all clinically relevant variants detected (class 3, 4, and 5). Clinical significance was determined according to the reference databases: ClinVar (15), LOVD3.0 (7), and UMD (16) as of March 2018.

For missense mutations not reported or reported with uncertain clinical significance (VUS), *in silico* programs were used to predict the change in protein function using software Align-GVGD (http://agvgd.iarc.fr/), SIFT (http://sift.bii.a-star.edu.sg), and Mutations Taster (http://www.mutationtaster.org/).

Large rearrangements were measured by MLPA using SALSA MLPA Probemix P002 and P045 provided by MRC-Holland, and Coffalyser.net software was used for data analysis; we confirm the positive results with P087 and P077 for *BRCA1* and *BRCA2*, respectively.

RESULTS

Participant's Characteristics

Overall, we include 279 patients (**Figure 1**) recruited among 2009–2017 as depicted in **Table 1**. Age range at counseling and genotyping was 20–87 years; 181 (174 females) had cancer diagnoses (mean age 48.3 ± 11.2 years) and 98 were healthy, cancer free high risk individuals (mean age at genotyping 47.7 ± 11.8). Of cancer cases the distribution was as follows: of 174 cancer affected women, breast, 145 (age range 28–74 years, mean 47.0 ± 9.6), breast and ovarian 7 (age range breast 35–64 years, mean 44.3 ± 9.8; age range ovary 41–64 years, mean 49.2 ± 8.9), ovarian 19 (age range 18–78 years, mean 52.5 ± 16.6) and one each with pancreas (68 years), endometrium (58 years) and melanoma (26 years). For the 7 cancer affected men, 6 had breast

TABLE 1 | Analysis of the study subjects.

Patients analyzed (Total = 279)	Patients with mutation (*n*)	Age range	Age mean ± *SD*
Healthy women	9	40–58	47.7 ± 5.3
Healthy men	2	47 & 76	61.5 ± 20.5
Breast cancer women	34	28–72	41.6 ± 9.3
Breast cancer men	3	59–75	65 ± 7.3
Ovary cancer	7	41–77	60.6 ± 13.3
Breast and ovary	6		
Breast		35–46	40.4 ± 5.2
Ovary		41–60	47.8 ± 7.2
Total with a mutation	61 (27.6%)*		

AJ mutations 55/279 (19.7%); Non-AJ mutations 6/76 (7.9%).

cancer (age range 59–75 years; mean 65.5 ± 5.7) and 1 with prostate cancer (61 years).

The 279 selected patients were first analyzed through the AJ founder mutation and, among those who tested normal, 76 patients were analyzed by the *BRCA1/2* comprehensive study; among the 6 patients with a non AJ mutation detected all but 2 were of full AJ origin, one was mixed Ashkenazi non Ashkenazi and the other mixed Ashkenazi and non-Jewish. The ages ranged from 20 to 87 years old; 181 of them were affected with a mean age of 48.3 ± 11.2 and 98 of them were healthy with a mean age of 47.7 ± 11.8. **Table 1** summarizes the age range and mean age ± SD for subjects with a mutation detected separated by diagnosis and gender. As eligible patients we included men or women selected by their AJ ethnicity, with two exceptions (**Figure 2**): proband [B: *BRCA1* c.5407-?_(*1_?)del] was half Sephardi, and proband [D: *BRCA2*: c.5351dup - p.(Asn1784Lysfs*3)] was half non-Jewish, although this side was not the side associated with inheritance of the syndrome.

Overall, 61/279 genotyped cases (21.8%) harbored a *BRCA1* mutation (*n* = 44) or a *BRCA2* mutation (*n* = 17). Of these mutations all but 3 in *BRCA1* and 3 in *BRCA2* were one of the predominant AJ mutations in both genes.

Table 1 summarizes age at genotyping and/or cancer diagnosis and type and gender of all mutation carriers and the specific unique non founder mutations.

Figure 1 summarizes the genotype analysis of 279 individuals with the DNA sequenced for the panel of the 3 founder Ashkenazi mutations and full sequence by NGS technique and MLPA.

Table 1 lists the patients with a mutation detected, to be remarked women with diagnosis of breast cancer were the youngest (range starts at 28 years), mean age non-statistically different from women with diagnosis of both, breast and ovary cancer.

In **Tables 2**, **3** are detailed the mutations detected in females and males respectively.

Figure 2 depicts the pedigrees for the six families with a non AJ founder mutation. Of the 6 mutations found, *BRCA2* c.7308del - p.(Asn2436Lysfs*33) has not been previously reported and is therefore novel while the other 5 have been reported in non-Jewish populations.
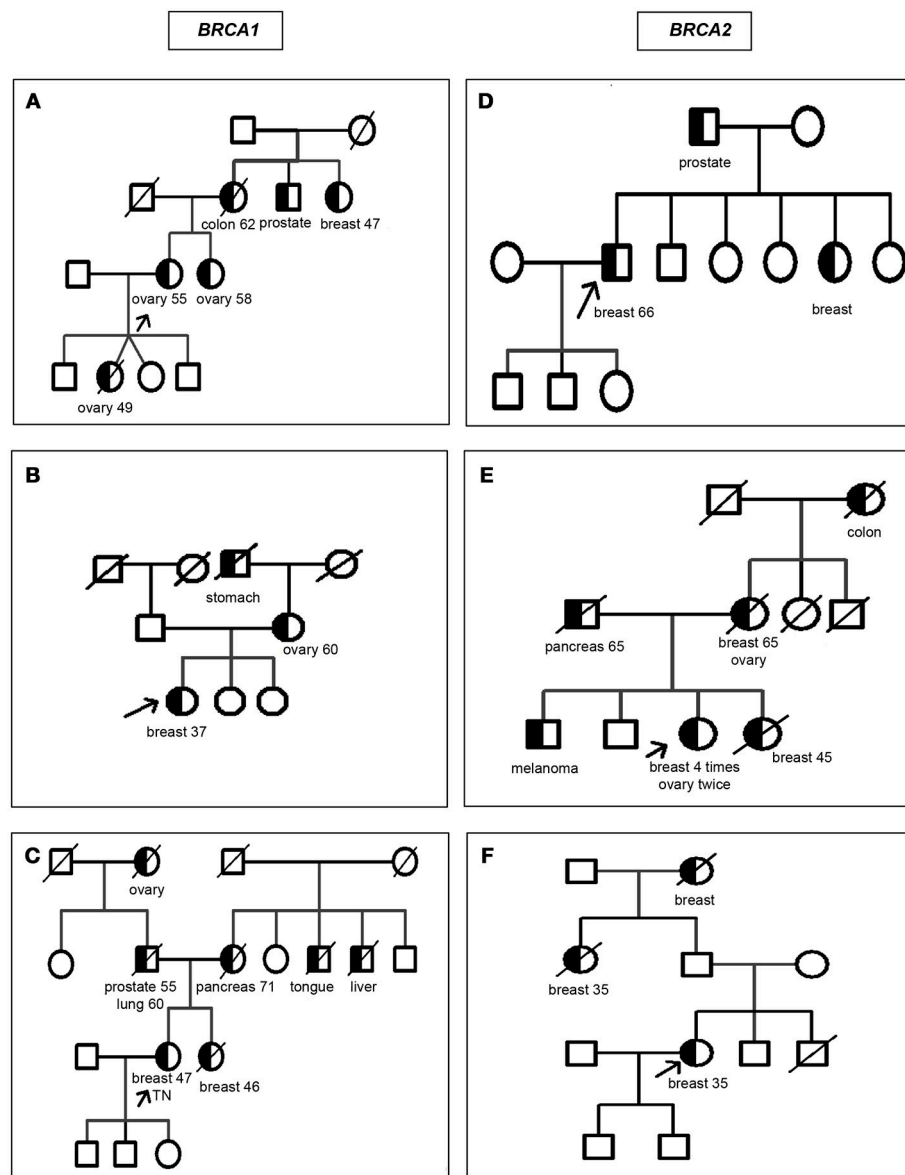
FIGURE 2 | Pedigrees of the families carrying mutations in *BRCA1/2* other than the Ashkenazi founder mutations. Circles: women; squares: men; half blackened symbols: individuals affected with cancer; white symbols: unaffected individuals; TN: triple negative breast cancer; slash diagonal line: deceased. The cancer and age at diagnosis are indicated below each individual when available. An arrow indicates the proband. **(A)** *BRCA1* c.2728C>T - p.(Gln910*); **(B)** *BRCA1* c.5407-?_(*1_?)del; **(C)** *BRCA1*: c.5445G>A-p.(Trp1815*); **(D)** *BRCA2*: c.5351dup - p.(Asn1784Lysfs*3); **(E)** *BRCA2* c.7308del - p.(Asn2436Lysfs*33); **(F)** *BRCA2* c.9026_9030del - p.(Tyr3009Serfs*7).

Regarding the families with a non-Ashkenazi mutation, the particular details of the family history of cancers related to *BRCA* was strong for the 6 probands, as shown in the pedigrees drawn in **Figure 2**. The description of the mutations are in **Table 4**, as follows: *BRCA1*: c.2728C>T - p.(Gln910*), rs397509004, c.5407-?_(*1_?)del, c.5445G>A - p.(Trp1815*), rs397509284 and *BRCA2*: c.5351dup - p.(Asn1784Lysfs*3), rs80359508, c.7308del - p.(Asn2436Lysfs*33), c.9026_9030del - p.(Tyr3009Serfs*7), rs80359741.

DISCUSSION

The likelihood that AJ high risk individuals who do not carry any of the predominant AJ mutations in *BRCA1* and *BRCA2* would harbor a unique *BRCA1* or *BRCA2* mutation in the present study was 7.9%. As the family D was self-reported as half-non AJ, the resulting prevalence for *BRCA1/2* non AJ mutation excluding this family results to be 5 out of 76 cases (6.6%). This rate is line with previous studies, although there are only a few that have reported such a focused analysis (10–13). In those previous

TABLE 2 | Analysis of mutations detected in female individuals.

Female patients mutation analysis	*n*	%	*n* (tumor type, age range or age@diag)	Healthy
Recurring Ashkenazi mutations				
Total	259			
Gene: Mutation	50	19.3		9
BRCA1: c.68_69del -p.(Glu23Valfs*17)	24	9.3	18, Br (30-72); 1 Br/Ov (46, 47); 2, Ov (59, 77)	3 (*)
BRCA1: c.5266dup -p.(Gln1756Profs*74)	15	5.8	8, Br (28-46); 4, Ov (41-72); 2, Br/Ov (35,41; 37,45)	1
BRCA2: c.5946del -p.(Ser1982Argfs*22)	11	4.2	4, Br (30-56); 1 Br & Fallopian tube (64); 1 Br/Ov (46)	5
Unique mutations (NGS)				
Total	73			
Gene: Mutation	5	6.8		
BRCA1: c.2728C>T - p.(Gln910*)	1		Ov (55)	0
BRCA1: c.5407-?_(*1_?)del	1		Br (37)	0
BRCA1: c.5445G>A - p.(Trp1815*)	1		Br (47)	
BRCA2: c.7308del - p.(Asn2436Lysfs*33)	1		Br (38,45,47,52) & Ov (60, 64)	0
BRCA2: c.9026_9030del - p.(Tyr3009Serfs*7)	1		Br (35)	0

n, number of probands; age@diag, age at diagnosis; Br, Breast Cancer; Ov, Ovarian Cancer.
(*) a patient was carrier of the mutation in the gene MSH2: c.1906G>C - p.(Ala636Pro), also of Ashkenazi origin (17).

TABLE 3 | Analysis of mutations detected in male individuals.

Male patients mutation analysis	*n*	(%)	age@diag	Healthy
Ashkenazi mutation				
Total	20			
A mutation found	5	25.0		2
BRCA1: c.68_69del -p.(Glu23Valfs*17)	2		Br (75)	1
BRCA1: c.5266dup -p.(Gln1756Profs*74)	0		0	0
BRCA2: c.5946del - p.(Ser1982Argfs*22)	3		Br (60); Br (59)	1
Non-Ashkenazi mutation (NGS):		0.0		0
Total	3			
A mutation found	1	33.3		
BRCA2: c.5351dup - p.(Asn1784Lysfs*3)	1		Br (66)	0

n, number of probands; age@diag, age at diagnosis; Br, Breast Cancer.

studies the rates were 4-5% of fully genotyped AJ cases. The rationale behind our two step sample analysis approach lies in the fact that non-routine tests in the Argentinian health care system requires approval by a specialized committee that tends to not approve oncogenetic testing for unaffected individuals despite having a significant family history of cancer.

A rather difficult task is to express the results of this study, as not all the patients without an Ashkenazi mutation were analyzed by NGS and MLPA. As a consequence, the percentages of mutations cannot be straightforwardly interpreted, which may constitute a limitation for expressing the current findings. However, our results still remain valid, as even if all samples normal for the Ashkenazi panel had been tested for the comprehensive *BRCA1/2* study and no other mutation would had been detected, the percentage obtained would still have been 3.3% (6 out of 181 affected patients).

The population targeted in this study was selected on the basis of clinically having features of inherited cancer syndrome. Overall, non AJ mutations were detected in 6.8% of female cancer cases; 4.1% in female breast cancer cases and 2.7% in ovarian cancer (alone or with breast cancer). These rates are higher than those previously reported for Ashkenazi population, likely due to different criteria used for patient selection (10, 11). The highest mutation rates were found in cases of ovarian cancer including breast cancer diagnosed in the same patient, 50%, with similar rates to what was previously published by our group in a description of 940 patients (6).

Worth highlighting, 11 out of these 13 mutations were among the Ashkenazi panel and two detected by full analysis. The two probands found to carry a mutation in *BRCA1/2* support the application of precision medicine: patients benefit from being considered for poly–(adenosine diphosphate–ribose) polymerase (PARP) inhibitor therapy (1, 2), while their families still qualify for prevention measures as the first goal in these genetic studies.

Family history of cancer in first degree relatives as a major selection criterion for selecting cases to be genotyped carries an inherent limitation, as gender specific cancer (e.g., ovarian cancer) cannot be used in cases where the mutation arises on the paternal side yet, the value of incorporating second or third degree relatives on either parental side has its own merits. Moreover there are cases where the family history is distinct or the paternal from the maternal sides and both should be taken into account and guide the genetic testing platform. A case that exemplifies this point is the patient who co-harbored the *MSH2* c.1906G>C - p.(Ala636Pro) (17) and *BRCA1* c.68_69del - p.(Glu23Valfs*17). These findings are of the utmost importance, as surveillance of the unaffected proband will focus on both Lynch and hereditary breast-ovary cancer syndromes, while

TABLE 4 | Families with a non-AJ founder mutation in *BRCA1/2* from pedigrees drawn in **Figure 2**.

Mutation (HGVS nomenclature)	Description
BRCA1: c.2728C>T - p.(Gln910*) rs397509004	This is a rare mutation reported in ClinVar and we deposited in LOVD (Genomic Variant #0000206572). The proband with age 71 at testing had ovarian cancer at 55 and her daughter died from an ovarian cancer at 49. There is no information of the twin. The sister, also affected, tested positive for the mutation.
BRCA1: c.5407-?_(*1_?)del	This large rearrangement has been described many times in the databases. As shown in **Figure 2**, this half Sephardi patient had family history of stomach and ovary cancer.
BRCA1: c.5445G>A - p.(Trp1815*) rs397509284	This mutation is deposited by different laboratories at least 5 times at the LOVD. The proband had triple negative breast cancer (the type of cancer most frequently associated with *BRCA1* mutations) at 47, was tested at 49 and had a sister who died of breast cancer at 46.
BRCA2: c.5351dup - p.(Asn1784Lysfs*3) rs80359508	We have detected this mutation 10 times and it is reported in LOVD at least 49 times (including 7 from our laboratory) (18). The family history is relatively scarce in information, with a sister with breast cancer and the father with prostate cancer. The daughter and one of the sons tested normal for the mutation. Considering the origin of this mutation it may be the case that the cancer diagnosis from the Jewish side could be phenocopies, and for this reason the prevalence for *BRCA2* non AJ was also calculated non including this case.
BRCA2: c.7308del - p.(Asn2436Lysfs*33)	Novel, not reported in the databases and we deposited at LOVD (Genomic Variant #0000206927). The proband, tested at 67, has history of breast cancer 4 times, starting at 38, 45, 47, and 52 years of age; she also developed ovary cancer at 60 and 64. Two relatives were tested, a niece of 38 was found to be a non-carrier and a nephew at 41 an unaffected carrier.
BRCA2: c.9026_9030del - p.(Tyr3009Serfs*7), rs80359741	This mutation was detected 3 times in our laboratory and at least 10 more times deposited at LOVD. This mutation was associated to Spanish ascendancy (19).

The following variant of uncertain significance (VUS) was detected in the samples analyzed by NGS: BRCA2: c.7232A>C - p.(Lys2411Thr), rs80358950 (we deposited at LOVD Genomic Variant #0000206714).

genetic counseling will be crucial for both paternal and maternal relatives.

The results of this study support the guidelines recommending genetic testing for the recurrent *BRCA* mutations in all AJ breast and ovarian cancer patients (NCCN Clinical Practice in Oncology, Version 1.2018). In addition, the value, albeit limited to <8% more mutations, of an extended comprehensive *BRCA* genotyping for high risk Ashkenazim who are negative for the 3 predominant mutations needs to be offered and discussed with the patients at genetic counseling, regardless of the cost. This is particularly important as having a *BRCA* mutation has therapeutic implications especially in ovarian cancer cases, who can benefit from PARP inhibitor treatment.

AUTHOR CONTRIBUTIONS

AS, FC, and EP discussed the conception of the work and take responsibility for data integrity and the accuracy of data analysis. FJ, NL, AM, and DF collected and organized patient samples.

PM, FJ, and DF analyzed the literature and evaluated statistics. All authors critically revised the manuscript for important intellectual content.

REFERENCES

1. Audeh MW, Carmichael J, Penson RT, Friedlander M, Powell B, Bell-McGuinn KM, et al. Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial. *Lancet* (2010) 376:245–51. doi: 10.1016/S0140-6736(10)60893-8
2. Ledermann J, Harter P, Gourley C, Friedlander M, Vergote I, Rustin G, et al. Olaparib maintenance therapy in platinum-sensitive relapsed ovarian cancer. *N Eng J Med.* (2012) 366:1382–92. doi: 10.1056/NEJMoa1105535
3. Zhong Q, Peng HL, Zhao X, Zhang L, Hwang WT. Effects of BRCA1- and BRCA2-related mutations on ovarian and breast

cancer survival: a meta-analysis. *Clin Cancer Res.* (2015) 21:211–20. doi: 10.1158/1078-0432.CCR-14-1816
4. Paluch-Shimon S, Cardoso F, Sessa C, Balmana J, Cardoso MJ, Gilbert F, et al. Prevention and screening in BRCA mutation carriers and other breast/ovarian hereditary cancer syndromes: ESMO Clinical Practice Guidelines for cancer prevention and screening. *Annal Oncol.* (2016) 27(Suppl. 5):v103–v10. doi: 10.1093/annonc/mdw327
5. Solano AR, Aceto GM, Delettieres D, Veschi S, Neuman MI, Alonso E, et al. BRCA1 And BRCA2 analysis of Argentinean breast/ovarian cancer patients selected for age and family history highlights a role for novel mutations of putative south-American origin. *SpringerPlus* (2012) 1:20. doi: 10.1186/2193-1801-1-20

6. Solano AR, Cardoso FC, Romano V, Perazzo F, Bas C, Recondo G, et al. Spectrum of BRCA1/2 variants in 940 patients from Argentina including novel, deleterious and recurrent germline mutations: impact on healthcare and clinical practice. *Oncotarget* (2017) 8:60487–95. doi: 10.18632/oncotarget.10814

7. Leiden Open Variation Database 3.0 Available online at: http://www.lovd.nl/ 3.0/home [Internet] (Accessed March 30, 2018).

8. Leiden Open Variation Database (Chapter for Argentina) Available online at: http://ar.lovd.org [Internet].

9. Lieberman S, Tomer A, Ben-Chetrit A, Olsha O, Strano S, Beeri R, et al. Population screening for BRCA1/BRCA2 founder mutations in Ashkenazi Jews: proactive recruitment compared with self-referral. *Genet Med.* (2017) 19:754–62. doi: 10.1038/gim.2016.182

10. Rosenthal E, Moyes K, Arnell C, Evans B, Wenstrup RJ. Incidence of BRCA1 and BRCA2 non-founder mutations in patients of Ashkenazi Jewish ancestry. *Breast Cancer Res Treat.* (2015) 149:223–7. doi: 10.1007/s10549-014-3 218-x

11. Walsh T, Mandell JB, Norquist BM, Casadei S, Gulsuner S, Lee MK, et al. Genetic Predisposition to Breast Cancer Due to Mutations Other Than BRCA1 and BRCA2 Founder Alleles Among Ashkenazi Jewish Women. *JAMA Oncol.* (2017) 3:1647–53. doi: 10.1001/jamaoncol. 2017.1996

12. Ganguly T, Dhulipala R, Godmilow L, Ganguly A. High throughput fluorescence-based conformation-sensitive gel electrophoresis (F-CSGE) identifies six unique BRCA2 mutations and an overall low incidence of BRCA2 mutations in high-risk BRCA1-negative breast cancer families. *Hum Genet.* (1998) 102:549–56.

13. Laitman Y, Borsthein RT, Stoppa-Lyonnet D, Dagan E, Castera L, Goislard M, et al. Germline mutations in BRCA1 and BRCA2 genes in ethnically diverse high risk families in Israel. *Breast Cancer Res Treat.* (2011) 127:489–95. doi: 10.1007/s10549-010-1217-0

14. Robson ME, Bradbury AR, Arun B, Domchek SM, Ford JM, Hampel HL, et al. American society of clinical oncology policy statement update: genetic and genomic testing for cancer susceptibility. *J Clin Oncol.* (2015) 33:3660–7. doi: 10.1200/JCO.2015.63.0996

15. ClinVar Available online at: http://www.ncbi.nlm.nih.gov/clinvar/ [Internet] (Accessed March 30, 2018).

16. Universal Mutation Database (UMD). Available online at: http://www.umd. be/ [Internet] (Accessed March 30, 2018).

17. Foulkes WD, Thiffault I, Gruber SB, Horwitz M, Hamel N, Lee C, et al. The founder mutation MSH2*1906G–>C is an important cause of hereditary nonpolyposis colorectal cancer in the Ashkenazi Jewish population. *Am J Hum Genet.* (2002) 71:1395–412. doi: 10.1086/345075

18. Janavicius R. Founder BRCA1/2 mutations in the Europe: implications for hereditary breast-ovarian cancer prevention and control. *EPMA J.* (2010) 1:397–412. doi: 10.1007/s13167-010-0037-y

19. Campos B, Diez O, Odefrey F, Domenech M, Moncoutier V, Martinez-Ferrandis JI, et al. Haplotype analysis of the BRCA2 9254delATCAT recurrent mutation in breast/ovarian cancer families from Spain. *Hum Mutat.* (2003) 21:452. doi: 10.1002/humu.9133

Complex Landscape of Germline Variants in Brazilian Patients With Hereditary and Early Onset Breast Cancer

Giovana T. Torrezan [1,2†], Fernanda G. dos Santos R. de Almeida [1†],
Márcia C. P. Figueiredo [1], Bruna D. de Figueiredo Barros [1], Cláudia A. A. de Paula [1],
Renan Valieris [3], Jorge E. S. de Souza [4,5,6], Rodrigo F. Ramalho [1], Felipe C. C. da Silva [1],
Elisa N. Ferreira [1,7], Amanda F. de Nóbrega [8], Paula S. Felicio [9], Maria I. Achatz [8,10],
Sandro J. de Souza [2,6,11], Edenir I. Palmero [9,12] and Dirce M. Carraro [1,2*]

[1] Laboratory of Genomics and Molecular Biology, International Research Center, CIPE/A.C. Camargo Cancer Center, São Paulo, Brazil, [2] National Institute for Science and Technology in Oncogenomics and Therapeutic Innovation, São Paulo, Brazil, [3] Laboratory of Bioinformatics and Computational Biology, International Research Center, CIPE/A.C. Camargo Cancer Center, São Paulo, Brazil, [4] Instituto de Bioinformática e Biotecnologia—2bio, Natal, Brazil, [5] Instituto Metrópole Digital, Federal University of Rio Grande do Norte, Natal, Brazil, [6] Bioinformatics Multidisciplinary Environment, Federal University of Rio Grande do Norte, Natal, Brazil, [7] Research and Development, Fleury Group, São Paulo, Brazil, [8] Oncogenetics Department, A.C. Camargo Cancer Center, São Paulo, Brazil, [9] Molecular Oncology Research Center, Barretos Cancer Hospital, São Paulo, Brazil, [10] Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, United States, [11] Brain Institute, Federal University of Rio Grande do Norte, Natal, Brazil, [12] Barretos School of Health Sciences, Dr. Paulo Prata – FACISB, Barretos, Brazil

Pathogenic variants in known breast cancer (BC) predisposing genes explain only about 30% of Hereditary Breast Cancer (HBC) cases, whereas the underlying genetic factors for most families remain unknown. Here, we used whole-exome sequencing (WES) to identify genetic variants associated to HBC in 17 patients of Brazil with familial BC and negative for causal variants in major BC risk genes (*BRCA1/2, TP53,* and *CHEK2* c.1100delC). First, we searched for rare variants in 27 known HBC genes and identified two patients harboring truncating pathogenic variants in *ATM* and *BARD1*. For the remaining 15 negative patients, we found a substantial vast number of rare genetic variants. Thus, for selecting the most promising variants we used functional-based variant prioritization, followed by NGS validation, analysis in a control group, cosegregation analysis in one family and comparison with previous WES studies, shrinking our list to 23 novel BC candidate genes, which were evaluated in an independent cohort of 42 high-risk BC patients. Rare and possibly damaging variants were identified in 12 candidate genes in this cohort, including variants in DNA repair genes (*ERCC1* and *SXL4*) and other cancer-related genes (*NOTCH2, ERBB2, MST1R,* and *RAF1*). Overall, this is the first WES study applied for identifying novel genes associated to HBC in Brazilian patients, in which we provide a set of putative BC predisposing genes. We also underpin the value of using WES for assessing the complex landscape of HBC susceptibility, especially in less characterized populations.

Keywords: cancer predisposition genes, hereditary breast cancer, whole-exome sequencing, germline pathogenic variants, cancer susceptibility, DNA repair genes

INTRODUCTION

Hereditary breast cancer (HBC) corresponds to ∼5–10% of all breast cancer cases (Honrado et al., 2005). The most common breast cancer predisposing syndrome is hereditary breast and ovarian cancer syndrome (HBOC) that is related to pathogenic germline variants in *BRCA1* (OMIM 113705) and *BRCA2* (OMIM 600185) genes (Anglian Breast Cancer Study, 2000). These genes correspond to ∼20–25% of all HBC (Anglian Breast Cancer Study, 2000; Kean, 2014; Silva et al., 2014). Besides *BRCA1/2* genes, pathogenic variants in other high- and moderate-risk genes, such as *TP53*, *CHEK2*, *ATM*, *STK11*, *PALB2*, among others, also lead to an increased breast cancer (BC) risk, revealing a high complexity in breast cancer predisposition (Elledge and Allred, 1998; Meijers-Heijboer et al., 2002; Walsh and King, 2007).

To date, over 35 genes have been suggested to carry high and/or moderate BC risk variants (OMIM, 2015[1]; Shiovitz and Korde, 2015). However, only a minority of these genes have an established significant association demonstrated by both stringent burden testing and statistical analyses (Easton et al., 2015). Moreover, despite extensive sequencing efforts, variants in known BC susceptibility genes are present in < 30% of BC cases with positive family history or an early age of onset (Shiovitz and Korde, 2015; Chandler et al., 2016), meaning that the underlying genetic factors for most HBC remain unknown.

In the past few years, advances in next-generation sequencing (NGS), specially whole-exome sequencing (WES), have led to the identification of causative variants in several rare familial syndromes, including hereditary cancer (Comino-Méndez et al., 2011; Seguí et al., 2015). Up to the present time, more than 16 different WES studies (both family-based and case studies) have been carried out for HBC, and a few novel BC susceptibility genes were identified: *XRCC2*, *RINT1*, *RECQL*, and *FANCM* (Chandler et al., 2016). Nevertheless, the small number of novel major BC autosomal dominant predisposing genes disclosed in these studies has pointed to the possible existence of very rare, or even particular, high and moderate penetrant variants. Conversely, other forms of inheritance, such as recessive and oligogenic transmission of cancer predisposition, cannot be discarded (Sokolenko et al., 2015). In this sense, further WES investigation in different families or populations is crucial for expanding the catalog of breast tumor predisposing genes.

In two previous studies of our group, we screened young BC women (Carraro et al., 2013) and women with clinical criteria of HBOC (Silva et al., 2014) for pathogenic variants in the complete coding sequence of *BRCA1*, *BRCA2*, and *TP53* genes, and for *CHEK2* c.1100delC point mutation, detecting 22–26% of pathogenic variant carriers. Both studies disclosed a large number of women negative for pathogenic variants in the most important genes associated with BC risk, claiming for the necessity of identifying rare and/or novel BC predisposing genes. Thus, the aim of the current study was to investigate, by WES, breast cancer patients with clinical criteria for HBOC and

without pathogenic variants in major breast cancer predisposing genes, using rigorous functional criteria for selection of detected variants, in order to identify the most promising new HBC-causing genes.

MATERIALS AND METHODS

Patients and Controls

WES was performed in 17 patients from A.C. Camargo Cancer Center (15 unrelated patients and two siblings) diagnosed with BC and fulfilling one or more of the following criteria of HBOC syndrome: early onset BC (<36 years); bilateral BC; breast plus another primary related tumor (ovary, fallopian tube or primary peritoneal tumors). These patients were selected from previous studies (Carraro et al., 2013; Silva et al., 2014) from our group and were negative for pathogenic variants in *BRCA1/2*, *TP53*, and *CHEK2* c.1100delC. Two patients (including the two sisters) were carriers of variants of uncertain clinical significance (VUS) in *BRCA1* gene. The detailed inclusion criteria from both studies were described previously (Carraro et al., 2013; Silva et al., 2014). One affected woman of one family participated in the cosegregation study for specific candidate variants.

Five germline *BRCA1*-mutation carriers that were submitted to WES in the same platform were included for variant filtering. For validation of selected variants, target NGS validation was applied in 25 healthy women without family history of cancer, considered here as a control group. Additionally, a selected number of candidate genes were screened in an independent group of 42 patients at risk for HBC from a distinct project, obtained from Barretos Cancer Hospital (Barretos, São Paulo, Brazil). **Figure 1** depicts the study design and workflow, describing the projects steps and the analysis performed in each patients and controls groups.

All participants signed an informed consent. This study was performed in accordance with the Helsinki Declaration and was approved by the A.C. Camargo Cancer Center (1754/13) and Barretos Hospital (916/2015) ethics committees.

DNA Isolation

Genomic DNA was obtained from A.C. Camargo Cancer Center Biobank. In brief, DNA was extracted from peripheral leukocytes by Puregene®-DNA purification Kit (Qiagen, Hilden, Germany), according to manufacturer's instructions. DNA concentration, purity and integrity were assessed by spectrophotometry (Nanodrop 2000—Thermo Fisher Scientific, Waltham, MA) and fluorometry (Qubit—Life Technologies, Foster City, CA, USA).

Whole Exome Sequencing

For the 17 patients of the discovery set, WES was performed using the SOLiD and/or Ion Proton platforms. For SOLiD exomes, libraries were prepared using SOLiD™ Fragment Library Barcoding Kit (Life Technologies) and SureSelect Human All Exon V4 Kit 50 Mb (Agilent Technologies), according to the manufacturer's instructions. Sequencing of paired-end libraries (50 X 75 bp) was performed in a Solid 5500XL System (Life Technologies). For Ion Proton exomes, libraries were prepared using Ion Xpress™ Plus Fragment Library Kit and Ion TargetSeq™ Exome Kit (Thermo Fisher Scientific),

---

[1]Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 2018. Available online at: https://omim.org/

FIGURE 1 | Variants selection workflow. WES data from 17 breast cancer patients were analyzed using quality, frequency, and functional based filters, resulting in 186 breast cancer predisposing candidate genes. A final 23 candidate genes were entirely investigated for LOF or possibly pathogenic variants in 42 additional BC patients suspected of HBC and negative for mutations in the major predisposing genes, resulting in 12 final HBC candidate genes. FS, FisherStrand; QD, QualByDepth; PPV, probably pathogenic variant; PV, pathogenic variant; HBOC, Hereditary Breast and Ovarian Cancer; LOF, loss of function; SA, segregation analysis; WES, whole exome sequencing.

according to the manufacturer's instructions. Each Ion Proton exome library was sequenced on Ion Proton instrument using Ion PI Sequencing 200 Kit v3 and Ion PI Chip v3 (Thermo Fisher Scientific). The resulting sequences were mapped to the reference genome (GRCh37/hg19). Base Calling and alignment were performed by SOLiD™ BioScope 1.2™ Software (Life Technologies) (SOLID data) and by Torrent Suite v4.2 server (Ion Proton data). Variant calling and annotation were done by GATK (Genome Analysis Toolkit) pipeline made available by the Broad Institute. The data obtained in this study is available at Sequence Read Archive (SRP120031).

Variants Selection and Prioritization

For variant filtering, identified variants were annotated with VarSeq (Golden Helix) against reference databases (RefSeq, 1000Genomes, ESP6500, ExAC, dbSNP, and ClinVar). First, for quality filtering, we selected variants with QD > 2 (QD = variant call confidence normalized by depth of sample reads supporting a variant), FS < 6 (FS = strand bias estimated by GATK using Fisher's Exact Test), base coverage $\geq$ 10x, variant allele frequency (VAF) > 0.25. For four patients with data from both Solid and Ion Proton, only variants detected in both platforms were selected. For one patient with data exclusively from Ion Proton, variants occurring in regions of homopolymer > 4 bases were excluded.

Qualified variants were excluded if present in five *BRCA1*-mutation carriers patients analyzed by WES in Solid 5500, and variants present in population databases with frequency > 1% (minor allele frequency [MAF] > 0.01), as well as variants present in more than three unrelated patients. Finally, a recently public available Brazilian database of WES from 609 healthy individuals (Abraom—Brazilian genomic variants; http://abraom.ib.usp.br/) was also used for manually excluding population-specific variants (MAF > 0.01).

Next, for a function-based prioritization, we selected variants leading to loss of function in any gene (frameshift indels, stop codon, and canonical splice site variants) and missense or in-frame indels variants in 832 genes of interest. These genes were selected from commercial panels targeting somatic and germline cancer mutated genes, consensus cancer genes previously described (Futreal et al., 2004) and genes from DNA repair pathways (from KEGG and Putnam et al., 2016) (Supplementary Table 1). For the two related patients, any shared missense or in-frame indels variants in these 832 genes were selected. For the 15 unrelated patients, we selected only variants predicted to be damaging in at least four out of six variant effect prediction software. For these analyses, the results from the following tools were obtained using VarSeq: SIFT, Polyphen v2, Functional Analysis through Hidden Markov

Models (FATHAMM and FATHAMM-MKL), MutationAssessor and MutationTaster. Additionally, we analyzed the potential effect on splicing of the selected LOF and missense variants using dbscSNV annotations (cut-off > 0.6 in ADA and/or RF scores).

Sanger Validation

Two pathogenic variants (PV) or probably pathogenic variants (PPV) in *BARD1* and *ATM* were validated by Sanger sequencing. Briefly, 50 ng of leukocyte DNA was submitted to PCR performed with GoTaq Green Master Mix (Promega), cleaned with ExoSAP-IT (USB Corporation) and sequenced in both directions with BigDye Terminator v3.1 (Life Technologies) using an ABI 3130xl DNA sequencer (Life Technologies), according to manufacturer's instructions. The sequencing results were aligned using CLCBio Genomics Workbench Software (CLCBio, Qiagen). Primer sequences are available under request.

Targeted NGS Validation

A subset of 139 variants (Supplementary Table 2) selected from exome data were validated by multiplex targeted NGS using a custom Ion AmpliSeq panel. Primers were designed using Ion AmpliSeq Designer v3.0.1 (Life Technologies). Libraries were prepared with 20 ng of DNA from each patient using Ion AmpliSeq™ Library Kit 2.0 (Life Technologies). Sequencing was performed using either Ion PGM or Ion Proton platforms, according to the manufacturer's instructions. Sequencing reads mapped to the human genome reference (hg19) using Torrent Suite Browser 4.0.1. On average 166,697 mapped reads were obtained per sample, yielding a mean targeted base coverage of 156X (ranging from 54 to 450). Variants were identified using the VariantCaller v4.0.r73742 plugin and confirmed using CLC Genomics Workbench software (Qiagen). The identified variants were considered if base coverage was ≥10x and VAF > 25%.

To filter out genetic variants common in Brazilian population, the validated variants were evaluated in control group of 25 healthy women by using the same panel. For that, pools of five equimolar genomic DNA samples were prepared by containing 4 ng of each patient (five patients per pool). Libraries preparation, sequencing and mapping were performed as described above. On average 928,194 mapped reads were obtained per pool (mean targeted base coverage 1114X; ranging from 990 to 1,314). Variant calls were obtained using the VariantCaller v4.0.r73742 plugin applying the following filter parameters: VAF > 2%; variant coverage ≥10X.

Cosegregation Analysis

For one family in which a segregation analysis was feasible, DNA from one additional affected individual was obtained. The cosegregation study of specific variants was performed using the same custom gene panel and protocol described previously or with amplicon based library construction and sequencing in Ion Proton platform.

Independent Cohort Validation

For screening the HBC predisposing candidate genes selected in this study an independent cohort comprised of 42 breast cancer patients at risk for HBC from Barretos Cancer Hospital was used.

These samples were analyzed through WES in a parallel study using Nextera Rapid Capture Expanded Exome and NextSeq 500 System (Illumina, San Diego, CA). In these data, we assessed the entire coding regions of the 23 genes disclosed in this study for the presence of rare and possibly pathogenic variants, using the same criteria as in our discovery cohort.

RESULTS

In this study we used WES to disclose variants contributing to BC increased risk in patients fulfilling stringent clinical criteria indicating a genetic predisposition to BC and that were negative for pathogenic variants in four major BC genes (*BRCA1/2*, *TP53*, and *CHEK2* 1100delC). The clinical features and family history of cancer for the 17 selected patients are described in Supplementary Table 3.

For the WES, an average of 46,307,427 sequence reads was obtained for each patient and 75.7% (average) of the target bases were covered by 10 or more reads (Supplementary Table 4). More than 200,000 variants were identified in these patients. To prioritize the identified variants, we applied several filters focusing on quality, frequency and function of the identified alterations. The workflow of the variant prioritization is depicted in **Figure 1** and the details of used filters are described in the Materials and Methods section.

Regarding frequency filters, we excluded variants with a minor allele frequency (MAF) >1% in public databases or those present in five germline *BRCA1*-mutation carriers sequenced in our facility, assuming that these variants represent benign or low-penetrance variants. Following these initial data filtering, 25,412 were identified.

Variants in Moderated and High Penetrance Breast Cancer Genes

Initially, we used WES data to search for rare variants in 27 well-established and emerging HBC predisposing genes (the four previously evaluated genes (*BRCA1/2*, *TP53*, and *CHEK2* c.1100delC) and 23 additional genes): *ATM, BARD1, BLM, BRCA1, BRCA2, BRIP1, CDH1, CHEK2, FANCC, FANCM, MLH1, MSH2, MUTYH, NBN, NF1, PALB2, PMS2, PTEN, RAD51C, RAD51D, STK11, TP53, FAM175A, MRE11, RAD51B, RECQL,* and *RINT1* (Nielsen et al., 2016). In this analysis, we identified two patients harboring frameshift indel variants (one in *ATM* and one in *BARD1*) and five patients (including the two sisters) with variants of uncertain clinical significance (VUS) (**Table 1**). In three patients (MJ2037 and MJ2007/2012) we confirmed the *BRCA1* VUS previously detected by Sanger sequencing. All variants detected in these genes were classified according to the ACMG guidelines (Richards et al., 2015).

The *ATM* p.(Tyr2334Glnfs*4) variant is described as pathogenic in ClinVar database. The *BARD1* p.(Tyr739Leufs*2) is not described in any database and was classified as probably pathogenic, since it is a rare truncating variant leading to partial loss of the second BRCT domain and the phosphobinding region. These two variants were confirmed by Sanger sequencing

TABLE 1 | Pathogenic and VUS detected in 27 known HBC genes.

Patient	Gene	HGVS nomenclature	N of 6 Damaging	dbSNP	MAF (ExAC/ Abraom)	Clinical significance (ClinVar)	Clinical significance (ACMG)
SM001.049	*ATM*	c.7000_7003delTACA; p.(Tyr2334Glnfs*4)	–	rs786203421	ND/ND	Pathogenic	Pathogenic
MJ1007	*BARD1*	c.2215dupT; p.(Tyr739Leufs*2)	–	ND	ND/ND	ND	Probably Pathogenic
MJ2003	*RINT1*	c.961T>A; p.(Phe321Ile)	5 of 6	ND	ND/ND	ND	VUS
MJ2001	*RAD51B*	c.728A>G; p.(Lys243Arg)	4 of 6	rs34594234	0.007/0.005	ND	VUS
MJ2037	*BRCA1*	c.5006C>T p.(Ala1669Val)	5 of 6	ND	ND/ND	ND	VUS
MJ2007/2012[#]	*BRCA1*	c.4963T>C; p.(Ser1655Pro)	6 of 6	ND	ND/ND	ND	VUS

[#]Sisters; N of 6 Damaging: predictions considered as damaging in 6 pathogenicity predicting software; MAF, minor allele frequency; ND, not described; VUS, variant of unknown clinical significance. RefSeq reference number of transcripts are described at Supplementary Table 2.

in the proband and, for *ATM,* also in one affected relative (Supplementary Figure 1).

Four rare missense variants identified in our patients were classified as probably damaging by at least four prediction software, and three of them are not described in any population database. Three of them are located in recognized functional domains of the affected proteins: *BRCA1* p.Ala1699Val and p.Ser1655Pro are located at the C-terminal BRCT domain, responsible for BRCA1 interaction with others DNA repair proteins and *RINT1* p.Phe321Ile is located at the functional TIP20 domain.

Candidate Selection for Novel Breast Cancer Predisposing Genes

Next, for the 15 patients without any probable pathogenic variant (excluding *ATM* and *BARD1* mutated patients) we applied a functional-based variant prioritization. Candidate variants were selected according to the predicted impact in the protein function and affected gene, including all loss-of-function variants (nonsense, frameshift indels, and splice site) as well as missense and in-frame indels occurring in a list of 832 cancer-related genes (DNA repair and cancer related genes—Supplementary Table 1). For the two sisters (MJ2007 and MJ2012), all variants shared between the two were selected as candidates. For the 13 unrelated patients, we selected missense variants predicted to be damaging by at least 4 out of 6 prediction software.

After filtering, we obtained a total of 208 variants, including 125 LOF and 83 missenses (Supplementary Table 2). In order to technically validate our variant selection workflow, a subset of these 208 variants (133 out of 208) was submitted for technical validation by targeted NGS in the same WES samples and, of these, 126 were validated (95%) (Supplementary Table 4). Using this same custom panel, we evaluated 25 control samples of healthy Brazilian women without cancer for filtering common polymorphisms in our population. Eight variants were detected in at least one control sample and where then excluded from our

candidates list, resulting in 193 candidate variants (118 validated and 75 not evaluated).

For the family of the two affected sisters, one additional affected aunt diagnosed with ovarian cancer at age 45 was available for segregation analysis (**Figure 2**). We analyzed 17 variants that were shared between the two sisters and 8 variants were also present in the aunt, including the VUS variant in *BRCA1* (**Table 2**).

Then, the remaining 186 genes prioritized in our study were compared to candidate genes reported in eight previous WES studies of HBC (Snape et al., 2012; Thompson et al., 2012; Gracia-Aznarez et al., 2013; Hilbers et al., 2013; Kiiski et al., 2014; Wen et al., 2014; Noh et al., 2015; Kim et al., 2017) and 12 common genes were identified, 9 of them presenting LOFs variants in at least one study (**Table 3**). For two genes the same LOF variants were identified in our and at a second study (*PZP* p.Arg680* and *KRT76* p.Glu276*).

Thus, from the 193 final candidate variants, we selected 23 candidate genes of BC predisposition: 7 novel candidate genes segregating in the 3 members of the MJ2007/2012 family (*SLC22A16, ROS1, IL33, PTPRD, ARHGEF12, ERBB2, POLA1*), five cancer-related genes harboring LOF variants (*GALNT3, RAF1, PICALM, KL, ERCC1*) and 12 genes overlapping with candidate genes identified in other studies (*CAPN9, KRT76, PZP, DNAH7, MST1R, LAMB4, NIN, MSH3, SLX4, DDX1, NOTCH2,* and *ROS1—ROS1* was also selected in the segregating genes list). The entire coding region of the 23 genes were evaluated in an independent Brazilian cohort.

Assessing 23 Candidate Genes in an Independent Cohort of Patients at Risk for HBC

To select the most promising candidate genes, we analyzed the 23 candidate genes disclosed in our study in an independent cohort of 42 Brazilian women at risk for HBC. These patients were all negative for pathogenic variants in *BRCA1/2, TP53,* and *ATM* genes. In these data, we assessed the entire coding regions
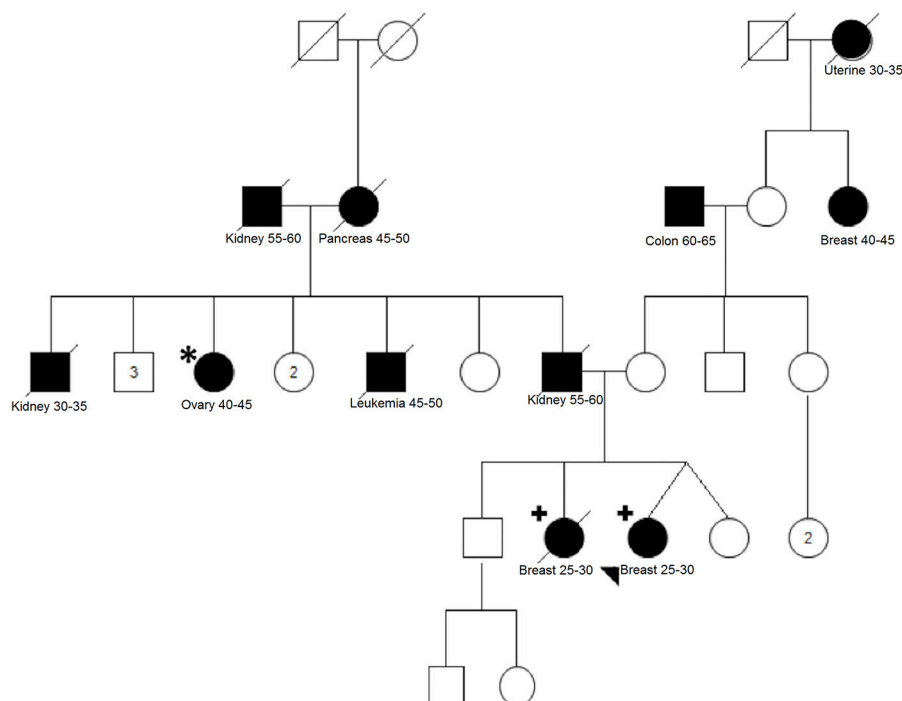
FIGURE 2 | Pedigree of family MJ2007/MJ20012. Two breast cancer affected sisters (indicated with the plus sign) were analyzed using WES. Variants shared between the sisters were screened in one affected paternal aunt (indicated with an asterisk). Filled-in symbols indicate individuals affected by cancer. Cancer type, cancer age of onset or age appear underneath each individual. Numbers inside the symbols indicate the number of unaffected siblings not shown individually in the pedigree.

of the selected genes for the presence of rare (MAF < 1%) and possibly pathogenic variants, selecting all LOF variants and missense variants predicted to be pathogenic in at least 3 out of 6 algorithms.

In this cohort, we detected 16 variants in 12 of the 23 candidate genes (**Table 4**). *NOTCH2* gene was the one with more variants, harboring three missense; *ERBB2* and *DNAH7* harbored two missenses each. Only one LOF variant was detected, affecting *ERCC1* gene, which was the same variant detected in our discovery cohort (c.875G>A; p.Trp292*). The remaining genes presented one rare missense variant each.

DISCUSSION

Recently, the use of WES in clinical genetics has been proven to be an effective alternative for establishing the genetic basis of Mendelian diseases, particularly in diseases where multiple genes can be affected (Trujillano et al., 2016). Moreover, in both clinical and research settings, WES has been applied to elucidate the genetic cause of cancer predisposition. In this sense, WES offers the opportunity to concomitantly investigate several known cancer risk genes as well as to identify novel cancer predisposing genes. Thus, in this study we used WES to disclose variants contributing to BC increased risk in patients that were negative for pathogenic variants in three major BC genes—*BRCA1/2* and *TP53* genes—and the most common point mutation in *CHEK2*

gene (c.1100delC). For this, we used stringent clinical criteria for selecting patients with strong indicative of harboring a genetic predisposition to BC, such as early onset BC (<36 years); bilateral BC; or the presence of a second primary related tumor.

First, by evaluating known BC predisposing genes, we could establish the causative variants in two probands. One of them harbored an *ATM* truncating pathogenic variant and the other a novel *BARD1* truncating variant, considered as probably pathogenic. The *BARD1* p.(Tyr739Leufs*2) variant is predicted to cause partial loss of the second functional BRCT domain and the phosphobinding region. Several studies suggest that both BRCT repeats are necessary for BARD1 normal function (Birrane et al., 2007; Irminger-Finger et al., 2016) and truncating variants in this region have been previously reported in association with HBC (De Brakeleer et al., 2010). Additionally, compatible with the probable pathogenic role of this variant, our proband presented triple negative BC and *BARD1* pathogenic variants were recently described to be related to this molecular subtype (De Brakeleer et al., 2016).

Besides these LOF variants, we identified four rare missense VUS in three HBC genes (*BRCA1, RINT1,* and *RAD51B*). The identification of VUS in genetic testing represent a challenging concern for genetic counselors due to uncertainty in clinical decision making, which can lead to more intensive management than necessary in most of the times or, more rarely, in inappropriate prevention measures (Plon et al., 2011). The recently introduction of NGS gene panels in genetic testing

TABLE 2 | Cosegregation analysis of variants detected in the sisters MJ2007 and MJ2012.

								Cosegregation
				Shared variants of MJ2017 and MJ2012				
Chr:Pos	Ref/Alt	Gene names	Type	HGVS nomenclature	N of 6 Damaging	dbSNP	MAF (ExAC*/ Abraom)	Present in affected aunt (OV)
1:3328745	G/A	*PRDM16*	Missense	c.1984G>A; p.(Val662Met)	1 of 6	ND	ND/0.001	No
1:116670844	G/T	*MAB21L3*	Stop gained	c.739G>T; p.(Glu247*)	2 of 6	rs149122915	0.0002/ND	No
3:48716158	G/A	*NCKIPSD*	Missense	c.1804C>T; p.(Arg602Cys)	2 of 6	ND	0.000008/0.001	No
5:149514363	A/G	*PDGFRB*	Missense	c.581T>C; p.(Ile194Thr)	5 of 6	rs2229560	0.001/0.002	No
6:117622137	C/T	*ROS1*	Missense	c.6733G>A; p.(Gly2245Ser)	2 of 6	rs142264513	0.0008/0.002	No
7:116380062	A/G	*MET*	Missense	c.1451A>G; p.(His484Arg)	2 of 6	ND	0.00005/ND	No
8:17815082	T/G	*PCM1*	Missense	c.1838T>G; p.(Ile613Ser)	0 of 6	rs181777656	0.003 (OT 0.01)/ 0.002	No
14:55467701	T/C	*WDHD1*	Missense	c.703A>G; p.(Ile235Val)	1 of 6	rs139440460	0.004/0.004	No
15:40897315	A/G	*KNL1*	Missense	c.43A>G; p.(Ile15Val)	0 of 6	ND	0.00003/ND	No
6:110778048	C/-	*SLC22A16*	Frameshift	c.226delG; p.(Ala76fs*66)	–	ND	ND/ND	Yes
6:117715381	A/G	*ROS1*	Missense	c.1108T>C; p.(Ser370Pro)	2 of 6	rs56274823	0.002/ND	Yes
9:6255967	G/C	*IL33*	Splice acceptor	c.613-1G>C; p.(spl?)	–	rs146597587	0.002/0.001	Yes
9:8501026	G/A	*PTPRD*	Missense	c.1856C>T; p.(Thr619Ile)	2 of 6	ND	ND/0.001	Yes
11:120298916	C/T	*ARHGEF12*	Missense	c.545C>T; p.(Ser182Phe)	3 of 6	rs147982337	0.002/0.002	Yes
17:37865694	G/A	*ERBB2*	Missense	c.563G>A; p.(Arg188His)	3 of 6	ND	0.00002/0.002	Yes
17:41222968	A/G	*BRCA1*	Missense	c.4963T>C; p.(Ser1655Pro)	6 of 6	ND	ND/ND	Yes
X:24861673	T/C	*POLA1*	Missense	c.3908T>C; p.(Met1303Thr)	2 of 6	ND	ND/ND	Yes

*Chr, chromosome; Pos, position; Ref, reference allele; Alt, alternate allele; N of 6 Damaging, predictions considered as damaging in 6 pathogenicity predicting software; ND, not described; MAF, minor allele frequency; OV, ovary cancer; OT, others. *Variants in ExAc that had a MAF >1% in any ethnic group are underlined and the highest ExAc population MAF is shown inside parenthesis. RefSeq reference number of transcripts are described at Supplementary Table 2.*

have increased the number of patients diagnosed with VUS, emphasizing the urgent need for better pathogenicity predictions models and collaborative efforts to increase observational data that can aid a posteriori classification to variants, such as cosegregation analysis, personal and family history, co-occurrence with pathogenic variants, and histological and molecular features of tumors (Spurdle et al., 2012).

In the 15 patients without known pathogenic variants, we could identify more than 25,000 novel or rare variants (MAF < 1%), thus several filtering strategies were applied to prioritize those more likely to be related to HBC. Since the majority of hereditary cancer predisposing genes harbor an excess of loss-of-function variants, we focused on this type of overtly deleterious variants, regardless of the affected gene. Furthermore, most BC risk genes are involved in DNA repair and genomic integrity pathways (Shiovitz and Korde, 2015; Nielsen et al., 2016), and prioritizing variants in these genes is a rational approach that have been used successfully in previous studies (Mantere et al., 2016). As so, we have also focused on missense variants in a defined set of cancer-related and DNA repair genes. By doing that, we were able to reduce our candidate genes list to a few hundreds.

Importantly, for one family with two sisters affected by BC at young ages (29 years), we could improve the selection by retaining only shared variants and also perform segregation analysis of the candidate variants in an aunt affected by ovarian cancer. From this analysis, eight cosegregating variants emerged, including a *BRCA1* VUS. Besides *BRCA1* gene, only *ERBB2* has been previously implicated in BC predisposition, although

with conflicting data about the increased risk conferred by some alleles (Breyer et al., 2009; Wang et al., 2013). Regarding the two LOF variants found to be cosegregating in this family (genes *SLC22A16* and *IL33*), no relation between both genes and BC could be recognized in the literature.

One possible explanation for the results observed in this family and that could also be responsible for the cancer predisposition in other patients of our study is the polygenic model. In this model, which has been suggested and reviewed by different authors (Oldenburg et al., 2007; Shiovitz and Korde, 2015), moderate and low penetrance alleles would act in synergy and play a predominant role. Additionally, the high number of affected relatives with different tumor types in both maternal and paternal sides of this family can be a confounding factor for understanding the phenotypes and cosegregation results. Unfortunately, most affected family members of this family were deceased, limiting additional investigations and the interpretation of our findings.

To gain further insight on the relevance of our identified candidate genes, we evaluated the most promising ones in an independent cohort comprising 42 Brazilian HBC women. Several rare and possibly damaging variants were identified in this cohort, providing additional evidence of the potential role in BC predisposition of some new genes. Of those, we highlight four genes related to cancer development and progression (*NOTCH2*, *ERBB2*, *MST1R*, and *RAF1*) and two DNA repair genes (*ERCC1* and *SLX4*). Interestingly, *ERCC1* and *SLX4* are partners that act in the repair of interstrand cross-links and are also required for homology-directed repair of DNA double-strand breaks.

TABLE 3 | Overlapping genes selected as candidates in other WES studies.

| | | | 15 patients (current study) | | | Other studies | | | | |
Patient	Gene	Type	HGVS nomenclature	dbSNP	MAF (ExAC*/Abraom)	Number/type	HGVS nomenclature	dbSNP	MAF (ExAC*/Abraom)	Study
SM001.040	NOTCH2	missense	c.2292T>A; p.(Asn764Lys)	ND	0.00002/ND	2 missense	c.854G>A; p.(Arg285His) c.6178C>T; p.(Arg2060Cys)	rs782452794 rs746551843	0.000008/ND 0.000008/ND	Wen et al., 2014
MJ2013S	CAPN9	splice site	c.1657+2T>G; p.(spl?)	rs143145032	0.001/0.002	1 frameshift del	c.1976_1982delAGAATG; p.(Glu659Glyfs*20)	ND	0.002/ND	Thompson et al., 2012
MJ2014S and MJ2016S	DNAH7	nonsense	c.10359T>G; p.(Tyr3453*)	ND	0.000008/ND	1 frameshift ins	c.4787dupA; p.(Tyr1596*)	rs573013205	0.002/ND	Thompson et al., 2012
SM001.06	MST1R	splice site	c.1231-1G>C; p.(spl?)	ND	0.00008/ND	1 nonsense	c.3322C>T; p.(Arg1108*)	rs150876558	0.0001/ND	Thompson et al., 2012
SM001.088	MSH3	missense	c.2659G>A; p.(Asp887Asn)	ND	0.00002/ND	1 in frame del	c.162_179del; p.(Ala57_Ala62del)	ND	ND	Thompson et al., 2012
MJ2007/ MJ2012	ROS1	missense	c.1108T>C; p.(Ser370Pro)	rs56274823	0.002/ND	1 splice site 1 nonsense	c.5079+2T>C; p.(spl?) c.3303G>A; p.(Trp1101*)	ND rs200145587	ND 0.00009/ND	Thompson et al., 2012
MJ2004S	LAMB4	splice site	c.34+1G>A; p.(spl?)	rs7788865	0.001 (AF 0.0166)/0.003	1 frameshift del	c.5265delA; p.(Lys1755Asnfs*11)	rs568834649	0.008/0.004	Thompson et al., 2012
SM001.040	DDX10	missense	c.1088G>A; p.(Arg363His)	ND	0.0001/ND	1 nonsense	c.973G>T; p.(Glu325*)	ND	ND	Kiiski et al., 2014
MJ2016S	PZP	nonsense	c.2038C>T; p.(Arg680*)	rs145240281	0.005/ND	1 nonsense	c.2038C>T; p.(Arg680*)	rs145240281	0.005 (EU 0.01013)/ND	Thompson et al., 2012
MJ2015S	KRT76	nonsense	c.826G>T; p.(Glu276*)	rs149868801	0.007/0.006	1 nonsense	c.826G>T; p.(Glu276*)	rs149868801	0.007/0.006	Thompson et al., 2012
MJ2014S	NIN	in frame del	c.1736_1738delAAG; p.(Glu579del)	ND	0.0003/0.005	1 frameshift del	c.4261delG; p.(Glu1421Lysfs*18)	ND	ND	Thompson et al., 2012
MJ2013S	SLX4	missense	c.4766G>A; p.(Arg1589His)	rs746314060	0.00004/0.0008	1 missense	c.2484G>C; p.(Glu828Asp)	rs199656607	0.0001/ND	Kiiski et al., 2014

ND, not described; MAF, minor allele frequency; AF, Africans; EU, European (Non-Finnish). *Variants in ExAc that had a MAF >1% in any ethnic group are underlined and the highest ExAc population MAF is shown inside parenthesis. RefSeq reference number of transcripts are described at Supplementary Table 2.

TABLE 4 | Variants detected in 12 candidate genes at the discovery cohort and at an independent cohort.

			15 patients (discovery cohort)				42 patients (validation cohort)				
Patient	Gene	Type	HGVS nomenclature	N of 6 Damaging	dbSNP	MAF (ExAC*/Abraom)	Number/type	HGVS nomenclature	N of 6 Damaging	dbSNP	MAF (ExAC/Abraom)
SM001.040	NOTCH2	missense	c.2292T>A; p.(Asn764Lys)	6 of 6	ND	0.00002/ND	3 missense	c.2579T>G; p.(Leu860Trp) c.3625T>G; p.(Phe1209Val) c.7223T>A; p.(Leu2408His)	3 of 6 5 of 6 4 of 6	ND rs147223770 rs35586704	ND/ND 0.003/0.002 0.002/0.007
MJ2014S and MJ2016S	DNAH7	nonsense	c.10355T>G; p.(Tyr3453*)	–	ND	0.000008/ND	2 missense	c.3265C>T; p.(Pro1089Ser) c.11947C>T; p.(Arg3983Trp)	5 of 6 5 of 6	ND rs114621989	ND/ND 0.009/0.006
MJ2016S	RAF1	frameshift	c.1241dupA; p.(Asp415fs)	–	ND	ND/ND	1 missense	c.923C>T; p.(Pro308Leu)	3 of 6	rs5746220	0.002/0.008
SM001.06	MST1R	splice site	c.1231-1G>C; p.(spl?)	–	ND	0.00008/ND	1 missense	c.4180T>A; p.(Ser1394Thr)	4 of 6	rs141338964	0.0002/ND
MJ2004S	LAMB4	splice site	c.34+1G>A; p.(spl?)	–	rs7788865	0.001/0.003	1 missense	c.1843C>A; p.(Pro615Thr)	5 of 6	rs201909531	0.0003/0.0008
MJ2014S	NIN	in frame del	c.1736_1738delAAG; p.(Glu579del)	–	ND	0.0003/0.005	1 missense	c.848C>T; p.(Ser283Leu)	3 of 6	rs763293400	0.0001/ND
MJ2013S	SLX4	missense	c.4766G>A; p.(Arg1589His)	5 of 6	rs74631406	0.00004/0.0008	1 missense	c.3368C>A; p.(Ser1123Tyr)	3 of 6	rs144647122	0.0003/0.0008
MJ2037S	ERCC1	nonsense	c.875G>A; p.(Trp292*)	–	rs116640350	0.002 (AF 0.02627/0.0008	1 nonsense	c.875G>A; p.(Trp292*)	–	rs116640350	0.002/0.0008
MJ2007/2012	SLC22A16	frameshift	c.226delG; p.(Ala76fs*66)	–	ND	ND/ND	1 missense	c.599C>T; p.(Ala200Val)	5 of 6	rs61729086	0.003/0.005
MJ2007/2012	PTPRD	missense	c.1856C>T; p.(Thr619Ile)	2 of 6	ND	ND/0.0008	1 missense	c.2585G>T; p.(Arg862Leu)	5 of 6	rs142397137	0.0001/ND
MJ2007/2012	ARHGEF12	missense	c.545C>T; p.(Ser182Phe)	3 of 6	rs147982337	0.002/0.002	1 missense	c.3986G>A; p.(Arg1329Gln)	3 of 6	ND	ND/ND
MJ2007/2012	ERBB2	missense	c.563G>A; p.(Arg188His)	3 of 6	ND	0.00002/0.002	2 missense	c.236A>C; p.(Glu79Ala) c.1586T>C; p.(Val529Ala)	6 of 6 3 of 6	rs61737968 ND	0.0009/0.005 ND/ND

N of 6 Damaging, predictions considered as damaging in 6 pathogenicity predicting software; ND, not described; MAF, minor allele frequency; AF, Africans. *Variants in ExAc that had a MAF >1% in any ethnic group are underlined and the highest ExAc population MAF is shown inside parenthesis. RefSeq reference number of transcripts are described at Supplementary Table 4.

Additionally, *ERCC1* is also involved in the nucleotide excision repair pathway (McNeil and Melton, 2012). Both genes have been investigated regarding BC susceptibility, with some common *ERCC1* variants being identified as risk alleles in Chinese population (Yang et al., 2013) and rare truncating and possibly damaging variants in *SLX4* being described in some high risk HBOC patients (Bakker et al., 2013; Shah et al., 2013). Remarkably, in the *ERCC1* gene we identified the same nonsense variant in both discovery and validation cohorts (p.Trp292*), while in *SLX4* one of the rare missense identified in our cohorts (p.Ser1123Tyr) was previously described in one HBC patient (Shah et al., 2013).

Some limitations of our study are inherent to WES method since predisposition variants can be located in non-coding or not captured regions of the genome, such as promoter or deep intronic pathogenic variants. Moreover, although the strategic filtering applied here is necessary to reduce the number of proposed candidates, it can result in the omission of the causative variant (for example, by excluding protein-impacting synonymous variants). Additionally, large genomic rearrangements have been implicated in HBC, and even though specific bioinformatics pipelines can be applied in WES data to extract these results, these analyses were not performed in our study. Finally, when it comes to interpreting the potential effect of our candidate variants in splicing, both coding as well as splice site variants can cause splicing alterations that lead to in-frame functional proteins instead of frameshift truncated ones, and functional assays would be necessary to validate bioinformatics predictions.

Considering the evidence presented here, we can neither conclude that these variants identified in the 15 patients negative for known pathogenic variant are the definitive cause of BC predisposition nor determine the magnitude of the risk that these genes could present. Nevertheless, our results provide a set of novel putative BC predisposing genes and reinforce WES as useful tool for assessing the complex landscape of HBC predisposition. Importantly, this represents the first WES data of a HBC cohort from South America and the analysis of an admixed population such as the Brazilian can reveal unique features compared to other Western populations. In this sense, the WES data generated in our study, as well as other

previous and future studies, can be reanalyzed in the future and possibly identify genetic overlaps between families, aiding to gene discoveries (Chandler et al., 2016). Finally, the assignment of a novel gene or specific variant as a true BC predisposition factor requires solid phenotypic evidence from cosegregation analysis, *in vitro* and *in vivo* functional assays and genotyping large series of case and controls from distinct populations. The efforts for discovery and validation of novel HBC genes will continue to provide insights into disease mechanisms, eventually leading to the development of more effective therapies and improved management of affected families.

AUTHOR CONTRIBUTIONS

GT, FdS, EF, and DC: conceived and designed the experiments; GT, FdA, MF, BB, CdP, and EF: performed and analyzed the experiments; RV, JdS, RR, and SdS: performed bioinformatics analysis; AdN, MA, PF, and EP: assessed clinical data, selected, and recruited the patients; SdS, EP, and DC: contributed reagents, materials, and analysis tools; GT, FdA, and DC: wrote and edited the paper. All authors have read and approved the final manuscript.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00161/full#supplementary-material

REFERENCES

Anglian Breast Cancer Study (2000). Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group. *Br. J. Cancer* 83, 1301–1308. doi: 10.1054/bjoc.2000.1407

Bakker, J. L., van Mil, S. E., Crossan, G., Sabbaghian, N., De Leeneer, K., Poppe, B., et al. (2013). Analysis of the novel fanconi anemia gene SLX4/FANCP in familial breast cancer cases. *Hum. Mutat.* 34, 70–73. doi: 10.1002/humu.22206

Birrane, G., Varma, A. K., Soni, A., and Ladias, J. A. (2007). Crystal structure of the BARD1 BRCT domains. *Biochemistry* 46, 7706–7712. doi: 10.1021/bi700323t

Breyer, J. P., Sanders, M. E., Airey, D. C., Cai, Q., Yaspan, B. L., Schuyler, P. A., et al. (2009). Heritable variation of ERBB2 and breast cancer risk. *Cancer Epidemiol. Biomarkers Prev.* 18, 1252–1258. doi: 10.1158/1055-9965.EPI-08-1202

Carraro, D. M., Koike Folgueira, M. A., Garcia Lisboa, B. C., Ribeiro Olivieri, E. H., Vitorino Krepischi, A. C., de Carvalho, A. F., et al. (2013). Comprehensive analysis of BRCA1, BRCA2 and TP53 germline mutation and tumor characterization: a portrait of early-onset breast cancer in Brazil. *PLoS ONE* 8:e57581. doi: 10.1371/journal.pone.0057581

Chandler, M. R., Bilgili, E. P., and Merner, N. D. (2016). A Review of whole-exome sequencing efforts toward hereditary breast cancer susceptibility gene discovery. *Hum. Mutat.* 37, 835–846. doi: 10.1002/humu.23017

Comino-Méndez, I., Gracia-Aznárez, F. J., Schiavi, F., Landa, I., Leandro-García, L. J., Letón, R., et al. (2011). Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma. *Nat. Genet.* 43, 663–667. doi: 10.1038/ng.861

De Brakeleer, S., De Grève, J., Desmedt, C., Joris, S., Sotiriou, C., Piccart, M., et al. (2016). Frequent incidence of BARD1-truncating mutations in germline

DNA from triple-negative breast cancer patients. *Clin. Genet.* 89, 336–340. doi: 10.1111/cge.12620

De Brakeleer, S., De Grève, J., Loris, R., Janin, N., Lissens, W., Sermijn, E., et al. (2010). Cancer predisposing missense and protein truncating BARD1 mutations in non-BRCA1 or BRCA2 breast cancer families. *Hum. Mutat.* 31, 1175–1185. doi: 10.1002/humu.21200

Easton, D. F., Pharoah, P. D., Antoniou, A. C., Tischkowitz, M., Tavtigian, S. V., Nathanson, K. L., et al. (2015). Gene-panel sequencing and the prediction of breast-cancer risk. *N. Engl. J. Med.* 372, 2243–2257. doi: 10.1056/NEJMsr1501341

Elledge, R. M., and Allred, D. C. (1998). Prognostic and predictive value of p53 and p21 in breast cancer. *Breast Cancer Res. Treat.* 52, 79–98. doi: 10.1023/A:1006163101948

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. (2004). A census of human cancer genes. *Nat. Rev. Cancer* 4, 177–183. doi: 10.1038/nrc1299

Gracia-Aznarez, F. J., Fernandez, V., Pita, G., Peterlongo, P., Dominguez, O., de la Hoya, M., et al. (2013). Whole exome sequencing suggests much of non-BRCA1/BRCA2 familial breast cancer is due to moderate and low penetrance susceptibility alleles. *PLoS ONE* 8:e55681. doi: 10.1371/journal.pone.0055681

Hilbers, F. S., Meijers, C. M., Laros, J. F., van Galen, M., Hoogerbrugge, N., Vasen, H. F. A., et al. (2013). Exome sequencing of germline DNA from non-BRCA1/2 familial breast cancer cases selected on the basis of aCGH tumor profiling. *PLoS ONE* 8:e55734. doi: 10.1371/journal.pone.0055734

Honrado, E., Benítez, J., and Palacios, J. (2005). The molecular pathology of hereditary breast cancer: genetic testing and therapeutic implications. *Mod. Pathol.* 18, 1305–1320. doi: 10.1038/modpathol.3800453

Irminger-Finger, I., Ratajska, M., and Pilyugin, M. (2016). New concepts on BARD1: regulator of BRCA pathways and beyond. *Int. J. Biochem. Cell Biol.* 72, 1–17. doi: 10.1016/j.biocel.2015.12.008

Kean, S. (2014). The " other " breast cancer genes. *Science* 343, 1457–1459. doi: 10.1126/science.343.6178.1457

Kiiski, J. I., Pelttari, L. M., Khan, S., Freysteinsdottir, E. S., Reynisdottir, I., Hart, S. N., et al. (2014). Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. *Proc. Natl. Acad. Sci. U.S.A.* 111, 15172–15177. doi: 10.1073/pnas.1407909111

Kim, Y. C., Soliman, A. S., Cui, J., Ramadan, M., Hablas, A., Abouelhoda, M., et al. (2017). Unique features of germline variation in five Egyptian familial breast cancer families revealed by exome sequencing. *PLoS ONE* 12:e0167581. doi: 10.1371/journal.pone.0167581

Mantere, T., Winqvist, R., Kauppila, S., Grip, M., Jukkola-Vuorinen, A., Tervasmäki, A., et al. (2016). Targeted next-generation sequencing identifies a recurrent mutation in MCPH1 associating with hereditary breast cancer susceptibility. *PLoS Genet.* 12:e1005816. doi: 10.1371/journal.pgen.1005816

McNeil, E. M., and Melton, D. W. (2012). DNA repair endonuclease ERCC1-XPF as a novel therapeutic target to overcome chemoresistance in cancer therapy. *Nucleic Acids Res.* 40, 9990–10004. doi: 10.1093/nar/gks818

Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., et al. (2002). Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat. Genet.* 31, 55–59. doi: 10.1038/ng879

Nielsen, F. C., van Overeem Hansen, T., and Sørensen, C. S. (2016). Hereditary breast and ovarian cancer: new genes in confined pathways. *Nat. Rev. Cancer* 16, 599–612. doi: 10.1038/nrc.2016.72

Noh, J. M. , Kim, J., Cho, D. Y., Choi, D. H., Park, W., and Huh, S. J. (2015). Exome sequencing in a breast cancer family without BRCA mutation. *Radiat. Oncol. J.* 33, 149–154. doi: 10.3857/roj.2015.33.2.149

Oldenburg, R. A., Meijers-Heijboer, H., Cornelisse, C. J., and Devilee, P. (2007). Genetic susceptibility for breast cancer: how many more genes to be found? *Crit. Rev. Oncol. Hematol.* 63, 125–149. doi: 10.1016/j.critrevonc.2006.12.004

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

Putnam, C. D., Srivatsan, A., Nene, R. V., Martinez, S. L., Clotfelter, S. P., Bell, S. N., et al. (2016). A genetic network that suppresses genome rearrangements in Saccharomyces cerevisiae and contains defects in cancers. *Nat. Commun.* 11256, 1–14. doi: 10.1038/ncomms11256

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., 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, 405–423. doi: 10.1038/gim.2015.30

Seguí, N., Mina, L. B., Lázaro, C., Sanz-Pamplona, R., Pons, T., Navarro, M., et al. (2015). Germline mutations in FAN1 cause hereditary colorectal cancer by impairing DNA repair. *Gastroenterology* 149, 563–566. doi: 10.1053/j.gastro.2015.05.056

Shah, S., Kim, Y., Ostrovnaya, I., Murali, R., Schrader, K. A., Lach, F. P., et al. (2013). Assessment of SLX4 mutations in hereditary breast cancers. *PLoS ONE* 8:e66961. doi: 10.1371/journal.pone.0066961

Shiovitz, S., and Korde, L. A. (2015). Genetics of breast cancer: a topic in evolution. *Ann. Oncol.* 26, 1291–1299. doi: 10.1093/annonc/mdv022

Silva, F. C., Lisboa, B. C. G., Figueiredo, M. C. P., Torrezan, G. T., Santos, É. M. M., Krepischi, A. C., et al. (2014). Hereditary breast and ovarian cancer: assessment of point mutations and copy number variations in Brazilian patients. *BMC Med. Genet.* 15:55. doi: 10.1186/1471-2350-15-55

Snape, K., Ruark, E., Tarpey, P., Renwick, A., Turnbull, C., Seal, S., et al. (2012). Predisposition gene identification in common cancers by exome sequencing: insights from familial breast cancer. *Breast Cancer Res. Treat.* 134, 429–433. doi: 10.1007/s10549-012-2057-x

Sokolenko, A. P., Suspitsin, E. N., Kuligina, E. S., Bizin, I. V., Frishman, D., and Imyanitov, E. N. (2015). Identification of novel hereditary cancer genes by whole exome sequencing. *Cancer Lett.* 369, 274–288. doi: 10.1016/j.canlet.2015.09.014

Spurdle, A. B., Healey, S., Devereau, A., Hogervorst, F. B. L., Monteiro, A. N., Nathanson, K. L., et al. (2012). ENIGMA–evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Hum. Mutat.* 33, 2–7. doi: 10.1002/humu.21628

Thompson, E. R., Doyle, M. A., Ryland, G. L., Rowley, S. M., Choong, D. Y. H., Tothill, R. W., et al. (2012). Exome sequencing identifies rare deleterious mutations in DNA repair genes FANCC and BLM as potential breast cancer susceptibility alleles. *PLoS Genet.* 8:e1002894. doi: 10.1371/journal.pgen.1002894

Trujillano, D., Bertoli-Avella, A. M., Kumar Kandaswamy, K., Weiss, M. E., Köster, J., Marais, A., et al. (2016). Clinical exome sequencing: results from 2819 samples reflecting 1000 families. *Eur. J. Hum. Genet.* 25, 176–182. doi: 10.1038/ejhg.2016.146

Walsh, T., and King, M.-C. (2007). Ten genes for inherited breast cancer. *Cancer Cell* 11, 103–105. doi: 10.1016/j.ccr.2007.01.010

Wang, H., Liu, L., Lang, Z., Guo, S., Gong, H., Guan, H., et al. (2013). Polymorphisms of ERBB2 and breast cancer risk: a meta-analysis of 26 studies involving 35,088 subjects. *J. Surg. Oncol.* 108, 337–341. doi: 10.1002/jso.23386

Wen, H., Kim, Y. C., Snyder, C., Xiao, F., Fleissner, E. A., Becirovic, D., et al. (2014). Family-specific, novel, deleterious germline variants provide a rich resource to identify genetic predispositions for BRCAx familial breast cancer. *BMC Cancer* 14:470. doi: 10.1186/1471-2407-14-470

Yang, Z., Fang, X., Pei, X., and Li, H. (2013). Polymorphisms in the ERCC1 and XPF genes and risk of breast cancer in a Chinese population. *Genet. Test Mol. Biomarkers* 17, 700–706. doi: 10.1089/gtmb.2013.0122

Prognostic Genes of Breast Cancer Identified by Gene Co-expression Network Analysis

Jianing Tang[1], Deguang Kong[2], Qiuxia Cui[1], Kun Wang[3], Dan Zhang[3], Yan Gong[4]* and Gaosong Wu[1]*

[1] Department of Thyroid and Breast Surgery, Zhongnan Hospital of Wuhan University, Wuhan, China, [2] Department of General Surgery, Zhongnan Hospital of Wuhan University, Wuhan, China, [3] Department of Thyroid and Breast Surgery, Tongji Hospital, Huazhong University of Science and Technology, Wuhan, China, [4] Department of Biological Repositories, Zhongnan Hospital of Wuhan University, Wuhan, China

Breast cancer is one of the most common malignancies. The molecular mechanisms of its pathogenesis are still to be investigated. The aim of this study was to identify the potential genes associated with the progression of breast cancer. Weighted gene co-expression network analysis (WGCNA) was used to construct free-scale gene co-expression networks to explore the associations between gene sets and clinical features, and to identify candidate biomarkers. The gene expression profiles of GSE1561 were selected from the Gene Expression Omnibus (GEO) database. RNA-seq data and clinical information of breast cancer from TCGA were used for validation. A total of 18 modules were identified via the average linkage hierarchical clustering. In the significant module ($R^2 = 0.48$), 42 network hub genes were identified. Based on the Cancer Genome Atlas (TCGA) data, 5 hub genes (CCNB2, FBXO5, KIF4A, MCM10, and TPX2) were correlated with poor prognosis. Receiver operating characteristic (ROC) curve validated that the mRNA levels of these 5 genes exhibited excellent diagnostic efficiency for normal and tumor tissues. In addition, the protein levels of these 5 genes were also significantly higher in tumor tissues compared with normal tissues. Among them, CCNB2, KIF4A, and TPX2 were further upregulated in advanced tumor stage. In conclusion, 5 candidate biomarkers were identified for further basic and clinical research on breast cancer with co-expression network analysis.

Keywords: breast cancer, weighted gene co-expression network analysis (WGCNA), prognosis, GEO, TCGA

INTRODUCTION

Breast cancer is the most frequently diagnosed malignancy and the second leading cause of cancer death in females worldwide, accounting for 30% of cancer diagnoses and 14% of cancer death. In 2017, it was estimated that nearly 252,710 new cases were diagnosed in the United States, with ∼40,610 deaths (1). Therapeutic strategies of breast cancer have been markedly improved. A number of treatments such as surgery, chemotherapy, radiotherapy, hormone therapy, and

targeted therapy are available for breast cancer (2). However, the patients with distant metastases were usually diagnosed with a late stage and nearly incurable (3). Moreover, 30% patients diagnosed with early stage were easy to recur in distant organs even after surgery of removing the primary tumor (4). The classification of breast cancer affects treatment decision and prognosis: hormone-based therapy for ER+ patients; targeted therapy for HER2+ patients; and poorly differentiated cancer often has the worse prognosis (5–7).

Inheritance plays an important role in the development of breast cancer. BRCA1 and BRCA2 are 2 biomarkers which are currently used clinically to assess the familial breast cancer risk. BRCA-associated breast cancer has relatively distinct pathologic characteristics. Up to 20% women with triple-negative breast cancer present BRCA mutations, while BRCA mutations occur less common in general population (8, 9). HER2 expression was found to be upregulated in over 30% patients with breast cancer (10). Previous data suggested that high HER2 levels not only indicated prognostic value, but also affected treatment decisions. Lapatinib and trastuzumab presented dramatically therapeutic effects in patients with HER2-positive breast cancer (11, 12).

Expression levels of hormone receptors (ER/PR) predicted the efficacy of endocrine therapies, and their upregulation was often associated with a favorable prognosis (13). Ki-67 was reported to be associated with disease-free survival (14). High CXCR4 levels were associated with lymph node metastasis and distant metastasis (15). Despite the substantial improvements in the treatment of breast cancer, to date, the ability to treat the advanced ones is still limited due to the lack of precise molecular targets for breast cancer (16). Therefore, it is important to explore the molecule mechanisms involved in the occurrence and development of breast cancer. More novel candidate genes are needed to improve the early diagnosis and treatment decisions.

Co-expression analysis is a powerful technique to construct free-scale gene co-expression networks. The weighted gene co-expression network analysis (WGCNA) was widely used to analyze large-scale data sets and to find modules of highly correlated genes. WGCNA was successfully used to explore the associations between gene sets and clinical features, and to identify candidate biomarkers (17). Thus, we described the correlation patterns among genes through a systematic biology

FIGURE 1 | Flow chart of data preparation, processing, analysis, and validation.
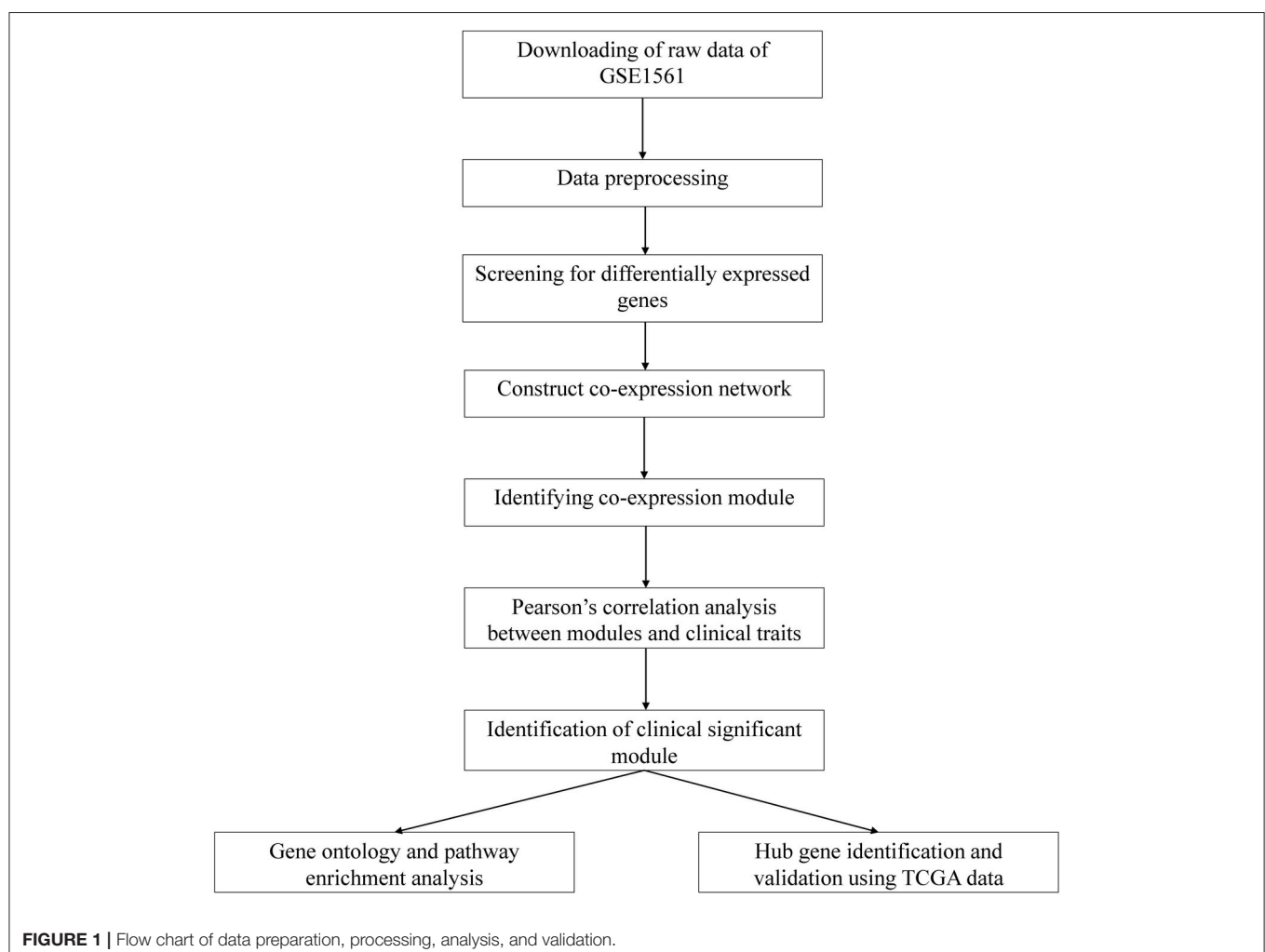
method based on WGCNA and identified novel biomarkers associated with breast cancer prognosis.

MATERIALS AND METHODS

Data Procession

A workflow of this study was indicated in **Figure 1**. The gene expression profiles of GSE1561 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1561) submitted by Richard Iggo et al. was downloaded from the Gene Expression Omnibus (GEO) database. The GSE1561 was an expression profiling based on GPL96 platform (Affymetrix Human Genome U133A Array) and contained 49 samples. Most patients had 2 trucut biopsies taken, and both biopsies were analyzed from 2 tumors to test the reproducibility of the technique. Repeat amplifications and

duplicate biopsies clustered together suggested that biological variation was greater than technical variation in this data set. The results of immunohistochemistry (IHC) also suggested the high quality of this data set ([18](#)). Robust Multi-array Average (RMA) algorithm in affy package within Bioconductor (http://www.bioconductor.org) in R was used to preprocess the gene expression profile data. After background correction, quantile normalization and probe summarization, the data set with 12,413 genes was further processed, and the top 50% most variant genes by analysis of variance (6,206 genes) were selected for WGCNA analysis.

Co-expression Network Construction

After validation, the expression data profile of these 6,206 genes were constructed to a gene co-expression network using
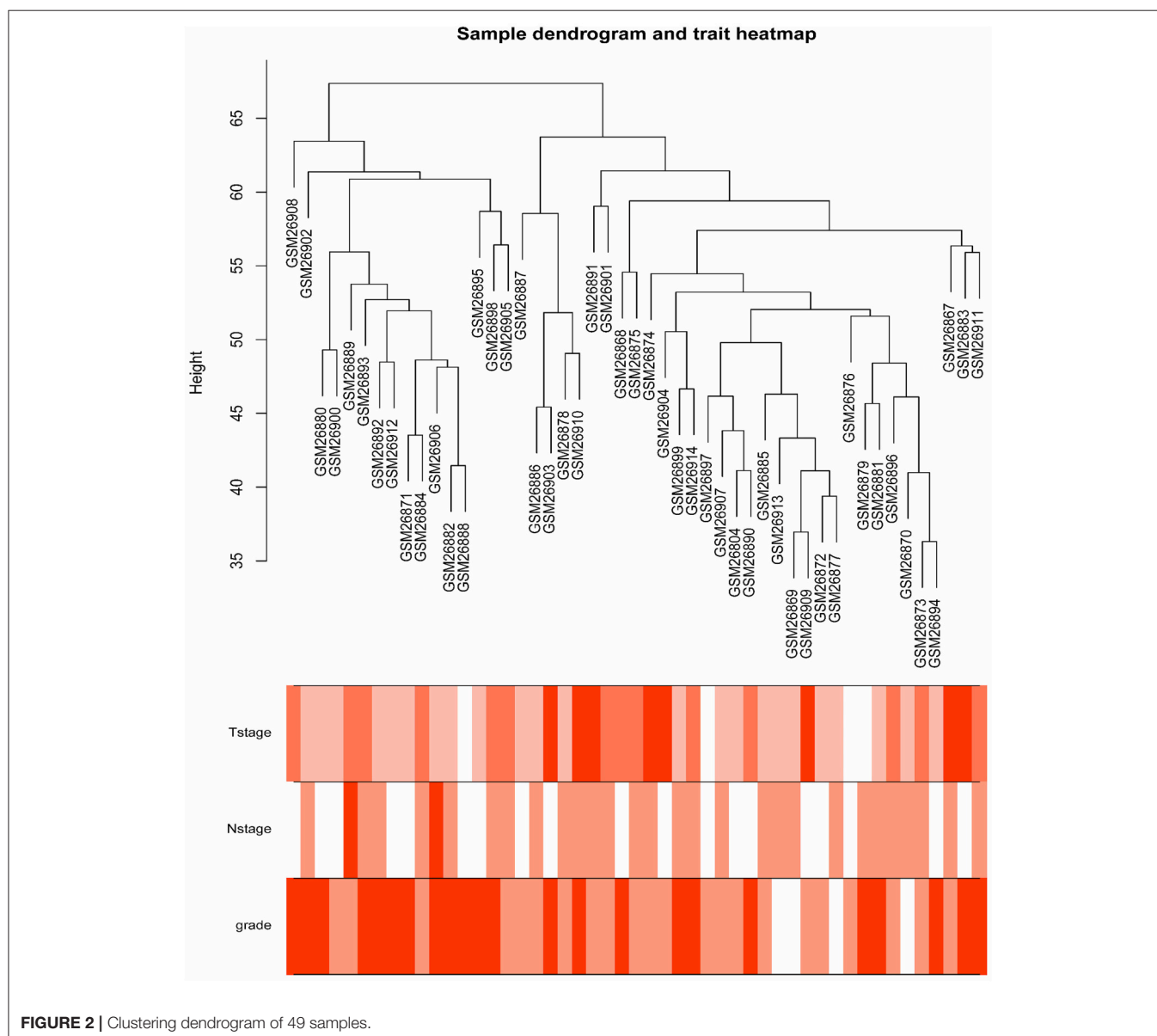
FIGURE 2 | Clustering dendrogram of 49 samples.

WGCNA package in R (**Supplementary Data Sheet 1**) (17). The analysis was performed as described previously (17).

The adjacency matrix a_{ij} which calculated the connection strength between each pair of nodes was calculated as follows:

$$s_{ij} = |cor(x_i, x_j)| a_{ij} = S_{ij}{}^\beta$$

Where X_i and X_j were vectors of expression value for gene i and j, s_{ij} represented the Pearson's correlation coefficient of gene i and gene j, a_{ij} encoded the network connection strength between gene i and gene j. In the presented study, the power of $\beta = 9$ (scale free $R^2 = 0.95$) was selected as the soft-thresholding parameter to ensure a scale-free network. In the co-expression network, genes with high absolute correlations were clustered into the same module. WGCNA method not only considers the association between the 2 connected genes, but also takes associated genes into account. Modules were also identified via hierarchical clustering of the weighting coefficient matrix. To further identify functional modules in the co-expression network with these 6,206 genes, the topological overlap measure (TOM) representing the overlap in shared neighbors, was calculated using the adjacency matrix.

$$TOM_{i,j} = \frac{\sum_{K=1}^{N} A_{i,k} \cdot A_{k,j} + A_{i,j}}{\min(K_i, K_j) + 1 - A_{i,j}}$$

Where A is the weighted adjacency matrix given by $A_{ij} = |cor(x_i, x_j)|^\beta$ and $\beta = 9$ is the soft thresholding power. According to the TOM-based dissimilarity measure with a minimum size (gene group) of 30 for the gene dendrogram, average linkage hierarchical clustering was conducted, and genes with similar expression profiles were classified into the same gene modules using the DynamicTreeCut algorithm.

Identification of Clinical Significant Modules

Two approaches were used to identify modules associated with clinical information of breast cancer. First, module eigengenes (MEs) were defined as the first principal component of each gene module and the expression of MEs was considered as a representative of all genes in a given module. The correlation between MEs and clinical trait was calculated to identify the clinical significant module. In addition, the gene significance (GS) was defined as mediated p-value of each gene (GS = lgP) in the linear regression between gene expression and the clinical traits. Then, the module significance (MS) were defined as the average GS of all the genes involved in the module. MS was measured to incorporate clinical information into the co-expression network. Module significance (MS) was defined as the average absolute gene significance measured for all genes in a given module.

Gene Ontology and Pathway Enrichment Analysis

DAVID (http://david.abcc.ncifcrf.gov/) is a database for annotation, visualization and integrated discovery. Gene Ontology (GO) and KEGG pathway analysis of differentially expressed mRNAs were carried out using DAVID (version 6.8) online tools: functional annotation. The ontology contains three categories: biological process (BP), molecular function (MF), and cellular component (CC). Enriched GO terms and KEGG pathways were identified according to the cut-off criterion of adjusted $P < 0.001$.

Hub Gene Identification and Validation

The connectivity of genes was measured by absolute value of the Pearson's correlation. Genes with high within-module

FIGURE 3 | Determination of soft-thresholding power in the WGCNA. **(A)** Analysis of the scale-free fit index for various soft-thresholding powers (β). **(B)** Analysis of the mean connectivity for various soft-thresholding powers. **(C)** Checking the scale free topology when β = 9.
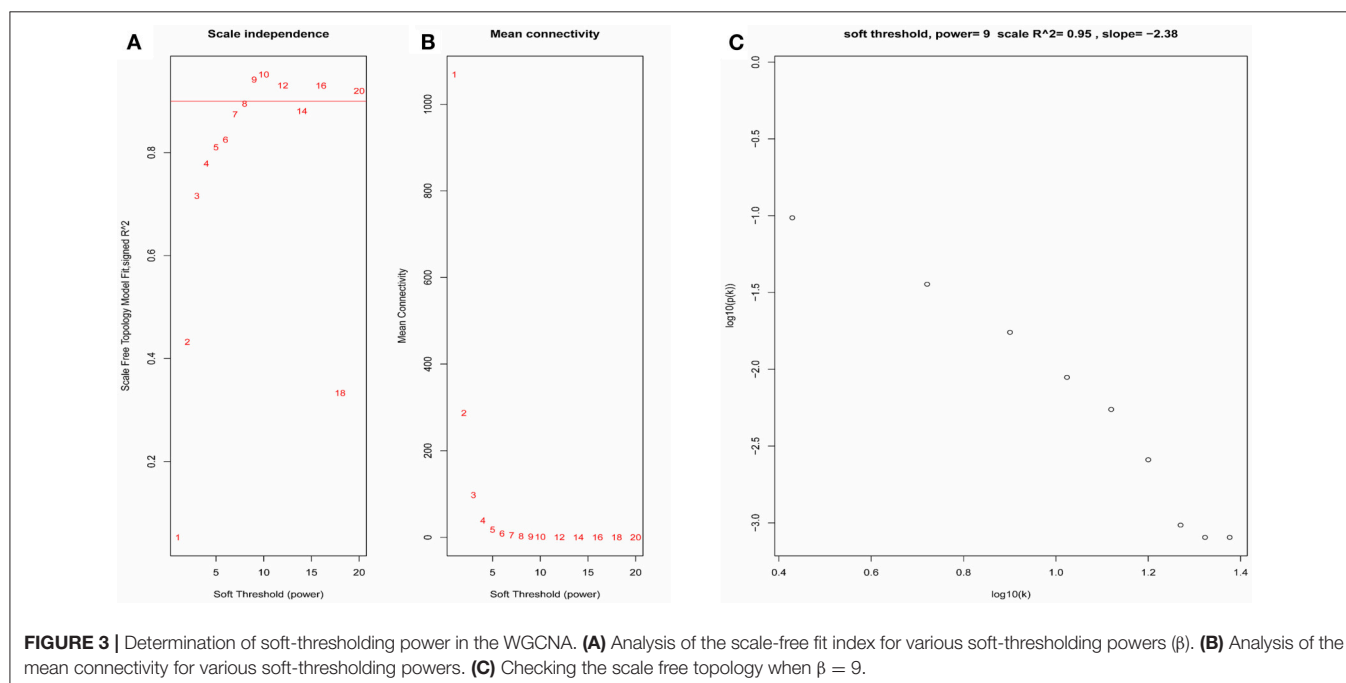
connectivity were considered as hub genes of the modules (cor.geneModuleMembership > 0.8). Hub genes inside a given module tended to have a strong correlation with certain clinical trait, which was measured by absolute value of the Pearson's correlation (cor.geneTraitSignificance > 0.2). To validate the hub genes, the clinical information and RNA sequencing data of breast cancer were obtained from the Cancer Genome Atlas

Project database (TCGA, https://cancergenome.nih.gov/). The mRNA sequencing data was normalized using edgeR package in R language. The Human Protein Atlas (http://www.proteinatlas. org) was also used to validate the immunohistochemistry of candidate hub genes. The direct link to these images in the human protein atlas are as follows: http://www.proteinatlas. org/ENSG00000112029-FBXO5/tissue/breast#img (FBXO5 in
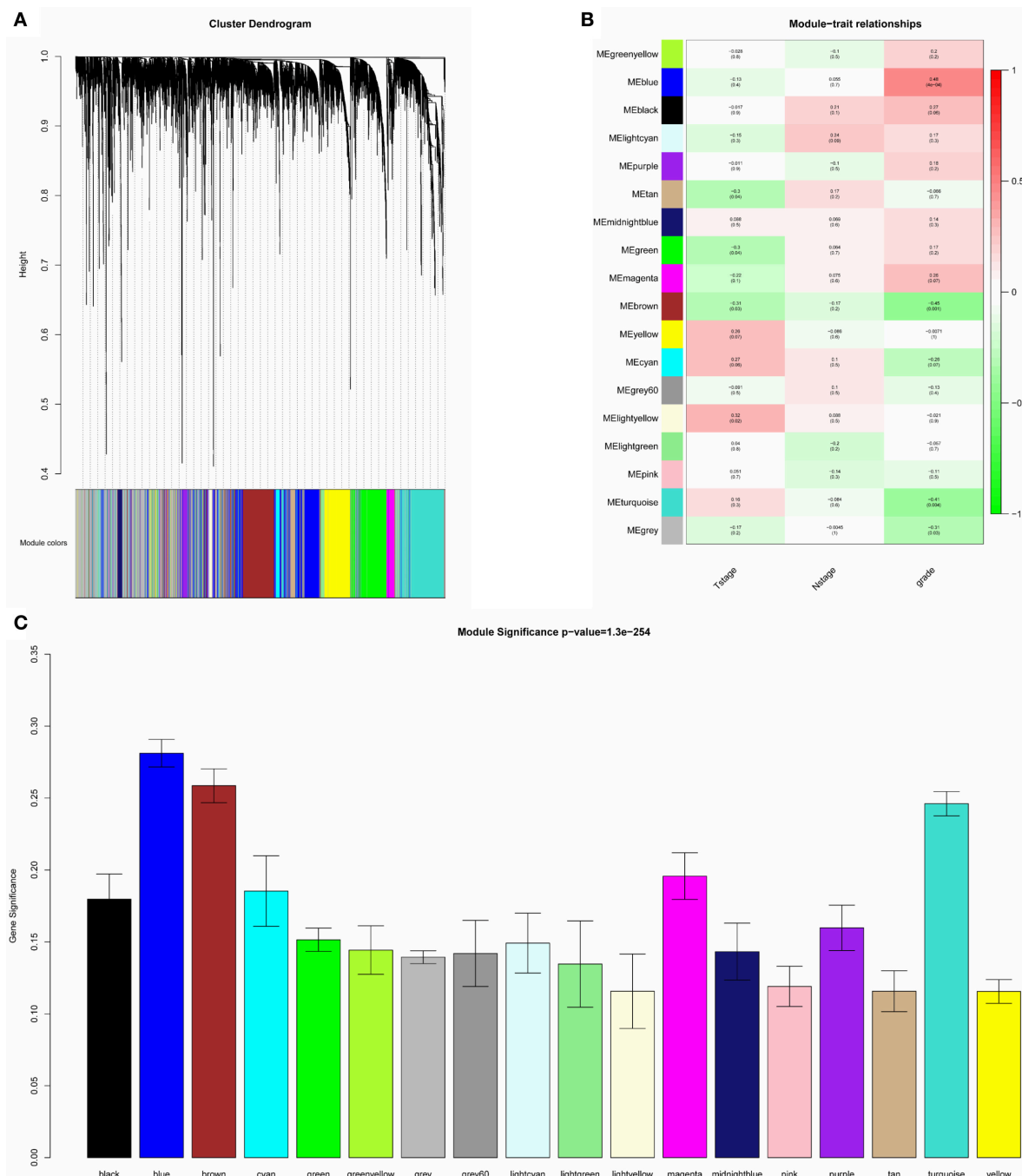
FIGURE 4 | Identification of modules associated with the clinical traits of breast cancer. **(A)** Dendrogram of all differentially expressed genes clustered based on a dissimilarity measure (1-TOM). **(B)** Heatmap of the correlation between module eigengenes and clinical traits of breast cancer. **(C)** Distribution of average gene significance and errors in the modules associated with tumor grades of breast cancer.

normal tissue); http://www.proteinatlas.org/ENSG00000112029-FBXO5/pathology/tissue/breast$+$cancer#img (FBXO5 in tumor tissue); http://www.proteinatlas.org/ENSG00000157456-CCNB2/tissue/breast#img (CCNB2 in normal tissue); http://www.proteinatlas.org/ENSG00000157456-CCNB2/pathology/tissue/breast$+$cancer#img (CCNB2 in tumor tissue); http://www.proteinatlas.org/ENSG00000090889-KIF4A/tissue/breast#img (CCNB2 in normal tissue); http://www.proteinatlas.org/ENSG00000090889-KIF4A/pathology/tissue/breast$+$cancer#img (CCNB2 in tumor tissue); http://www.proteinatlas.org/ENSG00000065328-MCM10/tissue/breast#img (MCM10 in normal tissue); http://www.proteinatlas.org/ENSG00000065328-MCM10/pathology/tissue/breast$+$cancer#img (MCM10 in tumor tissue); http://www.proteinatlas.org/ENSG00000088325-TPX2/tissue/breast#img (TPX2 in normal tissue); http://www.proteinatlas.org/ENSG00000088325-TPX2/pathology/tissue/breast$+$cancer#img (TPX2 in tumor tissue). Survival analysis of hub genes were performed using Kaplan Meier-plotter (www.kmplot.com) (19).

RESULTS

Weighted Co-expression Network Construction and Key Modules Identification

The samples of GSE1561 were clustered using average linkage method and Pearson's correlation method (**Figure 2**). The co-expression analysis was carried out to construct the co-expression network. In this study, the power of β = 9 (scale free $R^2 = 0.95$) was selected as the soft-thresholding parameter to ensure a scale-free network (**Figure 3**). A total of 18 modules were identified via the average linkage hierarchical clustering. Blue module was found to have the highest association with tumor grade (**Figure 4**), and this module was selected as the clinical significant module for further analysis.

Gene Ontology and Pathway Enrichment Analysis

The genes in the clinical significant module were categorized into 3 functional groups (BP, CC, and MF). Clinical significant module genes in the BP group were mainly enriched in cell division, DNA replication, sister chromatid cohesion, mitotic nuclear division, and DNA replication initiation; The genes in the MF group were mainly enriched in protein binding, poly(A) RNA binding, RNA binding, and ATP binding; the genes in the CC group were significantly enriched in nucleoplasm, nucleus, nucleolus, cytosol, and cytoplasm (**Figure 5**). According to Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, our results demonstrated that these genes were mainly involved in cell cycle, DNA replication, spliceosome, ribosome biogenesis in eukaryotes and RNA transport. These results indicated that the clinical significant module genes were mainly involved in mitotic cell cycle process.
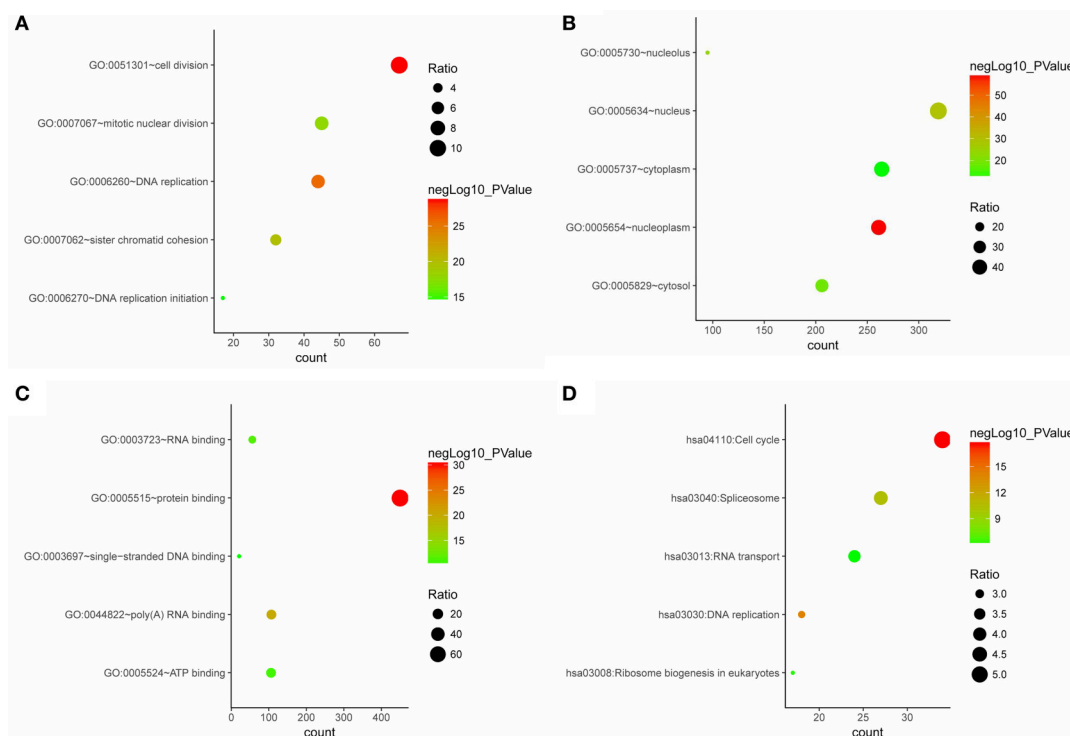
FIGURE 5 | Gene ontology and pathway enrichment analysis of blue module genes. **(A)** Biological process analysis. **(B)** Cellular component analysis. **(C)** Molecular function analysis. **(D)** KEGG pathway analysis.

Identification and Validation of Hub Genes

Based the cut-off criteria (|MM| > 0.8 and |GS| > 0.2), 42 genes with high connectivity in the clinical significant module were identified as hub genes. Among them, CCNB2, FBXO5, KIF4A, MCM10, and TPX2 were negatively associated with the overall survival and relapse free survival (**Figures 6, 7**). Moreover, based on the TCGA data, the expression levels of these 5 genes were significantly higher in tumor tissues, especially in the triple negative breast cancers. The

expression of CCNB2, KIF4A, and TPX2 were upregulated in the advanced tumor stages. ROC curve indicated that CCNB2, FBXO5, KIF4A, MCM10, and TPX2 exhibited excellent diagnostic efficiency for normal and tumor tissues (**Figures 8, 9**). In addition, the protein levels of these 5 genes were significantly higher in tumor tissues compared with normal tissues based on the Human Protein Atlas database (**Figure 10**). Since these 5 genes were all hub genes in the clinical significant module, they might have a tendency to co-express.
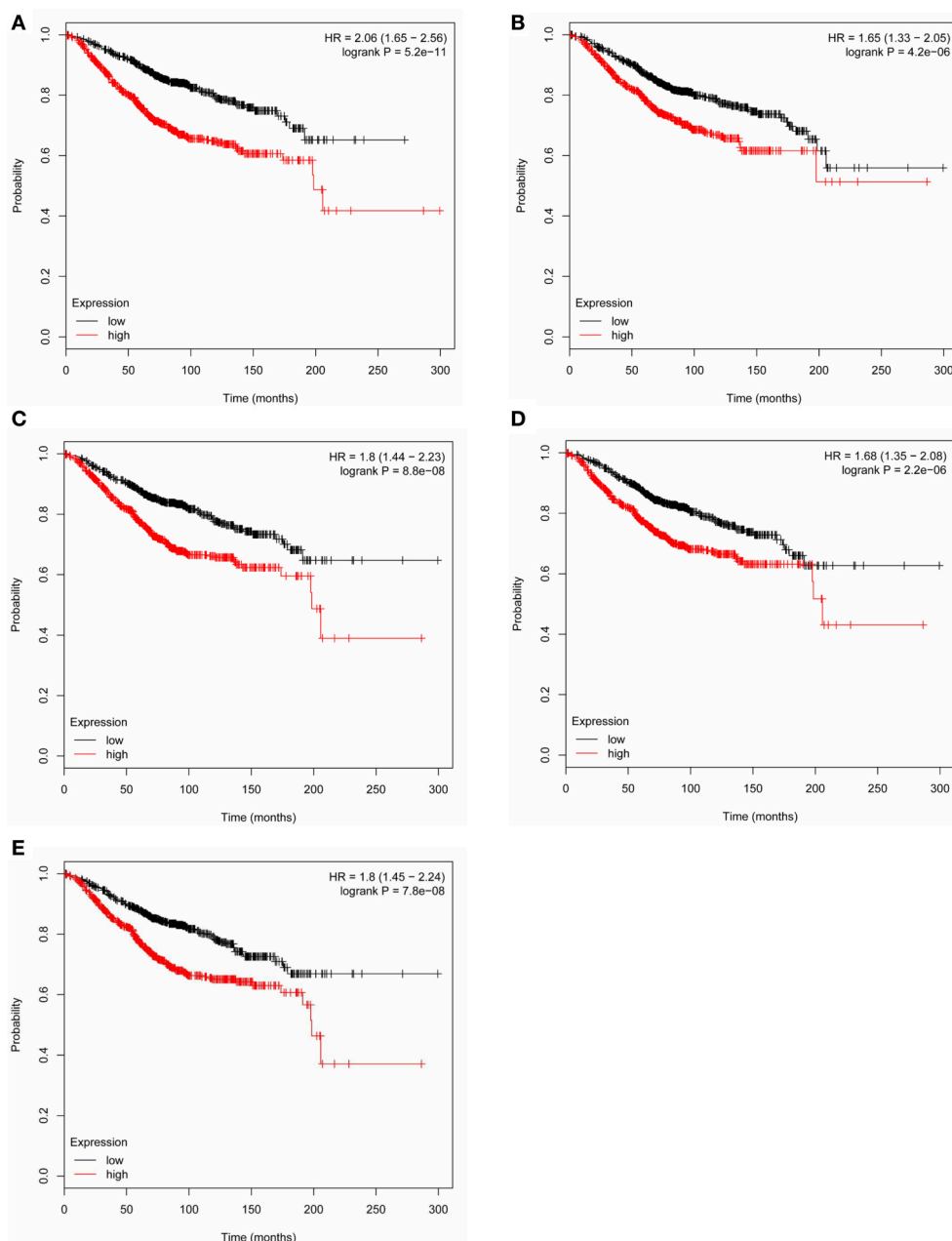
FIGURE 6 | Overall survival of the five hub genes in breast cancer based on Kaplan Meier-plotter. The patients were stratified into high-level group and low-level group according to median expression. **(A)** CCNB2. **(B)** FBXO5. **(C)** KIF4A. **(D)** MCM10. **(E)** TPX2.

Our results of correlation analysis demonstrated a strong correlation of mRNA expression levels between KIF4A and TPX2 (**Supplementary Data Sheet 2**).

DISCUSSION

Breast cancer seriously endangers female health, and it is easy to recur even after combined therapy. Although the treatment of breast cancer was improved during the last decades, the ability to treat the advanced ones is still limited due to the lack of precise molecular targets for breast cancer. Therefore, it is important to explore the molecule mechanisms involved in the occurrence and development of breast cancer. Better biomarkers for cancer specific prognosis and progression are highly demanded. In the presented study, we used gene expression datasets from GEO database to screen potential biomarkers related to the progression and prognosis of breast cancer. We also obtained the clinical
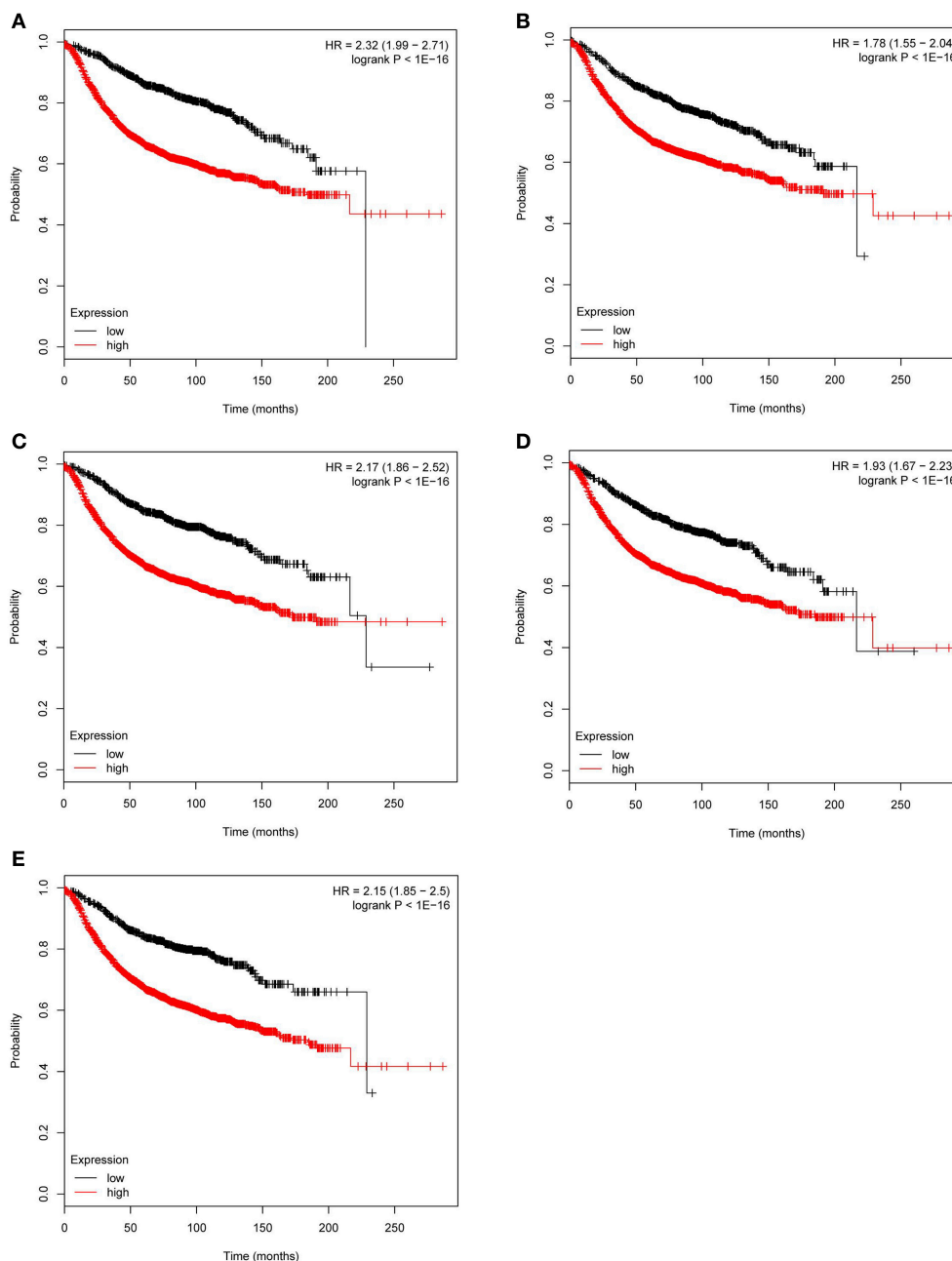
FIGURE 7 | Relapse free survival analysis of the five hub genes in breast cancer based on Kaplan Meier-plotter. The patients were stratified into high-level group and low-level group according to median expression **(A)** CCNB2. **(B)** FBXO5. **(C)** KIF4A. **(D)** MCM10. **(E)** TPX2.

information and RNA sequencing data of breast cancer from TCGA database for validation.

WGCNA was performed to explore gene co-expression modules associated with progression of breast cancer. A total of 6,206 most variant genes were used to construct co-expression network and 18 modules were identified. Blue module was found to have the highest association with tumor grades and 42 genes with high connectivity were screened out from the module. Among them, CCNB2, FBXO5, KIF4A, MCM10, and TPX2 were negatively associated with the overall survival (**Figure 6**).

CCNB2, also known as cyclin B2, is a member of cyclin family. CCNB2 was reported to regulate cell cycle by activating CDC2 kinase in eukaryotes, and inhibition of CCNB2 induced cell cycle arrest. CCNB2 was overexpressed in multiple tumors, including bladder cancer, uterine corpus endometrial carcinoma, prostate cancer, and gastric cancer (20–23). In addition, compared with normal controls, the levels of serum circulating CCNB2 are higher in digestive tract cancer and lung cancer patients, and they are found to be significantly associated with tumor stage and metastasis status (24). In invasive breast carcinoma, cytoplasmic CCNB2 protein levels

were significantly correlated with a poor disease specific survival. CCNB2 expression level was reported to be an independent prognostic factor for the disease specific survival of breast cancer (25). Our results indicated that CCNB2 was upregulated in breast cancer tissues compared to normal tissues, and that its expression was significantly associated with molecular subtypes of breast cancer and tumor stages (**Figure 8**). The underlying mechanisms of CCNB2 on tumor progression need to be further clarified.

F-Box Protein 5 (FBXO5) is a key cell cycle regulatory gene which regulates the progression to S phase and mitosis by inhibiting the anaphase promoting complex (APC). FBXO5 is overexpressed in various solid tumors. In the G0 and early G1 phases, the expression of FBXO5 is low, while in the S phase it is upregulated. In ovarian clear cell carcinoma, FBXO5 accumulation was related to mitotic errors with centrosome overduplication and abnormal spindle formation. These findings demonstrated that it might be involved in human cell cycle disorders and genomic stability to promote tumor growth (26–28). In breast carcinoma tissues, FBXO5 induced proliferation through the PI3K/Akt pathway. Overexpression of FBXO5 was

FIGURE 8 | Validation of CCNB2, FBXO5, KIF4A, MCM10, and TPX2. **(A)** The correlation of CCNB2 **(A)**, FBXO5 **(B)**, KIF4A **(C)**, MCM10 **(D)**, and TPX2 **(E)** expression with breast cancer molecular subtypes. **(F)** The correlation of CCNB2 expression with pathological stage. **(G)** The correlation of KIF4A expression with pathological stage. **(H)** The correlation of TPX2 expression with pathological stage.*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; ****$P < 0.0001$. One-way analysis of variance (ANOVA) was used to evaluate the statistical significance of differences.

reported to correlate with poor prognosis. In addition, PI3K inhibitor reduced FBXO5 expression (29).

The protein encoded by Kinesin family member 4A (KIF4A) was reported to be involved in the intracellular transport of membranous organelles and chromosome integrity during mitosis. In patients with colorectal cancer, KIF4A was upregulated, and downregulation of KIF4A reduced cell proliferation in colorectal cancer cells (30). In hepatocellular carcinoma (HCC) patients, KIF4A overexpression was associated with poorer overall and disease-free survival. In HCC cells,

higher levels of KIF4A dramatically increased cellular clonogenic abilities and proliferation, while KIF4A depletion caused a significant augmentation of apoptosis (31). In breast cancer, high KIF4A levels were associated with poor relapse-free survival of ER-positive patients. In tamoxifen-resistant and sensitive breast cancer cells, KIF4A knockdown significantly impeded cellular proliferation and induced apoptosis (32).

Mini-chromosome maintenance complex component 10 (MCM10) is one of the highly conserved mini-chromosome maintenance proteins. MCM10 is bound to chromatin through
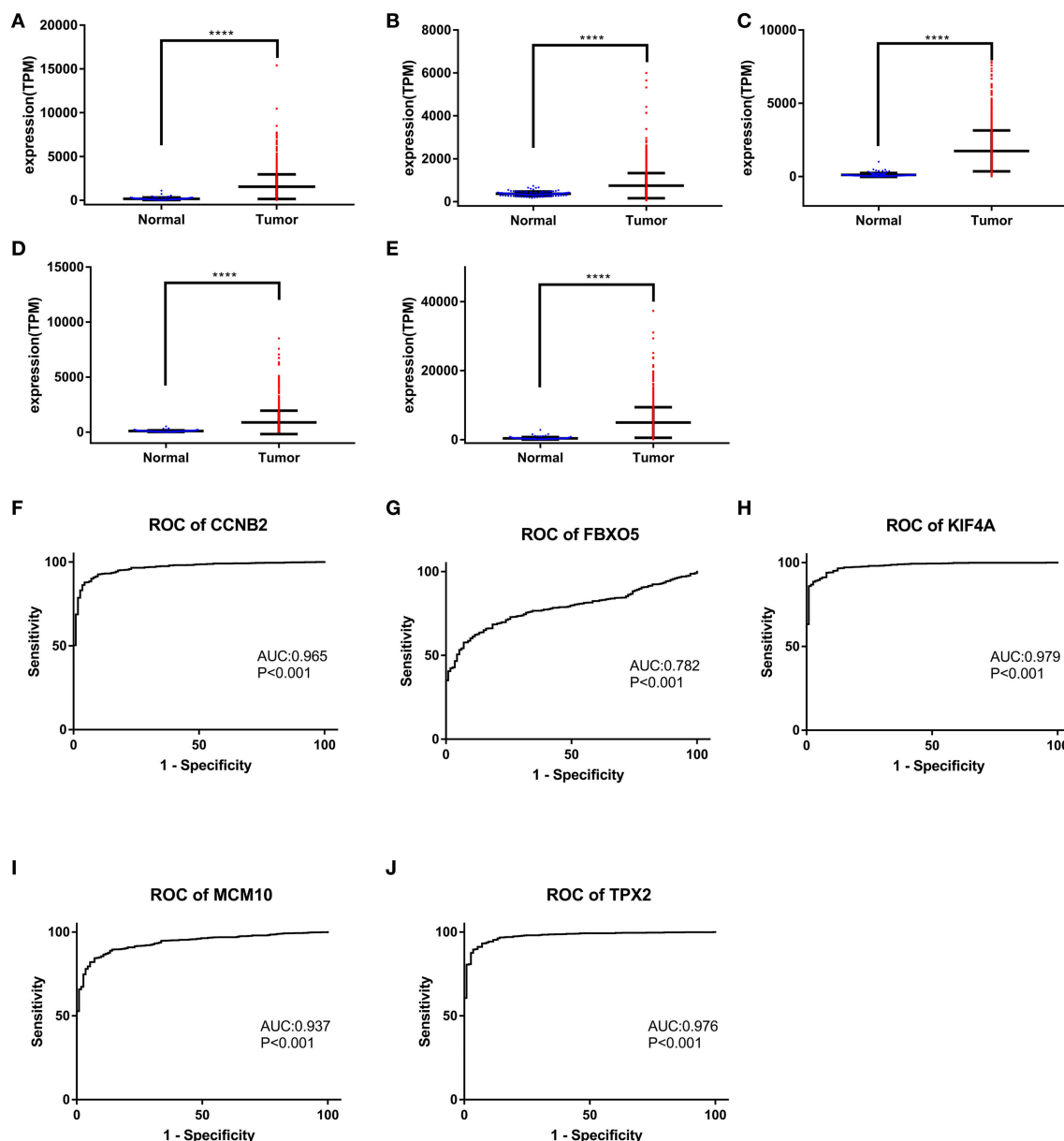
FIGURE 9 | Gene expression levels of CCNB2, FBXO5, KIF4A, MCM10, and TPX2 between normal breast and tumor samples. The mRNA levels of CCNB2 **(A)**, CCNB2 **(B)**, FBXO5 **(C)**, KIF4A **(D)**, and TPX2 **(E)**. ROC curve of CCNB2 **(F)**, FBXO5 **(G)**, KIF4A **(H)**, MCM10 **(I)**, and TPX2 **(J)**. **(A–E)** *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001. Two-tailed Student's t-tests was used to evaluate the statistical significance of differences.
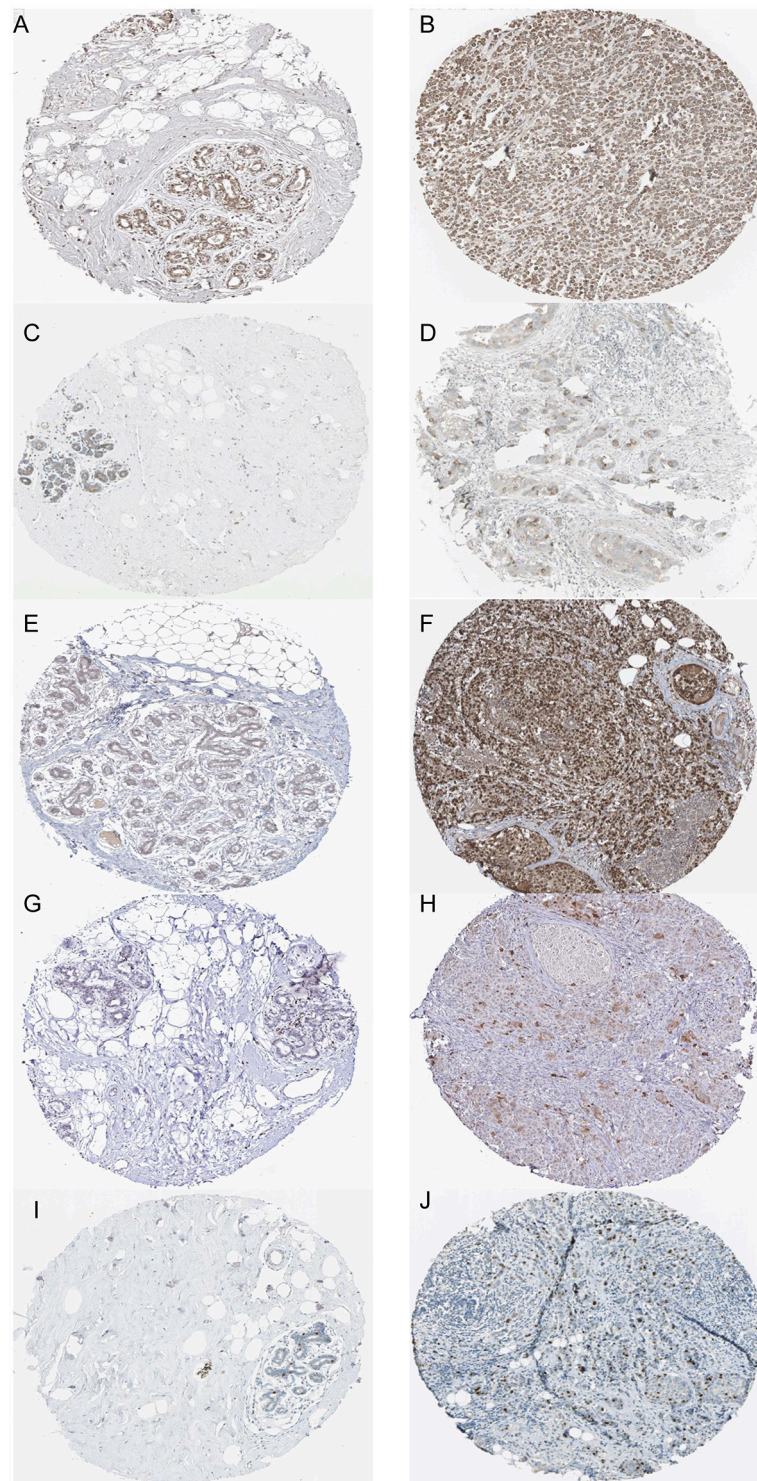
FIGURE 10 | Immunohistochemistry of the five hub genes based on the Human Protein Atlas. **(A)** Protein levels of FBXO5 in normal tissue (staining: medium; intensity: moderate; quantity: >75%). **(B)** Protein levels of FBXO5 in tumor tissue (staining: high; intensity: strong; quantity: >75%). **(C)** Protein levels of CCNB2 in normal tissue (staining: low; intensity: moderate; quantity: <25%). **(D)** Protein levels of CCNB2 in tumor tissue (staining: medium; intensity: strong; quantity: <25%). **(E)** Protein levels of KIF4A in normal tissue (staining: low; intensity: weak; quantity: 25–75%). **(F)** Protein levels of KIF4A in tumor tissue (staining: high; intensity: strong; quantity: >75%). **(G)** Proteins level of MCM10 in normal tissue (staining: not detected; intensity: weak; quantity: <25%). **(H)** Protein levels of MCM10 in tumor tissue (staining: low; intensity: moderate; quantity:<25%). **(I)** Protein levels of TPX2 in normal tissue (staining: medium; intensity: strong; quantity: <25%). **(J)** Protein levels of TPX2 in tumor tissue (staining: medium; intensity: strong; quantity: <25%).

the interaction with MCM2-7, and plays crucial roles both in initiation and elongation during eukaryotic genome replication (33). For urothelial carcinoma, high MCM10 levels were significantly correlated with advanced tumors stages, vascular invasion, and nodal status. MCM10 overexpression also predicted poor disease-specific survival and inferior metastasis-free survival (34). In our analysis of GSE1561, MCM10 was one of the hub genes in the blue module which was significantly associated with tumor grade (**Figure 3**). In the validation dataset of TCGA, our results indicated that MCM10 was significantly upregulated in breast tumor tissues, and even higher in the triple negative breast cancer (**Figures 8**, **9**).

Targeting protein for Xenopus kinesin-like protein 2 (TPX2) plays a critical role in chromosome segregation machinery during mitosis (35). It was reported to be overexpressed in multiple tumors: lung cancer, kidney renal clear cell carcinoma, hepatocellular Carcinoma, prostate cancer, and breast cancer (36). TPX2 activates PI3K/Akt pathway and upregulates matrix metalloproteinases (MMP) family members in colon cancer. Previous studies showed that TPX2 expression promoted proliferation, migration, and invasion of liver cancer and breast cancer cells via upregulating expressions of MMP2 and MMP9 (37, 38). In patients with HCC, overexpression TPX2 was correlated with worse prognosis. In addition, knockdown TPX2 in HCC cells strongly reduced cellular proliferation, induced apoptosis and inhibited EMT (39).

Co-expression analysis is a powerful technique for multigene analysis of large-scale data sets. In cancer research, co-expression analyses revealed the mRNA and microRNA expression network in multiple cancers. In the present study, we used WGCNA to construct a gene co-expression network, to measure the relationships between genes and modules, and to explore the relationships between modules and clinical traits. We also screened out a clinical significant module which was associated with the progression of breast cancer. KEGG pathway analysis demonstrated that this module was mostly involved in cell cycle. In addition, 5 hub genes, CCNB2, FBXO5, KIF4A, MCM10, and TPX2 were identified and validated to be associated with the progression and worse prognosis of breast cancer. Our results provided valuable indication for basic and clinical research on breast cancer. The underlying concept of gene co-expression analysis is guilt-by-association. The groups of genes known as co-expression modules were found to maintain a consistent expression relationship independent of phenotype, and might share a common biological role. Similar to the limitations of most other data mining methods, our results of WGCNA can be biased or invalid when dealing with technical artifacts or tissue contaminations (6). To increase the credibility of WGCNA results, TCGA RNA-seq data and IHC data from the Human Protein Atlas database were used for validation. While due to the limitation of the database, the related IHC of each sample can't be found, tumor and normal samples were from different patients.

AUTHOR CONTRIBUTIONS

JT, YG, and GW reviewed relevant literature and drafted the manuscript. DK, KW, DZ, and QC conducted all statistical analyses. All authors read and approved the final manuscript.

SUPPLEMENTARY MATERIAL

REFERENCES

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin.* (2011) 61:69–90. doi: 10.3322/caac.20107
2. Shi H, Zhang L, Qu Y, Hou L, Wang L, Zheng M. Prognostic genes of breast cancer revealed by gene co-expression network analysis. *Oncol Lett.* (2017) 14:4535–42. doi: 10.3892/ol.2017.6779
3. Redig AJ, McAllister SS. Breast cancer as a systemic disease: a view of metastasis. *J Intern Med.* (2013) 274:113–26. doi: 10.1111/joim.12084
4. McAllister SS, Gifford AM, Greiner AL, Kelleher SP, Saelzler MP, Ince TA, et al. Systemic endocrine instigation of indolent tumor growth requires osteopontin. *Cell* (2008) 133:994–1005. doi: 10.1016/j.cell.2008.04.045
5. Zhang Y, Lv F, Yang Y, Qian X, Lang R, Fan Y, et al. Clinicopathological features and prognosis of metaplastic breast carcinoma: experience of a major Chinese cancer center. *PLoS ONE* (2015) 10:e0131409. doi: 10.1371/journal.pone.0131409
6. Clarke C, Madden SF, Doolan P, Aherne ST, Joyce H, O'Driscoll L, et al. Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis* (2013) 34:2300–8. doi: 10.1093/carcin/bgt208
7. Krishnamurti U, Silverman JF. HER2 in breast cancer: a review and update. *Adv Anat Pathol.* (2014) 21:100–7. doi: 10.1097/PAP.0000000000000015
8. Trainer AH, Thompson E, James PA. BRCA and beyond: a genome-first approach to familial breast cancer risk assessment. *Discov Med.* (2011) 12:433–43.
9. Weitzel JN. The genetics of breast cancer: what the surgical oncologist needs to know. *Surgical Oncol Clin N Am.* (2015) 24:705–32. doi: 10.1016/j.soc.2015.06.011
10. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* (1987) 235:177–82.
11. Romond EH, Perez EA, Bryant J, Suman VJ, Geyer CE Jr, Davidson NE, et al. Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *N Engl J Med.* (2005) 353:1673–84. doi: 10.1056/NEJMoa052122
12. Geyer CE, Forster J, Lindquist D, Chan S, Romieu CG, Pienkowski T, et al. Lapatinib plus capecitabine for HER2-positive advanced breast cancer. *N Engl J Med.* (2006) 355:2733–43. doi: 10.1056/NEJMoa064320
13. M Braden A, V Stankowski R, M Engel J, A Onitilo A. Breast cancer biomarkers: risk assessment, diagnosis, prognosis, prediction of treatment efficacy and toxicity, and recurrence. *Curr Pharm Design* (2014) 20:4879–98. doi: 10.2174/1381612819666131125145517
14. Kontzoglou K, Palla V, Karaolanis G, Karaiskos I, Alexiou I, Pateras I, et al. Correlation between Ki67 and breast cancer prognosis. *Oncology* (2013) 84:219–25. doi: 10.1159/000346475
15. Zhao Q, Song W, He DY, Li Y. Identification of key gene modules and pathways of human breast cancer by co-expression analysis. *Breast Cancer* (2017) 25:213–23. doi: 10.1007/s12282-017-0817-5

16. Meng L, Xu Y, Xu C, Zhang W. Biomarker discovery to improve prediction of breast cancer survival: using gene expression profiling, meta-analysis, and tissue validation. *Oncotargets Ther.* (2016) 9:6177–85. doi: 10.2147/OTT.S113855

17. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* (2008) 9:559. doi: 10.1186/1471-2105-9-559

18. Farmer P, Bonnefoi H, Becette V, Tubiana-Hulin M, Fumoleau P, Larsimont D, et al. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* (2005) 24:4660–71. doi: 10.1038/sj.onc.1208561

19. Lanczky A, Nagy A, Bottai G, Munkácsy G, Szabó A, Santarpia L, et al. miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients. *Breast Cancer Res Treat.* (2016) 160:439–46. doi: 10.1007/s10549-016-4013-7

20. Shen L, Liu M, Liu W, Cui J, Li C. Bioinformatics analysis of RNA sequencing data reveals multiple key genes in uterine corpus endometrial carcinoma. *Oncol Lett.* (2018) 15:205–12. doi: 10.3892/ol.2017.7346

21. Han Y, Jin X, Zhou H, Liu B. Identification of key genes associated with bladder cancer using gene expression profiles. *Oncol Lett.* (2018) 15:297–303. doi: 10.3892/ol.2017.7310

22. Huang CG, Li FX, Pan S, Xu CB, Dai JQ, Zhao XH. Identification of genes associated with castrationresistant prostate cancer by gene expression profile analysis. *Mol Med Rep.* (2017) 16:6803–13. doi: 10.3892/mmr.2017.7488

23. Shi Q, Wang W, Jia Z, Chen P, Ma K, Zhou C. ISL1, a novel regulator of CCNB1, CCNB2 and c-MYC genes, promotes gastric cancer cell proliferation and tumor growth. *Oncotarget* (2016) 7:36489–500. doi: 10.18632/oncotarget.9269

24. Mo ML, Chen Z, Li J, Li HL, Sheng Q, Ma HY, et al. Use of serum circulating CCNB2 in cancer surveillance. *Int J Biol Markers* (2010) 25:236–42. doi: 10.5301/JBM.2010.6088

25. Shubbar E, Kovacs A, Hajizadeh S, Parris E, Nemes S, Gunnarsdóttir K, et al. Elevated cyclin B2 expression in invasive breast carcinoma is associated with unfavorable clinical outcome. *BMC Cancer* (2013) 13:1. doi: 10.1186/1471-2407-13-1

26. Guan C, Zhang J, Zhang J, Shi H, Ni R. Enhanced expression of early mitotic inhibitor-1 predicts a poor prognosis in esophageal squamous cell carcinoma patients. *Oncol Lett.* (2016) 12:114–20. doi: 10.3892/ol.2016.4611

27. Vaidyanathan S, Cato K, Tang L, Pavey S, Haass NK, Gabrielli BG, et al. *In vivo* overexpression of Emi1 promotes chromosome instability and tumorigenesis. *Oncogene* (2016) 35:5446–55. doi: 10.1038/onc.2016.94

28. Gutgemann I, Lehman NL, Jackson PK, Longacre TA. Emi1 protein accumulation implicates misregulation of the anaphase promoting complex/cyclosome pathway in ovarian clear cell carcinoma. *Mod Pathol.* (2008) 21:445–54. doi: 10.1038/modpathol.3801022

29. Liu X, Wang H, Ma J, Xu J, Sheng C, Yang S, et al. The expression and prognosis of Emi1 and Skp2 in breast carcinoma: associated with PI3K/Akt pathway and cell proliferation. *Med Oncol.* (2013) 30:735. doi: 10.1007/s12032-013-0735-0

30. Matsumoto Y, Saito M, Saito K, Saito, K, Kanke, Y, Watanabe, Y, et al. Enhanced expression of KIF4A in colorectal cancer is associated with

lymph node metastasis. *Oncol Lett.* (2018) 15:2188–94. doi: 10.3892/ol.2017.7555

31. Huang Y, Wang H, Lian Y, Wu X, Zhou L, Wang, J, et al. Upregulation of kinesin family member 4A enhanced cell proliferation via activation of Akt signaling and predicted a poor prognosis in hepatocellular carcinoma. *Cell Death Dis.* (2018) 9:141. doi: 10.1038/s41419-017-0114-4

32. Zou JX, Duan Z, Wang J, Sokolov A, Xu J, Chen CZ, et al. Kinesin family deregulation coordinated by bromodomain protein ANCCA and histone methyltransferase MLL for breast cancer cell growth, survival, and tamoxifen resistance. *Mol Cancer Res.* (2014) 12:539–49. doi: 10.1158/1541-7786.MCR-13-0459

33. Izumi M, Mizuno T, Yanagi KI, Sugimura K, Okumura K, Imamoto N, et al. The Mcm2-7-interacting domain of human mini-chromosome maintenance 10 (Mcm10) protein is important for stable chromatin association and origin firing. *J Biol Chem.* (2017) 292:13008–21. doi: 10.1074/jbc.M117.779371

34. Li WM, Huang CN, Ke HL, Li CC, Wei YC, Yeh HC, et al. MCM10 overexpression implicates adverse prognosis in urothelial carcinoma. *Oncotarget* (2016) 7:77777–92. doi: 10.18632/oncotarget.12795

35. Pan HW, Su HH, Hsu CW, Huang GJ, Wu TT. Targeted TPX2 increases chromosome missegregation and suppresses tumor cell growth in human prostate cancer. *Oncotargets Ther.* (2017) 10:3531–43. doi: 10.2147/OTT.S136491

36. Geiger TR, Ha NH, Faraji F, Michael HT, Rodriguez L, Walker RC, et al. Functional analysis of prognostic gene expression network genes in metastatic breast cancer models. *PLoS ONE* (2014) 9:e111813. doi: 10.1371/journal.pone.0111813

37. Liu Q, Tu K, Zhang H, Zheng X, Yao Y, Liu Q. TPX2 as a novel prognostic biomarker for hepatocellular carcinoma. *Hepatol Res.* (2015) 45:906–18. doi: 10.1111/hepr.12428

38. Yang Y, Li DP, Shen N, Yu XC, Li JB, Song Q, et al. TPX2 promotes migration and invasion of human breast cancer cells. *Asian Pac J Trop Med.* (2015) 8:1064–70. doi: 10.1016/j.apjtm.2015.11.007

39. Liang B, Jia C, Huang Y, He H, Li J, Liao, HE, et al. TPX2 level correlates with hepatocellular carcinoma cell proliferation, apoptosis, and EMT. *Digest Dis Sci.* (2015) 60:2360–72. doi: 10.1007/s10620-015-3730-9

Contribution of *MUTYH* Variants to Male Breast Cancer Risk: Results From a Multicenter Study in Italy

Piera Rizzolo [1], Valentina Silvestri [1], Agostino Bucalo [1], Veronica Zelli [1], Virginia Valentini [1], Irene Catucci [2], Ines Zanna [3], Giovanna Masala [3], Simonetta Bianchi [4], Alessandro Mauro Spinelli [5], Stefania Tommasi [6], Maria Grazia Tibiletti [7], Antonio Russo [8], Liliana Varesco [9], Anna Coppa [10], Daniele Calistri [11], Laura Cortesi [12], Alessandra Viel [13], Bernardo Bonanni [14], Jacopo Azzollini [15], Siranoush Manoukian [15], Marco Montagna [16], Paolo Radice [17], Domenico Palli [3], Paolo Peterlongo [2] and Laura Ottini [1]*

[1] Department of Molecular Medicine, Sapienza University of Rome, Rome, Italy, [2] Genome Diagnostics Program, IFOM - The FIRC Institute of Molecular Oncology, Milan, Italy, [3] Cancer Risk Factors and Lifestyle Epidemiology Unit, Institute for Cancer Research, Prevention and Clinical Network (ISPRO), Florence, Italy, [4] Division of Pathological Anatomy, Department of Surgery and Translational Medicine, University of Florence, Florence, Italy, [5] Institute for Maternal and Child Health IRCCS Burlo Garofolo, Trieste, Italy, [6] Molecular Genetics Laboratory, Istituto Tumori Giovanni Paolo II, Bari, Italy, [7] Dipartimento di Patologia, ASST Settelaghi and Centro di Ricerca per lo studio dei tumori eredo-familiari, Università dell'Insubria, Varese, Italy, [8] Section of Medical Oncology, Department of Surgical and Oncological Sciences, University of Palermo, Palermo, Italy, [9] IRCCS Ospedale Policlinico San Martino, Genoa, Italy, [10] Department of Experimental Medicine, Sapienza University of Rome, Rome, Italy, [11] Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori (IRST), Meldola, Italy, [12] Department of Oncology and Haematology, University of Modena and Reggio Emilia, Modena, Italy, [13] Unità di Oncogenetica e Oncogenomica Funzionale, Centro di Riferimento Oncologico di Aviano (CRO), IRCCS, Aviano, Italy, [14] Division of Cancer Prevention and Genetics, European Institute of Oncology (IEO), IRCCS, Milan, Italy, [15] Unità di Genetica Medica, Dipartimento di Oncologia Medica ed Ematologia, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan, Italy, [16] Immunology and Molecular Oncology Unit, Veneto Institute of Oncology IOV - IRCCS, Padua, Italy, [17] Unità di Ricerca Medicina Predittiva: Basi molecolari Rischio genetico e Test genetici, Dipartimento di Ricerca, Fondazione IRCCS Istituto Nazionale Tumori (INT), Milan, Italy

Inherited mutations in *BRCA1*, and, mainly, *BRCA2* genes are associated with increased risk of male breast cancer (MBC). Mutations in *PALB2* and *CHEK2* genes may also increase MBC risk. Overall, these genes are functionally linked to DNA repair pathways, highlighting the central role of genome maintenance in MBC genetic predisposition. *MUTYH* is a DNA repair gene whose biallelic germline variants cause MUTYH-associated polyposis (MAP) syndrome. Monoallelic *MUTYH* variants have been reported in families with both colorectal and breast cancer and there is some evidence on increased breast cancer risk in women with monoallelic variants. In this study, we aimed to investigate whether *MUTYH* germline variants may contribute to MBC susceptibility. To this aim, we screened the entire coding region of *MUTYH* in 503 *BRCA1/2* mutation negative MBC cases by multigene panel analysis. Moreover, we genotyped selected variants, including p.Tyr179Cys, p.Gly396Asp, p.Arg245His, p.Gly264Trpfs*7, and p.Gln338His, in a total of 560 MBC cases and 1,540 male controls. Biallelic *MUTYH* pathogenic variants (p.Tyr179Cys/p.Arg241Trp) were identified in one MBC patient with phenotypic manifestation of adenomatous polyposis. Monoallelic pathogenic variants were identified in 14 (2.5%) MBC patients, in particular, p.Tyr179Cys was detected in seven cases, p.Gly396Asp in five cases, p.Arg245His and p.Gly264Trpfs*7 in one case each. The majority of MBC cases with *MUTYH* pathogenic variants had family history of cancer including breast, colorectal, and gastric cancers. In the case-control study, an association between

the variant p.Tyr179Cys and increased MBC risk emerged by multivariate analysis [odds ratio (OR) = 4.54; 95% confidence interval (CI): 1.17–17.58; p = 0.028]. Overall, our study suggests that *MUTYH* pathogenic variants may have a role in MBC and, in particular, the p.Tyr179Cys variant may be a low/moderate penetrance risk allele for MBC. Moreover, our results suggest that MBC may be part of the tumor spectrum associated with MAP syndrome, with implication in the clinical management of patients and their relatives. Large-scale collaborative studies are needed to validate these findings.

INTRODUCTION

Male Breast Cancer (MBC) is a rare disease whose etiology appears to be associated with genetic factors. Inherited mutations in *BRCA1* and, mainly, *BRCA2*, predispose to MBC and account for up to 13% of all cases in the Italian population (1). Even though there is evidence supporting an association between increased MBC risk and pathogenic variants in *PALB2* and *CHEK2* (2–4), these two genes are unlikely to account for a substantial fraction of MBC cases. Thus, additional genes that may contribute to MBC genetic susceptibility need to be investigated.

BRCA1, *BRCA2*, *PALB2,* and *CHEK2* belong to or are functionally linked to the Homologous Recombination (HR) mechanism, one of the most important DNA Double-Strand Break (DSB) repair pathways, highlighting the central role of genome maintenance in MBC predisposition (5). Overall, the maintenance of genomic integrity is achieved by a coordinated interplay of different mechanisms of DNA repair, including Mismatch Repair (MMR), Nucleotide Excision Repair (NER) and Base Excision Repair (BER), in addition to DSB repair (6, 7). While dysregulation of DSB repair is known to play a relevant role in breast cancer (BC) pathogenesis, the involvement of other DNA repair pathways in BC is much less established.

MUTYH encodes a DNA glycosylase involved in BER, preventing 8-oxo-G:A mispairs generated by oxidative damage (8). Oxidative DNA damage, including 8-oxoG, may be due to hormonal metabolism and may contribute to BC susceptibility (9, 10). In this context, it is noteworthy that *BRCA1* and *BRCA2* are also involved in 8-oxoG repair (11), thus further supporting a possible role of BER and, more specifically, *MUTYH* in BC pathogenesis.

Biallelic (homozygous or compound heterozygous) *MUTYH* variants occur in 0.01–0.04% of European descent populations and cause MUTYH-associated polyposis syndrome (MAP), which predisposes patients to develop colorectal polyps and colorectal cancer (12–19). Monoallelic (heterozygous) *MUTYH*

variants occur in 1–2% of European descent populations and are associated with an increased risk of colorectal cancer (14, 16–21). Several studies on extracolonic cancers in carriers of *MUTYH* variants have been performed (21–26). The association of *MUTYH* variants with malignancies other than colon cancer is less robust, especially when establishing cancer risks in heterozygous *MUTYH* individuals. Increased risks of bladder and ovarian cancers have been reported for biallelic mutation carriers, while slightly increased risks of gastric, hepatobiliary, endometrial, and breast cancer have been observed in monoallelic mutation carriers (27).

Overall, the association between *MUTYH* mutations and BC risk remains controversial, some studies have shown an increased BC risk among *MUTYH* mutation carriers, while others have not (22–26, 28–30). An increased risk of BC associated with biallelic and monoallelic variants of *MUTYH* has been reported in *BRCA1/2* mutation negative individuals (21–23, 26). A higher frequency of monoallelic *MUTYH* mutations in families with both breast and colorectal cancer has been also reported compared to general population (21). Recently, an increased BC risk has been also reported for women with the common p.Gln338His variant (31).

To date the possible association between *MUTYH* variants and MBC risk has not been investigated. MBC is recognized as being primarily a hormone-dependent malignancy and is widely accepted as an estrogen-driven disease specifically related to hyperestrogenism (32) thus, oxidative DNA damage, due to hormonal metabolism, may particularly contribute to BC susceptibility in men. In this context, impairment of MUTYH activity due to inactivating/pathogenic variants may contribute to increase MBC risk.

To assess if *MUTYH* germline variants may contribute to MBC susceptibility, we screened a large series of *BRCA1/2* mutation negative MBC patients by sequencing the entire *MUTYH* coding region. Furthermore, to explore whether *MUTYH* variants were significantly associated with MBC risk, we performed a case-control study of selected *MUTYH* variants.

PATIENTS AND METHODS

Study Population

A total of 560 *BRCA1/2* mutation negative MBC cases and 1,540 male controls, enrolled in the frame of the ongoing Italian Multicenter Study on MBC (33), were included in the

---

present study. For each MBC case, information on the main clinical-pathologic characteristics were collected as previously described (33, 34). Controls were male individuals without personal history of cancer, enrolled under research or clinical protocols, or blood donors. All controls were recruited in the same geographical area of cases. For each study participant, samples of blood or DNA from peripheral blood leukocytes were collected. DNA from blood samples was extracted and quantified as previously described (35). The study was approved by Local Ethical Committee (Sapienza University of Rome, Prot. 669/17) and informed consent for using information and biological samples was obtained from all participants to the study.

MUTYH Gene Sequencing

A total of 503 MBC cases underwent next generation sequencing (NGS) of a custom panel of 50 cancer susceptibility genes including MUTYH. Briefly, paired-end libraries were prepared using the Nextera Rapid Capture Custom Enrichment kit (Illumina, San Diego, California, USA), pooled and loaded into a MiniSeq system (Illumina) for automated cluster generation, sequencing, and data analysis, including variant calling. Variant annotation and filtering was performed with Illumina Variant Studio Software version 2.2 against the human reference genome GRCh37. Variants were classified as pathogenic or likely pathogenic (collectively termed, pathogenic) according to the American College of Medical Genetics and Genomics (ACMG) recommendations (36). Briefly, variants were classified as pathogenic if they had a truncating, initiation codon or splice donor/acceptor effect or if pathogenicity was demonstrated by functional studies supportive of a damaging effect on the gene or gene product. All pathogenic variants were confirmed by double-stranded Sanger Sequencing (primer sequences are available upon request). Variants were named according to Human Genome Variation Society nomenclature (HGVS, hpp://www.hgvs.org).

Genotyping Analysis

Genotyping analysis of five MUTYH variants, rs34612342 (c.536A>G; p.Tyr179Cys), rs36053993 (c.1187G>A; p.Gly396Asp), rs140342925 (c.734G>A; p.Arg245His), rs587780751 (c.933+3A>C; p.Gly264Trpfs*7), and rs3219489 (c.1014G>C; p.Gln338His), identified by NGS and selected because previously proposed to be associated with increased risk of extracolonic cancer, including BC, was performed by allelic discrimination real-time PCR, in an ABI 7500 fast real-time PCR instrument (Life Technologies, Carlsbad, California, USA), using commercially available TaqMan SNP genotyping assays (Life Technologies) and according to the manufacturer's instructions. The specific assay IDs used are: C_32911941_10 (rs36512342), C_27860250_10 (rs36053993), C_166223223_10 (rs140342925), C_362043726_10 (rs587780751), and C_27504565_10 (rs3219489). In each experiment, positive (cases for which genotype was confirmed by Sanger Sequencing) and negative (water) controls were always included. A total of 560 MBC cases, including the 503 cases analyzed by NGS, and 1,540 male controls were genotyped.

Statistical Analysis

Chi-square test was performed in a case-case analysis in order to evaluate potential associations between pathogenic variants and specific clinical-pathologic characteristics.

The genotype frequency for each variant was evaluated in both series of cases and controls. The association between each variant and overall MBC risk was measured by the odds ratio (OR) and its corresponding 95% confidence interval (CI) by univariate logistic regression, and also by a multivariate analysis including adjustment for age, center and type of enrolment. A p-value <0.05 was considered statistically significant. All the analyses were performed using STATA version 13.1 statistical program.

RESULTS

Clinical-Pathologic Characteristics of MBC Cases

The study population consisted of 560 BRCA1/2 mutation negative MBC cases, enrolled in the frame of the ongoing Italian Multicenter Study on MBC. Overall, mean age at first BC diagnosis was 61.8 years (range 22–91 years); 91 cases (16.2%) reported first-degree family history of breast and/or ovarian cancer (BC/OC), 247 cases (44.1%) had first-degree family history of cancer and 101 cases (18%) had a personal history of cancer in addition to BC, mostly colorectal and prostate cancer. The majority of male breast tumors were invasive ductal carcinomas (85.9%), estrogen receptor positive (ER+, 94.2%), progesterone receptor positive (PR+ 88.4%), and HER2 negative (79.2%).

MUTYH Gene Sequencing in MBC Cases

The entire coding region of MUTYH was screened in 503 BRCA1/2 mutation negative MBC cases, by a custom multigene panel using NGS technologies. MUTYH variants detected are shown in **Table 1**. p.Tyr179Cys and p.Gly396Asp variants were the most frequently detected pathogenic variants and were identified in 1.6 and 1.0% of the MBC cases, respectively. The common variant p.Gln338His was identified in 41.7% of the MBC cases (**Table 1**).

Overall, pathogenic variants were identified in 15 (3.0%) MBC cases (**Table 2**), 14 cases were carriers of monoallelic (heterozygous) pathogenic variants and one case was carrier of the biallelic p.Tyr179Cys/p.Arg241Trp (compound heterozygous) pathogenic variants. The majority of MBC cases with MUTYH pathogenic variants had family history of cancer including breast, colorectal, and gastric cancers (**Table 2**). In particular, the biallelic MUTYH pathogenic variant carrier was a man diagnosed with BC at 51 years of age who developed colon cancer, with phenotypic manifestation of adenomatous polyposis, at early age (41 years) and had a first-degree relative affected by melanoma at young age (26 years). With the exception of this case, clinical features of the other MBC patients with MUTYH pathogenic variants did not suggest a MAP phenotype.

Overall, comparison of the clinical-pathologic characteristics between MUTYH pathogenic variant carriers and non-carriers did not show any statistically significant differences (**Table 3**).

TABLE 1 | *MUTYH* variants detected by NGS in 503 *BRCA1/2* mutation negative MBC cases [a].

[b]Location	Nucleotide change	Protein change	Variant type	dbSNP ID	Frequency *N* (%)
Exon 2	c.37 G>A	p.Ala13Thr	Missense	rs375349172	1 (0.2%)
Exon 2	c.64G>A	p.Val22Met	Missense	rs3219484	40 (8.0%)
Exon 7	**c.536A>G**	**p.Tyr179Cys**	**Missense**	**rs34612342**	**8 (1.6%)**
Exon 9	c.694A>T	p.Thr232Ser	Missense	rs587782351	1 (0.2%)
Exon 9	**c.721C>T**	**p.Arg241Trp**	**Missense**	**rs34126013**	**1 (0.2%)**
Exon 9	**c.734G>A**	**p.Arg245His**	**Missense**	**rs140342925**	**1 (0.2%)**
Exon 10	c.919C>T	p.Arg307Trp	Missense	rs759822330	1 (0.2%)
IVS 10	**c.933+3A>C** [c]	**p.Gly264Trpfs*7**	**Frameshift**	**rs587780751**	**1 (0.2%)**
Exon 12	c.1014G>C	p.Gln338His	Missense	rs3219489	210 (41.7%)
Exon 12	c.1037C>G	p.Ser346Trp	Missense	rs587778538	1 (0.2%)
Exon 13	**c.1187G>A**	**p.Gly396Asp**	**Missense**	**rs36053993**	**5 (1.0%)**
Exon 13	c.1258C>A	p.Leu420Met	Missense	rs144079536	4 (0.8%)
Exon 13	c.1276C>T	p.Arg426Cys	Missense	rs150792276	2 (0.4%)
Exon 16	c.1544 C>T	p.Ser515Phe	Missense	rs140118273	6 (1.2%)

[a]NGS, Next Generation sequencing; MBC, Male Breast Cancer.
[b]Pathogenic variants are shown in bold text.
[c]This variant affects a splicing site and causes the skipping of exon 10 that leads to a premature stop codon.

TABLE 2 | Personal and family history of cancer in MBC cases with germline *MUTYH* pathogenic variants[a].

ID	Variant[b]	Personal history of cancer (age)	First-degree family history of cancer (age)
#40	c.1187G>A (p.Gly396Asp)	Breast (55)	Prostate (85)
#61	c.1187G>A (p.Gly396Asp)	Breast (70)	Colorectal (56)
#138	c.1187G>A (p.Gly396Asp)	Breast (66)	Breast (50)
#153	c.1187G>A (p.Gly396Asp)	Breast (51)	–
#321	c.1187G>A (p.Gly396Asp)	Breast (75)	–
#146	c.536A>G (p.Tyr179Cys)	Breast (80)	Breast (45); Gastric (54)
#236	c.536A>G (p.Tyr179Cys)	Breast (45)	Breast (58); Colorectal (58)
#317	c.536A>G (p.Tyr179Cys)	Breast (67); Prostate (68)	–
#341	c.536A>G (p.Tyr179Cys)	Breast (72)	Breast (70); Esophageus (76)
#352	c.536A>G (p.Tyr179Cys)	Breast (63)	–
#376	c.536A>G (p.Tyr179Cys)	Breast (61)	Gastric (43); Liver (67)
#478	c.536A>G (p.Tyr179Cys)	Breast (82)	Colon (50)
#257	c.933+3A>C p.Gly264Trpfs*7)	Breast (72)	3 Breast (72,76, na)[a]
#358	c.734G>A (p.Arg245His)	Breast (79)	Breast (65); Gastric (69)
#227	c.536A>G (p.Tyr179Cys); c.721C>T (p.Arg241Trp)	Breast (51); Colon (41)	Melanoma (26)

[a]MBC, Male Breast Cancer; na, not available.
[b]Variants nomenclature in according to RefSeq NM_001128425.1, NP_001121897.1.

Genotyping Analysis of Selected *MUTYH* Variants in MBC Cases and Controls

MUTYH pathogenic variants, including p.Tyr179Cys (rs34612342), p.Gly396Asp (rs36053993), p.Arg245His (rs140342925), p.Gly264Trpfs*7 (rs587780751), and the common variant p.Gln338His (rs3219489), were genotyped in 560 cases and 1,540 male controls. Overall, pathogenic variants were detected at significantly higher frequency ($p = 0.04$) in MBC cases (15/560 2.7%) than in controls (21/1540, 1.3%).

The distribution of genotype frequencies and the estimates for the association between each genotyped variant and overall MBC risk are summarized in **Table 4**. Significant differences in the distribution of genotypes between MBC cases and controls emerged for p.Tyr179Cys (rs34612342) variant. The analysis of the genotype-specific risks showed that men with heterozygous genotype for *MUTYH* p.Tyr179Cys variant were at increased BC risk both in the univariate (OR = 5.56; 95%CI:1.67–18.55; $p = 0.005$) and in the multivariate analysis (OR = 4.54; 95%CI:1.17–17.58; $p = 0.028$). No statistically significant differences in genotype distribution between case and controls emerged for the other variants analyzed.

TABLE 3 | Clinical-pathologic characteristics of *MUTYH* pathogenic variant carriers and non-carriers.

Characteristics[a]	*MUTYH* variant carriers (*N* = 15)		Non-carriers (*N* = 488)		*p*-value
	N	%	*N*	%	
Mean age at diagnosis ± SD (range)	65.9 ± 11.4 (45–82)		61.7 ± 12 (22–91)		0.2
FIRST-DEGREE FAMILY HISTORY OF BC/OC[b]					
Negative	13	86.6	411	84.4	
Positive	2	13.4	76	15.6	1
FIRST-DEGREE FAMILY HISTORY OF CANCER					
Negative	6	40.0	274	56.3	
Positive	9	60.0	213	43.7	0. 3
PERSONAL HISTORY OF CANCER IN ADDITION TO BC					
Negative	12	80.0	396	81.1	
Positive	3	20.0	92	8.9	1
TUMOR HISTOTYPE					
Invasive ductal carcinoma	13	92.9	342	83	
In situ ductal carcinoma	1	7.1	36	8.7	
Other	–	–	34	8.3	0.9
TNM STAGE					
0	1	7.1	33	9.8	
1	6	42.9	152	45.4	
2	6	42.9	95	28.4	
3–4	1	7.1	55	16.4	0.7
HISTOLOGIC GRADE					
1	1	7.7	44	13.3	
2	9	69.2	198	59.8	
3	3	23.1	89	26.9	0.9
LYMPH NODE STATUS					
Negative	8	61.5	213	63.2	
Positive	5	38.5	124	36.8	1
ER[b] STATUS					
Negative	1	8.3	23	6.1	
Positive	11	91.7	353	93.9	0.5
PR[b] STATUS					
Negative	1	9.1	43	11.5	
Positive	10	90.9	331	88.5	1
HER2[b] STATUS					
Negative	11	91.7	236	80.0	
Positive	1	8.3	59	20.0	0.5
Ki67/MIB1 STATUS					
Low	2	28.6	172	58.9	
High	5	71.4	120	41.1	0.1

[a]*Some data for each pathologic characteristic are not available.*
[b]*BC, breast cancer; OC, ovarian cancer; ER, Estrogen receptor; PR, Progesterone receptor; HER2, human epidermal growth factor receptor 2.*

DISCUSSION

In this study, we aimed to evaluate the contribution of *MUTYH* variants in MBC susceptibility. To this purpose, we obtained NGS data of the entire coding region of *MUTYH* from a large series of *BRCA1/2* mutation negative MBC cases, from the ongoing Italian Multicenter Study on MBC, and further genotyped selected variants in a case-control study. To date, there is contrasting evidence on the impact of *MUTYH* pathogenic variants on risk

of BC in women and, to the best of our knowledge, no study has been performed in MBC.

By NGS, we identified 15 MBC patients (3.0%) with germline *MUTYH* pathogenic variants, including one biallelic and 14 monoallelic variant carriers. The MBC patient with biallelic *MUTYH* pathogenic variants was affected by colorectal cancer at early age with phenotypic manifestation of adenomatous polyposis. Thus, our results allowed a molecular diagnosis of MAP. To the best of our knowledge, to date, only another

TABLE 4 | Distribution of 560 *BRCA1/2* negative MBC cases and 1,540 controls according to genotype frequencies and MBC risk estimates for selected *MUTYH* variants[a].

Variant	Genotype	Cases N (%)	Controls N (%)	Univariate analysis		Multivariate analysis[b]	
				OR (95% CI)	p-value[c]	OR (95% CI)	p-value[c]
p.Tyr179Cys (rs34612342)	AA	552 (98.57)	1536 (99.74)				
	AG	8 (1.43)	4 (0.26)	**5.56 (1.67–18.55)**	**0.005**	**4.54 (1.17–17.58)**	**0.028**
p.Arg245His (rs140342925)	GG	559 (99.8)	1540 (100)				
	GA	1 (0.2)	–	–	–	–	–
p.Gly264Trpfs*7 (rs587780751)	AA	559 (99.8)	1539 (99.94)				
	AC	1 (0.2)	1 (0.06)	2.75 (0.17–44)	0.455	0.94(0.04–19.95)	0.97
p.Gln338His (rs3219489)	GG	327 (58.4)	931(60)				
	GC	203 (36.2)	526 (34.2)	1 (0.89–1.34)	0.36	1.2 (0.95–1.5)	0.12
	CC	30 (5.4)	83 (5.4)	2.1 (0.66–1.59)	0.89	1.2 (0.76–2)	0.37
p.Gly396Asp (rs36053993)	GG	555 (99.1)	1524 (99)				
	GA	5 (0.9)	16 (1)	0.86 (0.31–2.35)	0.77	0.58 (0.16–2.14)	0.42

[a]MBC, Male breast Cancer; OR, Odds Ratio; 95% CI, 95% confidence interval.
[b]ORs and 95% CI for specific genotypes were calculated using logistic regression models adjusted for age, center and type of enrolment.
[c]p-values <0.05 in bold text.

MBC case has been reported with MAP syndrome (23). Taking into account the rarity of both MBC and MAP, the occurrence of MBC in MAP patients may underline a possible common genetic pathway and suggest that MBC could be considered a MAP-related malignancy.

Overall, *MUTYH* monoallelic pathogenic variants, including p.Tyr179Cys, p.Gly396Asp, p.Arg245His, and p.Gly264Trpfs*7, were found with a frequency of 2.8% in our MBC series. p.Tyr179Cys and p.Gly396Asp were the most frequently variants detected and were identified in 2.4% of the cases. Published data showed that these two variants are the most frequent pathogenic variants in populations of European origin and account for 50 to 90% of *MUTYH* pathogenic variants identified in MAP patients (13, 14, 37, 38). The p.Arg245His variant was identified in a MBC patient with family history of breast and gastric cancers. This variant has been reported strongly associated with familial colorectal cancer (23, 39), and has also been identified in patients with suspected Lynch Syndrome and in a patient with gastric cancer (23, 40). The p.Gly264Trpfs*7 variant was identified in a MBC patient, from North-East of Italy, where it occurs as a founder mutation accounting for about 15.0% of the *MUTYH* pathogenic variants identified in MAP patients (41). By contrast, this variant has been reported with lower frequency, ranging from 1.0 to 8.0%, in MAP patients from other populations of Caucasian ethnicity (23, 41–47).

To investigate whether MBC arising in *MUTYH* pathogenic variant carriers may be characterized by specific features, we compared clinical-pathologic characteristics between carriers and non-carriers. No statistically significant association emerged for any of the clinical features tested. However, the great majority of MBC patients with *MUTYH* pathogenic variants had family history of cancer, including, breast, colorectal, and gastric cancers. These findings, if confirmed by additional data, may be useful in decisions concerning clinical management of patients and their families.

To further investigate the role of *MUTYH* in MBC, we evaluated the risk of MBC associated with selected *MUTYH* variants previously proposed to be associated with increased cancer risk, including BC risk (21, 27, 31), by performing a case-control study. Among the pathogenic variants examined, the p.Tyr179Cys variant was associated with an increased MBC risk (OR = 4.54, 95%CI = 1.17–17.58). A higher frequency of p.Tyr179Cys has been reported in families with both breast and colorectal cancer compared to the general population (21), but an association between p.Tyr179Cys variant and increased BC risk has not been observed (25, 26, 28, 30). Our results, suggest that p.Tyr179Cys variant may be a low/moderate penetrance risk allele for BC in men. This variant, located at 8-oxo-G binding site, causes major structural protein changes and a reduction in functionality (48, 49). Thus, oxidative DNA damage due to hormonal metabolism, like estrogen-induced 8-oxo-dG generation, may particularly contribute to MBC susceptibility, as BC in men is primarily a hormone-dependent tumor, specifically related to hyperestrogenism. Furthermore, it can be hypothesized that MBC, unencumbered by the many confounding factors that exist in female BC (i.e., reproductive factors and high frequency) might facilitate the identification of genetic factors and molecular mechanisms that may influence BC risk in general (50).

We also assessed whether the common p.Gln338His variant, reported to increase BC risk in women (31), was associated with MBC risk. We did not observe any significant differences in p.Gln338His genotypes distribution between MBC cases and controls inconsistent with a possible role of this variant in MBC risk. The other common variant, p.Val22Met, has not been reported to be associated with cancer risk (51–53) and was not examined in this study.

Overall, we observed that the majority of MBC patients with pathogenic *MUTYH* variants have first-degree family history of cancers. This raises the question of whether *MUTYH* variants, especially the Tyr179Cys variant, may be associated with MBC

risk only, or with the risk of familial or multi-syndromic diseases, including MBC. Further clinical/phenotype assessments and detailed statistical analyses would be useful in future studies to answer this question.

In conclusion, our study suggests that *MUTYH* pathogenic variants may have a role in MBC, in particular, p.Tyr179Cys variant may be a low/moderate penetrance risk allele for MBC. Our findings also suggest that MBC may be part of the tumor spectrum associated with MAP syndrome, with implications in the clinical management of the patients and their relatives.

Although we have a large series of MBC cases, this study may be underpowered to detect smaller risk effects and large-scale collaborative studies are needed to investigate any possible association with rarer variants and to have a more comprehensive examination and characterization of the link between *MUTYH* variants and MBC risk.

DATA AVAILABILITY STATEMENT

Datasets are available on request. The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHORS CONTRIBUTIONS

REFERENCES

1. Rizzolo P, Silvestri V, Tommasi S, Pinto R, Danza K, Falchetti M, et al. Male breast cancer: genetics, epigenetics, and ethical aspects. *Ann Oncol.* (2013) 24(Suppl 8):viii75–82. doi: 10.1093/annonc/mdt316

2. Falchetti M, Lupi R, Rizzolo P, Ceccarelli K, Zanna I, Calò V, et al. BRCA1/BRCA2 rearrangements and CHEK2 common mutations are infrequent in Italian male breast cancer cases. *Breast Cancer Res Treat.* (2008) 110:161–7. doi: 10.1007/s10549-007-9689-2

3. Silvestri V, Zelli V, Valentini V, Rizzolo P, Navazio AS, Coppa A, et al. Whole-exome sequencing and targeted gene sequencing provide insights into the role of PALB2 as a male breast cancer susceptibility gene. *Cancer* (2017) 123:210–8. doi: 10.1002/cncr.30337

4. Giordano SH. Breast cancer in men. *N Engl J Med.* (2018) 378:2311–20. doi: 10.1056/NEJMra1707939

5. Nielsen FC, van Overeem Hansen T, Sørensen CS. Hereditary breast and ovarian cancer: new genes in confined pathways. *Nat Rev Cancer* (2016) 16:599–612. doi: 10.1038/nrc.2016.72

6. Tham KC, Kanaar R, Lebbink JH. Mismatch repair and homeologous recombination. *DNA Repair* (2016) 38:75–83. doi: 10.1016/j.dnarep.2015.11.010

7. Tubbs A, Nussenzweig A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell* (2017) 168:644–56. doi: 10.1016/j.cell.2017.01.002

8. Mazzei F, Viel A, Bignami M. Role of MUTYH in human cancer. *Mutat Res.* (2013) 743–744:33–43. doi: 10.1016/j.mrfmmm.2013.03.003

9. Li D, Zhang W, Zhu J, Chang P, Sahin A, Singletary E, et al. Oxidative DNA damage and 8-hydroxy-2-deoxyguanosine DNA glycosylase/apurinic lyase in human breast cancer. *Mol Carcinog.* (2001) 31:214–23. doi: 10.1002/mc.1056

10. Dziaman T, Huzarski T, Gackowski D, Rozalski R, Siomek A, Szpila A, et al. Elevated level of 8-oxo-7,8-dihydro-2′-deoxyguanosine in leukocytes of BRCA1 mutation carriers compared to healthy controls. *Int J Cancer* (2009) 125:2209–13. doi: 10.1002/ijc.24600

11. Le Page F, Randrianarison V, Marot D, Cabannes J, Perricaudet M, Feunteun J, et al. BRCA1 and BRCA2 are necessary for the transcription-coupled repair of the oxidative 8-oxoguanine lesion in human cells. *Cancer Res.* (2000) 60:5548–52.

12. Al-Tassan N, Eisen T, Maynard J, Bridle H, Shah B, Fleischmann C, et al. Inherited variants of MYH associated with somatic G:C–>T:A mutations in colorectal tumors. *Nat Genet.* (2002) 30:227. doi: 10.1038/ng828

13. Sieber OM, Lipton L, Crabtree M, Heinimann K, Fidalgo P, Phillips RK, et al. Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. *N Engl J Med.* (2003) 348:791–9. doi: 10.1056/NEJMoa025283

14. Cleary SP, Cotterchio M, Jenkins MA, Kim H, Bristow R, Green R, et al. Germline MutY human homologue mutations and colorectal cancer: a multisite case–control study. *Gastroenterology* (2009) 136:1251–60. doi: 10.1053/j.gastro.2008.12.050

15. Lubbe SJ, Di Bernardo MC, Chandler IP, Houlston RS. Clinical implications of the colorectal cancer risk associated with MUTYH mutation. *J Clin Oncol.* (2009) 27:3975–80. doi: 10.1200/JCO.2008.21.6853

16. Theodoratou E, Campbell H, Tenesa A, Houlston R, Webb E, Lubbe S, et al. A large-scale meta-analysis to refine colorectal cancer risk estimates associated with MUTYH variants. *Br J Cancer* (2010) 103:1875–84. doi: 10.1038/sj.bjc.6605966

17. Win AK, Dowty JG, Cleary SP, Kim H, Buchanan DD, Young JP, et al. Risk of colorectal cancer for carriers of mutations in MUTYH, with and without a family history of cancer. *Gastroenterology* (2014) 146:1208–11. doi: 10.1053/j.gastro.2014.01.022

18. Peterlongo P, Mitra N, Sanchez de Abajo A, de la Hoya M, Bassi C, Bertario L, et al. Increased frequency of disease-causing MYH mutations in colon cancer families. *Carcinogenesis* (2006) 27:2243–9. doi: 10.1093/carcin/bgl093

19. Peterlongo P, Mitra N, Chuai S, Kirchhoff T, Palmer C, Huang H, et al. Colorectal cancer risk in individuals with biallelic or monoallelic mutations of MYH. *Int J Cancer* (2005) 114:505–7. doi: 10.1002/ijc.20767

20. Jones N, Vogt S, Nielsen M, Christian D, Wark PA, Eccles D, et al. Increased colorectal cancer incidence in obligate carriers of heterozygous mutations in MUTYH. *Gastroenterology* (2009) 137:489–94. doi: 10.1053/j.gastro.2009.04.047

21. Wasielewski M, Out AA, Vermeulen J, Nielsen M, van den Ouweland A, Tops CM, et al. Increased MUTYH mutation frequency among Dutch families with breast cancer and colorectal cancer. *Breast Cancer Res Treat* (2010) 124:635–41. doi: 10.1007/s10549-010-0801-7

22. Nielsen M, Franken PF, Reinards TH, Weiss MM, Wagner A, van der Klift H, et al. Multiplicity in polyp count and extracolonic manifestations in 40 Dutch patients with MYHassociated polyposis coli (MAP). *J Med Genet.* (2005) 42:e54. doi: 10.1136/jmg.2005.033217

23. Vogt S, Jones N, Christian D, Engel C, Nielsen M, Kaufmann A, et al. Expanded extracolonic tumor spectrum in MUTYH-associated polyposis. *Gastroenterology* (2009) 137:1976–85.e1-10. doi: 10.1053/j.gastro.2009.08.052

24. Win AK, Cleary SP, Dowty JG, Baron JA, Young JP, Buchanan DD, et al. Cancer risks for monoallelic MUTYH mutation carriers with a family history of colorectal cancer. *Int J Cancer* (2011) 129:2256–62. doi: 10.1002/ijc.25870

25. Out AA, Wasielewski M, Huijts PE, van Minderhout IJ, Houwing-Duistermaat JJ, Tops CM, et al. MUTYH gene variants and breast cancer in a Dutch case–control study. *Breast Cancer Res Treat.* (2012) 134:219–27. doi: 10.1007/s10549-012-1965-0

26. Rennert G, Lejbkowicz F, Cohen I, Pinchev M, Rennert HS, Barnett-Griness O. MutYH mutation carriers have increased breast cancer risk. *Cancer* (2012) 118:1989–93. doi: 10.1002/cncr.26506

27. Win AK, Reece JC, Dowty JG, Buchanan DD, Clendenning M, Rosty C, et al. Risk of extracolonic cancers for people with biallelic and monoallelic mutations in MUTYH. *Int J Cancer* (2016) 139:1557–63. doi: 10.1002/ijc.30197

28. Beiner M, Zhang W, Zhang S, Gallinger S, Sun P, Narod SA. Mutations of the MYH gene do not substantially contribute to the risk of breast cancer. *Breast Cancer Res Treat.* (2009) 114:575–8. doi: 10.1007/s10549-008-0042-1

29. Zhu M, Chen X, Zhang H, Xiao N, Zhu C, He Q, et al. AluYb8 insertion in the MUTYH gene and risk of early-onset breast and gastric cancers in the Chinese population. *Asian Pac J Cancer Prev.* (2011) 12:1451–5.

30. Boesaard EP, Vogelaar IP, Bult P, Wauters CA, van Krieken JH, Ligtenberg MJ. Germline MUTYH gene mutations are not frequently found in unselected patients with papillary breast carcinoma. *Hered Cancer Clin Pract.* (2014) 12:21. doi: 10.1186/1897-4287-12-21

31. Kappil M, Terry MB, Delgado-Cruzata L, Liao Y, Santella RM. Mismatch repair polymorphisms as markers of breast cancer prevalence in the breast cancer family registry. *Anticancer Res.* (2016) 36:4437–41. doi: 10.21873/anticanres.10987

32. Brinton LA, Cook MB, McCormack V, Johnson KC, Olsson H, Casagrande JT, et al. Anthropometric and hormonal risk factors for male breast cancer: male breast cancer pooling project results. *J Natl Cancer Inst.* (2014) 106:djt465. doi: 10.1093/jnci/djt465

33. Silvestri V, Rizzolo P, Zelli V, Valentini V, Zanna I, Bianchi S, et al. A possible role of FANCM mutations in male breast cancer susceptibility: results from a multicenter study in Italy. *Breast* (2018) 38:92–7. doi: 10.1016/j.breast.2017.12.013

34. Ottini L, Silvestri V, Rizzolo P, Falchetti M, Zanna I, Saieva C, et al. Clinical and pathologic characteristics of BRCA-positive and BRCA-negative male breast cancer patients: results from a collaborative multicenter study in Italy. *Breast Cancer Res Treat.* (2012) 134:411–8. doi: 10.1007/s10549-012-2062-0

35. Ottini L, Silvestri V, Saieva C, Rizzolo P, Zanna I, Falchetti M, et al. Association of low-penetrance alleles with male breast cancer risk and clinicopathological characteristics: results from a multicenter study in Italy. *Breast Cancer Res Treat.* (2013) 138:861–8. doi: 10.1007/s10549-013-2459-4

36. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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.* (2015) 17(5):405–24. doi: 10.1038/gim.2015.30

37. Kanter-Smoler G, Björk J, Fritzell K, Engwall Y, Hallberg B, Karlsson G, et al. Novel findings in Swedish patients with MYH-associated polyposis: mutation detection and clinical characterization. *Clin Gastroenterol Hepatol.* (2006) 4:499–506. doi: 10.1016/j.cgh.2006.01.005

38. Aretz S, Tricarico R, Papi L, Spier I, Pin E, Horpaopan S, et al. MUTYH-associated polyposis (MAP): evidence for the origin of the common European mutations p.Tyr179Cys and p.Gly396Asp by founder events. *Eur J Hum Genet.* (2014) 22:923–9. doi: 10.1038/ejhg.2012.309

39. Ali M, Kim H, Cleary S, Cupples C, Gallinger S, Bristow R. Characterization of mutant MUTYH proteins associated with familial colorectal cancer. *Gastroenterology* (2008) 135:499–507. doi: 10.1053/j.gastro.2008.04.035

40. Yurgelun MB, Allen B, Kaldate RR, Bowles KR, Judkins T, Kaushik P, et al. Identification of a variety of mutations in cancer predisposition genes in patients with suspected lynch syndrome. *Gastroenterology* (2015) 149(3):604–13.e20. doi: 10.1053/j.gastro.2015.05.006

41. Pin E, Pastrello C, Tricarico R, Papi L, Quaia M, Fornasarig M, et al. MUTYH c.933+3A>C, associated with a severely impaired gene expression, is the first Italian founder mutation in MUTYH-Associated Polyposis. *Int J Cancer* (2013) 132:1060–9. doi: 10.1002/ijc.27761

42. Sampson JR, Dolwani S, Jones S, Eccles D, Ellis A, Evans, et al. Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of MYH. *Lancet* (2003) 362:39–41. doi: 10.1016/S0140-6736(03)13805-6

43. Eliason K, Hendrickson BC, Judkins T, Norton M, Leclair B, Lyon E, et al. The potential for increased clinical sensitivity in genetic testing for polyposis colorectal cancer through the analysis of MYH mutations in North American patients. *J Med Genet.* (2005) 42:95–6. doi: 10.1136/jmg.2004.025973

44. Aretz S, Uhlhaas S, Goergens H, Siberg K, Vogel M, Pagenstecher C, et al. MUTYH-associated polyposis: 70 of 71 patients with biallelic mutations present with an attenuated or atypical phenotype. *Int J Cancer* (2006) 119:807–14. doi: 10.1002/ijc.21905

45. O'Shea AM, Cleary SP, Croitoru MA, Kim H, Berk T, Monga N, et al. Pathological features of colorectal carcinomas in MYH-associated polyposis. *Histopathology* (2008) 53:184–94. doi: 10.1111/j.1365-2559.2008.03071.x

46. Nielsen M, Joerink-van de Beld MC, Jones N, Vogt S, Tops CM, Vasen HF, et al. Analysis of MUTYH genotypes and colorectal phenotypes in patients With MUTYH - associated polyposis. *Gastroenterology* (2009) 136:471–6. doi: 10.1053/j.gastro.2008.10.056

47. Morak M, Laner A, Bacher U, Keiling C, Holinski-Feder E. MUTYH-associated polyposis - variability of the clinical phenotype in patients with biallelic and monoallelic MUTYH mutations and report on novel mutations. *Clin Genet.* (2010) 78:353–63. doi: 10.1111/j.1399-0004.2010.01478.x

48. Fromme JC, Banerjee A, Huang SJ, Verdine GL. Structural basis for removal of adenine mispaired with 8-oxoguanine by MutY adenine DNA glycosylase. *Nature* (2004) 427:652–6. doi:10.1038/nature02306

49. D'Agostino VG, Minoprio A, Torreri P, Marinoni I, Bossa C, Petrucci TC, et al. Functional analysis of MUTYH mutated proteins associated with familial adenomatous polyposis. *DNA Repair* (2010) 9:700–7. doi: 10.1016/j.dnarep.2010.03.008

50. Ottini L. Male breast cancer: a rare disease that might uncover underlying pathways of breast cancer. *Nat Rev Cancer* (2014) 14:643. doi: 10.1038/nrc3806

51. Görgens H, Krüger S, Kuhlisch E, Pagenstecher C, Höhl R, Schackert HK, et al. Microsatellite stable colorectal cancers in clinically suspected hereditary nonpolyposis colorectal cancer patients without vertical transmission of disease are unlikely to be caused by biallelic germline mutations in MYH. *J Mol Diagn.* (2006) 8:178–82. doi: 10.2353/jmoldx.2006.050119

52. Shin EJ, Chappell E, Pethe V, Hersey K, van der Kwast T, Fleshner N, et al. MYH mutations are rare in prostate cancer. *J Cancer Res Clin Oncol.* (2007) 133:373–8. doi: 10.1007/s00432-006-0181-x

53. Ashton KA, Proietto A, Otton G, Symonds I, Scott RJ. Genetic variants in MUTYH are not associated with endometrial cancer risk. *Hered Cancer Clin Pract.* (2009) 7:3. doi: 10.1186/1897-4287-7-3.

Elucidating the Underlying Functional Mechanisms of Breast Cancer Susceptibility Through Post-GWAS Analyses

Mahdi Rivandi[1,2], John W. M. Martens[1,3] and Antoinette Hollestelle[1]*

[1] Department of Medical Oncology, Erasmus MC Cancer Institute, Rotterdam, Netherlands, [2] Department of Modern Sciences and Technologies, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran, [3] Cancer Genomics Centre, Utrecht, Netherlands

Genome-wide association studies (GWAS) have identified more than 170 single nucleotide polymorphisms (SNPs) associated with the susceptibility to breast cancer. Together, these SNPs explain 18% of the familial relative risk, which is estimated to be nearly half of the total familial breast cancer risk that is collectively explained by low-risk susceptibility alleles. An important aspect of this success has been the access to large sample sizes through collaborative efforts within the Breast Cancer Association Consortium (BCAC), but also collaborations between cancer association consortia. Despite these achievements, however, understanding of each variant's underlying mechanism and how these SNPs predispose women to breast cancer remains limited and represents a major challenge in the field, particularly since the vast majority of the GWAS-identified SNPs are located in non-coding regions of the genome and are merely tags for the causal variants. In recent years, fine-scale mapping studies followed by functional evaluation of putative causal variants have begun to elucidate the biological function of several GWAS-identified variants. In this review, we discuss the findings and lessons learned from these post-GWAS analyses of 22 risk loci. Identifying the true causal variants underlying breast cancer susceptibility and their function not only provides better estimates of the explained familial relative risk thereby improving polygenetic risk scores (PRSs), it also increases our understanding of the biological mechanisms responsible for causing susceptibility to breast cancer. This will facilitate the identification of further breast cancer risk alleles and the development of preventive medicine for those women at increased risk for developing the disease.

Keywords: breast cancer, susceptibility loci, post-GWAS analysis, fine-scale mapping, functional analysis

INTRODUCTION

Breast cancer, the second deadliest cancer among women worldwide, is still the most frequently diagnosed malignancy among females (Fitzmaurice et al., 2017). Different risk factors, related to the development of breast cancer, have been identified with genetic predisposition playing a pivotal role. About 10–15% of the women who develop breast cancer have a familial background of the disease and several genes have been identified that increase breast cancer risk when mutated in the

germline (Collaborative Group on Hormonal Factors in Breast Cancer, 2001; Stratton and Rahman, 2008; Hollestelle et al., 2010b). Moreover, a large amount of non-coding germline variants have been identified that not only contribute to the breast cancer risk observed in individuals with a familial background, but also significantly in the general population (Lilyquist et al., 2018).

Currently identified breast cancer susceptibility genes and alleles can be stratified by their conferred risk in high, moderate and low-penetrant categories. *BRCA1* and *BRCA2* are the two most commonly mutated high-penetrance genes and about 15–20% of the familial breast cancer risk is attributable to germline mutations in one of these two genes (Miki et al., 1994; Wooster et al., 1995; Stratton and Rahman, 2008). Although germline mutations in *PTEN*, *TP53*, *STK11*, and *CDH1* also confer a high breast cancer risk, they are very rare and mostly found within the context of the cancer syndromes they cause. Hence, mutations in these genes explain no more than 1% of the familial breast cancer risk (Stratton and Rahman, 2008). A more intermediate risk of developing breast cancer is conferred by germline mutations in the genes *CHEK2*, *ATM*, *PALB2*, and *NBS1*, which are, in the general population, more prevalent than mutations in the high risk breast cancer genes. Together they explain another 5% of the familial breast cancer risk (Meijers-Heijboer et al., 2002; Vahteristo et al., 2002; Renwick et al., 2006; Steffen et al., 2006; Rahman et al., 2007; Hollestelle et al., 2010b). Interestingly, all high and moderate-risk genes identified so far have been implicated in the DNA damage response pathway (Hollestelle et al., 2010b).

Lastly, more than 170 low penetrant breast cancer susceptibility alleles have been identified through large-scale GWAS, which explain about 18% of the familial breast cancer risk (Michailidou et al., 2017). The vast majority of these GWAS-identified SNPs are, however, located outside coding regions (www.genome.gov/gwastudies). It is therefore not immediately obvious how these SNPs confer an increased risk to develop breast cancer. Moreover, since a GWAS design takes advantage of the linkage disequilibrium (LD) structure of the human genome and thus includes only SNPs tagging a particular locus, GWAS-identified SNPs usually do not represent the causal risk variants. Post-GWAS analyses are therefore imperative to identify the underlying causal SNP(s) and discern their mechanism of action. Since these causal SNPs are expected to display a stronger association with breast cancer risk than the original GWAS-identified SNPs (Spencer et al., 2011), their identification not only improves our estimates of the explained familial breast cancer risk by these SNPs, it also improves PRSs that aid in the identification of women at risk to develop breast cancer. In this review, we summarize the findings from post-GWAS analyses to date and discuss lessons learned with respect to design of these studies and the results that they have produced.

GWAS-IDENTIFIED SNPs

Since 2007, when one of the first large GWASs for breast cancer was published, multiple GWASs have been performed in order to identify those SNPs associated with the development of breast cancer (Easton et al., 2007; Hunter et al., 2007; Stacey et al., 2007, 2008; Gold et al., 2008; Ahmed et al., 2009; Thomas et al., 2009; Zheng et al., 2009; Turnbull et al., 2010; Cai et al., 2011a, 2014; Fletcher et al., 2011; Haiman et al., 2011; Ghoussaini et al., 2012; Kim et al., 2012; Long et al., 2012; Siddiq et al., 2012; Garcia-Closas et al., 2013; Michailidou et al., 2013, 2015, 2017; Purrington et al., 2014; Couch et al., 2016; Han et al., 2016; Milne et al., 2017). To date, 172 SNPs have been identified that associate with breast cancer risk. One of the major driving forces behind this success is the establishment of large international research consortia such as BCAC, which facilitated large sample sizes for breast cancer GWAS. Additionally, the cooperation between different large association consortia for breast, ovarian, prostate, lung and colon cancer (i.e., BCAC, CIMBA, OCAC, PRACTICAL, GAME-ON), which led to the development of the iCOGS array and the OncoArray has also been critical. In this respect, the iCOGS array facilitated the identification of 41 and 15 new breast cancer susceptibility loci, while the latest OncoArray facilitated identification of another 65 (Michailidou et al., 2013, 2015, 2017). Although the latest GWAS on the OncoArray has identified the most novel risk loci to date, the GWAS-identified variants were responsible for only 4% of familial breast cancer risk, suggesting that increasing samples sizes are allowing the identification of SNPs that confer smaller risks (Michailidou et al., 2017). Up to now, GWAS-identified SNPs collectively explain 18% of the familial breast cancer risk, but it is estimated that this is only 44% of the familial breast cancer risk that can be explained by all imputable SNPs combined (Michailidou et al., 2017). Identification of those SNPs as breast cancer susceptibility alleles will require even larger GWAS sample sizes, but also enrichment of phenotypes associated with breast cancer risk, as SNPs underlying ER-negative breast cancer are currently underrepresented.

In this respect, GWAS has also shown that estrogen receptor (ER)-positive and ER-negative breast cancer share a common etiology as well as a partly distinct etiology. Twenty loci were identified to associate specifically with ER-negative breast cancer, where a further 105 SNPs also associate with overall breast cancer (Milne et al., 2017). Furthermore, there is a common shared etiology for ER-negative breast cancer and breast cancers arising in *BRCA1* mutation carriers as well as overall breast cancer and breast cancer in *BRCA2* mutation carriers (Lilyquist et al., 2018).

Although the risks associated with single GWAS-identified SNPs are low, combining these SNPs in PRSs has shown to be useful for identifying women at high risk for developing breast cancer. In fact, based on a 77-SNP PRS developed by Mavaddat et al. 1% of women with the highest PRS have an estimated 3.4-fold higher risk of developing breast cancer as compared with the women in the middle quintile (Mavaddat et al., 2015). Moreover, PRSs were shown to be particularly useful for risk prediction within carriers of *BRCA1*, *BRCA2*, and *CHEK2* germline mutations as well as in addition to clinical risk prediction models (Dite et al., 2016; Kuchenbaecker et al., 2017; Muranen et al., 2017).

In summary, GWAS has allowed the research community to be very successful in the identification of risk loci that are associated with genetic predisposition to breast cancer. To date, more than 170 low-risk breast cancer susceptibility alleles have

been identified. Unfortunately, for the vast majority of the GWAS-identified risk loci, the causal variant(s), target gene(s) and their functional mechanism(s) have not yet been elucidated (Fachal and Dunning, 2015). Despite the development of tools and strategies for fine-scale mapping and functional analyses, the effort is still huge to characterize each GWAS-identified risk locus and reveal its underlying biology in breast tumorigenesis (Edwards et al., 2013; Fachal and Dunning, 2015; Spain and Barrett, 2015). However, for those 22 breast cancer risk that have been analyzed in more detail, this has provided already significant insight into the, sometimes complex, mechanisms underlying breast cancer susceptibility (**Table 1**) (Meyer et al., 2008, 2013; Udler et al., 2009, 2010a; Ahmadiyeh et al., 2010; Stacey et al., 2010; Beesley et al., 2011; Cai et al., 2011b; Bojesen et al., 2013; French et al., 2013; Ghoussaini et al., 2014, 2016; Quigley et al., 2014; Darabi et al., 2015, 2016; Glubb et al., 2015; Guo et al., 2015; Lin et al., 2015; Orr et al., 2015; Dunning et al., 2016; Hamdi et al., 2016; Horne et al., 2016; Lawrenson et al., 2016; Shi et al., 2016; Sun et al., 2016; Wyszynski et al., 2016; Zeng et al., 2016; Betts et al., 2017; Helbig et al., 2017; Michailidou et al., 2017).

FINE-SCALE MAPPING OF GWAS-IDENTIFIED LOCI

GWAS-identified SNPs usually do not represent the causal risk variants. These are merely tags to a locus associated with risk for developing the disease. However, because each causal variant is located in a region containing an independent set of correlated highly associated variants (iCHAV) (Edwards et al., 2013), fine-scale mapping of GWAS-identified loci in large sample sizes is required in order to identify the causal variant from a background of non-functional highly correlated neighboring SNPs.

In order to fulfill successful fine-scale mapping, a complete list of all SNPs, including the causal variants, should be available for the risk locus of interest. Direct sequencing of the risk locus would be a good approach for achieving this, however, it is an expensive method. Particularly since successful fine-scale mapping requires sufficient statistical power and thus sample sizes up to 4-fold to that of the original GWAS (Udler et al., 2010b). In this respect, the 1000 genome project containing whole genome sequencing data of 2,504 individuals from 26 populations is a valuable resource (Auton et al., 2015; Zheng-Bradley and Flicek, 2017). A second prerequisite for successful fine-scale mapping is large sample sizes, which are usually only achieved within large consortia such as BCAC. Therefore, both the iCOGS array as well as the OncoArray, in addition to a GWAS backbone, additionally contained numerous SNPs for fine-scale mapping of previously GWAS-identified risk loci (Michailidou et al., 2013, 2017).

Once a dense set of SNPs for a given GWAS-identified risk locus has been genotyped statistical analyses are applied to reduce the number of candidate causal SNPs. Interestingly, it seems to be a common theme among GWAS-identified loci that the underlying risk is conferred by more than one iCHAV. For breast cancer risk loci at 1p11.2, 2q33, 4q24, 5p12, 5p15.33, 5q11.2, 6q25.1, 8q24, 9q31.2, 10q21, 10q26, 11q13, and 12p11 multiple

iCHAVs have been identified ranging from two to a maximum of five iCHAVs at 6q25.1 and 8q24 (**Table 1**) (Bojesen et al., 2013; French et al., 2013; Meyer et al., 2013; Darabi et al., 2015; Glubb et al., 2015; Guo et al., 2015; Lin et al., 2015; Orr et al., 2015; Dunning et al., 2016; Ghoussaini et al., 2016; Horne et al., 2016; Shi et al., 2016; Zeng et al., 2016). For this reason, the first step in the fine-scale mapping process is establishing how many iCHAVs are present at a particular GWAS-identified risk locus using forward conditional regression analysis (Edwards et al., 2013). Then for each iCHAV, the SNP displaying the strongest association with breast cancer risk is identified. Based on this SNP, other SNPs within the same iCHAV are excluded from being candidate causal variants when the likelihood ratio for that SNP is smaller than 1:100 in comparison with the SNP showing the strongest association (Udler et al., 2010b). The reduction in candidate causal variants that is achieved during this process not only depends on sample size, but also the LD structure of the GWAS-identified locus.

Importantly, the majority of GWAS-identified risk loci were discovered in populations of European ancestry. Because the LD structure of the European ancestry population shows larger LD blocks containing more highly correlated SNPs than Asian or African ancestry populations, this offers an advantage in GWAS studies since less tagging SNPs are needed to achieve genome-wide coverage. However, for fine-scale mapping this is disadvantageous since the large number of highly correlated variants within an iCHAV may not allow sufficient reduction of candidate causal variants (Edwards et al., 2013). Therefore, fine-scale mapping in additional populations besides the European ancestry population (i.e., Asian and African ancestry populations) can be an effective strategy to reduce the number of candidate causal variants from iCHAVs located at GWAS-identified regions and add validity to the remaining candidate causal SNPs (Stacey et al., 2010; Edwards et al., 2013). Requirements for success are sufficient sample sizes for all populations, different correlation patterns between the studied populations and the risk association must be detectable in the additional populations, which usually depends on the risk allele frequency in these populations (Edwards et al., 2013). Unfortunately, the LD structure at the GWAS-identified risk loci is not always favorable and multiple highly correlated candidate causal variants remain. In this respect, analysis of the haplotypes that are present in a particular population and evaluation of their association with breast cancer risk may provide another strategy for exclusion of non-causal SNPs within an iCHAV (Chatterjee et al., 2009).

The purpose of fine-scale mapping is to identify the number of iCHAVs underlying GWAS-identified risk loci and reducing the number of candidate causal variants in these iCHAVs to a minimum. In practice, this reduction does not directly lead to identification of the single causal variant responsible for this risk due to several of the reasons described above. Either way, whether only one, a few or many candidate causal SNPs remain, in the next phase the candidate causal variants need to be validated or further reduced by elucidating the functional mechanism through which these variants operate. First, overlap between the candidate causal variants and regulatory sequences

TABLE 1 | Overview of post-GWAS studies that have performed more extensive fine-scale mapping, in-silico prediction or functional analysis.

Locus	Putative casual SNPs	Target genes	DHSS	FAIRE	TFBS	Histone marks	3C	ChIA-PET	EMSA	eQTLs	Luciferase reporter assay	Other	References
1p11.2	2 iCHAVs: rs11249433, rs12134101; rs146784183	NOTCH2	rs11249433			H3K27Ac marks at rs11249433				No associations for rs11249433 or rs146784183			Horne et al., 2016
1p34	rs4233486, rs35054111, rs11804913, rs7554973	CITED4	rs4233486		JARID1B and FOXM1 bind rs4233486	H3K4Me1, H3K4Me2, H3K4Me3, H3K9Ac, H3K27Ac at rs4233486; H3K4Me1 and H3K27Ac at rs11804913; H3K27Ac at rs7554973	PRE1 and PRE2 interact with the CITED4 promoter				The rs4233486 risk allele in PRE1 enhances CITED4 promoter activity		Michailidou et al., 2017
1p36	rs2992756	KLHDC7A	rs2992756		ER, PBX1, POLR2A, SPDEF, JARID1B, EP300, FOXA1, GATA3, HIF1α, HIF1β	H3K4Me1, H3K4Me2, H3K4Me3, H3K9Ac, H3K27Ac					The KLHDC7A promoter containing the rs2992756 risk allele has reduced activity		Michailidou et al., 2017
2q33	4 iCHAVs: rs1830298, rs10197246; rs36043647; rs59278883; rs7558475	ALS2CR, CASP8, CFLAR	rs3769823 and rs3769821 in iCHAV1			H3K27Ac marks at rs3769823 and rs3769821 in iCHAV1				minor alleles of rs6754084 and rs6743068 in iCHAV1 decrease CASP8 expression			Lin et al., 2015
2q35	1 iCHAV: rs4442975, rs6721996	IGFBP5	rs4442975		FOXA1 is preferentially recruited to the common allele of rs4442975	H3K4Me1, H3K4Me2 marks near rs4442975	The common allele of rs4442975 interacts with the C allele of chr2:271,557,291 and the IGFBP5 promoter			The common allele of rs4442975 increases IGFBP5 expression in ER+ cell lines and normal breast tissue, estrogen induction increases IGFBP5 expression in cells carrying a common allele of rs4442975	PRE containing rs4442975 does not affect IGFBP5 expression		Ghoussaini et al., 2014
2q35	1 iCHAV: 14 SNPs including the 1.3kb enCNV				Enhanced ERα binding on the 1.3kb enCNV before and after estrogen treatment		1.3kb enCNV interacts with IGFBP5 promoter					Differential allelic binding of ERα at the 1.3kb enCNV reduces allele-specific IGFBP5 expression in response to estrogen	Wyszynski et al., 2016
4q21	89 SNPs	FAM175A, HELQ, MRPS18C, HSPE	rs6844460		MAX binds rs6844460	H3K9Ac marks at rs11099601; H3K4Me3, H3K9Ac and H3K27Ac marks at rs6844460		rs11099601 and rs6844460 interact with the MRPS18C promoter		The risk allele of rs11099601 associates with decreased HELQ and increased MRPS18C, FAM175A and HPSE expression, but this was inconsistent across data sets			Hamdi et al., 2016

(Continued)

TABLE 1 | Continued

Locus	Putative casual SNPs	Target genes	DHSS	FAIRE	TFBS	Histone marks	3C	ChIA-PET	EMSA	eQTLs	Luciferase reporter assay	Other	References
4q24	2 iCHAVs: 24 SNPs; 5 SNPs	TET2	rs62331150 and rs73838678 in iCHAV1		rs62331150 in iCHAV1 lies in a SP1, EGR1, NIFC and near at a P300 binding site, rs73838678 in iCHAV1 lies in a PR, EGR1, NIFC and near a P300 binding site	H3K4Me and H3K27Ac marks at rs62331150 and H3K27Ac marks are at rs73838678 in iCHAV1		rs62331150 and rs73838678 in iCHAV1 interact with the TET2 promoter		The risk allele of rs62331150 decreases TET2 expression			Guo et al., 2015
5p12	rs7716600	MRPS30	DHSS cluster at a locus 700 bp from rs7716600	Reduced euchromatic conditions at the MRPS30 promoter after estrogen stimulation for common homozygotes	2-fold increase in ERα binding at de MRPS30 promoter and DHSS locus, significant increase in CTCF binding at the MRPS30 promoter and CTCF locus in common homozygotes after estrogen stimulation	The risk allele of rs7716600 associates with decreased methylation of a probe located immediately 5' of MRPS30				rs7716600 risk allele upregulates MRPS30 expression	rs7716600 risk allele upregulates MRPS30 expression	In ER+ breast tumors MRPS30 expression correlated strongly with expression of genes in the estrogen signaling pathway. MRPS30 expression is increased in response to estrogen in MPE600 cells which are homozygous for the risk allele.	Quigley et al., 2014
5p12	3 iCHAVs: rs10941679; 38 SNPs; rs200229088	FGF10, MRPS30, HCN1			FOXA1 and OCT1 bind rs10941679, but not allele specific	None	FGF10 and MRPS30 promoter		allele-specific binding of rs10941679 by FOXA1, FOXA2, CEBPB and OCT1	rs10941679 risk allele upregulates FGF10 and MRPS30 expression	rs10941679 risk allele had no additional effect on the PRE enhancer activity for FGF10, MRPS30 and BRCAT54 promoters		Ghoussaini et al., 2016
5p15.33	rs2736108, rs2736109	TERT								None	rs2736108 and rs2736109 risk alleles combined reduce TERT promoter activity		Beesley et al., 2011
5p15.33	3 iCHAVs: 7 SNPs, 6 SNPs, 3 SNPs	TERT	None	rs2736108 and rs2736109 common alleles, but not rs3215401 and rs2853669 alleles associated with open chromatin	None	None					rs2736107, rs2736108 and rs2736109 minor alleles from iCHAV1 decreased transcription; including the rs705526 minor allele from iCHAV2 increases TERT promoter activity; the PRE and rs2242652, but not rs10069690 from iCHAV3 decrease TERT promoter activity	The rs10069690 minor allele associates with an alternatively spliced TERT isoform leading to a premature stop codon	Bojesen et al., 2013
5p15.33	rs3215401, rs2853669	TERT			SP2, ZTTB7A for rs3215401 and ETS, MYC, MIXL1, RBPJ, SIN3A, ZNF143, EP300 for rs2853669; ChIP for GABPA and MYC, but not ETS2, ELF1 or E2F1 led to preferential isolation of the rs2853669 risk allele		None				rs3215401 and rs2853669 risk alleles reduce TERT promoter activity, but not rs2736107, rs2736108, rs2736109 and rs145544133 risk alleles	Silencing of MYC, but not ETS2 downregulated TERT promoter activity, irrespective of rs2853669 genotype	Helbig et al., 2017

(Continued)

TABLE 1 | Continued

Locus	Putative causal SNPs	Target genes	DHSS	FAIRE	TFBS	Histone marks	3C	ChIA-PET	EMSA	eQTLs	Luciferase reporter assay	Other	References
5q11.2	3 iCHAVs: 15 SNPs; 90 SNPs, 66 SNPs; 5 SNPs	MAP3K1			FOXA1 binds PRE-B1, ERα binds PRE-C, GATA3 preferentially bind the risk allele of iCHAV3 rs17432750 in PRE-B3	H3K4Me1, H3K4Me2, H3K27Ac	MAP3K1 promoter	All 4 PREs interact with the MAP3K1 promoter		No association	In iCHAV1 PRE-A downregulates MAP3K1 and PRE-B1 and PRE-C upregulate MAP3K1, rs74345699 and rs62355900 risk alleles in PRE-C further upregulate MAP3K1 in the presence of estrogen; in iCHAV2a PRE-D upregulates MAP3K1, which is further enhanced by the rs16886397 risk allele; in iCHAV2b PRE-2B upregulates MAP3K1, which is further enhanced by the rs62355881 risk allele; in iCHAV3 PRE-B3 downregulates MAP3K1, which is further reduced by the rs17432750 risk allele	siRNA against GATA3 reduced the enhancer activity of PRE-B3 containing the risk allele of rs17432750	Glubb et al., 2015
6q25.1	rs9397435, rs77275268	ESR1, PGR			None for rs9397435, CTCF for rs77275268	H3K4Me1, H3K4Me2, H3K9Ac for rs9397435				rs9397435 risk homozygotes show increased ESR1 and PGR expression		rs77275268 is located in a partially methylated CpG sequence	Stacey et al., 2010
6q25.1	rs6913578, rs7763637				None				The risk allele of rs6913578 significantly altered DNA-protein complex intensity, no detectable interaction of rs7763637 with nuclear proteins		Transcription activation was significantly increased for common alleles of rs6913578 and rs7763637		Cai et al., 2011b
6q25.1	rs7763637	AKAP12, ESR1, RMND1, ZBTB2			ZNF217, FOS, KAP1, JUND, FOSL2, JUN, MYC	H3K4Me3, H3K4Me1, H3K27Ac				rs7763637 risk allele upregulates AKAP12 expression in adjacent normal breast tissue and breast tumors, but downregulates ESR1, RMND1 and ZBTB2 in breast tumors			Sun et al., 2016

(Continued)

TABLE 1 | Continued

Locus	Putative causal SNPs	Target genes	DHSS	FAIRE	TFBS	Histone marks	3C	ChIA-PET	EMSA	eQTLs	Luciferase reporter assay	Other	References
6q25.1	5 iCHAVs: 10 SNPs; 3 SNPs; 4 SNPs; 3 SNPs; 6 SNPs	ESR1, RMND1, CCDC170	19 of the 26 candidate causal SNPs	19 of the 26 candidate causal SNPs	GATA3 binds the risk allele of iCHAV3 SNP rs851982; CTCF binds the risk allele of iCHAV3 SNP rs851983 and the common allele of iCHAV4 SNP rs1361024; MYC binds the common allele of iCHAV5 SNP rs9910416	19 of the 26 candidate causal SNPs were associated with enhancer enriched histone marks; H3K27Ac marks were enriched at rs20462010, rs12173570 and rs851984	iCHAV1-2 elements interact with ESR1, RMND1-ARMT1 and CCDC170 promoters; iCHAV3-5 elements interact with ESR1 and RMND1-ARMT1 promoters; the common allele of iCHAV4 SNP rs1361024 increases looping to ESR1 and RMND1 promoters		11 of the 19 causal candidate SNPs associated with PREs, altered the binding activity of transcription factors of which 7 fell within promoter -specific interactions as identified by 3C	Risk alleles of iCHAV1 reduced ER expression; risk alleles of iCHAV1, 3 and 5 increased ESR1 expression in ER+ tumors compared with normal, tumor-adjacent tissue; Imbalance in allele-specific expression in ESR1 for SNPs in iCHAV1-3, in CCDC170 for iCHAV2 SNP rs9397437 and in RMND1 for iCHAV3 SNP rs851983; risk alleles of iCHAV3 associate with CCDC170 expression	iCHAV1 SNP rs6557160, iCHAV2 SNP rs17081533 and iCHAV5 SNP rs9910416 reduce ESR1 and RMND1 promoter activity, although inclusion of the iCHAV1 haplotype reduced ESR1, RMND1 and CCDC170 promoter activity; iCHAV3 SNP rs851982 increased ESR1 promoter activity		Dunning et al., 2016
7q22	rs13229095, rs6979850, rs6961094, rs7796917, rs71559437, rs11972884	CUX1, RASA4, PRKRIP1	rs4233486		CEBPB, ER, FOXA1, FOXM1, E2F1, MAX, P300, PBX1, SIN3A, MYC, SPDEF, FOSL2, GATA3, NR2F2, RARA, TCF7L2, POLR2A, REST, RIP140	H3K4Me1, H3K4Me2, H3K4Me3, H3K9Ac, H3K27Ac	PRE1 interacts with the CUX1 and RASA4 promoter; PRE2 interacts with the RASA4 and PRKRIP1 promoter; the risk haplotype associated with chromatin looping						Michailidou et al., 2017
8q24						H3K4Me2	MYC						Ahmadiyeh et al., 2010
8q24	5 iCHAVs: rs35961416; rs13281615; rs7815245; rs2033101; rs1121948		rs7815245; rs1121948		The rs7815245 risk allele alters the TCF12 binding motif and is located in an ESR1 and close to a FOXA1 binding site; rs1121948 is located in a GATA3 and MAX binding site	H3K4Me1, H3K27Ac marks at rs1121948				The rs7815245 risk allele downregulates POU5F1B; the rs1121948 risk allele downregulates PVT1 and MYC			Shi et al., 2016

(Continued)

TABLE 1 | Continued

Locus	Putative casual SNPs	Target genes	DHSS	FAIRE	TFBS	Histone marks	3C	ChIA-PET	EMSA	eQTLs	Luciferase reporter assay	Other	References
9q31.2	3 iCHAVs: 28 SNPs; rs10816625; rs13294895	KLF4	iCHAV1: rs662694, rs471467, rs5899787		CTCF binds iCHAV1 rs662694 and rs471467, ERα, FOXA1, FOXM1 and GATA3 bind iCHAV1 rs5899787; ERα, FOXA1, FOXM1, GATA3, HDAC2, Max, NR2F2, P300 and Sin3A bind iCHAV2 rs10816625 and iCHAV3 rs13294895	iCHAV1 rs5899787 site is enriched for H3K27me3 marks; iCHAV2 rs10816625 and iCHAV3 rs13294895 localize to a PRE marked by H3K27Ac and H3K4Me1	iCHAV1 rs10816625 and iCHAV3 rs13294895 interact with the KLF4 promoter				iCHAV2 rs10816625 and iCHAV3 rs13294895 decrease KLF4 expression		Orr et al., 2015
10q21	4 iCHAVS: 12 SNPs; 17 SNPs; 18 SNPs; rs9971363, rs7090365	ZNF365 NRBF2	PRE1 and PRE2 in iCHAV2			H3K4Me1 and H3K4Me2 marks are enriched at PRE1 and PRE2 in iCHAV2	iCHAV2 interacts with ZNF365 and NRBF2 promoters			No association	iCHAV2 protective haplotype downregulates NRBF2 and ZNF365 expression		Darabi et al., 2015
10q26	rs7895676, rs2981578	FGFR2			C/EBPβ, RUNX2				C/EBPβ binds rs7895676 minor allele, RUNX2 binds rs2981578 minor allele	Minor alleles of rs7895676 and rs2981578 upregulate FGFR2	No significant transcription activation for minor allele of rs7895676, but rs2981578; 2-5 fold higher transcription activation for minor allele of rs2981578		Meyer et al., 2008
10q26	3 iCHAVs: rs35054928, rs34032268, rs2981579, rs2912779, rs2912780; rs45631563; rs2981578; rs45631539	FGFR2	rs35054928, rs2981579, rs2912779; rs45631563; rs2981578	Increased chromatin accessibility of the risk allele	E2F1 preferentially binds rs35054928 minor allele, no enrichment for SP1; Ser5P-Pol II, FOXA1 and ERα preferentially bind rs2981578 minor allele, low enrichment for RUNX2		FGFR2 promoter	FGFR2 promoter	E2F1 and SP1 bind rs35054928, ERα binds rs2981579, an unidentified protein binds to rs2912779; an unidentified nuclear protein binds rs45631563; OCT1, RUNX2 and FOXA1 bind rs2981578	No association between rs35054928 or rs2981578 genotypes and FGFR2, ATE1, NSMCE4A or TACC2 expression		siRNA against FOXA1 downregulates FGFR2, siRNA against E2F1 had little effect on FGFR2 but upregulated FOXA1	Meyer et al., 2013
10q26	rs7895676, rs10736303, rs2912781, rs2912778, rs2981578		rs2912778, rs2981578										Udler et al., 2009
11p15	19 SNPs	PIDD1	chr11:801630_ ATG, rs7104785, rs7484123, rs7484068, rs11246313, rs11246314		ER, NR2F2, RIP140, SPDEF, CTCF, E2F4, POLR2A, EGR1, GABPA, E2F1, JARID1B, PML, FOXM1, EGLN2, HIF1α, HIF1β, NRF1	H3K4Me1, H3K4Me2, H3K4Me3, H3K9Ac, H3K27Ac					The PIDD1 promoter containing the risk haplotype has increased activity		Michailidou et al., 2017

(Continued)

TABLE 1 | Continued

Locus	Putative casual SNPs	Target genes	DHSS	FAIRE	TFBS	Histone marks	3C	ChIA-PET	EMSA	eQTLs	Luciferase reporter assay	Other	References
11q13	3 iCHAVs: rs661204, rs78540526, rs554219, rs657686; rs75915166; rs494406, rs585568, rs5933679, rs679162	CCND1			The 4 iCHAV1 SNPs fall in PRE1 which binds ERα and FOXA1, allele-specific binding of ELK4 to rs554219; iCHAV2 SNP rs75915166 falls in PRE2	PRE1 is flanked by H3K4Me1 and H3K4Me2 marks	PRE1 interacts with the CCND1 promoter and an enhancer of CCND1 located in the CCND1 terminator region; PRE2 interacts with the CCND1 promoter; PRE1 interacts with PRE2	PRE1 interacts with the CCND1 promoter and an enhancer of CCND1 located in the CCND1 terminator region	The common alleles of rs661204 and rs78540526 preferentially bind USF1 and USF2, the common allele of rs554219 is bound by ELK4 and GABPA; the minor allele of rs75915166 interacts specifically with GATA3	homozygotes for the rs554219 risk allele have reduced cyclin D1 expression	Risk alleles of rs78540526 and rs554219 abolish PRE1 enhancer activity and decrease CCND1 promoter activity, PRE1 is estrogen inducible irrespective of the SNPs genotypes; minor allele of rs75915166 increases strength of PRE2 silencer	siRNA against ELK4 reduce enhancer activity of wild type PRE1, but not PRE1 containing the risk allele of rs554219; siRNA against GATA3 increases transcription in the presence of the rs75915166 risk allele, but not the common allele	French et al., 2013
11q13		CCND1, CUPID1, CUPID2					the risk alleles of rs661204 and rs78540526 abolish interaction of PRE1 with the predicted promoter of CUPID1 and CUPID2				risk alleles of rs661204 and rs78540526 reduced PRE1 enhancer activity on the CUPID1 promoter	silencing PRE1 by dCas9-KRAB reduced CUPID1, CUPID2 and CCND1 expression, estrogen induction of the CUPID1 and CUPID2 promoter depended on PRE1 but not the risk SNPs, CUPID1 and CUPID2 regulated genes affect DNA repair and recombination, tumors with low CCND1, CUPID1 or CUPID2 expression had similar mutational signatures as BRCA1 and BRCA2 deficient tumors, silencing of CUPID1 and CUPID2 impaired end resection and NHEJ could compensate for the lack of HRR	Betts et al., 2017
12p11	4 iCHAVs: 4 SNPs; 74 SNPs; 376 SNPs; 2 SNPs	CCDC91, PTHLH			iCHAV1 rs812020 disrupts a E2F3 binding site; iCHAV2 rs788463 is in a C/EBP binding site and rs10843066 disrupts a HNF1B binding site; iCHAV3 rs10843110 disrupts a PPARG binding site and rs11049453 disrupts a PAX binding site	H3K4Me3 and H3K27Ac at iCHAV1-4		multiple iCHAV1 and iCHAV2 SNPs interact with the PTHLH promoter		The iCHAV3 rs11049453 risk allele increases PTHLH and decreases CCDC91 expression			Zeng et al., 2016

(Continued)

TABLE 1 | Continued

Locus	Putative casual SNPs	Target genes	DHSS	FAIRE	TFBS	Histone marks	3C	ChIA-PET	EMSA	eQTLs	Luciferase reporter assay	Other	References
16q12	14 SNPs	TOX3, LOC643714	rs12930156, rs3095604, rs45538731, rs28463809, rs4784226		The risk allele of rs4784227 creates a C/EBPα binding site					No association between rs3803662 genotypes and TOX3 expression; rs3803662 genotypes associated with RBL2 expression in lymphocytes, but not breast tumors			Udler et al., 2010b
17q22	28 SNPs	STXBP4	rs244353, rs2787481, rs244371			None		No interactions		rs2787481 genotypes associate with COX11 expression; rs2787481, rs244317 and rs11658717 genotypes downregulate STXBP4 and upregulate a short STXBP4 isoform, rs244353 genotypes downregulate STXBP4 expression		rs244353 is located in an enhancer predicted to target the STXBP4 gene and an enhancer predicted to target the HLF gene	Darabi et al., 2016
19p13.1	1 iCHAV; 13 SNPs	ABHD8, ANKLE1			rs55924783 and rs56069439 coincided with CTCF binding sites	rs56069439 and rs4808616 coincided with H3K4Me1 marks	rs4808075, rs10419397, rs56069439 and rs4808076 interact with the ABHD8 promoter			The risk allele of 13 SNPs associate with increased ABHD8 expression; the risk allele of rs56069439 associates with greater allele-specific expression of ABHD8	PRE-A, B and C upregulate ABHD8, which is further enhanced by the risk alleles of rs56069439, rs1132992211, rs67397200, rs61494113, rs4808616 and rs55924783; PRE-A silences ANKLE1; PRE-C upregulates ANKLE1, which is reduced by the risk alleles of rs4808616 and rs55924783	CRISPR/Cas9 deletion of a 57 bp region containing rs56069439 reduced ANKLE1, but not ABHD8 or BABAM1 expression; overexpression of ABHD8 reduced cell migration and invasion and caused expression changes in cancer-related pathways, overexpression of ANKLE1 caused expression changes in cancer-associated and cell growth/proliferation pathways	Lawrenson et al., 2016

DHSS, DNase I hypersensitivity sites; FAIRE, Formaldehyde-assisted isolation of regulatory elements; TFBS, Transcription factor binding sites; 3C, Chromatin conformation capture; ChIA-PET, Chromatin interaction analysis by paired-end tag sequencing; EMSA, Electrophoretic mobility shift assay; eQTLs, Expression quantitive trait loci; Ref, reference; iCHAV, Independent set of correlated highly associated variants; enCNV, Enhancer copy number variation; PRE, Putative regulatory element; NHEJ, Non-homologous end joining; HRR, homologous recombination repair.

such as transcription factor (TF) binding sites, histone marks or regions of open chromatin is evaluated *in silico*. In addition, expression quantitative trait loci (eQTL) studies are performed in order to identify the genes that are deregulated by the candidate causal variants. The hypotheses for the functional mechanisms by which the candidate causal SNPs confer breast cancer risk are then further tested by molecular experiments in *in-vitro* model systems.

IN-SILICO PREDICTION OF FUNCTIONAL MECHANISMS

The vast majority of GWAS-identified SNPs are not protein-coding and are located in intronic or intragenic regions, or even in gene deserts (www.genome.gov/gwastudies). Their underlying causal variants usually have a regulatory role by modulating the expression of target genes or non-coding RNAs (ncRNAs). Therefore, causal variants usually coincide with regulatory regions associated with open chromatin, TF binding sites, sites of histone modification or chromatin interactions (**Table 1**) (Meyer et al., 2008, 2013; Stacey et al., 2010; Udler et al., 2010a; Beesley et al., 2011; Cai et al., 2011a; Bojesen et al., 2013; French et al., 2013; Ghoussaini et al., 2014, 2016; Quigley et al., 2014; Darabi et al., 2015, 2016; Glubb et al., 2015; Guo et al., 2015; Lin et al., 2015; Orr et al., 2015; Dunning et al., 2016; Hamdi et al., 2016; Lawrenson et al., 2016; Shi et al., 2016; Sun et al., 2016; Wyszynski et al., 2016; Zeng et al., 2016; Betts et al., 2017; Helbig et al., 2017; Michailidou et al., 2017). Mining public data for these regulatory features can be an effective way to narrow down the list of candidate causal variants after fine-scale mapping. Furthermore, to determine which candidate causal SNPs affect gene expression, eQTLs can be evaluated. Besides narrowing down the list of candidate causal variants, these *in silico* predictions, additionally, provide clues about the functional mechanisms involved, which will guide the design of molecular experiments.

Regulatory Features

A wealth of data is publically available regarding regulatory features throughout the genome. Via ENCODE (https://www.encodeproject.org/), data on locations of open chromatin, TF binding sites, DNA methylation, RNA expression and histone modifications can be retrieved (Djebali et al., 2012; ENCODE Project Consortium, 2012; Neph et al., 2012; Sanyal et al., 2012; Thurman et al., 2012). The NIH Roadmap Epigenomics project (http://www.roadmapepigenomics.org/) contains data on locations of open chromatin, DNA methylation and histone modifications (Kundaje et al., 2015; Zhou et al., 2015). In addition, Nuclear Receptor Cistrome (http://cistrome.org/NR_Cistrome/index.html) also has information on TF binding locations. Using FunctiSNP (http://www.bioconductor.org/packages/release/bioc/html/FunciSNP.html), RegulomeDB (http://www.regulomedb.org/) and HaploReg (http://archive.broadinstitute.org/mammals/haploreg/haploreg.php) these sources of information can be mined allowing the prediction of putative regulatory regions (PREs) within an iCHAV (Boyle

et al., 2012; Coetzee et al., 2012; Ward and Kellis, 2012). The long range chromatin interactions that these PREs may establish can subsequently be assessed via GWAS3D (http://jjwanglab.org/gwas3d) and the 3D Genome Browser (http://promoter.bx.psu.edu/hi-c/) providing clues about the target genes or ncRNAs that could be deregulated (Li et al., 2013a; Yardimci and Noble, 2017).

Interestingly, several regulatory features appear to be enriched among GWAS-identified breast cancer risk loci, such as TF binding sites for ERα, FOXA1, GATA3, E2F1, and TCF7L2, but also H3K4Me1 histone marks as well as regions of open chromatin marked by DNAse I hypersensitivity sites (DHSSs) (Cowper-Sal lari et al., 2012; Michailidou et al., 2017). It is important to keep in mind, however, that despite of the wealth of data available, these data sources harbor information for only a fraction of the TFs present in the human proteome. This means that other regulatory features, which we are currently unable to evaluate, may also play an important role in mediating the susceptibility to breast cancer. Moreover, TFs, as well as histone marks and chromatin interactions, are highly tissue specific and it will therefore be crucial to evaluate these regulatory features in the proper tissue type or cell line to prevent either false positive or false negative associations. In order to obtain a more comprehensive understanding of the mechanisms underlying breast cancer predisposition, we thus need cistrome data on more TFs from more tissue types.

Still, mining of the currently available data has facilitated the identification of causal variants and/or functional mechanisms for several of the identified GWAS-identified loci (Meyer et al., 2008, 2013; Udler et al., 2010a; French et al., 2013; Ghoussaini et al., 2014, 2016; Quigley et al., 2014; Darabi et al., 2015; Glubb et al., 2015; Guo et al., 2015; Orr et al., 2015; Dunning et al., 2016; Hamdi et al., 2016; Lawrenson et al., 2016; Shi et al., 2016; Zeng et al., 2016; Helbig et al., 2017; Michailidou et al., 2017). Combining information on regulatory features from candidate causal variants with eQTLs will further narrow down the list of candidate variants, identify target genes and provide a starting point for subsequent *in-vitro* molecular experiments.

eQTLs

eQTLs are variants that control gene expression levels and are therefore found in regulatory regions in the genome. Evidence for a candidate causal variant to be associated with gene expression can be obtained from eQTL studies. In an eQTL study, the presence of a correlation between expression levels of potential target genes and the genotypes of the candidate causal variants is evaluated in an unbiased manner. Two types of eQTL studies are generally distinguished based on the distance of the gene from the candidate SNP. In cis-eQTL studies, the target genes being evaluated are in close proximity to the candidate causal variant, usually within 1 to 2 megabases. For trans-eQTL studies, all genes outside this region, thus also on other chromosomes, are subjected to evaluation (Cheung and Spielman, 2009). Far more genes are thus tested for correlation with candidate causal variants in trans-eQTL analyses than cis-eQTL analyses and, consequently, trans-eQTL studies require far more statistical power than cis-eQTL studies. It is therefore that in most of

the post-GWAS analyses only cis-eQTL analysis is performed. Moreover, besides gene expression, eQTLs can also influence the expression of ncRNAs, mRNA stability, differences in allelic expression and differential isoform expression (Ge et al., 2009; Lalonde et al., 2011; Pai et al., 2012; Kumar et al., 2013).

SNPs that are located in regulatory regions of genome show a higher tissue specificity and it is therefore no surprise that eQTLs in GWAS-identified regions also display high tissue specificity (Dimas et al., 2009; Fu et al., 2012). Consequently, choice of tissue type in an eQTL study is critical to prevent false positive or false negative associations. The most obvious choice is the target tissue under investigation. For breast cancer, this can be either normal breast tissue or breast tumor tissue. In this respect, the cancer genome atlas (TCGA; https://cancergenome.nih.gov/), Molecular Taxonomy of Breast Cancer International Consortium (METABRIC; http://www.ebi.ac.uk/ega/) and Genotype Tissue Expression (GTEx; https://gtexportal.org/home/) are valuable resources (Cancer Genome Atlas Network, 2012; Curtis et al., 2012; Battle et al., 2017). However, eQTL studies in breast cancer tissue are confounded by the presence of copy number variation, somatic mutations and differential methylation that influence gene expression levels. Therefore, eQTLs are ideally evaluated in normal breast tissue. Unfortunately, availability of both genotyping and gene expression data for normal breast tissue is limited as compared with breast tumor tissue, resulting in lower statistical power in eQTL analyses. Alternatively, for breast tumor analyses, gene expression data could also be adjusted for somatic CNVs and methylation variation (Li et al., 2013b). In addition, it should also be considered that the tumor micro-environment plays an important role in the development of breast cancer and that expression levels deregulated in stroma or immune cells might also be relevant.

It is important to treat the identification of eQTLs with some caution. False positives and false negatives could be a result from choosing the incorrect tissue type. In six post-GWAS studies to date an eQTL association was observed and an attempt was made to validate these results with luciferase reporter assays (Meyer et al., 2008; French et al., 2013; Ghoussaini et al., 2014, 2016; Dunning et al., 2016; Lawrenson et al., 2016). For GWAS-identified risk loci at 2q35 and 5p12, luciferase reporter assays did not confirm the eQTL association, whilst this was the case for eQTL associations at 6q25.1, 10q26, 11q13, and 19q13.1 (**Table 1**). In addition, when evaluating cis-eQTLs, false negative results could also imply that more distant eQTLs are involved. Moreover, since causal variants from different iCHAVs within a GWAS-identified region can influence the same target gene (Bojesen et al., 2013; French et al., 2013; Glubb et al., 2015; Dunning et al., 2016; Lawrenson et al., 2016), eQTLs may remain undetected. For example, in the post-GWAS study by Glubb et al. at the 5q11.2 locus, PRE-A downregulated *MAP3K1*, whereas PRE-B1 and PRE-C upregulated *MAP3K1* expression although no eQTL associations were identified (Glubb et al., 2015). Similarly, Lawrenson et al. studied the GWAS-identified breast cancer risk locus at 19p13.1 and noticed PRE-A downregulating *ANKLE1* and PRE-C upregulating *ANKLE1* expression, while no eQTL association was detected. Interestingly, at this same locus three PREs regulating *ABHD8* all upregulated its expression

and consistent with this 13 eQTL associations were detected of which one was allele-specific (Lawrenson et al., 2016). Thus, absence of an association does not necessarily imply trans-eQTL associations. For the above reasons, additional *in vitro* molecular experiments are necessary to confirm the results from eQTL studies, but also from the *in silico* predictions of regulatory features and chromatin interactions.

A recently developed tool that is also of interest to predict target genes from GWAS-identified breast cancer risk loci is INQUISIT (integrated expression quantitative trait and *in silico* prediction of GWAS targets) which combines both regulatory features and eQTL data from publically available resources (Michailidou et al., 2017). Interestingly, INQUISIT predicted target genes for 128 out of 142 GWAS-identified breast cancer risk loci and among the 689 target genes a strong enrichment was observed for breast cancer drivers. Furthermore, pathway analysis of these genes revealed involvement of fibroblast growth factor, platelet-derived growth factor and Wnt signaling pathways to be involved in genetic predisposition to breast cancer as well as the ERK1/2 cascade, immune response and cell cycle pathways (Michailidou et al., 2017). However, the expression of breast cancer driver genes is not necessarily deregulated in the same direction by the germline variants as by somatic mutations. For example, *MAP3K1* is upregulated and *CCND1* and *TERT* are downregulated in the germline. This is in contrast with breast tumors, where *MAP3K1* is downregulated and *CCND1* and *TERT* are upregulated by somatic mutations (Bojesen et al., 2013; French et al., 2013; Glubb et al., 2015).

IN-VITRO FUNCTIONAL EXPERIMENTS

After *in silico* prediction of regulatory features and the identification of putative target genes, results should be validated by molecular experiments and the working hypotheses of the mechanistic model should be tested. The model system for these molecular experiments are commonly normal breast or breast cancer cell lines. This is because cell lines can easily be maintained and manipulated. Furthermore, they represent an unlimited source of cells and are generally well characterized (Hollestelle et al., 2010a). The advantage of breast cancer cell lines is that many are available with different characteristics, however, as with eQTL analysis, CNVs, somatic mutations and methylation may be confounding the results of the experiments. Furthermore, for studying the effects of germline variants in breast cancer predisposition and considering that these are likely early events in tumorigenesis, normal breast cell lines seem the obvious choice. Currently two normal breast cell lines have been used in post-GWAS analysis, MCF10A and Bre-80 (Darabi et al., 2015; Glubb et al., 2015; Dunning et al., 2016; Ghoussaini et al., 2016; Lawrenson et al., 2016; Betts et al., 2017; Helbig et al., 2017). Both normal breast cell lines are, however, ER-negative which may not be the best model system for studying candidate causal variants in iCHAVs that are only associated with ER-positive breast cancer. Because of tissue specificity the compromise would therefore be to at least use one normal breast cancer cell line and two breast cancer cell lines, one ER-positive and one ER-negative.

Chip Assays and EMSA

In order to validate the *in silico* predictions of regulatory functions, such as TF binding to a candidate causal SNP or PRE, but also its allele-specific binding, two different techniques can be used. The first is a chromatin immunoprecipitation (ChIP) assay in which antibodies are used to enrich DNA fragments bound by one specific protein. The ChIP is subsequently followed by either sequencing, a qPCR or an allele-specific PCR to identify where a particular TF binds and whether this is allele-specific (Collas, 2010). The second is an electrophoretic mobility shift assay (EMSA) in which a protein or protein extract is mixed with a particular DNA fragment and incubated to allow binding. This mixture is subsequently separated by gel electrophoresis and compared to the length of the probe without protein. When protein binds to the DNA fragment, this results in an upward shift of the gel band. Although this does not provide any clue about the proteins involved in binding the DNA fragment, this assay can be adapted to a super shift assay by adding antibodies against TFs of interest to the protein-DNA mixtures (Hellman and Fried, 2007).

The advantage of ChIP assays is that they produce reliable results for assessing allele-specific binding of TF, in contrast to EMSAs. However, ChIP assays are relatively expensive and the resolution for determining the binding site is low (Edwards et al., 2013). In the post-GWAS analysis at 6q25.1 by Dunning et al. both EMSAs and ChIP assays were performed (**Table 1**). In this study, a total of five iCHAVs were identified containing 26 candidate causal variants using fine-scale mapping. *In silico* analyses showed that 19 of these candidate causal variants were located in DHSSs. Then, using EMSAs, 11 of these 19 variants were shown to alter the binding affinity of TFs *in vitro*. In the end, the TF identity for four of these candidate causal variants could be established and they appeared to be GATA3, CTCF, and MYC. With ChIP, the authors then confirmed GATA3 binding to iCHAV3 SNP rs851982. Moreover, CTCF binding was enriched at the common allele of iCHAV4 rs1361024, suggesting allele-specific binding of CTCF at this locus (Dunning et al., 2016).

3C and ChIA-PET

To validate *in silico* predictions of chromatin interactions or to confirm results from eQTL studies, molecular experiments such as chromatin confirmation capture (3C) can be performed. Using 3C, loci that are physically associated through chromatin loops are ligated together and these ligation products can subsequently be quantified using qPCR (Dekker et al., 2002). In addition, the ligation products can also be sequenced. This way, allele-specific chromatin interactions can be identified. For validating specific chromatin interactions, 3C is a very suitable technique as shown by its wide use in post-GWAS studies (**Table 1**). However, there are of course also some disadvantages to 3C. One of these is that the background is high at short distances between the two interacting loci. Consequently the two loci under evaluation should be further than 10 kb apart (Monteiro and Freedman, 2013). For instance, in the post-GWAS study at the 19p13 region by Lawrenson et al., only five from the 13 candidate causal variants could be evaluated due to the close proximity of these variants to their target gene, *ANKLE1* (Lawrenson et al., 2016). Usually, this however does not present a problem, since three

quarters of distal PREs influences a gene that is not the nearest one (Sanyal et al., 2012).

Another technique that is important to mention in this respect is chromatin-interaction analysis by paired-end tag sequencing (ChIA-PET). This is an adaptation of the original 3C technique allowing the detection of chromatin interactions bound by a specific protein, using an antibody (Fullwood et al., 2009). Usually, ChIA-PET experiments are not specifically performed for each separate post-GWAS study. Because the data is genome-wide, it is usually mined from databases containing interactomes for the most common TFs and histone marks such as ER, CTCF, RNA polymerase II and H3K4Me2. As with the publically available data from cistromes, as discussed earlier, having ChIA-PET data from more cell types and more TFs will improve upon the value of these data for the research community.

Luciferase Reporter Assays and CRISPR/Cas9 Genome Editing

By now, having compiled all *in silico* data and data from molecular experiments, a working hypothesis should be established of how the candidate causal variants confer breast cancer risk. This model includes which candidate causal variant via what TF can modulate gene expression of that particular gene via chromatin interaction. The last step is then usually to conduct luciferase reporter assays in order to confirm this hypothesis and assess what impact the candidate causal variants have on the promoter of that target gene, either enhancing or repressive.

In luciferase reporter assays, PREs are cloned into a reporter construct that expresses the *luciferase* cDNA when the promoter of interest is activated (Gould and Subramani, 1988; Williams et al., 1989; Fan and Wood, 2007). It is common to first establish a baseline for luciferase expression from the wild-type PREs. After that, PREs containing the risk allele or risk haplotype for one or more candidate causal variants are assessed, usually per PRE or per iCHAV. Depending on the levels of luciferase expression after introduction of the risk allele(s), an enhancing or repressive effect can be determined. Moreover, by varying the size of the PREs in subsequent experiments the boundaries of the PRE can be better defined. As discussed before, again the choice of cell type is also relevant here as well as the choice of promoter to use.

For most of the post-GWAS breast cancer risk loci, luciferase reporter assays were performed to confirm the working hypothesis for the functional model (**Table 1**) (Meyer et al., 2008; Beesley et al., 2011; Cai et al., 2011b; Bojesen et al., 2013; French et al., 2013; Ghoussaini et al., 2014, 2016; Darabi et al., 2015; Orr et al., 2015; Dunning et al., 2016; Lawrenson et al., 2016; Betts et al., 2017; Helbig et al., 2017; Michailidou et al., 2017). However, at the 2q35 locus in the study by Ghoussaini *et al.*, the PRE did not influence *IGFBP5* expression despite positive 3C and eQTL results (Ghoussaini et al., 2014). Similarly, at 5q12, the risk allele of a candidate causal variant had no effect on expression of predicted target genes *FGF10* and *MRPS30* (Ghoussaini et al., 2016).

An alternative method to study the effects of a (candidate causal variant in a) PRE is the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/CRISPR associated (Cas)9

gene editing system, which was first discovered in bacteria (Wiedenheft et al., 2012). Using CRISPR/Cas9 it has now become possible to, reliably and efficiently, introduce precise mutations in the human genome (Jinek et al., 2012). This gene editing technique makes use of a guide RNA (gRNA) that is complementary to the genomic region to be edited and a Cas9 enzyme that is guided by the gRNA to generate a double strand break (DSB) at this genomic region. The generated DSB can subsequently be repaired by either the non-homologous end joining pathway, which generally produces random insertions or deletions or by the homologous recombination repair pathway when a homology arm with the mutation of interest is co-transfected into the cells (Salsman and Dellaire, 2017). The latter pathway is able to generate specifically targeted mutations. At the 19p13.1 breast cancer locus this technique was used to generate a 57 base pair deletion containing the candidate causal SNP rs56069439. Lawrenson et al. showed a reduced *ANKLE1*, but not *ABHD8* or *BABAM1* expression as a result of this deletion (Lawrenson et al., 2016). A modified version of the Cas9 enzyme was used in the post-GWAS study by Betts et al. to silence PRE1 at 11q13, resulting in reduced *CUPID1*, *CUPID2* and *CCND1* expression (Betts et al., 2017). This nuclease-deficient Cas9 (dCas9) enzyme binds the target genomic region, but does not cleave the DNA. By fusion of dCas9 to various effector domains, CRISPR/Cas9 can be modified to a gene silencing or activation tool (Dominguez et al., 2016).

Interestingly, an average PRE has been predicted to regulate two or three different target genes (Sanyal et al., 2012). From the post-GWAS studies to date, evidence has now been presented for this at only 4 out of the 22 GWAS-identified breast cancer risk loci: 6q25.1, 10q21, 11q13, 19p13.1 (French et al., 2013; Darabi et al., 2015; Dunning et al., 2016; Lawrenson et al., 2016; Betts et al., 2017), which might suggest that maybe not all target genes have been identified yet at every locus investigated so far. Also considering the GWAS-identified breast cancer risk loci for which no post-GWAS analysis has been performed yet, there is still much work ahead.

Although the majority of the post-GWAS studies have followed this general pipeline for elucidating the functional mechanisms, one important step is still missing. Namely, evaluating of the tumorgenicity of the causal variants and the target genes in *in vitro* and *in-vivo* model systems, such as normal breast cancer cells or mice. Discovery of the genome-editing technique CRISPR/Cas9 has greatly enhanced our capabilities for taking this next step. Not only, because of the precision of this gene editing tool, but also because it allows for simultaneous genome-edits (Cho et al., 2013). However, there are certainly some challenges on this path and simply showing that the target gene is tumorigenic in an *in vitro* or *in vivo* model system is not sufficient, as it does not tie the germline variant to breast tumorgenicity. More subtle gene editing is necessary, and the question remains, whether this will always give a phenotype, since cancer risks conferred by these germline variants is low. This will probably be one of the biggest issues besides choosing the appropriate model system or animal.

DISCUSSION

In addition to the more than 170 GWAS-identified loci associated with breast cancer risk, 22 of these loci have been studied in more detail by post-GWAS analysis (**Table 1**). So far, the functional mechanism that candidate causal variants seem to make use of are mainly on the transcriptional level and deregulating target genes. In addition, the target genes involved do not seem to be specifically involved in DNA damage repair, like for high- and moderate-penetrant breast cancer risk genes, instead, somatic breast cancer drivers also appear to be enriched (Michailidou et al., 2017). Furthermore, the mechanisms that these causal variants use to confer breast cancer risk, are probably more complex than we anticipated, with often several iCHAVs at a GWAS-identified locus and some of them being able to regulate multiple target genes or ncRNAs (**Table 1**). Although we are not even half way this challenge, the availability of data on regulatory features, chromatin interactions and gene expression as well as the development of bioinformatics tools is definitely accelerating the process. However, in the future we could still benefit from more cistrome and interactome data on more TFs and on different cell types, especially normal breast cells. To facilitate more effective fine-scale mapping, more and larger case-control studies from African ancestry are necessary to benefit from the more structured LD in this population. Finally, we could also benefit from more paired genotype and gene expression data from normal breast samples for eQTL analysis as well as a variety of different normal breast epithelial cell-type models.

Regarding the GWAS-identified loci itself, it is obvious that more lower-risk variants predisposing to breast cancer risk still exist (Michailidou et al., 2017), however, again, larger sample sizes, especially for ER-negative breast cancer, as well as new statistical models to asses GWAS SNPs tagging causal variants with lower allele frequencies and smaller effect sizes are necessary (Fachal and Dunning, 2015). Interestingly, at the same time researchers are making use of alternative methods to identify novel breast cancer risk loci, which are mostly based on the same regulatory features that are also involved in exerting their biological function. Some of these features are gene expression, methylation and TF binding (Shenker et al., 2013; Xu et al., 2013; Anjum et al., 2014; Severi et al., 2014; van Veldhoven et al., 2015; Ambatipudi et al., 2017; Hoffman et al., 2017; Liu et al., 2017; Wu et al., 2018). In fact, the risk allele at 4q21 identified by Hamdi et al. was not discovered from GWAS, but from mapping SNPs associated with allele-specific gene expression in cancer-related pathway genes. The SNPs which were discovered in one dataset then act as proxies for allele specific expression and were evaluated for association with breast cancer risk in a second large GWAS study. Because the number of SNPs evaluated is reduced significantly as compared with GWAS, these type of analyses have more power and could thus identify lower risk alleles (Hamdi et al., 2016). These studies are called transcriptome-, epigenome- and phenome-wide association studies (TWAS, EWAS, and PheWAS) for gene expression features, methylation features and phenotypic features respectively. Interestingly, in the largest breast cancer TWAS to date, the expression levels of 48 genes were shown to be associated with breast cancer risk, of

which 14 were novel and 34 were associated with known loci. However, 23 of these 34 genes were not previously identified as targets of GWAS-identified risk loci (Wu et al., 2018). This demonstrates that these types of studies are capable of identifying novel breast cancer risk loci, as well as validating previous GWAS-identified loci. EWASs, however, have not yet been very successful in identifying breast cancer risk loci associated with epigenetic changes, which is most likely a result of small sample sizes in these studies (Johansson and Flanagan, 2017). Finally, a recent PheWAS on multiple cancers, including breast cancer, has shown that using trait-specific PRS instead of single variants leads to improvement of the trait prediction power (Fritsche et al., 2018). In addition to these approaches, pathway-based analyses created to identify SNP-SNP interactions also open new avenues for identifying novel breast cancer risk SNPs and their interactors (Wang et al., 2017).

In this review, we have discussed the findings and lessons learned from post-GWAS analyses of 22 GWAS-identified risk loci. Identifying the true causal variants underlying breast cancer susceptibility provides better estimates of the explained familial relative risk thereby improving polygenetic risk scores (PRSs).

Further stratification of their risk and contribution according the different subtypes of breast cancer and different populations will, however, be necessary. Moreover, unraveling the function of the causal variants involved in susceptibility to breast cancer increases our understanding of the biological mechanisms responsible for causing susceptibility to breast cancer, which will facilitate the identification of further breast cancer risk alleles and the development of preventive medicine for those women at risk for developing the disease.

AUTHOR CONTRIBUTIONS

MR and AH designed the article and all authors wrote the article and approved of the final manuscript.

ACKNOWLEDGMENTS

REFERENCES

Ahmadiyeh, N., Pomerantz, M. M., Grisanzio, C., Herman, P., Jia, L., Almendro, V., et al. (2010). 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9742–9746. doi: 10.1073/pnas.0910668107

Ahmed, S., Thomas, G., Ghoussaini, M., Healey, C. S., Humphreys, M. K., Platte, R., et al. (2009). Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat. Genet.* 41, 585–590. doi: 10.1038/ng.354

Ambatipudi, S., Horvath, S., Perrier, F., Cuenin, C., Hernandez-Vargas, H., Le Calvez-Kelm, F., et al. (2017). DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility. *Eur. J. Cancer* 75, 299–307. doi: 10.1016/j.ejca.2017.01.014

Anjum, S., Fourkala, E. O., Zikan, M., Wong, A., Gentry-Maharaj, A., Jones, A., et al. (2014). A BRCA1-mutation associated DNA methylation signature in blood cells predicts sporadic breast cancer incidence and survival. *Genome Med.* 6:47. doi: 10.1186/gm567

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Battle, A., Brown, C. D., Engelhardt, B. E., and Montgomery, S. B. (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204–213. doi: 10.1038/nature24277

Beesley, J., Pickett, H. A., Johnatty, S. E., Dunning, A. M., Chen, X., Li, J., et al. (2011). Functional polymorphisms in the TERT promoter are associated with risk of serous epithelial ovarian and breast cancers. *PLoS ONE* 6:e24987. doi: 10.1371/journal.pone.0024987

Betts, J. A., Moradi Marjaneh, M., Al-Ejeh, F., Lim, Y. C., Shi, W., Sivakumaran, H., et al. (2017). Long Noncoding RNAs CUPID1 and CUPID2 Mediate Breast Cancer Risk at 11q13 by Modulating the Response to DNA Damage. *Am. J. Hum. Genet.* 101, 255–266. doi: 10.1016/j.ajhg.2017.07.007

Bojesen, S. E., Pooley, K. A., Johnatty, S. E., Beesley, J., Michailidou, K., Tyrer, J. P., et al. (2013). Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat. Genet.* 45, 371–384, 384e371–372. doi: 10.1038/ng.2566

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1797. doi: 10.1101/gr.137323.112

Cai, Q., Long, J., Lu, W., Qu, S., Wen, W., Kang, D., et al. (2011a). Genome-wide association study identifies breast cancer risk variant at 10q21.2: results

from the Asia Breast Cancer Consortium. *Hum. Mol. Genet.* 20, 4991–4999. doi: 10.1093/hmg/ddr405

Cai, Q., Wen, W., Qu, S., Li, G., Egan, K. M., Chen, K., et al. (2011b). Replication and functional genomic analyses of the breast cancer susceptibility locus at 6q25.1 generalize its importance in women of chinese, Japanese, and European ancestry. *Cancer Res.* 71, 1344–1355. doi: 10.1158/0008-5472.CAN-10-2733

Cai, Q., Zhang, B., Sung, H., Low, S. K., Kweon, S. S., Lu, W., et al. (2014). Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. *Nat. Genet.* 46, 886–890. doi: 10.1038/ng.3041

Collaborative Group on Hormonal Factors in Breast Cancer (2001). Familial breast cancer: collaborative reanalsis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet* 358, 1389–1399. doi: 10.1016/S0140-6736(01)06524-2

Chatterjee, N., Chen, Y. H., Luo, S., and Carroll, R. J. (2009). Analysis of case-control association studies: SNPs, imputation and haplotypes. *Stat. Sci.* 24, 489–502. doi: 10.1214/09-STS297

Cheung, V. G., and Spielman, R. S. (2009). Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat. Rev. Genet.* 10, 595–604. doi: 10.1038/nrg2630

Cho, S. W., Kim, S., Kim, J. M., and Kim, J. S. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* 31, 230–232. doi: 10.1038/nbt.2507

Coetzee, S. G., Rhie, S. K., Berman, B. P., Coetzee, G. A., and Noushmehr, H. (2012). FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Res.* 40:e139. doi: 10.1093/nar/gks542

Collas, P. (2010). The current state of chromatin immunoprecipitation. *Mol. Biotechnol.* 45, 87–100. doi: 10.1007/s12033-009-9239-8

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Couch, F. J., Kuchenbaecker, K. B., Michailidou, K., Mendoza-Fandino, G. A., Nord, S., Lilyquist, J., et al. (2016). Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. *Nat. Commun.* 7:11375. doi: 10.1038/ncomms11375

Cowper-Sal lari, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., et al. (2012). Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* 44, 1191–1198. doi: 10.1038/ng.2416

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346–352. doi: 10.1038/nature10983

Darabi, H., Beesley, J., Droit, A., Kar, S., Nord, S., Moradi Marjaneh, M., et al. (2016). Fine scale mapping of the 17q22 breast cancer locus using dense SNPs, genotyped within the Collaborative Oncological Gene-Environment Study (COGs). *Sci. Rep.* 6:32512. doi: 10.1038/srep32512

Darabi, H., McCue, K., Beesley, J., Michailidou, K., Nord, S., Kar, S., et al. (2015). Polymorphisms in a putative enhancer at the 10q21.2 breast cancer risk locus regulate NRBF2 expression. *Am. J. Hum. Genet.* 97, 22–34. doi: 10.1016/j.ajhg.2015.05.002

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science* 295, 1306–1311. doi: 10.1126/science.1067799

Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., et al. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325, 1246–1250. doi: 10.1126/science.1174148

Dite, G. S., MacInnis, R. J., Bickerstaffe, A., Dowty, J. G., Allman, R., Apicella, C., et al. (2016). Breast cancer risk prediction using clinical models and 77 independent risk-associated SNPs for women aged under 50 years: australian breast cancer family registry. *Cancer Epidemiol. Biomarkers Prev.* 25, 359–365. doi: 10.1158/1055-9965.EPI-15-0838

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., et al. (2012). Landscape of transcription in human cells. *Nature* 489, 101–108. doi: 10.1038/nature11233

Dominguez, A. A., Lim, W. A., and Qi, L. S. (2016). Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. *Nat. Rev. Mol. Cell Biol.* 17, 5–15. doi: 10.1038/nrm.2015.2

Dunning, A. M., Michailidou, K., Kuchenbaecker, K. B., Thompson, D., French, J. D., Beesley, J., et al. (2016). Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nat. Genet.* 48, 374–386. doi: 10.1038/ng.3521

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D., Thompson, D., Ballinger, D. G., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 447, 1087–1093. doi: 10.1038/nature05887

Edwards, S. L., Beesley, J., French, J. D., and Dunning, A. M. (2013). Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.* 93, 779–797. doi: 10.1016/j.ajhg.2013.10.012

Fachal, L., and Dunning, A. M. (2015). From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. *Curr. Opin. Genet. Dev.* 30, 32–41. doi: 10.1016/j.gde.2015.01.004

Fan, F., and Wood, K. V. (2007). Bioluminescent assays for high-throughput screening. *Assay Drug Dev. Technol.* 5, 127–136. doi: 10.1089/adt.2006.053

Fitzmaurice, C., Allen, C., Barber, R. M., Barregard, L., Bhutta, Z. A., Brenner, H., et al. (2017). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. *JAMA Oncol.* 3, 524–548. doi: 10.1001/jamaoncol.2016.5688

Fletcher, O., Johnson, N., Orr, N., Hosking, F. J., Gibson, L. J., Walker, K., et al. (2011). Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. *J. Natl. Cancer Inst.* 103, 425–435. doi: 10.1093/jnci/djq563

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

Fritsche, L. G., Gruber, S. B., Wu, Z., Schmidt, E. M., Zawistowski, M., Moser, S. E., et al. (2018). Association of polygenic risk scores for multiple cancers in a phenome-wide study: results from the michigan genomics initiative. *Am. J. Hum. Genet.* 102, 1048–1061. doi: 10.1016/j.ajhg.2018.04.001

Fu, J., Wolfs, M. G., Deelen, P., Westra, H. J., Fehrmann, R. S., Te Meerman, G. J., et al. (2012). Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.* 8:e1002431. doi: 10.1371/journal.pgen.1002431

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 462, 58–64. doi: 10.1038/nature08497

Garcia-Closas, M., Couch, F. J., Lindstrom, S., Michailidou, K., Schmidt, M. K., Brook, M. N., et al. (2013). Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat. Genet.* 45, 392–398. doi: 10.1038/ng.2561

Ge, B., Pokholok, D. K., Kwan, T., Grundberg, E., Morcos, L., Verlaan, D. J., et al. (2009). Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.* 41, 1216–1222. doi: 10.1038/ng.473

Ghoussaini, M., Edwards, S. L., Michailidou, K., Nord, S., Cowper-Sal Lari, R., Desai, K., et al. (2014). Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. *Nat. Commun.* 4:4999. doi: 10.1038/ncomms5999

Ghoussaini, M., Fletcher, O., Michailidou, K., Turnbull, C., Schmidt, M. K., Dicks, E., et al. (2012). Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nat. Genet.* 44, 312–318. doi: 10.1038/ng.1049

Ghoussaini, M., French, J. D., Michailidou, K., Nord, S., Beesley, J., Canisus, S., et al. (2016). Evidence that the 5p12 Variant rs10941679 confers susceptibility to estrogen-receptor-positive breast cancer through FGF10 and MRPS30 regulation. *Am. J. Hum. Genet.* 99, 903–911. doi: 10.1016/j.ajhg.2016.07.017

Glubb, D. M., Maranian, M. J., Michailidou, K., Pooley, K. A., Meyer, K. B., Kar, S., et al. (2015). Fine-scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating MAP3K1. *Am. J. Hum. Genet.* 96, 5–20. doi: 10.1016/j.ajhg.2014.11.009

Gold, B., Kirchhoff, T., Stefanov, S., Lautenberger, J., Viale, A., Garber, J., et al. (2008). Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc. Natl. Acad. Sci. U.S.A.* 105, 4340–4345. doi: 10.1073/pnas.0800441105

Gould, S. J., and Subramani, S. (1988). Firefly luciferase as a tool in molecular and cell biology. *Anal. Biochem.* 175, 5–13. doi: 10.1016/0003-2697(88)90353-3

Guo, X., Long, J., Zeng, C., Michailidou, K., Ghoussaini, M., Bolla, M. K., et al. (2015). Fine-scale mapping of the 4q24 locus identifies two independent loci associated with breast cancer risk. *Cancer Epidemiol. Biomarkers Prev.* 24, 1680–1691. doi: 10.1158/1055-9965.EPI-15-0363

Haiman, C. A., Chen, G. K., Vachon, C. M., Canzian, F., Dunning, A., Millikan, R. C., et al. (2011). A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat. Genet.* 43, 1210–1214. doi: 10.1038/ng.985

Hamdi, Y., Soucy, P., Adoue, V., Michailidou, K., Canisius, S., Lemaçon, A., et al. (2016). Association of breast cancer risk with genetic variants showing differential allelic expression: Identification of a novel breast cancer susceptibility locus at 4q21. *Oncotarget* 7, 80140–80163. doi: 10.18632/oncotarget.12818

Han, M. R., Long, J., Choi, J. Y., Low, S. K., Kweon, S. S., Zheng, Y., et al. (2016). Genome-wide association study in East Asians identifies two novel breast cancer susceptibility loci. *Hum. Mol. Genet.* 25, 3361–3371. doi: 10.1093/hmg/ddw164

Helbig, S., Wockner, L., Bouendeu, A., Hille-Betz, U., McCue, K., French, J. D., et al. (2017). Functional dissection of breast cancer risk-associated TERT promoter variants. *Oncotarget* 8, 67203–67217. doi: 10.18632/oncotarget.18226

Hellman, L. M., and Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat. Protoc.* 2, 1849–1861. doi: 10.1038/nprot.2007.249

Hoffman, J. D., Graff, R. E., Emami, N. C., Tai, C. G., Passarelli, M. N., Hu, D., et al. (2017). Cis-eQTL-based trans-ethnic meta-analysis reveals novel genes associated with breast cancer risk. *PLoS Genet.* 13:e1006690. doi: 10.1371/journal.pgen.1006690

Hollestelle, A., Nagel, J. H., Smid, M., Lam, S., Elstrodt, F., Wasielewski, M., et al. (2010a). Distinct gene mutation profiles among luminal-type and basal-type breast cancer cell lines. *Breast Cancer Res. Treat.* 121, 53–64. doi: 10.1007/s10549-009-0460-8

Hollestelle, A., Wasielewski, M., Martens, J. W., and Schutte, M. (2010b). Discovering moderate-risk breast cancer susceptibility genes. *Curr. Opin. Genet. Dev.* 20, 268–276. doi: 10.1016/j.gde.2010.02.009

Horne, H. N., Chung, C. C., Zhang, H., Yu, K., Prokunina-Olsson, L., Michailidou, K., et al. (2016). Fine-mapping of the 1p11.2 breast cancer susceptibility locus. *PLoS ONE* 11:e0160316. doi: 10.1371/journal.pone.0160316

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., et al. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* 39, 870–874. doi: 10.1038/ng2075

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–821. doi: 10.1126/science.1225829

Johansson, A., and Flanagan, J. M. (2017). Epigenome-wide association studies for breast cancer risk and risk factors. *Trends Cancer Res.* 12, 19–28.

Kim, H. C., Lee, J. Y., Sung, H., Choi, J. Y., Park, S. K., Lee, K. M., et al. (2012). A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. *Breast Cancer Res.* 14:R56. doi: 10.1186/bcr3158

Kuchenbaecker, K. B., McGuffog, L., Barrowdale, D., Lee, A., Soucy, P., Dennis, J., et al. (2017). Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. *J. Natl. Cancer Inst.* 109:djw302. doi: 10.1093/jnci/djw302

Kumar, V., Westra, H. J., Karjalainen, J., Zhernakova, D. V., Esko, T., Hrdlickova, B., et al. (2013). Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet.* 9:e1003201. doi: 10.1371/journal.pgen.1003201

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. doi: 10.1038/nature14248

Lalonde, E., Ha, K. C., Wang, Z., Bemmo, A., Kleinman, C. L., Kwan, T., et al. (2011). RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res.* 21, 545–554. doi: 10.1101/gr.111211.110

Lawrenson, K., Kar, S., McCue, K., Kuchenbaeker, K., Michailidou, K., Tyrer, J., et al. (2016). Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast-ovarian cancer susceptibility locus. *Nat. Commun.* 7:12675. doi: 10.1038/ncomms12675

Li, M. J., Wang, L. Y., Xia, Z., Sham, P. C., and Wang, J. (2013a). GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nucleic Acids Res.* 41, W150–W158. doi: 10.1093/nar/gkt456

Li, Q., Seo, J. H., Stranger, B., McKenna, A., Pe'er, I., Laframboise, T., et al. (2013b). Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* 152, 633–641. doi: 10.1016/j.cell.2012.12.034

Lilyquist, J., Ruddy, K. J., Vachon, C. M., and Couch, F. J. (2018). Common genetic variation and breast cancer risk-past, present, and future. *Cancer Epidemiol. Biomarkers Prev.* 27, 380–394. doi: 10.1158/1055-9965.EPI-17-1144

Lin, W. Y., Camp, N. J., Ghoussaini, M., Beesley, J., Michailidou, K., Hopper, J. L., et al. (2015). Identification and characterization of novel associations in the CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk. *Hum. Mol. Genet.* 24, 285–298. doi: 10.1093/hmg/ddu431

Liu, Y., Walavalkar, N. M., Dozmorov, M. G., Rich, S. S., Civelek, M., and Guertin, M. J. (2017). Identification of breast cancer associated variants that modulate transcription factor binding. *PLoS Genet.* 13:e1006761. doi: 10.1371/journal.pgen.1006761

Long, J., Cai, Q., Sung, H., Shi, J., Zhang, B., Choi, J. Y., et al. (2012). Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer. *PLoS Genet.* 8:e1002532. doi: 10.1371/journal.pgen.1002532

Mavaddat, N., Pharoah, P. D., Michailidou, K., Tyrer, J., Brook, M. N., Bolla, M. K., et al. (2015). Prediction of breast cancer risk based on profiling with common genetic variants. *J. Natl. Cancer Inst.* 107:djv036. doi: 10.1093/jnci/djv036

Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., et al. (2002). Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat. Genet.* 31, 55–59. doi: 10.1038/ng879

Meyer, K. B., Maia, A. T., O'Reilly, M., Teschendorff, A. E., Chin, S. F., Caldas, C., et al. (2008). Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS Biol.* 6:e108. doi: 10.1371/journal.pbio.0060108

Meyer, K. B., O'Reilly, M., Michailidou, K., Carlebur, S., Edwards, S. L., French, J. D., et al. (2013). Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1. *Am. J. Hum. Genet.* 93, 1046–1060. doi: 10.1016/j.ajhg.2013.10.026

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., et al. (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* 47, 373–380. doi: 10.1038/ng.3242

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., et al. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* 45, 353–361, 361e351–352. doi: 10.1038/ng.2563

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature* 551, 92–94. doi: 10.1038/nature24284

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 266, 66–71. doi: 10.1126/science.7545954

Milne, R. L., Kuchenbaecker, K. B., Michailidou, K., Beesley, J., Kar, S., Lindström, S., et al. (2017). Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat. Genet.* 49, 1767–1778. doi: 10.1038/ng.3785

Monteiro, A. N., and Freedman, M. L. (2013). Lessons from postgenome-wide association studies: functional analysis of cancer predisposition loci. *J. Intern. Med.* 274, 414–424. doi: 10.1111/joim.12085

Muranen, T. A., Greco, D., Blomqvist, C., Aittomäki, K., Khan, S., Hogervorst, F., et al. (2017). Genetic modifiers of CHEK2*1100delC-associated breast cancer risk. *Genet. Med.* 19, 599–603. doi: 10.1038/gim.2016.147

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489, 83–90. doi: 10.1038/nature11212

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

Orr, N., Dudbridge, F., Dryden, N., Maguire, S., Novo, D., Perrakis, E., et al. (2015). Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. *Hum. Mol. Genet.* 24, 2966–2984. doi: 10.1093/hmg/ddv035

Pai, A. A., Cain, C. E., Mizrahi-Man, O., De Leon, S., Lewellen, N., Veyrieras, J. B., et al. (2012). The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS Genet.* 8:e1003000. doi: 10.1371/journal.pgen.1003000

Purrington, K. S., Slager, S., Eccles, D., Yannoukakos, D., Fasching, P. A., Miron, P., et al. (2014). Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer. *Carcinogenesis* 35, 1012–1019. doi: 10.1093/carcin/bgt404

Quigley, D. A., Fiorito, E., Nord, S., Van Loo, P., Alnæs, G. G., Fleischer, T., et al. (2014). The 5p12 breast cancer susceptibility locus affects MRPS30 expression in estrogen-receptor positive tumors. *Mol. Oncol.* 8, 273–284. doi: 10.1016/j.molonc.2013.11.008

Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., et al. (2007). PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.* 39, 165–167. doi: 10.1038/ng1959

Renwick, A., Thompson, D., Seal, S., Kelly, P., Chagtai, T., Ahmed, M., et al. (2006). ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat. Genet.* 38, 873–875. doi: 10.1038/ng1837

Salsman, J., and Dellaire, G. (2017). Precision genome editing in the CRISPR era. *Biochem. Cell Biol.* 95, 187–201. doi: 10.1139/bcb-2016-0137

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature* 489, 109–113. doi: 10.1038/nature11279

Severi, G., Southey, M. C., English, D. R., Jung, C. H., Lonie, A., McLean, C., et al. (2014). Epigenome-wide methylation in DNA from peripheral blood as a marker of risk for breast cancer. *Breast Cancer Res. Treat.* 148, 665–673. doi: 10.1007/s10549-014-3209-y

Shenker, N. S., Polidoro, S., van Veldhoven, K., Sacerdote, C., Ricceri, F., Birrell, M. A., et al. (2013). Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. *Hum. Mol. Genet.* 22, 843–851. doi: 10.1093/hmg/dds488

Shi, J., Zhang, Y., Zheng, W., Michailidou, K., Ghoussaini, M., Bolla, M. K., et al. (2016). Fine-scale mapping of 8q24 locus identifies multiple independent risk variants for breast cancer. *Int. J. Cancer* 139, 1303–1317. doi: 10.1002/ijc.30150

Siddiq, A., Couch, F. J., Chen, G. K., Lindström, S., Eccles, D., Millikan, R. C., et al. (2012). A meta-analysis of genome-wide association studies of breast cancer

identifies two novel susceptibility loci at 6q14 and 20q11. *Hum. Mol. Genet.* 21, 5373–5384. doi: 10.1093/hmg/dds381

Spain, S. L., and Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Hum. Mol. Genet.* 24, R111–119. doi: 10.1093/hmg/ddv260

Spencer, C., Hechter, E., Vukcevic, D., and Donnelly, P. (2011). Quantifying the underestimation of relative risks from genome-wide association studies. *PLoS Genet.* 7:e1001337. doi: 10.1371/journal.pgen.1001337

Stacey, S. N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S. A., et al. (2007). Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.* 39, 865–869. doi: 10.1038/ng2064

Stacey, S. N., Manolescu, A., Sulem, P., Thorlacius, S., Gudjonsson, S. A., Jonsson, G. F., et al. (2008). Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.* 40, 703–706. doi: 10.1038/ng.131

Stacey, S. N., Sulem, P., Zanon, C., Gudjonsson, S. A., Thorleifsson, G., Helgason, A., et al. (2010). Ancestry-shift refinement mapping of the C6orf97-ESR1 breast cancer susceptibility locus. *PLoS Genet.* 6:e1001029. doi: 10.1371/journal.pgen.1001029

Steffen, J., Nowakowska, D., Niwinska, A., Czapczak, D., Kluska, A., Piatkowska, M., et al. (2006). Germline mutations 657del5 of the NBS1 gene contribute significantly to the incidence of breast cancer in Central Poland. *Int. J. Cancer* 119, 472–475. doi: 10.1002/ijc.21853

Stratton, M. R., and Rahman, N. (2008). The emerging landscape of breast cancer susceptibility. *Nat. Genet.* 40, 17–22. doi: 10.1038/ng.2007.53

Sun, Y., Ye, C., Guo, X., Wen, W., Long, J., Gao, Y. T., et al. (2016). Evaluation of potential regulatory function of breast cancer risk locus at 6q25.1. *Carcinogenesis* 37, 163–168. doi: 10.1093/carcin/bgv170

Thomas, G., Jacobs, K. B., Kraft, P., Yeager, M., Wacholder, S., Cox, D. G., et al. (2009). A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat. Genet.* 41, 579–584. doi: 10.1038/ng.353

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., et al. (2012). The accessible chromatin landscape of the human genome. *Nature* 489, 75–82. doi: 10.1038/nature11232

Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M., et al. (2010). Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* 42, 504–507. doi: 10.1038/ng.586

Udler, M. S., Ahmed, S., Healey, C. S., Meyer, K., Struewing, J., Maranian, M., et al. (2010a). Fine scale mapping of the breast cancer 16q12 locus. *Hum. Mol. Genet.* 19, 2507–2515. doi: 10.1093/hmg/ddq122

Udler, M. S., Meyer, K. B., Pooley, K. A., Karlins, E., Struewing, J. P., Zhang, J., et al. (2009). FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. *Hum. Mol. Genet.* 18, 1692–1703. doi: 10.1093/hmg/ddp078

Udler, M. S., Tyrer, J., and Easton, D. F. (2010b). Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. *Genet. Epidemiol.* 34, 463–468. doi: 10.1002/gepi.20504

Vahteristo, P., Bartkova, J., Eerola, H., Syrjäkoski, K., Ojala, S., Kilpivaara, O., et al. (2002). A CHEK2 genetic variant contributing to a substantial fraction of familial breast cancer. *Am. J. Hum. Genet.* 71, 432–438. doi: 10.1086/341943

van Veldhoven, K., Polidoro, S., Baglietto, L., Severi, G., Sacerdote, C., Panico, S., et al. (2015). Epigenome-wide association study reveals decreased average methylation levels years before breast cancer diagnosis. *Clin. Epigenetics* 7:67. doi: 10.1186/s13148-015-0104-2

Wang, W., Xu, Z. Z., Costanzo, M., Boone, C., Lange, C. A., and Myers, C. L. (2017). Pathway-based discovery of genetic interactions in breast cancer. *PLoS Genet.* 13:e1006973. doi: 10.1371/journal.pgen.1006973

Ward, L. D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40, D930–D934. doi: 10.1093/nar/gkr917

Wiedenheft, B., Sternberg, S. H., and Doudna, J. A. (2012). RNA-guided genetic silencing systems in bacteria and archaea. *Nature* 482, 331–338. doi: 10.1038/nature10886

Williams, T. M., Burlein, J. E., Ogden, S., Kricka, L. J., and Kant, J. A. (1989). Advantages of firefly luciferase as a reporter gene: application to the interleukin-2 gene promoter. *Anal. Biochem.* 176, 28–32. doi: 10.1016/0003-2697(89)90267-4

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., et al. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature* 378, 789–792. doi: 10.1038/378789a0

Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., et al. (2018). A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.* doi: 10.1038/s41588-018-0132-x. [Epub ahead of print].

Wyszynski, A., Hong, C. C., Lam, K., Michailidou, K., Lytle, C., Yao, S., et al. (2016). An intergenic risk locus containing an enhancer deletion in 2q35 modulates breast cancer risk by deregulating IGFBP5 expression. *Hum. Mol. Genet.* 25, 3863–3876. doi: 10.1093/hmg/ddw223

Xu, Z., Bolick, S. C., DeRoo, L. A., Weinberg, C. R., Sandler, D. P., and Taylor, J. A. (2013). Epigenome-wide association study of breast cancer using prospectively collected sister study samples. *J. Natl. Cancer Inst.* 105, 694–700. doi: 10.1093/jnci/djt045

Yardimci, G. G., and Noble, W. S. (2017). Software tools for visualizing Hi-C data. *Genome Biol.* 18:26. doi: 10.1186/s13059-017-1161-y

Zeng, C., Guo, X., Long, J., Kuchenbaecker, K. B., Droit, A., Michailidou, K., et al. (2016). Identification of independent association signals and putative functional variants for breast cancer risk through fine-scale mapping of the 12p11 locus. *Breast Cancer Res.* 18, 64. doi: 10.1186/s13058-016-0718-0

Zheng-Bradley, X., and Flicek, P. (2017). Applications of the 1000 genomes project resources. *Brief. Funct. Genomics* 16, 163–170. doi: 10.1093/bfgp/elw027

Zheng, W., Long, J., Gao, Y. T., Li, C., Zheng, Y., Xiang, Y. B., et al. (2009). Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.* 41, 324–328. doi: 10.1038/ng.318

Zhou, X., Li, D., Zhang, B., Lowdon, R. F., Rockweiler, N. B., Sears, R. L., et al. (2015). Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nat. Biotechnol.* 33, 345–346. doi: 10.1038/nbt.3158

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

GEMO, a National Resource to Study Genetic Modifiers of Breast and Ovarian Cancer Risk in *BRCA1* and *BRCA2* Pathogenic Variant Carriers

Fabienne Lesueur[1]*, *Noura Mebirouk*[1], *Yue Jiao*[2], *Laure Barjhoux*[3], *Muriel Belotti*[2], *Maïté Laurent*[2], *Mélanie Léone*[4], *Claude Houdayer*[2], *Brigitte Bressac-de Paillerets*[5], *Dominique Vaur*[6], *Hagay Sobol*[7], *Catherine Noguès*[7], *Michel Longy*[8], *Isabelle Mortemousque*[9], *Sandra Fert-Ferrer*[10], *Emmanuelle Mouret-Fourme*[2], *Pascal Pujol*[11], *Laurence Venat-Bouvet*[12], *Yves-Jean Bignon*[13], *Dominique Leroux*[14], *Isabelle Coupier*[11], *Pascaline Berthet*[6], *Véronique Mari*[15], *Capucine Delnatte*[16], *Paul Gesta*[17], *Marie-Agnès Collonge-Rame*[18], *Sophie Giraud*[4], *Valérie Bonadona*[19,20], *Amandine Baurand*[21], *Laurence Faivre*[21], *Bruno Buecher*[2], *Christine Lasset*[19,20], *Marion Gauthier-Villars*[2], *Francesca Damiola*[3], *Sylvie Mazoyer*[22], *Sandrine M. Caputo*[2], *Nadine Andrieu*[1], *Dominique Stoppa-Lyonnet*[2,23] and GEMO Study Collaborators*

[1] INSERM, U900, Institut Curie, PSL Research University, Mines ParisTech, Paris, France, [2] Service de Génétique, Institut Curie, Paris, France, [3] Biopathologie, Centre Léon Bérard, Lyon, France, [4] Hospices Civils de Lyon, Groupement Hospitalier EST, Bron, France, [5] Gustave Roussy, Université Paris-Saclay, Département de Biopathologie et INSERM U1186, Villejuif, France, [6] Département de Biopathologie, Centre François Baclesse, Caen, France, [7] Institut Paoli Calmette, Département d'Anticipation et de Suivi des Cancers, Oncogénétique, Faculté de Médecine, Université d'Aix-Marseille, Marseille, France, [8] Biopathologie, Institut Bergonié, Bordeaux, France, [9] Service de Génétique, Hôpital Bretonneau, Tours, France, [10] Service de Génétique, Centre Hospitalier de Chambéry, Chambéry, France, [11] Service de Génétique Médicale et Oncogénétique, Hôpital Arnaud de Villeneuve, CHU Montpellier, INSERM 896, CRCM Val d'Aurelle, Montpellier, France, [12] Service d'Oncologie Médicale, Hôpital Universitaire Dupuytren, Limoges, France, [13] Université Clermont Auvergne, INSERM, U1240, Centre Jean Perrin, Clermont-Ferrand, France, [14] Département de Génétique, CHU de Grenoble, Hôpital Couple-Enfant, Grenoble, France, [15] Unité d'Oncogénétique, Centre Antoine Lacassagne, Nice, France, [16] Unité d'Oncogénétique, Centre René Gauducheau, Nantes, France, [17] Service d'Oncogénétique Régional Poitou-Charentes, Niort, France, [18] Service Génétique et Biologie du Développement-Histologie, CHU Hôpital Saint-Jacques, Besançon, France, [19] Université Claude Bernard Lyon 1, Villeurbanne, France, [20] CNRS UMR 5558; Unité de Prévention et Epidémiologie Génétique, Centre Léon Bérard, Lyon, France, [21] Institut GIMI, CHU de Dijon et Centre de Lutte contre le Cancer Georges François Leclerc, Dijon, France, [22] INSERM, U1028, CNRS, UMR5292, Centre de Recherche en Neurosciences de Lyon, Lyon, France, [23] INSERM, U830, Université Paris Descartes, Paris, France

INTRODUCTION

Women carrying a pathogenic variant (PV) in the *BRCA1* or *BRCA2* (*BRCA1/2*) genes are at high lifetime risk of developing breast cancer (BC) and ovarian cancer (OC), but estimation of the cumulative risk of cancer to age 70 years varies substantially between studies and populations. Initial estimations were obtained from selected high-risk families with multiple cases, such as those ascertained through the Breast Cancer Linkage Consortium used to identify disease loci (1). In the first retrospective studies conducted on such families, estimates for BC ranged from 40 to 87% for *BRCA1* PV carriers and from 27 to 84% for *BRCA2* PV carriers and estimates for OC ranged from 16 to 68% for *BRCA1* PV carriers and from 11 to 27% for *BRCA2* PV carriers (1–4). Recently, the largest prospective cohort conducted to date reported cumulative risks of BC to age 80 years of 72% for *BRCA1* PV carriers and 69% for *BRCA2* PV carriers (5). In the same study, cumulative risks of OC to age 80 years were 44% for *BRCA1* PV carriers and 17% for *BRCA2* PV carriers. Variation in cancer risks within or between BRCA1/2 families,

with respect to age at diagnosis or type of cancer, can be explained by other genetic factors and/or lifestyle and reproductive factors (6–10). Genome-wide association studies (GWAS) conducted by the Breast Cancer Association Consortium (BCAC) have identified 172 common single-nucleotide polymorphisms (SNPs) associated with small increases in breast and/or ovarian cancer risk in the general population (11). A subset of these SNPs modifies the risk of breast and ovarian cancer risk for *BRCA1/2* PV carriers (12–14) but most of the variability has not been explained yet (15). Breast and ovarian cancer risks in *BRCA1/2* PV carriers might also vary according to the location of the variant and/or its origin (14, 16–19).

Genetic testing for *BRCA1* and *BRCA2* has been part of genetic counseling in European Union countries and North America since their discovery in the 90's, and has greatly improved recommendations about clinical management options and the most appropriate treatments. Nonetheless, both retrospective and prospective studies on large datasets of *BRCA1/2* PV carrier families are still very much needed to refine individual cancer risk estimates by considering other genetic and lifestyle/environmental factors, and they will also contribute to a better understanding of the correlation between mutant *BRCA1/2* alleles and phenotype. In particular, accurate age-specific risk estimates for the different types of cancer would be useful when choosing risk reduction strategies such as prophylactic bilateral mastectomy or salphingo-oophorectomy.

The Genetic Modifiers of *BRCA1* and *BRCA2* (GEMO) Group is the French multidisciplinary, collaborative framework for the investigation of genetic factors modifying cancer risk in Hereditary Breast and Ovarian cancer (HBOC) families segregating *BRCA1/2* PVs. Its primary aims are to contribute to large-scale national and international projects to identify genetic modifiers and to facilitate the translation of research results to the clinical setting. This is achieved by establishing a resource of blood DNA samples from individuals carrying a PV together with family and clinical data through the nation-wide network of cancer genetic clinics. Here we report on the progress of the GEMO study, the characteristics of the 5,303 actual participants and the prevalence and spectrum of *BRCA1/2* cancer-associated variants identified so far.

PARTICIPANTS AND METHODS

Organization of Cancer Predisposition Testing in France

GEMO investigators include molecular geneticists, clinicians, genetic counselors, and epidemiologists who are involved in the Genetic and Cancer Group (GGC), a consortium with support of UNICANCER whose objectives are to define optimal testing practices both in terms of genetic counseling and laboratory

---

Abbreviations: BC, Breast Cancer; BCAC, Breast Cancer Association Consortium; BRCA1, BReast CAncer 1; BRCA2, BReast CAncer 2; CRF, Case Report Form; CIMBA, Consortium of Investigators of Modifiers of BRCA1/2; GWAS, Genome-Wide Association Study; HBOC, Hereditary Breast and Ovarian Cancer; OC, Ovarian Cancer; PRS, Polygenic Risk Score; SNP, Single Nucleotide Polymorphism, VUS, Variant of Uncertain Clinical Significance.

techniques, and to contribute to the estimation of individual's cancer risks (http://www.unicancer.fr/en/cancer-and-genetic-group). GGC has contributed to the national development of *BRCA1/2* screening tests and genetic consultations and, therefore improved management of subjects at high-risk of cancer.

Currently, there are 145 cancer genetic counseling units and 17 laboratories performing *BRCA1/2* testing (or panel testing of multiple cancer susceptibility genes) in France (see **Supplementary Data** for methods used by laboratories for PV identification).

Eligibility criteria for *BRCA1/2* testing according to the current national clinical guidelines are (i) at least 3 first or second degree relatives affected with breast or ovarian cancer in the same family branch, (ii) 2 first-degree relatives with BC, one of them having been diagnosed before age 41, or one before age 51 and the other before age 71, (iii) 2 first-degree relatives with BC, one of them being a male, (iv) 1 BC case before age 36, or before age 51 if triple negative tumor, (v) 1 case with bilateral BC, the first one before age 50, (vi) 1 male BC, (vii) 1 OC before age 71, or at any age if high-grade serous OC.

By 2016, 17,821 probands (i.e., the first individual tested in the family) were tested for *BRCA1/2*, and 1,670 (9.4%) were found to carry a PV. A similar number of probands carried a variant of uncertain clinical significance (VUS). A total of 6,417 relatives (essentially first-degree relatives of probands) underwent targeted screening tests and about 39% of them were found to carry the PV identified in the proband (http://www.e-cancer.fr).

Ascertainment of GEMO Participants

GEMO participants are from HBOC families ascertained prospectively through family cancer clinics and tested positive for a confirmed PV in *BRCA1/2*. The GEMO study was initiated in 2006 and is still ongoing. Initially, only female PV carriers aged 18 or older, affected or unaffected with cancer were invited to participate in the study by geneticists. Adult male PV carriers have been invited to participate since 2013. Today, GEMO involves 32 clinics and the 17 diagnostics laboratories from the GGC.

Protocol, Data Collection, and Database

The GEMO coordinating center was located at Centre Léon Bérard (Lyon) until September 2015 and is currently held at Institut Curie (Paris). All data and biospecimens are stored without personal identifiers. The GEMO case report form (CRF) includes information on participants' family history, gyneco-obstetrics risk factors (age at menarche, number of pregnancies, age at menopause), preventive surgery and tumor pathology (histology, grade, tumor size, hormone receptors status). Data on socio-demographic variables (age at inclusion, sex, ethnicity/population ancestry) and medical history of cancer (laterality, other cancer prior recruitment into study) are also collected.

Geneticists invite *BRCA1/2* PV carriers, whether affected with cancer or not, to participate in GEMO during the consultation informing them of their *BRCA1/2* positive test results. After completing the CRF with the participant, the geneticist sends it to

the coordinating center, and requests that an aliquot of the blood DNA sample (at least 10 µg) that was used for genetic testing is shipped from the testing laboratory to the coordinating center. The study protocol is illustrated in **Supplementary Figure 1**.

Recently, an upgraded electronic database on FileMaker Pro 16 (FileMaker Inc., Santa Clara, California, USA) was developed to collate, manage and distribute core data and DNA samples, and to facilitate inter-operability with the GGC BRCA1/2 (ex-UMD-BRCA1/BRCA2) database (20) and that of the prospective cohort on *BRCA1/2* PV carriers GENEPSO (21).

Ethics

The study is performed in compliance with the Helsinki Declaration and received a favorable review of the French National Committees for personal data protection in medical research (CCTIRS N°07223 and CNIL agreement N°1245228). GEMO has human ethics approval at all the participating institutions where subjects are recruited. All research projects making use of data and/or materials collected by GEMO are required to have independent ethical approval from their host institutions. Participants give written informed consent during genetic counseling sessions and understand that as a result of participation, personal details will be recorded and stored in a coded format on a database. They consent to samples of DNA material prepared from blood cells being stored in a central location and to de-identified information and samples being made available for scientifically and ethically approved research projects. Informed consent agreements signed by participants are kept in the clinics.

Access to DNA Samples and to Family and Clinical Data

Investigators wishing to use the GEMO DNA collection and related clinical and family data submit a brief expression of interest to principal investigators (gemo@curie.fr) who then circulate the proposal to the GEMO steering committee with a 10-day opportunity given to highlight any major issues, especially duplication of, or complementarity to, existing projects. If favorably reviewed, a full application is then submitted and verified to ensure that sufficient resources to conduct the project exist, the amount of DNA requested is appropriate, and that the proposal has any required ethics approvals. When the project is accepted, a material transfer agreement and/or a data transfer agreement are signed between the coordinating center and the research institution of the applicant. DNA samples along with related data are sent to the applicants who commit to providing annual progress reports. To further enrich the GEMO resource, applicants are required to supply their research data to GEMO after publication, and/or 12 months after completion of their projects.

Variant Classification

The description of the genetic variants follows recommendations proposed by the Human Genome Variation Society (22). Variants are denoted using the cDNA reference sequences NM_007294.3 (*BRCA1*) and NM_000059.3 (*BRCA2*). Only carriers of a clear *BRCA1/2* PV are included in GEMO. PVs are defined as variants considered as pathogenic by the GGC (20), the Evidence-based Network for the Interpretation of Germline Mutant Alleles consortium (23), the Consortium of Investigators of Modifiers of *BRCA1/2* (CIMBA) (24) and/or published variants classified as pathogenic using multifactorial likelihood approaches (25, 26).

RESULTS

Collection of DNA Samples and Data

As of April 2018, 5,303 participants with available DNA sample had been enrolled in GEMO. Participants included 3,087 *BRCA1* PV carriers (2,877 women and 210 men) and 2,216 *BRCA2* PV carriers (2,005 women and 211 men) belonging to 2,190 and 1,544 families, respectively. The mean number of participants per family was 1.4 (range: 1–11). For 600 families, DNA samples were collected from three or more family members. While no individuals in the dataset carried more than a single PV, four families segregated two PV in two branches of the family (family 1: *BRCA1*:c.5137del and *BRCA2*:c.2808_2811del; family 2: *BRCA1*:c.1480C>T and *BRCA1*:c.3839_3843delinsAGGC; family 3: *BRCA1*:c.3841C>T and *BRCA2*:c.4889C>G; family 4: *BRCA1*:c.4391_4393delinsTT and *BRCA2*:c.7680dup).

Participants' Characteristics

At inclusion, 56.3% of *BRCA1* female PV carriers were diagnosed with BC (mean age at diagnosis: 41.3, range 22–81), 18.3% were diagnosed with OC or fallopian tube cancer (mean age at diagnosis: 51.9, range 16–92) and 33.2% were free of these cancers (mean age at inclusion: 40.5, range 18–101). With respect to *BRCA2*, 61.1% of female PV carriers had BC (mean age at diagnosis: 43.6, range 21–90), 10.1% had OC or fallopian tube cancer (mean age at diagnosis: 57.9, range 31–99) and 32.9% were free of these cancers (mean age at inclusion: 42.1, range 19–91). Among the 421 male participants, 2.9% of *BRCA1* PV carriers and 6.2% of *BRCA2* PV carriers were diagnosed with prostate cancer at inclusion (mean age at diagnosis for *BRCA1*: 61.5, range 48–71 and 64.1, range 50–78 for *BRCA2*). Ten percent of males carrying a *BRCA2* PV had BC (mean age at diagnosis: 58.8, range 44–77) vs. none in male *BRCA1* PV carriers. Detailed characteristics of participants (probands and relatives) according to their cancer status are shown in **Table 1**. Parity, age at menarche and age at menopause (natural or artificial) for female PV carriers are shown in **Supplementary Table 1**. Female participants reported an average number of live births of 1.7 and a mean age at menarche of 12.9 years. No difference in parity or age at menarche was observed between women affected and unaffected with cancer, and no differences were observed between probands and relatives. Mean age at menopause (natural or artificial) was 45.7 and 47.8 years in *BRCA1* and *BRCA2* PV carriers, respectively. Information on prophylactic mastectomy or salphingo-oophorectomy is not systematically recorded in GEMO. However, based on available data, we identified 600 out of 4,882 female participants (12.3%) who had had bilateral or unilateral mastectomy. For 50 of them mastectomy was prophylactic as they had not developed BC at inclusion (1.0%). Among the 1,496 women (30.6%) who had had bilateral oophorectomy at inclusion, 1,005 (20.5%) had not

TABLE 1 | Characterisitics of the GEMO subjects.

	BRCA1 families						BRCA2 families					
	Probands			Relatives			Probands			Relatives		
	N	Mean age at inclusion, y (range)	Mean age of onset for first cancer, y (range)	N	Mean age at inclusion, y (range)	Mean age of onset for first cancer (range)	N	Mean age at inclusion, y (range)	Mean age of onset for first cancer (range)	N	Mean age at inclusion, y (range)	Mean age of onset for first cancer (range)
INDIVIDUALS AFFECTED WITH CANCER												
Women (breast only)	1,154	46.8 (19–92)	40.3 (22–81)	240	47.5 (18–77)	41.1 (25–71)	949	49.4 (21–86)	42.5 (21–77)	193	52.4 (21–87)	45.3 (25–76)
Women (ovarian/fallopian tube[a])	445	57.2 (33–80)	51.8 (25–86)	81	55.8 (22–72)	51.9 (16–92)	177	62.0 (37–84)	57.6 (31–84)	26	67.3 (44–100)	60.2 (41–99)
Women (breast and ovarian/fallopian tube)	186	59.9 (37–80)	46.0 (26–71)	38	57.4 (39–70)	46.0 (31–71)	72	63.2 (50–84)	49.1 (25–76)	11	71.6 (46–100)	61.3 (36–90)
Men (breast only)	0	–	–	0	–	–	16	62.2 (49–79)	58.4 (47–77)	2	70.0 (69–71)	63.5 (63–64)
Men (prostate only)	0	–	–	6	68.7 (58–78)	61.5 (48–71)	3	64.0 (57–77)	56.5 (50–63)	7	69.6 (61–78)	64.8 (53–78)
Men (breast and prostate)	0	–	–	0	–	–	2	68.0 (67–69)	60.0 (60–60)	1	76.0	53.0
Women (other cancer[b])	4	46.5 (33–58)	41.8 (31–56)	3	34.7 (24–43)	32.0 (21–39)	6	57.3 (44–69)	50.7 (34–64)	5	55.2 (45–69)	43.6 (30–68)
Men (other cancer[b])	5	66.8 (49–80)	62.0 (45–72)	6	71.0 (64–79)	65.4 (62–70)	3	53.0 (40–73)	51.7 (39–71)	4	56.8 (45–73)	50.0 (28–65)
INDIVIDUALS UNAFFECTED WITH CANCER												
Women	500	39.9 (18–83)	–	449	38.6 (18–86)	–	334	42.3 (18–80)	–	314	39.8 (18–91)	–
Men	78	47.6 (21–72)	–	117	51.6 (23–89)	–	52	49.4 (19–80)	–	122	48.2 (19–79)	–

[a]May have had also breast cancer.
[b]Other than breast, ovary/fallopian tube, prostate cancer or basal cell carcinoma.

FIGURE 1 | Distribution and occurrence of pathogenic variants along *BRCA1* **(A)** and *BRCA2* **(B)** in probands in the GEMO study. BRCA1 domains are: RING domain, Coiled Coil domains, BRCT (BRCA1 C-terminal) domains (14, 27). BRCA2 domains are: BRC repeats, helical domain (HD), OB fold binding domains, tower alpha (14, 28). Breast cancer risk regions: LR1, lower risk region in *BRCA1*, LR2 lower risk region in *BRCA2*, HR2, higher risk region in *BRCA2* (29); BCCR, breast cancer cluster region (14); Ovarian cancer risk regions: OCCR, ovarian cancer cluster region (14).

developed OC or fallopian tube cancer and this surgery was likely prophylactic.

Only 26.9% of participants self-reported their population ancestry/ethnicity. Among them, 90.8% were European, 3.5% were African, 0.3% were Asian and 4.1% were of other or mixed origin. Ashkenazi Jewish (AJ) ancestry was reported by 1.3% of participants.

BRCA1 and *BRCA2* Variants

Currently, 506 *BRCA1* and 494 *BRCA2* unique PVs are described in the GEMO database. The number of families in which each PV was observed is shown in **Supplementary Table 2** and the distribution of PVs across the gene sequences is shown in **Figure 1**. The five most common PVs accounted for 21.3% of all PVs in *BRCA1* and 14.9% of all *BRCA2* PVs. The most common *BRCA1* PVs were c.5266dup (7.5%) and c.68_69del (3.9%), originally described as founder PVs in the AJ population [30], the c.3481_3491del founder PV from North-Eastern France (4.9%) [31, 32], and the two common European PVs c.4327C>T (2.7%) and c.3839_3843delinsAGGC (2.2%) [33]. The most common *BRCA2* PVs were c.2808_2811del (3.3%), c.5946del (3.2%), a Western European PV of AJ origin [34], c.4889C>G (2.2%), c.8364G>A (2.1%), c.5645C>A (1.9%), and c.7680dup (1.9%). There were 267 *BRCA1* PVs and 265 *BRCA2* PVs observed only once in GEMO.

Representativeness of the GEMO Population

The GGC database was designed to compile information on all *BRCA1/2* variants (pathogenic, neutral and VUS), except common polymorphisms, identified probands in the 17 French licensed laboratories [20]. This database is therefore considered as the reference database for *BRCA1/2* variants in France. In June 2018, it contained PV from 6,385 BRCA1 and 4,839 BRCA2 families (Sandrine Caputo, personal communication), and about one third of the population recorded in the GGC database had been enrolled in GEMO. The distribution of PVs along the genes sequence in GEMO and the GGC *BRCA1/2* database overlaps (**Supplementary Table 2**), although a few variants were under-represented in GEMO reflecting a recruitment bias in the study due to the absence of participating cancer clinics in some regions (e.g., *BRCA1*:c.5260G>T is identified mostly in families from Western France). Other differences can be attributable to a different dynamics between the GEMO and the national registry (some PVs observed in GEMO had not been yet recorded in the GGC database).

DISCUSSION

Over 5,300 participants have been enrolled in GEMO to date, which provides an overview of *BRCA1*/2 PVs in a well-characterized sample of French counseled HBOC families. The GEMO resource is available to internal and external researchers who can apply for blood DNA and data for use in ethically approved, peer reviewed collaborative and interdisciplinary projects on the genetic epidemiology of cancer in BRCA1/2 families. Its overall goal is to facilitate the translation of research results to the clinical setting.

As an example, GEMO contributes massively to the CIMBA effort involving centers on six continents that have recruited *BRCA1/2* PV carriers with associated clinical, risk factors, and genetic data [24]. GEMO is one of the three most important contributors to CIMBA projects in terms of number of samples, phenotypic and pathology data. In total, 2,868 subjects (53.9% of the GEMO population) had been genotyped using the iCOGS and/or the Oncoarray chips in the context of large-scale GWAS [35, 36]. In brief, these international initiatives led to the identification of 26 and 16 SNPs associated with BC risk for *BRCA1* and *BRCA2* PV carriers respectively, and the corresponding numbers for OC risk are 11 and 13 [15]. The combined effect of these SNPs, modeled as Polygenic Risk Scores (PRS) is currently being investigated to improve individualized cancer risk predictions. Other goals of the Consortium are to precise age-specific cancer risk estimates considering position and functional effects of the PV, family history of cancer and genetic and lifestyle/hormonal modifier risk factors in order to integrate findings on SNPs into the genetic counseling process. GEMO study collaborators co-authored 43 CIMBA publications. Publications and summary results for iCOGS SNPs are accessible via http://cimba.ccge.medschl.cam.ac.uk/.

At the national level, the GEMO group is aiming to develop specific PRS in the French counseled families in order to assess the clinical utility of incorporating such scores in risk prediction models. Indeed, improvement in the performance of such models for risk stratification and personalized decision-making (e.g., prophylactic mastectomy/salphingo-oophorectomy or frequency of BC screening) has important clinical implications. Efforts are also being made to render the GEMO database interoperable with other national databases including that of GENEPSO, which is a prospective cohort initiated in 1999, where *BRCA1*/2 PV carriers are followed over time to observed prospectively characteristics of subjects who develop either primary or secondary cancers [5]. To date, about 1,400 individuals have been enrolled in both GEMO and GENEPSO.

Clinical management of healthy women with a *BRCA1/2* PV involves a combination of frequent screening, especially of the breasts, risk-reducing surgeries and possibly chemoprevention [37]. For these women, important decisions include whether or not to undergo preventive mastectomy and the age at which to undergo risk-reducing salphingo-oophorectomy. These choices are invasive, have substantial side effects, and are associated with adverse psychological effects [38]. It is therefore important to have precise estimates of associated age-specific cancer risks to provide optimal advices to women carrying a PV. Hence, women at particularly high risk or with a high risk of disease at early ages may benefit from early intervention, and women at lower risk may opt to delay surgery or chemoprevention.

AUTHOR CONTRIBUTIONS

FL, FD, SM, and DS-L coordinated the GEMO study. CN, MiL, IM, SF-F, EM-F, PP, LV-B, Y-JB, DL, IC, PB, VM, CD, PG, M-AC-R, SG, VB, LF, BB, CL, and MG-V invited GEMO participants. CH, BB-dP, DV, HS, NM, LB, MéL, FD, and FL managed the DNA samples. NM, LB, MB, FD, FL, MaL, and YJ managed family and clinical data. NM, LB, MéL, FD, YJ, SMC, and FL curated the

variants databases. FL and YJ analyzed the data. FL and NA wrote the paper. All authors read and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2018.00490/full#supplementary-material

Supplementary Table 1 | Age at menarche, age at menopause and parity for women affected and unaffected with cancer.

Supplementary Table 2 | Unique pathogenic variants and number of families in which each variant was observed. Variants are called using the cDNA reference sequences NM_007294.3 (*BRCA1*) and NM_000059.3 (*BRCA2*).

Supplementary Figure 1 | Protocol of the study.

Supplementary Data | Methods used for identification of *BRCA1/2* pathogenic variants.

REFERENCES

1. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, et al. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. *Breast Cancer Linkage Consort Am J Hum Genet.* (1998) 62:676–89. doi: 10.1086/301749
2. Hopper JL, Southey MC, Dite GS, Jolley DJ, Giles GG, Mccredie MR, et al. Population-based estimate of the average age-specific cumulative risk of breast cancer for a defined set of protein-truncating mutations in BRCA1 and BRCA2. *Aust Breast Cancer Family Study Cancer Epidemiol Biomarkers Prev.* (1999) 8:741–7.
3. Antoniou AC, Cunningham AP, Peto J, Evans DG, Lalloo F, Narod SA, et al. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *Br J Cancer* (2008a) 98:1457–66. doi: 10.1038/sj.bjc.6604305
4. Brohet RM, Velthuizen ME, Hogervorst FB, Meijers-Heijboer HE, Seynaeve C, Collee MJ, et al. Breast and ovarian cancer risks in a large series of clinically ascertained families with a high proportion of BRCA1 and BRCA2 Dutch founder mutations. *J Med Genet.* (2014) 51:98–107. doi: 10.1136/jmedgenet-2013-101974
5. Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips KA, Mooij TM, Roos-Blom MJ, et al. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA* (2017) 317:2402–16. doi: 10.1001/jama.2017.7112
6. Narod SA. Modifiers of risk of hereditary breast cancer. *Oncogene* (2006) 25:5832–6. doi: 10.1038/sj.onc.1209870

7. Antoniou AC, Spurdle AB, Sinilnikova OM, Healey S, Pooley KA, Schmutzler RK, et al. Common breast cancer-predisposition alleles are associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers. *Am J Hum Genet.* (2008) 82:937–48. doi: 10.1016/j.ajhg.2008.02.008

8. Milne RL, Antoniou AC. Genetic modifiers of cancer risk for BRCA1 and BRCA2 mutation carriers. *Ann Oncol.* (2011) 22(Suppl. 1):i11–7. doi: 10.1093/annonc/mdq660

9. Kotsopoulos J, Lubinski J, Lynch HT, Kim-Sing C, Neuhausen S, Demsky R, et al. Oophorectomy after menopause and the risk of breast cancer in BRCA1 and BRCA2 mutation carriers. *Cancer Epidemiol Biomarkers Prev.* (2012) 21:1089–96. doi: 10.1158/1055-9965.EPI-12-0201

10. Friebel TM, Domchek SM, Rebbeck TR. Modifiers of cancer risk in BRCA1 and BRCA2 mutation carriers: systematic review and meta-analysis. *J Natl Cancer Inst.* (2014) 106:dju091. doi: 10.1093/jnci/dju091

11. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* (2017) 551:92–4. doi: 10.1038/nature24284

12. Couch FJ, Wang X, Mcguffog L, Lee A, Olswold C, Kuchenbaecker KB, et al. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genet.* (2013) 9:e1003212. doi: 10.1371/journal.pgen.1003212

13. Gaudet MM, Kuchenbaecker KB, Vijai J, Klein RJ, Kirchhoff T, Mcguffog L, et al. Identification of a BRCA2-specific modifier locus at 6p24 related to breast cancer risk. *PLoS Genet.* (2013) 9:e1003173. doi: 10.1371/journal.pgen.1003173

14. Rebbeck TR, Mitra N, Wan F, Sinilnikova OM, Healey S, Mcguffog L, et al. Association of type and location of BRCA1 and BRCA2 mutations with risk of breast and ovarian cancer. *JAMA* (2015) 313:1347–61. doi: 10.1001/jama.2014.5985

15. Milne RL, Antoniou AC. Modifiers of breast and ovarian cancer risks for BRCA1 and BRCA2 mutation carriers. *Endocr Relat Cancer* (2016) 23:T69–84. doi: 10.1530/ERC-16-0277

16. Thompson D, Easton D, Breast Cancer Linkage C. Variation in cancer risks, by mutation position, in BRCA2 mutation carriers. *Am J Hum Genet.* (2001) 68:410–9. doi: 10.1086/318181

17. Thompson D, Easton D, Breast Cancer Linkage Consortium. Variation in BRCA1 cancer risks by mutation position. *Cancer Epidemiol Biomarkers Prev.* (2002) 11:329–36. Available online at: http://cebp.aacrjournals.org/content/11/4/329.long

18. Lecarpentier J, Nogues C, Mouret-Fourme E, Gauthier-Villars M, Lasset C, Fricker JP, et al. Variation in breast cancer risk associated with factors related to pregnancies according to truncating mutation location, in the French National BRCA1 and BRCA2 mutations carrier cohort (GENEPSO). *Breast Cancer Res.* (2012) 14:R99. doi: 10.1186/bcr3218

19. Lecarpentier J, Nogues C, Mouret-Fourme E, Buecher B, Gauthier-Villars M, Stoppa-Lyonnet D, et al. Breast cancer risk associated with estrogen exposure and truncating mutation location in BRCA1/2 carriers. *Cancer Epidemiol Biomarkers Prev.* (2015) 24:698–707. doi: 10.1158/1055-9965.EPI-14-0884

20. Caputo S, Benboudjema L, Sinilnikova O, Rouleau E, Beroud C, Lidereau R. Description and analysis of genetic variants in French hereditary breast and ovarian cancer families recorded in the UMD-BRCA1/BRCA2 databases. *Nucleic Acids Res.* (2012) 40:D992–1002. doi: 10.1093/nar/gkr1160

21. Andrieu N, Easton DF, Chang-Claude J, Rookus MA, Brohet R, Cardis E, et al. Effect of chest X-rays on the risk of breast cancer among BRCA1/2 mutation carriers in the international BRCA1/2 carrier cohort study: a report from the EMBRACE, GENEPSO, GEO-HEBON, and IBCCS Collaborators' Group. *J Clin Oncol.* (2006) 24:3361–6. doi: 10.1200/JCO.2005.03.3126

22. Den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, Mcgowan-Jordan J, et al. HGVS recommendations for the description of sequence variants: 2016 update. *Hum Mutat.* (2016) 37:564–9. doi: 10.1002/humu.22981

23. Spurdle AB, Healey S, Devereau A, Hogervorst FB, Monteiro AN, Nathanson KL, et al. ENIGMA–evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Hum Mutat.* (2012) 33:2–7. doi: 10.1002/humu.21628

24. Chenevix-Trench G, Milne RL, Antoniou AC, Couch FJ, Easton DF, Goldgar DE. An international initiative to identify genetic modifiers of cancer risk in BRCA1 and BRCA2 mutation carriers: the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA). *Breast Cancer Res.* (2007) 9:104. doi: 10.1186/bcr1670

25. Goldgar DE, Easton DF, Deffenbaugh AM, Monteiro AN, Tavtigian SV, Couch FJ, et al. Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. *Am J Hum Genet.* (2004) 75:535–44. doi: 10.1086/424388

26. Easton DF, Deffenbaugh AM, Pruss D, Frye C, Wenstrup RJ, Allen-Brady K, et al. A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. *Am J Hum Genet.* (2007) 81:873–83. doi: 10.1086/521032

27. Millot GA, Carvalho MA, Caputo SM, Vreeswijk MP, Brown MA, Webb M, et al. A guide for functional analysis of BRCA1 variants of uncertain significance. *Hum Mutat.* (2012) 33:1526–37. doi: 10.1002/humu.22150

28. Yang H, Jeffrey PD, Miller J, Kinnucan E, Sun Y, Thoma NH, et al. BRCA2 function in DNA binding and recombination from a BRCA2-DSS1-ssDNA structure. *Science* (2002) 297:1837–48. doi: 10.1126/science.297.5588.1837

29. Lecarpentier J, Noguès C, Mouret-Fourme E, Stoppa-Lyonnet D, Lasset C, Caron O, et al. Variation in breast cancer risk with mutation position, smoking, alcohol, and chest X-ray history, in the French National BRCA1/2 carrier cohort (GENEPSO). *Breast Cancer Res Treat.* (2011) 130:927–38. doi: 10.1007/s10549-011-1655-3

30. Hamel N, Feng BJ, Foretova L, Stoppa-Lyonnet D, Narod SA, Imyanitov E, et al. On the origin and diffusion of BRCA1 c.5266dupC (5382insC) in European populations. *Eur J Hum Genet.* (2011) 19:300–6. doi: 10.1038/ejhg.2010.203

31. Stoppa-Lyonnet D, Laurent-Puig P, Essioux L, Pages S, Ithier G, Ligot L, et al. BRCA1 sequence variations in 160 individuals referred to a breast/ovarian family cancer clinic. *Inst Curie Breast Cancer Group Am J Hum Genet.* (1997) 60:1021–30.

32. Muller D, Bonaiti-Pellie C, Abecassis J, Stoppa-Lyonnet D, Fricker JP. BRCA1 testing in breast and/or ovarian cancer families from northeastern France identifies two common mutations with a founder effect. *Fam Cancer* (2004) 3:15–20. doi: 10.1023/B:FAME.0000026819.44213.df

33. Rebbeck TR, Friebel TM, Friedman E, Hamann U, Huo D, Kwong A, et al. Mutational spectrum in a worldwide study of 29,700 families with BRCA1 or BRCA2 mutations. *Hum Mutat.* (2018) 39:593–620. doi: 10.1002/humu.23406

34. Hall MJ, Reid JE, Burbidge LA, Pruss D, Deffenbaugh AM, Frye C, et al. BRCA1 and BRCA2 mutations in women of different ethnicities undergoing testing for hereditary breast-ovarian cancer. *Cancer* (2009) 115:2222–33. doi: 10.1002/cncr.24200

35. Kuchenbaecker KB, Neuhausen SL, Robson M, Barrowdale D, Mcguffog L, Mulligan AM, et al. Associations of common breast cancer susceptibility alleles with risk of breast cancer subtypes in BRCA1 and BRCA2 mutation carriers. *Breast Cancer Res.* (2014) 16:3416. doi: 10.1186/s13058-014-0492-9

36. Milne RL, Kuchenbaecker KB, Michailidou K, Beesley J, Kar S, Lindstrom S, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat Genet.* (2017) 49:1767–78. doi: 10.1038/ng.3785

37. Evans DG, Graham J, O'connell S, Arnold S, Fitzsimmons D. Familial breast cancer: summary of updated NICE guidance. *BMJ* (2013) 346:f3829. doi: 10.1136/bmj.f3829

38. Eccles DM. Identification of personal risk of breast cancer: genetics. *Breast Cancer Res.* (2008) 10(Suppl. 4):S12. doi: 10.1186/bcr2172

Dealing With BRCA1/2 Unclassified Variants in a Cancer Genetics Clinic: Does Cosegregation Analysis Help?

Roberta Zuntini, Simona Ferrari, Elena Bonora, Francesco Buscherini, Benedetta Bertonazzi, Mina Grippa, Lea Godino, Sara Miccoli and Daniela Turchetti*

UO Genetica Medica, Azienda Ospedaliero-Universitaria di Bologna Policlinico S.Orsola-Malpighi and Centro di Ricerca sui Tumori Ereditari, Dipartimento di Scienze Mediche e Chirurgiche, Universitá di Bologna, Bologna, Italy

Background: Detection of variants of uncertain significance (VUSs) in BRCA1 and BRCA2 genes poses relevant challenges for counseling and managing patients. VUS carriers should be managed similarly to probands with no BRCA1/2 variants detected, and predictive genetic testing in relatives is discouraged. However, miscomprehension of VUSs is common and can lead to inaccurate risk perception and biased decisions about prophylactic surgery. Therefore, efforts are needed to improve VUS evaluation and communication at an individual level.

Aims: We aimed at investigating whether cosegregation analysis, integrated with a careful review of available functional data and *in silico* predictions, may improve VUSs interpretation and counseling in individual families.

Methods: Patients with Breast Cancer (BC) and/or Ovarian Cancer (OC) fulfilling established criteria were offered genetic counseling and BRCA1/2 testing; VUSs identified in index cases were checked in other relatives affected by BC/OC whenever possible. As an alternative, if BC/OC clustered only in one branch of the family, the parental origin of the VUS was investigated. Public prediction tools and databases were used to collect additional information on the variants analyzed.

Results: Out of 1045 patients undergoing BRCA1/2 testing in the period October 2011–April 2018, 66 (6.3%) carried class 3 VUSs. Cosegregation analysis was performed for 13 VUSs in 11 kindreds. Seven VUSs (53.8%) did not cosegregate with breast/ovarian cancer in the family, which provided evidence against their role in cancer clustering in those families. Among the 6 cosegregating VUSs, for two (BRCA1 c.5152+2T>G and BRCA2 c.7975A>G) additional evidence exists from databases and *in silico* tools supporting their pathogenicity, which reinforces the hypothesis they may have had a predisposing effect in respective families. For the remaining four VUSs (31%), cosegregation analysis failed to provide relevant information.

Conclusion: Our findings suggest that cosegregation analysis in a clinical context may be helpful to improve test result interpretation in the specific family and, therefore, should be offered whenever possible. Besides, obtaining and sharing cosegregation data helps gathering evidence that may eventually contribute to VUS classification.

Keywords: BRCA1, BRCA2, VUS, breast cancer, ovarian cancer, hereditary cancer

INTRODUCTION

In recent years, the increasing requests for BRCA testing have led to increased identification of patients carrying Variants of Unknown Significance (VUSs) in these genes. Several international consortia have been established with the aim of classifying VUSs; since functional assays for BRCA1 and 2, unlike other genes, are of limited availability and accuracy, classification mainly relies on multifactorial analysis, which requires that a large amount of data from multiple families is collected (Goldgar et al., 2004; Spurdle et al., 2012). This implies that a long time is frequently needed before a variant is conclusively classified. Therefore, in the Cancer Genetics Clinics, the detection of a VUS poses substantial challenges for counseling and managing patients. In fact, according to the widely adopted variants classification in 5 categories, class 3 VUS are those for which available evidence, if any, fails to significantly support either a pathogenic or a neutral significance (Plon et al., 2008; Lindor et al., 2012). For carriers of variants falling in this category, the same management as for probands with no BRCA variants detected is recommended, and predictive genetic testing in relatives is discouraged (Plon et al., 2008; Lindor et al., 2012). However, miscomprehension of VUS has been reported to be common among counselees and referring physicians (Richter et al., 2013) and several studies have consistently shown that risk perception is significantly greater in VUS carriers, if compared to patients with uninformative results, with a higher rate of prophylactic surgery undertaken or considered (Vos et al., 2011, 2012; Culver et al., 2013; Richter et al., 2013; Welsh et al., 2017).

Therefore, the ongoing international initiatives aimed at classifying VUSs should be paralleled by efforts to improve VUSs interpretation and communication at an individual level.

In particular, aim of this study was to investigate whether cosegregation analysis, integrated with a careful review of available functional data and *in silico* predictions, may improve VUSs interpretation and counseling in individual families.

PATIENTS AND METHODS

BRCA1 and BRCA2 Testing

The Cancer Genetics Clinic in Bologna is one of the four Hubs of a Hub-and-Spoke Network established in 2012 in the Emilia-Romagna region (Northern Italy) with the aim of identifying and managing women at familial risk of breast and ovarian cancer.

In patients fulfilling criteria for BRCA testing according to the regional protocol (Servizio Sanità Pubblica and Regione Emilia-Romagna, 2016), informed consent was collected and a venous blood sample drawn during a genetic counseling session.

Genomic DNA was extracted from peripheral blood leukocytes using standard techniques. Complete sequence analysis of BRCA1 and BRCA2 genes was performed through Next Generation Sequencing technology using ION Torrent[TM] Oncomine[TM] BRCA Research Assay (Life Technologies).

Manual libraries preparation was generated from 20 ng of DNA per sample according to the manufacturer's instructions with barcode incorporation. Templates for DNA libraries were prepared using the Ion Personal Genome Machine (PGM) Hi-Q View OT2 200 Kit (Life Technologies) on the Ion One Touch 2 according to the manufacturer's instructions. Sequencing of 24 samples multiplexed templates was performed using the Ion Torrent PGM on Ion 318 chips using the Ion PGM Hi-Q[TM] View Sequencing Kit (Life Technologies) according to the manufacturer's instructions. Data analysis was performed using Torrent Suite (5.6) applying Oncomine BRCA Research Germline workflow. Any variant (either pathogenic or of unknown significance) was confirmed by Sanger sequencing. Moreover, analysis of BRCA1 deletions/duplications was performed by Multiplex Ligation Probe Amplification (MLPA) using the P002 kit of MRC Holland (Amsterdam, Netherlands), and data were analyzed using Coffalyser.net software. Mutation nomenclature follows the general recommendations of the Human Genome Variation Society (HGVS): cDNA and protein numbering were based on the reference sequence ID NM_007294.3 and NM_000059.3, respectively.

Variant Classification

All variants were evaluated through the retrieval of information in the following public databases: UMD[1], BRCA Exchange[2], ARUP Scientific Resource for Research and Education: BRCA Database[3], ClinVar[4], LOVD IARC[5], LOVD3[6]. All databases were last accessed 17 May 2018.

In silico Predictions

Potential cryptic splice sites and exonic splicing enhancers were investigated through Human Splicing Finder[7], and ESEfinder 3.0[8].

The evaluation of conservation of BRCA1/2 amino acids and related probability of pathogenicity was assessed according to the multiple-sequence alignments available on the Align GVGD Website[9] (Tavtigian et al., 2006).

Retrieval of Functional Data

The retrieval of results from functional assays for the VUS considered, if any, was made by querying the databases LOVD3[6] and UMD[1], and the recent neXtProt Cancer variant portal[10] (Cusin et al., 2018).

---

[1]http://www.umd.be/
[2]http://brcaexchange.org/
[3]http://arup.utah.edu/database/BRCA/
[4]https://www.ncbi.nlm.nih.gov/clinvar/
[5]http://priors.hci.utah.edu
[6]https://databases.lovd.nl/shared/genes
[7]http://www.umd.be/HSF3/index.html
[8]http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi
[9]http://agvgd.iarc.fr/
[10]https://www.nextprot.org/portals/breast-cancer

Cosegregation Analysis

The search for VUSs was extended to other relatives affected by BC or OC if they were available and consenting. As an alternative, when BC/OC clustering was observed in only one branch of the family, parental origin of the variant was defined by testing one or both parents, depending on their availability and willing. Quantitative cosegregation analysis was performed through the "Analyze my variant" website (Ranola and Shirts, 2017), which uses three different statistical methods: full-likelihood method for Bayes factors (FLB) (Thompson et al., 2003), co-segregation likelihood ratios (CSLR) (Mohammadi et al., 2009) and meiosis counting method (Jarvik and Browning, 2016).

RESULTS

BRCA Test Results

From 1st October 2011 to 30th April 2018, 1045 index cases underwent BRCA testing at our center. Among those, 188 (18%) were found to carry pathogenic variants: 104 (55%) in BRCA1 and 84 (45%) in BRCA2. Among the remaining patients, 744 (71.2% of the total) had no variants detected, while 113 (10.8%) carried VUS. Among the VUS detected (96 in total), 33 are classified as class 2, 59 as class 3 and 4 as class 4 (**Tables 1**, **2**). Overall, a total of 66 probands (6.3% of those tested) carried class 3 VUS.

TABLE 1 | Class 3 ("uncertain") variants identified in the population under study.

BRCA1			BRCA2		
Nucleotide change	**Predicted effect on protein**	**Number of families carrying the variant**	**Nucleotide change**	**Predicted effect on protein**	**Number of families carrying the variant**
c.1397G > A	p.Arg466Gln	1	c.2755G > A	p.Glu919Lys	1
c.1912G > A	p.Glu638Lys	1	c.2944A > C	p.Ile982Leu	1
c.1934C > A	p.Ser645Tyr	1	c.3519T > G	p.Ile1173Met	1
c.3613G > A	p.Gly1205Arg	1	c.3749A > G	p.Glu1250Gly	1
c.3783A > T	p.Leu1261Phe	1	c.4291G > A	p.Ala1431Thr	1
c.3878C > T	p.Ala1293Val	1	c.4603G > T	p.Ala1535Ser	1
c.4013A > G	p.Lys1338Arg	1	c.476T > C	p.Val159Ala	1
c.4054G > A	p.Glu1352Lys	1	c.5200G > A	p.Glu1734Lys	1
c.4223A > G	p.Gln1408Arg	1	c.5390C > G	p.Ala1797Gly	1
c.441+5A > G		2	c.5498A > G	p.Asn1833Ser	1
c.457A > G	p.Ser153Gly	1	c.5702A > T	p.Glu1901Val	1
c.4777A > T	p.Ile1593Leu	1	c.5705A > C	p.Asp1902Ala	1
c.5509T > C	p.Trp1837Arg	4	c.5885T > C	p.Ile1962Thr	2
c.569C > T	p.Thr190Ile	1	c.599C > T	p.Thr200Ile	1
c.767G > A	p.Arg256Lys	1	c.6062A > G	p.His2021Arg	1
c.889A > C	p.Met297Leu	1	c.6290C > T	p.Thr2097Met	2
c.556T > G	p.Ser186Ala	1	c.7007+5G > A		1
c.1027_1028AA > TG	p.Asn343Cys	2	c.7534C > T	p.Leu2512Phe	1
c.2281G > C	p.Glu761Gln	1	c.7769C > T	p.Ser2590Phe	1
c.2589T > G	p.Val863 =	1	c.7786G > A	p.Gly2596Arg	1
c.3823A > G	p.Ile1275Val	1	c.8262T > G	p.His2754Gln	1
c.4895T > G	p.Val1632Gly	1	c.8351G > A	p.Arg2784Gln	1
			c.8386C > T	p.Pro2796Ser	1
			c.9006A > T	p.Glu3002Asp	1
			c.9458G > C	p.Gly3153Ala	1
			c.7756C > T	p.Leu2587Phe	1
			c.1244A > G	p.His415Arg	1
			c.1342C > T	p.Arg448Cys	1
			c.1550A > G	p.Asn517Ser	1
			c.1991G > A	p.Gly664Glu	1
			c.5386G > T	p.Asp1796Tyr	1
			c.8704G > A	p.Ala2902Thr	1
			c.9409A > T	p.Thr3137Ser	1
			c.9986A > G	p.Asn3329Ser	1
			c.9218A > C	p.Asp3073Ala	1
			c.6562A > G	p.Lys2188Glu	1
			c.1996A > G	p.Ile666Val	1

TABLE 2 | Class 2 ("likely not pathogenic") and class 4 ("likely pathogenic") variants identified in the population under study.

	BRCA1				BRCA2			
	Nucleotide change	Predicted effect on protein	Number of families carrying the variant		Nucleotide change	Predicted effect on protein	Number of families carrying the variant	
class 2	c.522A > G	p.Gln174 =	1		c.1247T > G	p.Ile416Ser	2	class 2
	c.1137T > G	p.Ile379Met	1		c.1810A > G	p.Lys604Glu	1	
	c.1974G > C	p.Met658Ile	1		c.267G > A	p.Pro89Pro	2	
	c.2522G > A	p.Arg841Gln	1		c.5635G > A	p.Glu1879Lys	4	
	c.2883C > T	p.Asn961 =	1		c.6322C > T	p.Arg2108Cys	1	
	c.213-8A > C		2		c.68-7T > A		5	
	c.81-14C > T		1		c.7601C > T	p.Ala2534Val	1	
class 4	c.670+1G > A		1		c.3693T > C	p.Thr1231 =	1	
	c.4485-2A > C		1		c.8010G > A	p.Ser2670 =	1	
	c.5017_5019delCAC	p.His1673del	15		c.8972G > A	p.Arg2991His	1	
					c.9104A > C	p.Tyr3035Ser	1	
					c.9227G > T	p.Gly3076Val	1	
					c.9242T > C	p.Val3081Ala	1	
					c.9586A > G	p.Lys3196Glu	1	
					c.927A > G	p.Ser309 =	1	
					c.1395A > C	p.Val465 =	1	
					c.1514T > C	p.Ile505Thr	1	
					c.1820A > C	p.Lys607Thr	1	
					c.2817C > T	p.Thr939 =	1	
					c.4584C > T	p.Ser1528 =	2	
					c.6513G > T	p.Val2171 =	3	
					c.6927C > T	p.Ser2309 =	1	
					c.9285C > T	p.Asp3095 =	1	
					c.9396A > G	p.Lys3132 =	1	
					c.10121C > T	p.Thr3374Ile	1	
					c.4584A > G	p.Glu1518 =	1	
					c.8009C > T	p.Ser2670Leu	1	class 4

Cosegregation Analysis

Segregation of 13 VUS was assessed in 11 families (three families carried two VUS, while one VUS was detected in two unrelated kindreds). All the families were of Italian ancestry. Details of the variants, including current classification, *in silico* predictions and cosegregation analysis results are reported in **Table 3**. For 7 variants, cosegregation with the disease in the family was excluded; accordingly, cosegregation ratios in these families ranged from 0.0036 to 0.145.

Families Description

Pedigree 281-O-15 (**Figure 1**) The proband, a woman aged 50 at the time of counseling, had developed triple-negative breast cancer under the age of 40. Her mother had had surgery for high-grade ovarian cancer in her 50 s and the maternal grandmother was reported with possible ovarian cancer, as well. No significant history of cancer was reported in the father's side of the family. Genetic testing performed in another center had detected two **BRCA1** variants in the proband: **c.2522G>A** and **c.5152+2T>G**. When she came to our center for a second opinion, we proposed to check the presence of the two variants in the mother: she was found to carry the c.5152+2T>G, not the c.2522G>A variant. This finding, besides supporting the

hypothesis that the c.5152+2T>G variant, predicted to affect splicing, may be associated with cancer risk, excludes a role for c.2522G>A in cancer clustering in this family. Moreover, if c.5152+2T>G will be definitely classified as pathogenic in the future, the co-occurrence *in-trans* with c.2522G>A in our patient will provide evidence for conclusively classifying the latter (now class 2) as neutral. Its neutrality is supported by results of functional studies, as no splicing alterations was detected through the minigene assay (Anczuków et al., 2008), and no difference was found, in comparison to wild-type BRCA1, on cisplatin response in a resazurin cell viability assay (Bouwman et al., 2013).

Pedigree 50-O-14 (**Figure 2**) The proband developed hormone-responsive breast cancer under the age of 40 and experienced multisite relapse few years later; her mother, who had undergone hysteroannessiectomy for unspecified reasons, had developed post-menopausal hormone-responsive breast cancer. The **BRCA1** variant **c.4223A>G** was detected in the proband and then confirmed in her mother. This finding failed to provide any significant information on the clinical significance of the variant; however, the low prior probability of BRCA1 pathogenic variants and *in silico* predictions do not support its pathogenicity (**Table 3**). No functional data were available for this variant.

TABLE 3 | Description of families and variants.

Pedigree	Prior probability of pathogenic variant % BOADICEA	Prior probability of pathogenic variant % BRCAPRO	Gene	Variant	LOVD	ENIGMA	ARUP	ClinVar	UMD Share	A-GVGD (GV-GD score)	Predicted effect on splicing	ExAC Allele frequency	Y or N	FLB	CSLR	meioses	Notes
281-O-15	BRCA1: 67.4 BRCA2: 3.9	BRCA1: 84.37 BRCA2: 5.63	BRCA1	c.2522G > A p.Arg841Gln	NC	NR	NP	2	2	C0 (111.49-0.00)	none	0.00004121	N	0.014	0.145	0	*in-trans* with c.5152+2T > G
50-O-14	BRCA1: 0.4 BRCA2: 4	BRCA1: 0.1 BRCA2: 1.3	BRCA1	c.4223A > G p.Gln1408Arg	NP	NR	NP	3	NP	C0 (29.27-42.81)	none	NP	Y	1.86	1.63	1	
191-O-15	BRCA1: 11.1 BRCA2: 68.1	BRCA1: 3.79 BRCA2: 10.16	BRCA1	c.4895T > G p.Val1632Gly	NP	NR	NP	NP	3	C0 (262.18-0.00)	none	NP	N	0.0036	0.0099	0	
281-O-15	BRCA1: 67.4 BRCA2: 3.9	BRCA1: 84.37 BRCA2: 5.63	BRCA1	c.5152+2T > G	NP	NR	NP	NP	NP	NA	donor site broken	NP	Y	1.98	1.86	1	*in-trans* with c.2522G > A
357-O-17	BRCA1: 1.4 BRCA2: 4.1	BRCA1: 0.5 BRCA2: 2.34	BRCA1	c.5509T > C p.Trp1837Arg	NC	NR	NP	4	5	C65 (0.00-101.29)	none	NP	Y	1.86	1.62	1	
146-O-15	BRCA1: 38.8 BRCA2:1.2	BRCA1: 36.4 BRCA2: 2.8	BRCA2	c.476T > C p.Val159Ala	NP	NR	NP	3	NP	C0 (170.06-0.00)	none	NP	N	0.0048	0.0284	−1	*in-cis* with c.6290C > T
282-O-17	BRCA1: 1.4 BRCA2: 7.9	BRCA1: 0.31 BRCA2: 0.61	BRCA2	c.1247T > G p.Ile416Ser	NP	NR	NP	1	3	C0 (353.86-0.00)	none	NP	Y	1.75	1.36	1	
191-O-15	BRCA1: 11.1 BRCA2: 68.1	BRCA1: 3.79 BRCA2: 10.16	BRCA2	c.5386G > T p.Asp1796Tyr	NP	NR	NP	3	3	C0 (353.86-0.00)	none	NP	N	0.0078	0.0167	0	
368-O-17	BRCA1: 0.3 BRCA2: 2.8	BRCA1: 0.3 BRCA2: 1.36	BRCA2	c.5635G > A p.Glu1879Lys	NP	NR	NP	2	3	C0 (98.18-0.00)	none	0.0002157	N	0.0321	NA	NA	
275-O-14	BRCA1: 17.4 BRCA2: 2.0	BRCA1: 4.7 BRCA2: 0.8	BRCA2	c.6290C > T p.Thr2097Met	NC	NR	NP	2	3	C0 (244.82-0.00)	none	0.0001168	Y	1.96	1.94	0	
146-O-15	BRCA1: 38.8 BRCA2: 1.2	BRCA1: 36.4 BRCA2: 2.8											N	0.0048	0.0284	−1	*in-cis* with c.476T > C
18-B-16	BRCA1: 0.8 BRCA2: 6.0	BRCA1: 0.1 BRCA2: 1.2	BRCA2	c.7534C > T p.Leu2512Phe	NC	NR	NP	3	3	C0 (116.58-8.11)	none	0.00004136	N	0.2	0.77	0	
418-O-17	BRCA1: 0.6 BRCA2: 4.5	BRCA1: 0.99 BRCA2: 0.95	BRCA2	c.7975A > G p.Arg2659Gly	NP	NR	NP	3	3	C65 (0.00-125.13)	donor site broken	NP	Y	2.63	2.5	2	LOH in OC
115-O-13	BRCA1: 1.7 BRCA2: 0.9	BRCA1: 4.4 BRCA2: 9.0	BRCA2	c.8386C > T p.Pro2796Ser	NP	NR	NP	2	NP	C0 (145.74-0.00)	none	0.00001647	N	0.11	0.1	0	

NC, not classified; NP, not present; NR, not reviewed; NA, not applicable; Y, Yes; N, No; FLB, full-likelihood method for bayes factor; CSLR, co-segregation likelihood ratio; LOH, loss of heterozygosity; OC, ovarian cancer.
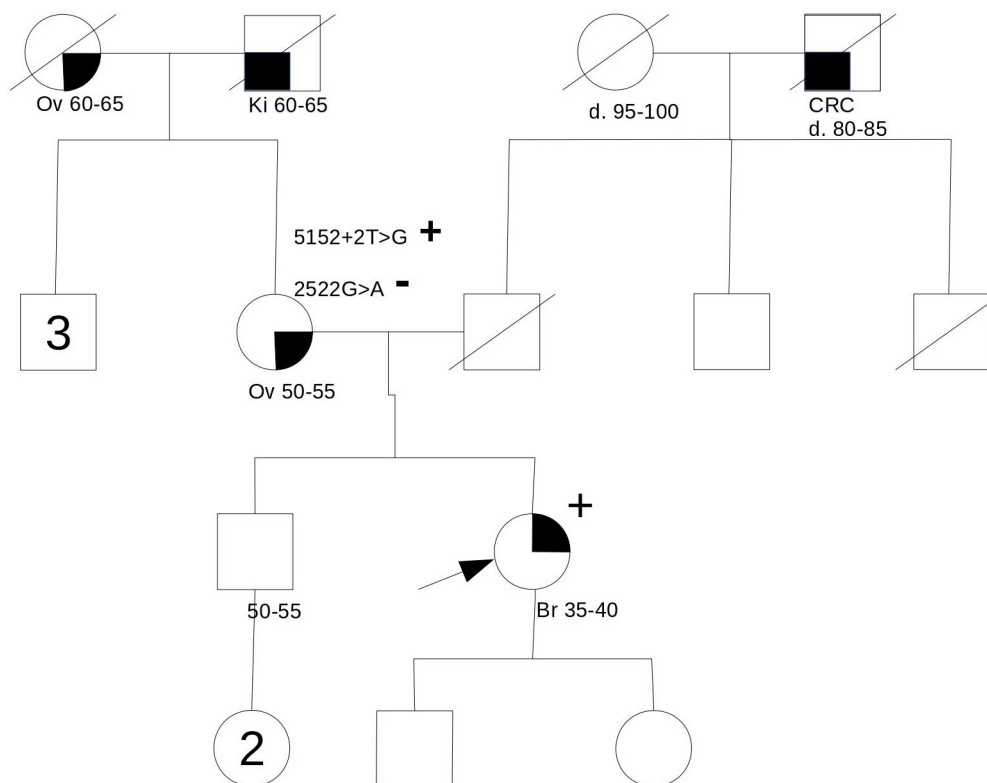
FIGURE 1 | Pedigree of Family 281-O-15 (Br: Breast Cancer. Ov: Ovarian Cancer. Ki: Kidney Cancer. CRC: Colorectal Cancer).

Pedigree 191-O-15 (Figure 3) The proband developed hormone-responsive breast cancer under age 35. Two paternal aunts had breast cancer diagnosed in their 40 s, with one developing contralateral breast cancer over 20 years later. Genetic testing performed in another center in the proband had detected the **BRCA1** variant **c.4895T>G** and the **BRCA2** variant **c.5386G>T**, both reported as class 3 in databases. When she came to our clinic with her mother for a second opinion, we proposed to check the parental origin of the variants by testing the mother, having clinical and family history negative for breast and ovarian cancer. The mother was found to carry both the variants, thus excluding a role for them in breast cancer clustering in the paternal side of the family; together with *in silico* predictions, cosegregation analysis supports the neutrality of both the variants.

Pedigree 357-O-17 (Figure 4) The proband had hormone-responsive breast cancer between 50 and 55 years of age. Her mother had developed ovarian cancer at the same age. BRCA testing in the proband led to the detection of the **BRCA1** variant **c.5509T>C**. This variant was subsequently tested in the mother, who was found to carry it, as well. This variant is currently reported as class 5 in the UMD database (class 4 in ClinVar) and its pathogenicity is supported by A-GVGD (C65). Indeed, functional studies have shown this variant to be associated with a severe folding defect, demonstrated through both a protease-based and a peptide-binding assay (Williams et al., 2003, 2004). Therefore, it is confirmed to be pathogenic and to explain the aggregation of breast and ovarian cancer in the family.

Pedigree 146-O-15 (Figure 5) The proband is an asymptomatic woman who requested an assessment of her
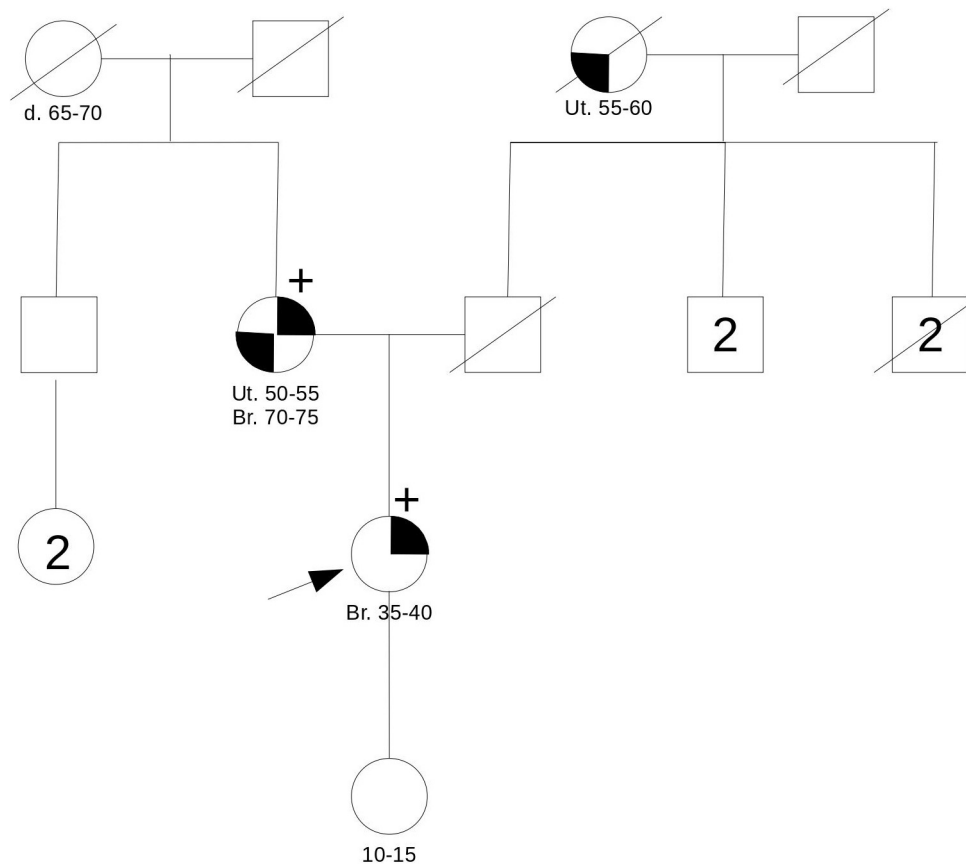
FIGURE 2 | Pedigree of Family 50-O-14 (Br: Breast Cancer. Ut: Uterine Cancer).

breast and ovarian cancer risk due to a strong history of both malignancies in the maternal side of the family. Indeed, the mother, a maternal aunt and the maternal grandmother had died for ovarian cancer diagnosed between 44 and 65 years of age, the other maternal aunt had a triple-negative breast cancer in her 70 s. Based on her high prior probability of BRCA pathogenic variants (36.4–38.8% for BRCA1; 1.2–2.8% for BRCA2), she was eligible for genetic testing even though she was asymptomatic. Genetic analysis revealed the presence of two **BRCA2** variants: **c.476T>C** and **c.6290C>T**. We proposed to test the father, who was found to carry both the variants. This allowed us to define that the variants were *in-cis* on the allele inherited from the father, thus excluding a role for them in the cancer aggregation in the mother's side. Together with *in silico* predictions, cosegregation analysis supports the neutrality of both the variants.

Pedigree 282-O-17 (**Figure 6**) The proband developed ductal *in situ* breast cancer under the age of 50 and experienced

local relapse (ductal infiltrating carcinoma) some years later; her mother developed hormone-responsive cancer of her right breast in the 7th decade of life and contralateral breast cancer 10 years later. The **BRCA2** variant **c.1847T>G** was detected in the proband and then confirmed in her mother. This finding failed to provide any significant information on the clinical significance of the variant. This finding failed to provide any significant information on the clinical significance of the variant; however, the low prior probability of BRCA2 pathogenic variants in the family and *in silico* predictions do not support its pathogenicity (**Table 3**). No functional data were available for this variant.

Pedigree 368-O-17 (**Figure 7**) The proband developed hormone-responsive breast cancer around the age of 40. Her mother had breast cancer in her 60 s. Short after we saw the proband for the first counseling session, a half-sister (same mother), was diagnosed with post-menopausal breast cancer. The proband was found to carry the **BRCA2** variant **c.5635G>A**,
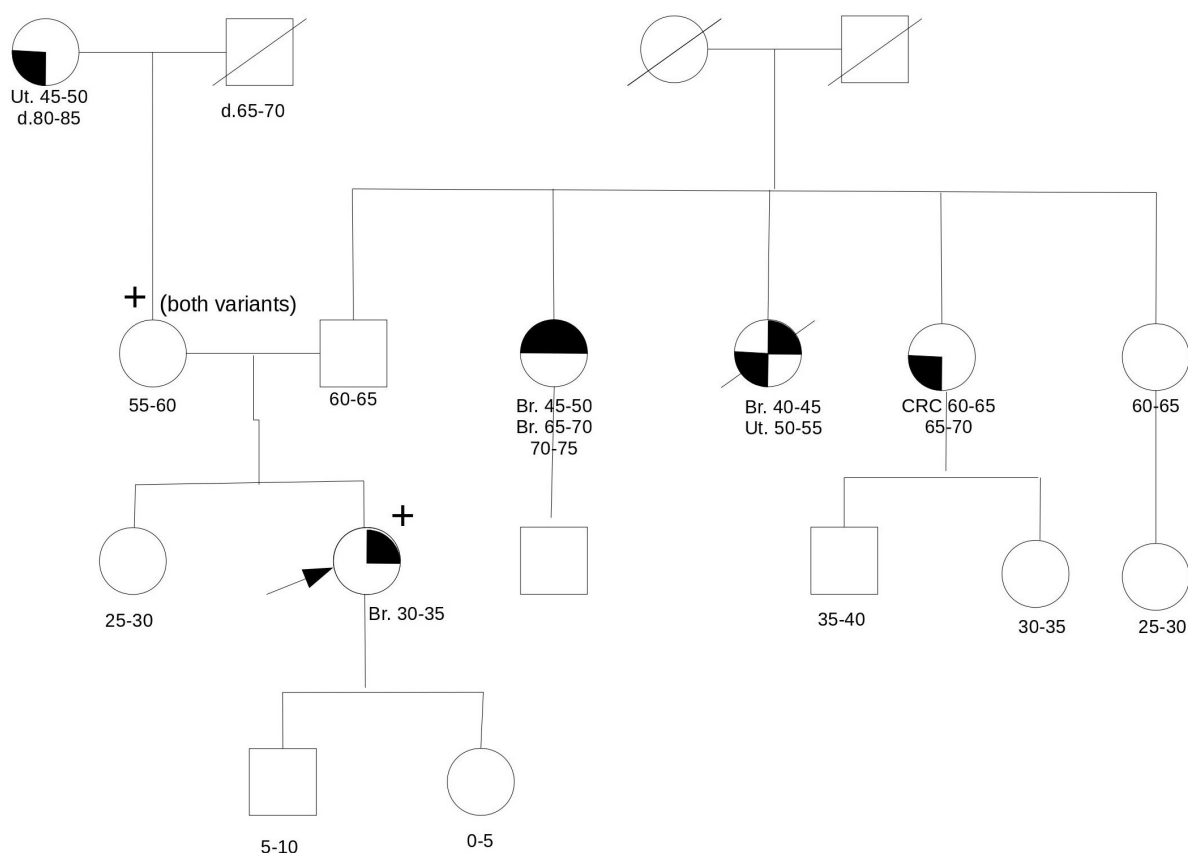
FIGURE 3 | Pedigree of Family 191-O-15 (Br: Breast Cancer. Ut: Uterine Cancer CRC: Colorectal Cancer).

which was subsequently excluded in the affected half-sister. One year later, we had the opportunity to analyze also the mother, who tested negative for the variant, thus excluding a role for it in breast cancer clustering in this family. Although no functional data are available for this variant, cosegregation analysis and *in silico* predictions provide data against its pathogenicity.

Pedigree 275-O-14 (**Figure 8**) The proband developed triple-negative breast cancer in her 30 s. A paternal first-degree cousin was reported to have died for breast cancer diagnosed at a similar age. Genetic testing revealed the **BRCA2** variant **c.6290C>T**. Through testing parents, we could define it had been inherited by the father. However, this finding fails to add relevant evidence on the significance of the variant, which in family 146-O-15 fails to cosegregate with the disease.

Pedigree 18-B-16 (**Figure 9**) The proband developed invasive lobular carcinoma around the age of 30. Her maternal

grandfather died for pancreatic cancer and two sisters of him died for post-menopausal breast cancer. Genetic testing detected the **BRCA2** variant **c.7534C>T**. When she came to our clinic with her father for post-test counseling, we proposed to check the parental origin of the variants by testing the father, whose clinical and family histories were negative for breast and ovarian cancer. The father was found to carry the variant, thus excluding a role for them in breast cancer clustering in the maternal side of the family. No functional data are available for this variant, however, cosegregation analysis and *in silico* predictions provide support against its pathogenicity.

Pedigree 418-O-17 (**Figure 10**) The proband is an asymptomatic 60-year-old woman who requested an assessment of her breast and ovarian cancer risk due to a history of both malignancies in her sisters. One of the sisters, affected with serous high grade ovarian carcinoma, had BRCA testing performed on
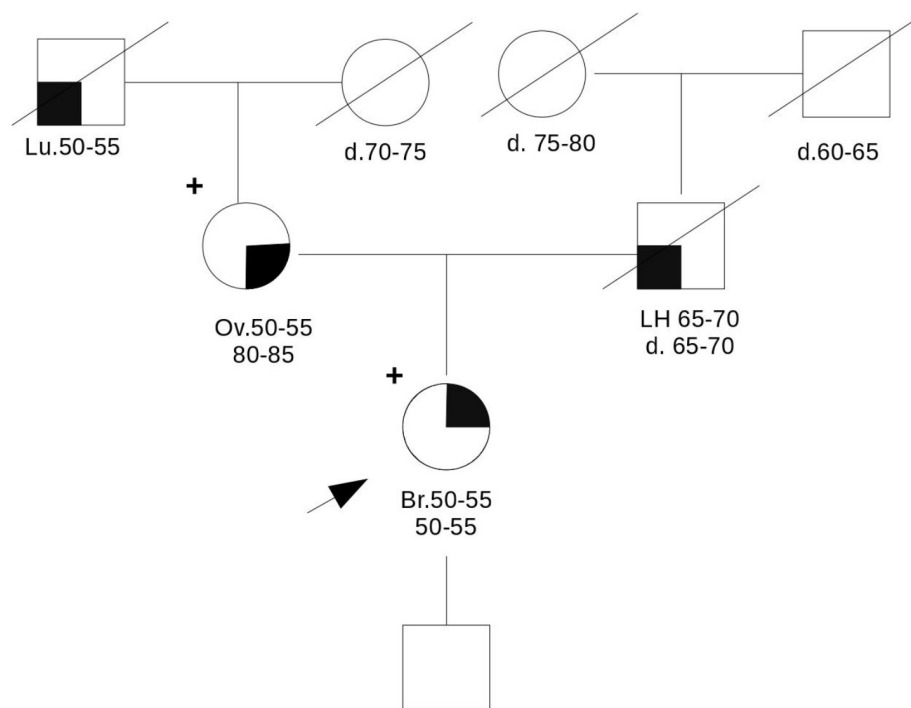
FIGURE 4 | Pedigree of Family 357-O-17 (Br: Breast Cancer. Ov: Ovarian Cancer. Lu: Lung Cancer. LH: Hodgkin Lymphoma).

cancer tissue in another center, with detection of the **BRCA2** variant **c.7975A>G**. Allelic load in tumor tissue was 80% and the variant was subsequently demonstrated in the germline. *In silico* evaluations supported a pathogenic effect: the amino acid Arg2659, which is substituted by a Glycine residue as a result of the variant, locates in the helical domain just prior to the OB1, with residues in this region fully conserved across all species, including the relatively distant pufferfish, *Tetraodon nigroviridis*. Although c.7975A>G is not described in variant classification databases, other changes of the same amino acid have been classified as definitely pathogenic (class 5). Indeed, the nucleotide c.7975 is located in a consensus splice site and bioinformatic tool predicted splice site alteration. Consistently, both Arg2659Thr and Arg2659Lys have been demonstrated to induce exon 17 skipping in patients' lymphocytes, (Hofmann et al., 2003; Farrugia et al., 2008). Moreover, allelic load in the tumor suggested Loss-Of-Heterozygosity, thus reinforcing the suspicion. Then, we proposed that the other affected sister, diagnosed with breast cancer in her 40 s, be tested for the variant. The sister was found to carry the variant as well. We then discussing with the proband about the added value of checking whether she carried the variant or not, and she consented to

be tested. She tested negative for the variant, which provides additional support to the hypothesis of pathogenicity; as shown in **Table 3**, cosegregation ratio in this family was ≥2 (2–2.63) with all the methods adopted, being the highest in the population under study.

Pedigree 115-O-13 (**Figure 11**) The proband developed breast cancer around the age of 40 and serous ovarian carcinoma 25 years later. Genetic testing revealed the **BRCA2** variant **c.8386C>T**. This variant was subsequently checked in the niece, who had developed breast cancer in her 30 s and was demonstrated not to carry the VUS. This result excluded the variant as a predisposition factor shared by the two women. No functional data are available for this variant, however, cosegregation analysis and *in silico* predictions support its neutrality.

DISCUSSION

VUSs in BRCA genes are reported in 5–20% of patients undergoing genetic testing (Lindor et al., 2012; Eccles et al., 2015). In line with those findings, we detected class 3
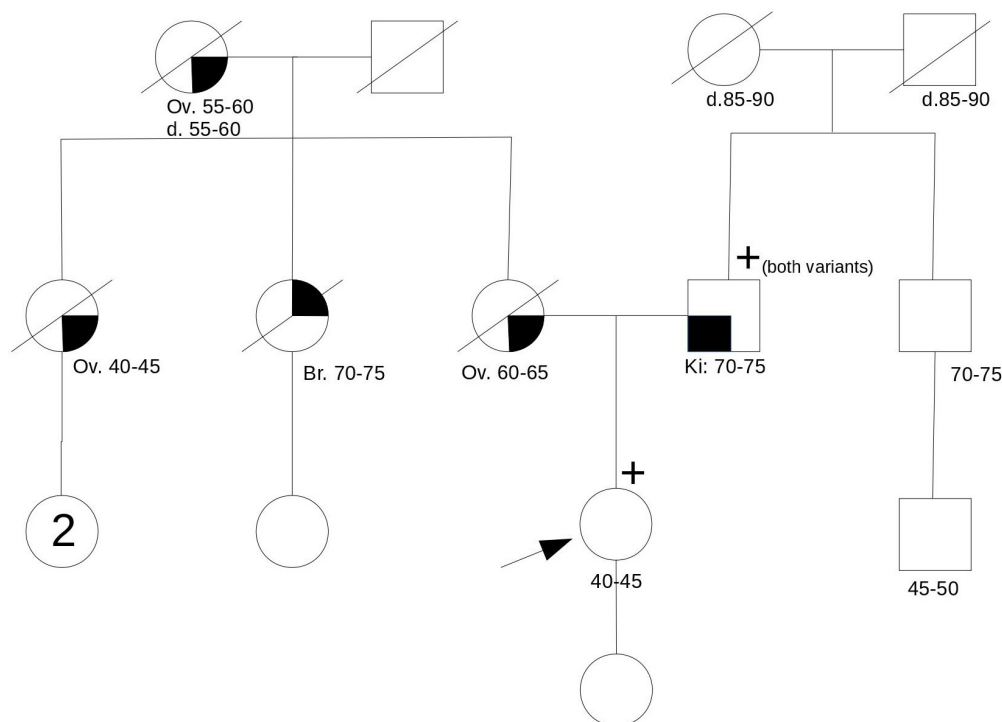
FIGURE 5 | Pedigree of Family 146-O-15 (Br: Breast Cancer. Ov: Ovarian Cancer. Ki: Kidney Cancer).

VUS in 6.3% of 1045 breast/ovarian cancer patients analyzed since 2011. Although VUSs are unanimously recognized as seriously challenging risk communication and perception, it is recommended that they are not used for predictive testing in other family members due to their uncertain clinical impact (Plon et al., 2008). Consequently, in most cancer genetics clinics, cosegregation of the variant with cancer in the family is not offered. In addition, quantitative cosegregation analysis performed in a clinical setting is unlikely to provide data significant enough to help classifying a variant, unless it is found in multiple large-size families (Ranola et al., 2018). Accordingly, in our experience, only for one VUS (1673delH in BRCA1), that had been found in 14 families (one very large), cosegregation analysis provided meaningful results to be incorporated in the multifactorial likelihood method, leading to a statistically significant ratio in favor of pathogenicity (Zuntini et al., 2017). All the other VUS were found in 1–5 families each, with pedigree size and structure impairing the significance of a cosegregation analysis. Nevertheless, here we show that cosegregation analysis in selected families may help understand

whether that variant may have played a role in cancer clustering in the specific kindred. Indeed, 7 out of 13 variants assessed failed to cosegregate with breast cancer in the family. Although this finding does not allow drawing any definite conclusion on the neutrality of the variant, it may promote a correct perception, by the counselees, about the scarce informativeness of that test result. In fact, many lines of evidence suggest that a VUS result is associated to higher levels of distress, anxiety and risk overestimation, if compared to true uninformative results (Vos et al., 2011, 2012; Culver et al., 2013; Richter et al., 2013). Consistently, bilateral prophylactic mastectomy was performed in 39% of asymptomatic VUS carriers attending the Mayo Clinic; of notice, among the VUSs subsequently reclassified in their experience, 95% were benign (Welsh et al., 2017). Probably, receiving a VUS result has an additive load to risk perception associated to family history: "I and many other women in my family have developed breast cancer AND I carry a BRCA variant: it is definitely genetic." Excluding that the variant is shared by the other cancer cases in the family is likely to remove a relevant factor of genetic risk overestimation.
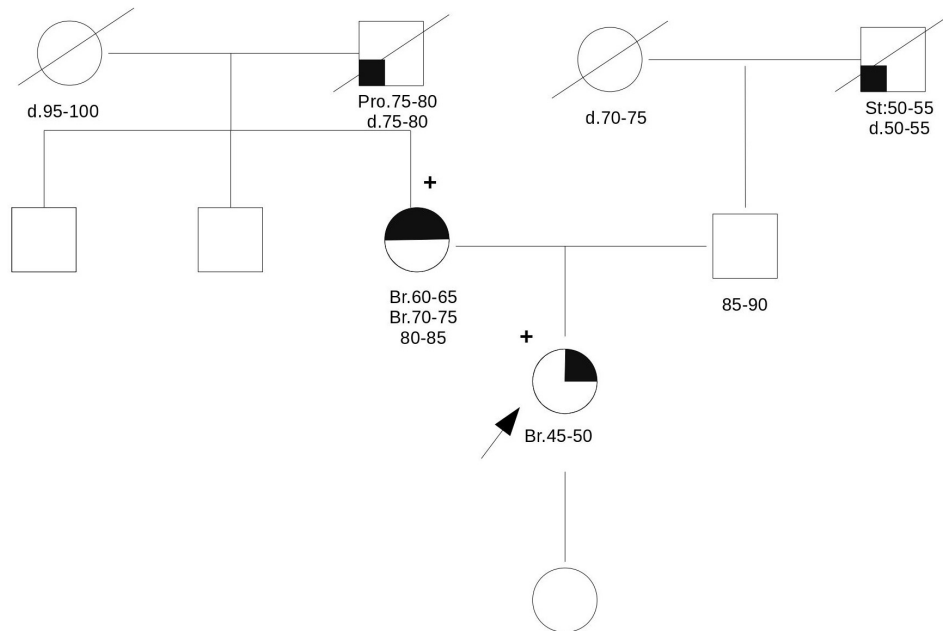
FIGURE 6 | Pedigree of Family 282-O-17 (Br: Breast Cancer. Pro: Prostate Cancer. St: Stomach Cancer).
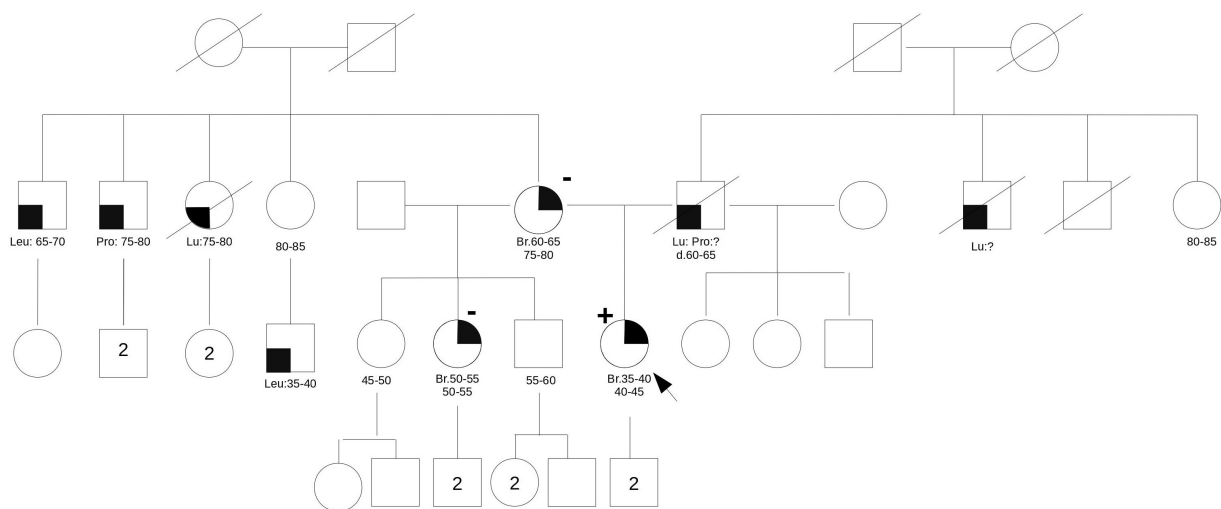
FIGURE 7 | Pedigree of Family 368-O-17 (Br: Breast Cancer. Lu: Lung Cancer. Pro: Prostate Cancer. Leu: Leukemia).

Pedigree 275-O-14

BRCA2 c.6290C>T (p.Thr2097Met)

FIGURE 8 | Pedigree of Family 275-O-14 (Br: Breast Cancer).

Pedigree 18-B-16

BRCA2 c.7534C>T (p.Leu2512Phe)

FIGURE 9 | Pedigree of Family 18-B-16 (Br: Breast Cancer. Pan: Pancreatic Cancer).

FIGURE 10 | Pedigree of Family 418-O-17 (Br: Breast Cancer. Ov: Ovarian Cancer. Pro: Prostate Cancer. St: Stomach Cancer).
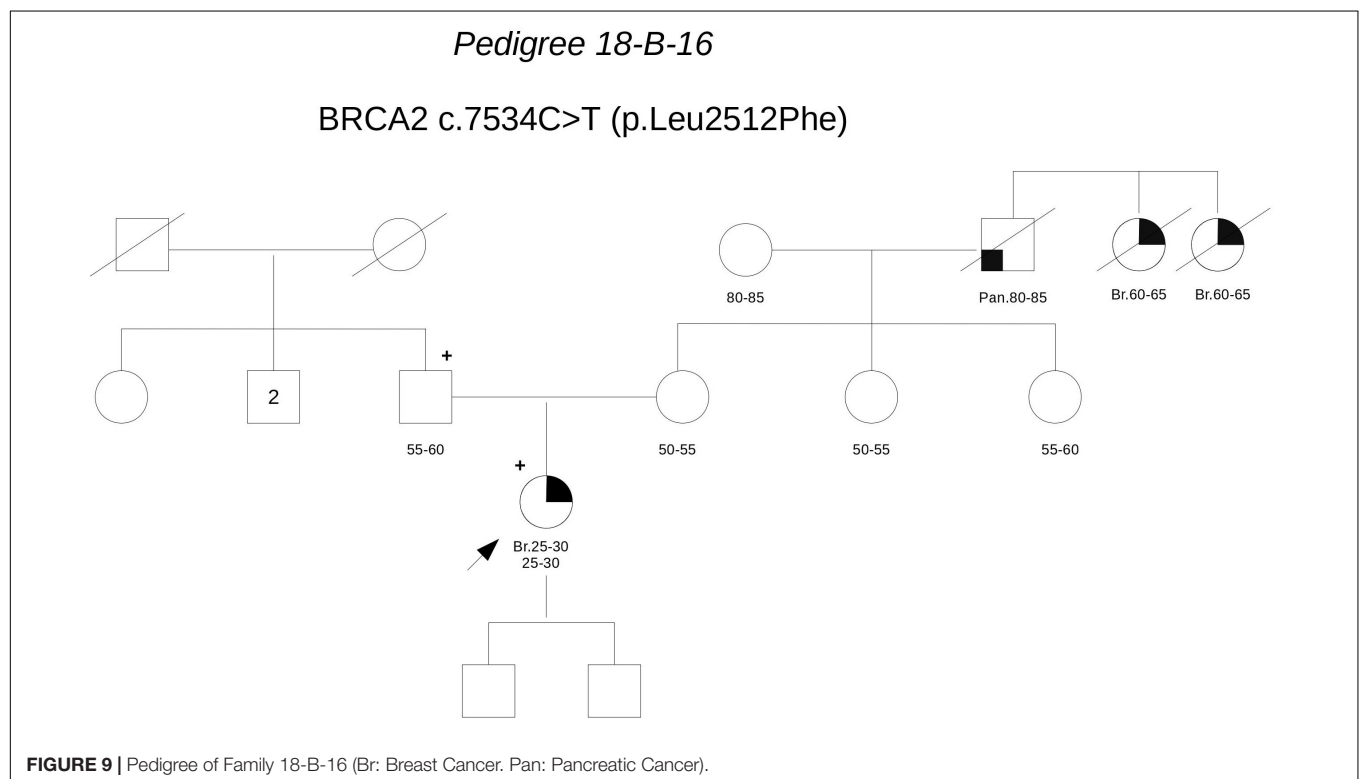

FIGURE 11 | Pedigree of Family 115-O-13 (Br: Breast Cancer. Ov: Ovarian Cancer. CRC: Colorectal Cancer).

However, an argument against such a "clinical cosegregation" approach may be that whenever the variant cosegregates with the disease, the false perception that it is causative may be reinforced. Actually, in our sample, among 6 variants cosegregating with the disease, two had additional evidence from literature and *in silico* predictions supporting their pathogenicity. In cases like these, we think that integrating pieces of information regarding the potential pathogenicity of the variant with the specific family situation, where cosegregation further supports its predisposing role, makes the communication process more accurate. To evaluate the actual impact of cosegregation analysis on risk perception, we plan to perform in these patients a qualitative study, using the same methods recently adopted on a different patient sample (Godino et al., 2018).

Finally, it is noteworthy that besides providing information potentially helpful for counseling patients, obtaining cosegregation data and sharing them within the scientific community is crucial to gather significant evidence that may eventually contribute to classify VUSs.

ETHICS STATEMENT

The study protocol conforms to the ethical guidelines of the WMA Declaration of Helsinki and was approved by the Ethical Board of Hospital S.Orsola-Malpighi, Bologna, Italy (Prot. 154/2010/O). All participants gave their informed consent to the analysis and signed the respective form.

REFERENCES

Anczuków, O., Buisson, M., Salles, M. J., Triboulet, S., Longy, M., Lidereau, R., et al. (2008). Unclassified variants identified in BRCA1 exon 11: consequences on splicing. *Genes Chromosomes Cancer* 47, 418–426. doi: 10.1002/gcc. 20546

Bouwman, P., van der Gulden, H., van der Heijden, I., Drost, R., Klijn, C. N., Prasetyanti, P., et al. (2013). A high-throughput functional complementation assay for classification of BRCA1 missense variants. *Cancer Discov.* 3, 1142–1155. doi: 10.1158/2159-8290.CD-13-0094

Culver, J. O., Brinkerhoff, C. D., Clague, J., Yang, K., Singh, K. E., Sand, S. R., et al. (2013). Variants of uncertain significance in BRCA testing: evaluation of surgical decisions, risk perception, and cancer distress. *Clin. Genet.* 84, 464–472. doi: 10.1111/cge.12097

Cusin, I., Teixeira, D., Zahn-Zabal, M., Rech de Laval, V., Gleizes, A., Viassolo, V., et al. (2018). A new bioinformatics tool to help assess the significance of BRCA1 variants. *Hum Genomics* 12:36. doi: 10.1186/s40246-018-0168-0

Godino, L., Jackson, L., Turchetti, D., Hennessy, C., and Skirton, H. (2018). Decision making and experiences of young adults undergoing presyntomatic genetic testing for familial cancer: a longitudinal grounded theory study. *Eur. J. Hum. Genet.* 26, 44–53. doi: 10.1038/s41431-017-0030-1

Goldgar, D. E., Easton, D. F., Deffenbaugh, A. M., Monteiro, A. N. A., Tavtigian, S. V., Couch, F. J., et al. (2004). Integrated evaluation of DNA sequence variants of unknown clinical significance. Application to BRCA1 and BRCA2. *Am. J. Hum. Genet.* 75, 535–544. doi: 10.1086/424388

Eccles, D. M., Mitchell, G., Monteiro, A. N. A., Schmutzler, R., Couch, F. J., Spurdle, A. B., et al. (2015). BRCA1 and BRCA2 genetic testing-pitfalls and recommendations for managing variants of uncertain clinical significance. *Ann. Oncol.* 26, 2057–2065. doi: 10.1093/annonc/mdv278

Farrugia, D. J., Agarwal, M. K., Pankratz, V. S., Deffenbaugh, A. M., Pruss, D., Frye, C., et al. (2008). Functional assays for classification of BRCA2 variants of uncertain significance. *Cancer Res.* 68, 3523–3531. doi: 10.1158/0008-5472. CAN-07-1587

AUTHOR CONTRIBUTIONS

FUNDING

ACKNOWLEDGMENTS

Hofmann, W., Horn, D., Huttner, C., Classen, E., and Scherneck, S. (2003). The BRCA2 variant 8204G > A is a splicing mutation and results in an in frame deletion of the gene. *J. Med. Genet.* 40:23. doi: 10.1136/jmg.40.3.e23

Jarvik, G. P., and Browning, B. L. (2016). Consideration of cosegregation in the pathogenicity classification of genomic variants. *Am. J. Hum. Genet.* 98, 1077–1081. doi: 10.1016/j.ajhg.2016.04.003

Lindor, N. M., Guidugli, L., Wang, X., Vallee, M. P., Monteiro, A. N. A., Tavtigian, S., et al. (2012). A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). *Hum. Mut.* 33, 8–21. doi: 10.1002/humu.21627

Mohammadi, L., Vreeswijk, M. P., Oldenburg, R., van den Ouweland, A., Oosterwijk, J. C., van der Hout, A. H., 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. doi: 10.1186/1471-2407-9-211

Plon, S. E., Eccles, D. M., Easton, D., Foulkes, W. D., Genuardi, M., Greenblatt, M. S., et al. (2008). Sequence variant classification and reporting: recommendations for improving the interpretazion of cancer susceptibility genetic test results. *Hum. Mut.* 29, 1282–1291. doi: 10.1002/humu. 20880

Ranola, J. M. O., Liu, Q., Rosenthal, E. A., and Shirts, B. H. (2018). A comparison of cosegregation analysis method for the clinical setting. *Fam. Cancer* 17, 295–302. doi: 10.1007/s10689-017-0017-7

Ranola, J. M. O., and Shirts, B. H. (2017). *Analyze my Variant Website*. Available at: http://www.analyze.myvariant.org

Richter, S., Haroun, I., Graham, T. C., Eisen, A., Kiss, A., and Warner, E. (2013). Variants of unknown significance in BRCA testing: impact on risk perception, worry, prevention and counseling. *Ann. Oncol.* 24, viii69–viii74. doi: 10.1093/annonc/mdt312

Servizio Sanità Pubblica and Regione Emilia-Romagna (2016). *Contributo n. 91/2016: "Protocollo Assistenziale Nelle Donne a Rischio Ereditario di Tumore Della Mammella e/o ovaio"*. Available at: http://salute.regione.emilia-romagna.it/documentazione/rapporti/contributi/contributi-n-91-protocollo-

assistenziale-nelle-donne-a-rischio-ereditario-di-tumore-della-mammella-e-o-ovaio-2016/view

Spurdle, A. B., Healey, S., Devereau, A., Hogervorst, F. B. L., Monteiro, A. N. A., Nathanson, K. L., et al. (2012). ENIGMA-evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Hum. Mutat.* 33, 2–7. doi: 10.1002/humu.21628

Tavtigian, S. V., Deffenbaugh, A. M., Yin, L., Judkins, T., Scholl, T., Samollow, P. B., et al. (2006). Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J. Med. Genet.* 43, 295–305. doi: 10.1136/jmg.2005.033878

Thompson, D., Easton, D. F., and Goldgar, D. E. (2003). A full-likelihood method for the evaluation of casuality of sequence variants from family data. *Am. J. Hum. Genet.* 73:652655. doi: 10.1086/378100

Vos, J., Jansen, A. M., Menko, F., van Asperen, C. J., Stiggelbout, A. M., and Tibben, A. (2011). Family communication matters: the impact of telling relatives about unclassified variants and uninformative DNA-test results. *Genet. Med.* 13, 333–341. doi: 10.1097/GIM.0b013e318204cfed

Vos, J., Oosterwijk, J. C., Gomez-Garcia, E., Menko, F. H., Collee, M. J., van Asperen, C. J., et al. (2012). Exploring the short-term impact of DNA-testing in breast cancer patients: the counselees' perception matters, but the actual BRCA1/2 result does not. *Patient Educ. Couns.* 86, 239–251. doi: 10.1016/j.pec.2011.04.017

Welsh, J. L., Hoskin, T. L., Day, C. N., Thomas, A. S., Cogswell, J. A., Couch, F. J., and Boughey, J. C. (2017). Clinical decision-making in patients with variant of uncertain significance in BRCA1 or BRCA2 genes. *Ann. Surg. Oncol.* 24, 3067–3072. doi: 10.1245/s10434-017-5959-3

Williams, R. S., Chasman, D. I., Hau, D. D., Hui, B., Lau, A. Y., and Glover, J. N. (2003). Detection of protein folding defects caused by BRCA1-BRCT truncation and missense mutations. *J. Biol. Chem.* 278, 53007–53016. doi: 10.1074/jbc.M310182200

Williams, R. S., Lee, M. S., Hau, D. D., and Glover, J. N. (2004). Structural basis of phosphopeptide recognition by the BRCT domain of BRCA1. *Nat. Struct. Mol. Biol.* 11, 519–525. doi: 10.1038/nsmb776

Zuntini, R., Cortesi, L., Calistri, D., Pippucci, T., Martelli, P. L., Casadio, R., et al. (2017). BRCA1 p.His1673del is a pathogenic mutation associated with a predominant ovarian cancer phenotype. *Oncotarget* 8, 22640–22648. doi: 10.18632/oncotarget.15151

Two Missense Variants Detected in Breast Cancer Probands Preventing BRCA2-PALB2 Protein Interaction

Laura Caleca[1], Irene Catucci[2], Gisella Figlioli[2], Loris De Cecco[3], Tina Pesaran[4], Maggie Ward[5], Sara Volorio[6,7], Anna Falanga[8], Marina Marchetti[7], Maria Iascone[9], Carlo Tondini[10], Alberto Zambelli[10], Jacopo Azzollini[11], Siranoush Manoukian[11], Paolo Radice[1] and Paolo Peterlongo[2]*

[1] Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy, [2] Genome Diagnostics Program, IFOM the FIRC Institute of Molecular Oncology, Milan, Italy, [3] Platform of Integrated Biology, Department of Applied Research and Technology Development, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy, [4] Ambry Genetics, Department of Clinical Diagnostics, Aliso Viejo, CA, United States, [5] Cancer Outreach and Risk Assessment, Via Christi Hospitals, Wichita, KS, United States, [6] IFOM, Fondazione Istituto FIRC di Oncologia Molecolare, Milan, Italy, [7] Cogentech Cancer Genetics Test Laboratory, Milan, Italy, [8] Department of Immunohematology and Transfusion Medicine, Azienda Ospedaliera Papa Giovanni XXIII, Bergamo, Italy, [9] USSD Laboratorio Genetica Medica, Azienda Ospedaliera Papa Giovanni XXIII, Bergamo, Italy, [10] Unit of Medical Oncology, Azienda Ospedaliera Papa Giovanni XXIII, Bergamo, Italy, [11] Unit of Medical Genetics, Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

PALB2 (partner and localizer of BRCA2) was initially identified as a binding partner of BRCA2. It interacts also with BRCA1 forming a complex promoting DNA repair by homologous recombination. Germline pathogenic variants in *BRCA1*, *BRCA2* and *PALB2* DNA repair genes are associated with high risk of developing breast cancer. Mutation screening in these breast cancer predisposition genes is routinely performed and allows the identification of individuals who carry pathogenic variants and are at risk of developing the disease. However, variants of uncertain significance (VUSs) are often detected and establishing their pathogenicity and clinical relevance remains a central challenge for the risk assessment of the carriers and the clinical decision-making process. Many of these VUSs are missense variants leading to single amino acid substitutions, whose impact on protein function is uncertain. Typically, VUSs are rare and due to the limited genetic, clinical, and pathological data the multifactorial approaches used for classification cannot be applied. Thus, these variants can only be characterized through functional analyses comparing their effect with that of normal and mutant gene products used as positive and negative controls. The two missense variants *BRCA2*:c.91T >G (p.Trp31Gly) and *PALB2*:c.3262C >T (p.Pro1088Ser) were detected in two breast cancer probands originally ascertained at Breast Cancer Units of Institutes located in Milan and Bergamo (Northern Italy), respectively. These variants were located in the BRCA2-PALB2 interacting domains, were predicted to be deleterious by *in silico* analyses, and were very rare and clinically not classified. Therefore, we initiate to study their functional effect by exploiting a green fluorescent protein (GFP)-reassembly *in vitro* assay specifically designed to test the BRCA2-PALB2 interaction. This functional assay proved to be easy to develop, robust and reliable. It also allows testing variants located in different genes. Results from these functional analyses showed

that the *BRCA2*:p.Trp31Gly and the *PALB2*:p.Pro1088Ser prevented the BRCA2-PALB2 binding. While caution is warranted when the interpretation of the clinical significance of rare VUSs is based on functional studies only, our data provide initial evidences in favor of the possibility that these variants are pathogenic.

INTRODUCTION

Approximately 20% of the familial aggregation of breast cancer is related to the presence of germline pathogenic variants in the tumor suppressor high-risk genes *BRCA1* (MIM#113705) and *BRCA2* (MIM#600185) [reviewed in (1)]. Additional germline variants in several other genes, including *PALB2* (partner and localizer of BRCA2) (MIM#610355) have also been implicated in increased predisposition to breast cancer (2, 3). Estimated cumulative breast cancer risk by age of 70 conferred by pathogenic variants in *BRCA1* and *BRCA2* is approximately 60 and 50%, respectively (4, 5). Loss of function *PALB2* pathogenic variants confer a breast cancer risk of 35% by age of 70, that is comparable to that conferred by *BRCA2* pathogenic variants (6). Sequencing of these genes has become a key step of the clinical management of breast cancer families as the carriers of a pathogenic variants may be offered appropriate surveillance programs or risk reducing options, whereas the non-carriers may be advised to follow the same recommendations offered to the general population (7).

The clinical utility and efficacy of genetic testing rely on the possibility to establish a correlation between the detected genetic variant and its protein functional effect. As an example, pathogenicity is generally inferred for variants introducing premature termination codons (PTCs), or affecting mRNA integrity and/or stability that give rise to functionally compromised proteins. However, the assessment of the clinical relevance of other variants, especially those that are rare, may not be equally straightforward. These are referred to as variants of uncertain significance (VUSs) and typically include missense variants, small in-frame deletions or insertions, exonic and intronic alterations potentially affecting the mRNA splicing, and variants in regulatory sequences (4, 8). Many of such variants located in the *BRCA1*, *BRCA2,* and *PALB2* genes have been deposited as "unclassified" in publicly available databases. The current approach to clinically classify a VUS is the multifactorial likelihood prediction model in which, data from epidemiological, genetic, pathological and clinical analyses are combined in order to derive a posterior likelihood of pathogenicity. However, reaching odds ratios in favor of or against causality requires such analyses to be based on several independent observations or to be carried out in large sample series which are usually difficult to obtain if a variant is rare (9, 10). This provides a compelling rationale to the inclusion in the multifactorial model of additional experimental evidences. As a possibility, VUSs —especially those located in the coding regions—can be studied using *in vitro* and functional assays that compare the effect of normal and mutant gene products.

At the molecular level, PALB2 was identified as a binding partner of BRCA2 and was subsequently shown to bridge, via direct protein-protein interaction, BRCA1 and BRCA2 at sites of DNA damage (11–13). Here, this complex promotes the repair by homologous recombination (HR) of the highly genotoxic DNA lesions, such as double-strand breaks (DSBs) or inter-strand crosslinks (ICLs) (14, 15). These BRCA1-PALB2-BRCA2 interactions are mediated via the coiled-coil domains located at the N-terminus of PALB2 (amino acids 9-44) and at the C-terminus of BRCA1 (amino acids 1,393–1,424), and by the seven-bladed β-propeller WD40 (tryptophan-aspartic acid rich) domain of the C-terminal end of PALB2 (amino acids 836–1,186) binding a domain in the N-terminal end of the BRCA2 (amino acids 21–39) (16, 17). Functional assays based on these domain bindings were used to study patient-derived missense variants in *BRCA1* and *BRCA2* to provide evidence in favor of or against pathogenicity. Three *BRCA2* missense variants, the c.73G>A (p.Gly25Arg), c.91T>C (p.Trp31Arg), and c.93G>T (p.Trp31Cys) were found to disrupt the BRCA2-PALB2 interaction, causing deficiencies in BRCA2 localization to the nucleus and in HR mediated DSB repair (16). Similarly, three *BRCA1* missense variants, the c.4198A>G (p.Met1400Val), c.4220T>C (p.Leu1407Pro), and c.4232T>C (p.Met1411Thr) abrogated or moderately impaired the BRCA1-PALB2 binding, causing reduced HR activity (17, 18). To date, only few patient-derived missense variants in the *PALB2* gene have been investigated for pathogenicity. Among these, the *PALB2*:c.104T>C (p.Leu35Pro), located in the coiled-coil domain, was found to co-segregate with two breast cancer cases in a family with a strong history for the disease, and was shown to abrogate the BRCA1-PALB2 binding and to completely prevent HR and resistance to DNA damaging agents. As a result, the p.Leu35Pro was suggested to be a pathogenic variant (19) and is to our knowledge the sole variant in *PALB2* to date suggested to be pathogenic. All these findings emphasize that functional assays on VUS located in the BRCA1-PALB2-BRCA2 interaction domains may provide clues on their pathogenicity and that other variants affecting such interactions may be associated with breast cancer susceptibility.

In the current study, we aimed to characterize functionally the two rare missense variants, *PALB2*:c.3262C>T (p.Pro1088Ser) and *BRCA2*:c.91T>G (p.Trp31Gly), that were initially identified in breast cancer families and that are located in the protein interaction domains. These two variants were tested for

pathogenicity using the green fluorescent protein (GFP)-reassembly *in vitro* assay that was recently developed for the study of protein-protein interactions (20, 21).

PATIENTS AND METHODS

Breast Cancer Probands and Genotyping Analysis

The two female Italian breast cancer probands included in this study were originally considered eligible for clinical genetic testing in breast cancer genes, based on criteria including age of onset for breast cancer and family history for the disease. One proband, recruited at the Genetics Unit of Fondazione IRCCS Istituto Nazionale dei Tumori in Milan (INT), was tested for mutations in the coding regions of *BRCA1* and *BRCA2* by massively parallel sequencing, using TruSeq Custom Amplicon v.1.2 (Illumina), and multiplex ligation-dependent probe amplification (MLPA) resulting carrier of the *BRCA2*:p.Trp31Gly. These tests were performed at Cogentech Cancer Genetic Test Laboratory (CGT Lab). The other proband, recruited at the Unit of Medical Oncology of the Ospedale Papa Giovanni XXIII in Bergamo (HPG23), was tested for mutations in the coding regions of *BRCA1* and *BRCA2* by Sanger sequencing and MLPA at Cogentech CGT Lab. No *BRCA1* or *BRCA2* mutations were detected, and so this probands was tested at Laboratorio Genetica Medica, HPG23 by massively parallel sequencing using TruSight Cancer assay (Illumina). No pathogenic or likely pathogenic variants were found and the only deleterious variant detected was the missense *PALB2*:p.Pro1088Ser variant.

Genotyping of the *PALB2*:p.Pro1088Ser was performed using a custom TaqMan assay (probes and experimental conditions are available upon request). This variant was tested in familial and consecutive breast cancer cases ascertained at HPG23, and in female blood donors used as controls recruited at the AVIS Bergamo.

All individuals included in this study and herein described signed an informed consent to the use of their biological samples and clinical data for research project. This study was approved by Ethical Committee of Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, and Ethical Committee of the Province of Bergamo.

Positive/Negative Controls for the GFP-Reassembly *in vitro* Assay

Six *BRCA2* or *PALB2* variants located in the protein interaction domain were included as positive and negative controls. The *PALB2*:c.2816T>G (p.Leu939Trp) variant was reported to be not associated with breast cancer risk and to not alter the protein DNA repair activity by HR (22, 23). The *BRCA2*:c.79A>G (p.Ile27Val) and *PALB2*:c.3064AT>GC (p.Met1022Ala) missense variants were functionally tested and resulted not disrupting the BRCA2-PALB2 interaction (16, 24). These three variants were used as positive controls. The three missense variants *BRCA2*:c.93G>T (p.Trp31Cys), *BRCA2*:c.91T>C (p.Trp31Arg) and *PALB2*:c. 3073G>A (p.Ala1025Arg) were reported to functionally prevent the BRCA2-PALB2 binding and were used

as negative controls (16, 24). All of these variants were patient-derived with the exception of the *PALB2*:p.Met1022Ala and *PALB2*:p.Ala1025Arg that were synthetically designed based on crystallography analyses.

Plasmid Construction

The pET11a-NfrGFP-Z and pMRBAD-Z-CfrGFP expression vectors, encoding anti-parallel leucine zipper motifs (Z) fused to the N-terminal or C-terminal fragment of the GFP Protein (NfrGFP and CfrGFP, respectively) (20) were kindly donated by TJ Magliery from the Ohio State University in Columbus (OH, USA). The DNA fragments encoding the N-terminal end of BRCA2 (amino acids 10–40) and the WD40 domain of PALB2 (amino acids 836-1186) were amplified from the cDNA of the 293T cells by PCR. The purified PCR products were subcloned into pET11a-NfrGFP between *XhoI* and *BamHI* restriction sites and pMRBAD-Z-CfrGFP between *NcoI* and *AatII* restriction sites, replacing the fragments encoding Z motifs. The *BRCA2* c.79A>G (p.Ile27Val), c.91T>C (p.Trp31Arg), c.91T>G (p.Trp31Gly), c.93G>T (p.Trp31Cys) and the *PALB2* c.2816T>G (p.Leu939Trp), c.3073G>A (p.Ala1025Arg), c.3266C>T (p.Pro1088Ser) variants were obtained by direct mutagenesis of pET11a-NfrGFP-BRCA2 and of pMRBAD-PALB2-CfrGFP using the QuickChange XL Site-directed Mutagenesis Kit (Stratagene) according to the manufacturer's instruction. The *PALB2* c.3064AT>GC (p.Met1022Ala) was obtained by the overlap extension PCR mutagenesis method (25). The presence of variants in recombinant clones was verified by DNA sequencing (Eurofins Genomics).

GFP-Fragment Reassembly Screening

Compatible pairs of plasmids (pET11a-NfrGFP-BRCA2 and pMRBAD-PALB2-CfrGFP, both as wild-type and mutant forms) were co-transformed into BL21 (DE3) *E. coli* competent cells by electroporation. Single colonies were then picked and used to inoculate 2 ml of LB broth medium containing ampicillin (100 μg/ml) and kanamycin (35 μg/ml). Following overnight incubation at 37°C, the cultured cells were diluted 1:1,000 and 100 μl were plated on inducing LB agar (LBA) plates supplemented with 20 μM Isopropyl β-D-1-tiogalattopiranoside (IPTG) and 0.2% L-arabinose, to promote the expression of recombinant proteins. The plates were incubated at 30°C for 24 h and then 3 days at room temperature (RT). Fluorescence was observed after excitation with long-wave (365 nm) UV light in combination with the short pass (SP) emission filter using a Syngene image capture system (SYNGENE) as specified by the manufacturer.

Purification of the Reassembled GFP Complexes

The pET11a-NfrGFP-BRCA2 (both wt and mutant forms) also encode a hexa histidine (H_6)-tag at the N-terminus of the NfrGFP useful for rapid purification by Immobilized metal affinity chromatography (IMAC) method of the H_6-tagged proteins. This method exploits the strong binding of H_6-tagged protein to metal ions as nickel, allowing them to be separated from other proteins that have lower or no affinity.
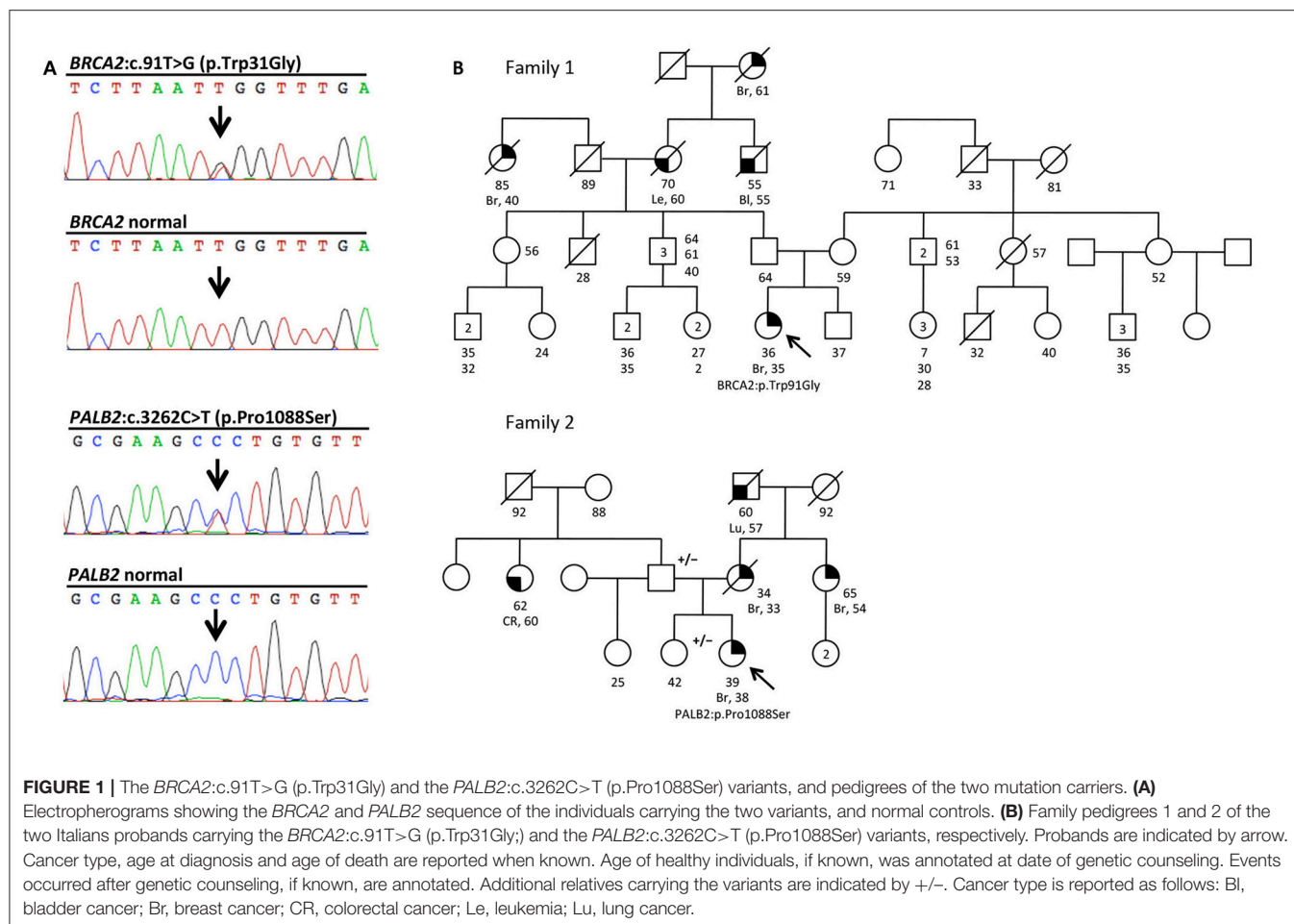
FIGURE 1 | The *BRCA2*:c.91T>G (p.Trp31Gly) and the *PALB2*:c.3262C>T (p.Pro1088Ser) variants, and pedigrees of the two mutation carriers. **(A)** Electropherograms showing the *BRCA2* and *PALB2* sequence of the individuals carrying the two variants, and normal controls. **(B)** Family pedigrees 1 and 2 of the two Italians probands carrying the *BRCA2*:c.91T>G (p.Trp31Gly;) and the *PALB2*:c.3262C>T (p.Pro1088Ser) variants, respectively. Probands are indicated by arrow. Cancer type, age at diagnosis and age of death are reported when known. Age of healthy individuals, if known, was annotated at date of genetic counseling. Events occurred after genetic counseling, if known, are annotated. Additional relatives carrying the variants are indicated by +/−. Cancer type is reported as follows: Bl, bladder cancer; Br, breast cancer; CR, colorectal cancer; Le, leukemia; Lu, lung cancer.

Co-transformed bacterial cells were recovered from inducing LBA media using a plate spreader and resuspended in two 1 ml-aliquots of 1X phosphate buffered saline (PBS). After centrifugation, each pellet was resuspended in 50 µl of 1xSDS loading buffer (whole cell extracts), or in 1 ml of lysis buffer (50 mM Tris-HCl, 300 mM NaCl, 0.1% v/v Triton X-100, 100 µM EDTA pH8.0, 0.5 mg/ml lysozime, 20 mM Imidazole, protease inhibitors, 5 µg/ml DNase and RNase) for IMAC purification using the nickel nitrilotriacetic (Ni-NTA) agarose resin (QIAGEN), following the manufacturer's instructions. The purified protein complexes were subjected to 13% SDS-PAGE and visualized by Western blotting using a polyclonal anti-GFP antibody (#600-101-215; Rockland). Whole cell extracts were similarly resolved and visualized, to detect expression levels of the all NfrGFP-BRCA2 and CfrGFP-PALB2 fusion peptides.

RESULTS

Identification of the *BRCA2*:c.91T>G (p.Trp31Gly) and the *PALB2*:c.3262C>T (p.Pro1088Ser) Variants

Part of our research activity stems from the collaboration with several Breast Cancer Units in which clinical genetic testing is routinely performed. One of our major interest is to functionally study and characterize VUSs in breast cancer genes. In this study, we report the identification and describe the initial functional analyses of the *BRCA2*:p.Trp31Gly and the *PALB2*:p.Pro1088Ser variants. The *BRCA2*:p.Trp31Gly and the *PALB2*:p.Pro1088Ser variants were originally identified in two different Italian breast cancer probands born in Milano and Bergamo, respectively and are located in the BRCA2-PALB2 interacting domains. Both these probands developed breast cancer at a young age and reported a close relative affected with early onset breast cancer (≤40 years). Unfortunately, we were not able to ascertain other family members to be genotyped in order to attempt co-segregation analyses (**Figure 1**). None of these two variants were reported in public databases such as GnomAD and 1000 genomes. However, the *BRCA2*:<underline>p.Trp31Gly was annotated in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) in a single individual submitted by Ambry Genetics and classified as a VUS. On the contrary, the *PALB2*:p.Pro1088Ser was not found in any of the clinical databases we searched. However, during the last annual meeting of the *PALB2* Interest Group (PIG; http://www.palb2. org/) colleagues from Ambry Genetics reported the finding of an additional carrier of the *PALB2*:p.Pro1088Ser variant. To our knowledge, to date, this is only the second proband found to carry this variant. For this reason, we report here his clinical phenotype

FIGURE 2 | Family pedigree of the additional proband carrying the *PALB2*:c.3262C>T (p.Pro1088Ser) variant. The proband had large multi gene panel test (67 genes) due to a family history of cancer. The analysis was performed at the Ambry Genetics using Next Generation Sequencing and Array CGA (https://www.ambrygen.com/clinician/genetic-testing/28/ oncology/cancernext-expanded). Proband is indicated by arrow. Cancer type, age at diagnosis and age of death are reported when known. Age of healthy individuals, if known, was annotated at date of genetic counseling. Events occurred after genetic counseling, if known, are annotated. Cancer type is reported as follows: BC, basal cells cancer; Bl, bladder cancer; Me, melanoma; Pa, pancreatic cancer; Pr, prostate cancer; Un, unknown.

and family cancer history. This proband was a male affected with colorectal polyps, type unknown, and from a family with cases of melanoma and pancreatic cancer but not breast cancers (**Figure 2**). Unfortunately, also in this case, no other samples were available for genotyping. We previously reported two different founder mutations, the *PALB2*:c.1027C>T (p.Gln343*) and the *BRCA2*:c.190T>C (p.Cys64Arg), originally identified in the Bergamo province where they have a carrier frequency approximately 10-fold higher than that of the Italian population (21, 26, 27). Hence, we genotyped the *PALB2*:p.Pro1088Ser in 126 familial and 477 consecutive breast cancer cases, and 1,074 controls all born in the province of Bergamo but no additional carriers were found.

The *BRCA2*:c.91T>G (p.Trp31Gly) and the *PALB2*:c.3262C>T (p.Pro1088Ser) Variants Prevent the BRCA2-PALB2 Binding

The *BRCA2*:c.91T>G and the *PALB2*:c.3262C>T variants were located within the protein domains mediating the BRCA2-PALB2 interaction (**Figure 3A**). To evaluate the effect of these variants on the BRCA2-PALB2 interaction, we exploited a bimolecular fluorescence complementation-based assay, the GFP-reassembly *in vitro* assay. In this assay, the GFP is dissected into two fragments, the N-terminal, NfrGFP and the C-terminal, CfrGFP which are fused to the N-terminal end of BRCA2 (amino acids 10-40) and the WD40 domain of PALB2 (amino acids 836-1186), respectively. These two plasmids are co-expressed in BL21 (DE3) *E. coli* cells and only if BRCA2-PALB2 interaction occurs, the GFP reassemble emitting cellular fluorescence after ultraviolet (UV) irradiation.

Bright GFP fluorescence was observed in bacterial cells co-expressing NfrGFP fused with normal BRCA2 and CfrGFP fused with either normal PALB2, the clinically neutral variant

PALB2:p.Leu939Trp or *PALB2*:p.Met1022Ala (positive controls). Similar results were observed in bacterial cells co-expressing CfrGFP fused with normal PALB2 and NfrGFP fused with *BRCA2*:p.Ile27Val (positive control). On the contrary, no fluorescence was observed in bacterial cells co-expressing NfrGFP fused with normal BRCA2 and CfrGFP fused with either *PALB2*:p.Pro1088Ser, or *PALB2*:p.Ala1025Arg (negative control). Similar results were observed in bacterial cells co-expressing CfrGFP fused with normal PALB2 and NfrGFP fused with either *BRCA2*:p.Trp31Gly, or the negative controls *BRCA2*:p.Trp31Cys, or *BRCA2*:p.Trp31Arg (**Figure 3B**).

To confirm that GFP-reassembly was effectively due to the BRCA2-PALB2 interaction, the IMAC purified reassembled complexes were analyzed by Western blotting using a polyclonal anti-GFP antibody. Two bands corresponding to the components of the GFP reassembled complexes were detected in lysates of the bacterial cells that resulted fluorescent in the GFP-reassembly screening. On the contrary, no or low intensity bands corresponding to the PALB2-CfrGFP fused domains were observed in lysates from bacterial cells for whom no fluorescence was detected (**Figure 3C**). In general, any mutations can cause the decrease or the complete loss of expression of the encoded GFP fused peptides. Thus, we wanted to confirm that the lack or the low intensity of the PALB2-CfrGFP bands was not due to loss of expression. To this aim, the whole cell extracts were analyzed by Western blotting using a polyclonal anti-GFP antibody as previously described. In this experiment, we showed that normal and mutated fusion peptides were expressed to a similar extent indicating that the loss of fluorescence observed in the GFP-reassembly *in vitro* assay, was attributable to the lack of binding between the proteins and not to poor expression of the mutants (**Figure 3D**). All these results provided experimental evidence that both the *BRCA2*:p.Trp31Gly and the *PALB2*:p.Pro1088Ser variants abrogate the BRCA2-PALB2 binding.

DISCUSSION

In clinical settings, VUSs in breast cancer genes represent a serious issue in the process of disease risk assessment in carriers. Typically, results from different sources such as epidemiological, genetic, and clinical analyses are combined together in order to derive a posterior likelihood of pathogenicity used to classify a VUS. While this multifactorial approach is successful to classify common VUSs, variants that are rare or unique can only be studied through functional analyses.

In the present study, we investigated the pathogenicity of the two *BRCA2*:p.Trp31Gly and *PALB2*:p.Pro1088Ser variants performing functional analyses. The *BRCA2*:p.Trp31Gly was previously reported in a single proband and annotated as VUS. To our knowledge, the *PALB2*:p.Pro1088Ser was never detected before. Both variants are located in the interaction domains of BRCA2 and PALB2. Large part of the BRCA2 functions in the repair of the DNA double strand breaks and inter-strand crosslinks by HR depends from its interaction with PALB2. Thus, we developed a GFP-reassembly assay based on the testing of this interaction speculating that this binding assay would be a predictor of the effect of the variants on the BRCA2 integrity.

FIGURE 3 | Detection of the BRCA2-PALB2 interaction. **(A)** Schematic representation of the specific domains mediating the BRCA2-PALB2 binding and of the location of variants analyzed in the present study. **(B)** GFP-reassembly *in vitro* assay. Fluorescence was recovered, under long-wave UV light (365 nm), after 24 h of growth at 30°C followed by 3 days of incubation at room temperature. The variants PALB2 L939W or M1022A and BRCA2 I27V were included as positive controls. The variants PALB2 A1025R and BRCA2 W31C or W31R were included as negative controls. **(C)** IMAC purification. The BRCA2-PALB2 reassembled complexes were purified from the co-transformed *E. coli* BL21 (DE3) cells, subjected to SDS-PAGE and visualized by Western blotting using a polyclonal anti-GFP antibody. **(D)** Analysis of expression of BRCA2-NfrGFP and PALB2-CfrGFP normal and mutant forms. Whole cell extracts from the co-transformed *E. coli* BL21 (DE3) cells were subjected to SDS-PAGE and visualized by Western blotting using a polyclonal anti-GFP antibody. Images of **(C, D)** were selected from original autoradiographic films shown in **Supplementary Figure 1**.

In the GFP-reassembly assay, we used six different *BRCA2* and *PALB2* missense variants as controls. Two patient-derived *BRCA2* variants, the p.Trp31Arg and p.Trp31Cys, and one synthetic *PALB2* variant, the p.Ala1025Arg, were known to prevent the BRCA2-PALB2 interaction. The patient derived *BRCA2*:p.Ile27Val and *PALB2*:Leu939Trp, and the synthetic *PALB2*:p.Met1022Ala were expected to not alter the binding of these proteins. For all of these variants, the results were concordant with their expected effect on the BRCA2-PALB2 binding.

The GFP-reassembly assay results indicated that both the *BRCA2*:p.Trp31Gly and *PALB2*:p.Pro1088Ser prevented the BRCA2-PALB2 interaction suggesting that in physiological conditions these alleles encode proteins that might be unable to interact with PALB2 and BRCA2, respectively. To our knowledge, the *PALB2*:p.Pro1088Ser is the first missense variant in the gene that was functionally shown to abrogate the binding with BRCA2. As the correct formation of the BRCA1-PALB2-BRCA2 complex is necessary for DNA repair by HR, our results provide

evidences in favor of the hypothesis that the *BRCA2*:p.Trp31Gly and *PALB2*:p.Pro1088Ser are pathogenic variants. While this assumption is at present most likely—in example vs. the possibility that the variants are neutral—other aspects need to be considered for a clearer picture of the effect of these variants on breast cancer risk. Firstly, Foo and colleagues showed that both the *PALB2*:p.Leu35Pro and p.Tyr28Cys caused the loss of the interaction with BRCA1; however, only the p.Leu35Pro completely abrogated the HR activity and the p.Tyr28Cys caused a loss of approximately 65% of the HR activity (19). Hence, *PALB2* missense variants causing the loss of the binding with BRCA1 might confer different risk magnitude for breast cancer. To be conservative, we cannot exclude that this might be true as well for the PALB2 variants abrogating the binding with BRCA2. As a second point, caution should be taken when inferring on the nature of a missense variant on the bases of functional studies only. Park and colleagues reported that the *PALB2*:p.Leu939Trp variant might be pathogenic based on the fact that it resulted in altered BRCA2-PALB2 binding, decreased

HR capacity, and increased sensitivity to ionizing radiation (28). However, we provided strong evidences deriving from additional functional studies and very large case-control studies that the *PALB2*:p.Leu939Trp is a neutral variant (23). As a final consideration, it should be noted that of the many missense variants that were functionally proved to prevent the BRCA1-PALB2-BRCA2 complex formation (16, 19, 24), all, with the only exception of the *PALB2*:p.Leu939Trp that is consider benign or likely benign, are annotated or should be treated clinically as VUS.

In conclusion, we report here results from functional studies indicating that the *BRCA2*:c.91T>G (p.Trp31Gly) and the *PALB2*:c.3262C>T (p.Pro1088Ser) missense variants abrogate the BRCA2-PALB2 protein binding. These data provide initial evidences corroborating the hypothesis that these variants are pathogenic. Importantly, novel data are warranted to progress in the clinical classification of these variants. The search for additional variant carriers and collection of their family members is crucial to provide genetic or pathological data; however, as the variants in study are very rare, we expect that not many variant carriers will be found in the near future. On the contrary, additional functional studies (i.e. testing specific protein functions in eukaryotic cells) could be immediately performed. While caution is warranted when clinical classification of a VUS is based on *in vitro* assays only, these results will provide additional evidences to better clarify the functional effect of the variants in study.

AUTHOR CONTRIBUTIONS

PP and PR designed and supervised the study. TP, MW, SV, AF, MM, MI, CT, AZ, JA, and SM provided samples and data. LC, IC, and LD performed experiments. LC, IC, GF, PP, and PR analyzed data. LC, IC, PP, and PR wrote the manuscript. All authors contributing to, critically revised and approved the manuscript.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2018.00480/full#supplementary-material

REFERENCES

1. Mavaddat N, Antoniou AC, Easton DF, Garcia-Closas M. Genetic susceptibility to breast cancer. *Mol Oncol.* (2010) 4:174–91. doi: 10.1016/j.molonc.2010.04.011

2. Easton DF, Pharoah PD, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, et al. Gene-panel sequencing and the prediction of breast-cancer risk. *N Engl J Med.* (2015) 372:2243–57. doi: 10.1056/NEJMsr1501341

3. Couch FJ, Shimelis H, Hu C, Hart SN, Polley EC, Na J, et al. Associations between cancer predisposition testing panel genes and breast cancer. *JAMA Oncol.* (2017) 3:1190–6. doi: 10.1001/jamaoncol.2017.0424

4. Couch FJ Nathanson KL, Offit K. Two decades after BRCA: setting paradigms in personalized cancer care and prevention. *Science* (2014) 343:1466. doi: 10.1126/science.1251827

5. Foulkes WD. BRCA1 and BRCA2—Update and implications on the genetics of breast cancer: a clinical perspective. *Clin Genet.* (2014) 85:1–4. doi: 10.1111/cge.12291

6. Antoniou AC, Foulkes WD, Tischkowitz M. Breast-cancer risk in families with mutations in PALB2. *N Engl J Med.* (2014) 371:1651–2. doi: 10.1056/NEJMc1410673

7. Girardi F, Barnes DR, Barrowdale D, Frost D, Brady AF, Miller C, et al. Risks of breast or ovarian cancer in BRCA1 or BRCA2 predictive test negatives: findings from the EMBRACE study. *Genet Med.* (2018). doi: 10.1038/gim.2018.44

8. Radice P, De Summa S, Caleca L, Tommasi S. Unclassified variants in BRCA genes: guidelines for interpretation. *Ann Oncol.* (2011) 22 (Suppl 1):i18–23. doi: 10.1093/annonc/mdq661

9. Spurdle AB. Clinical relevance of rare germline sequence variants in cancer genes: evolution and application of classification models. *Curr Opin Genet Dev.* (2010) 20:315–23. doi: 10.1016/j.gde.2010.03.009

10. Lindor NM, Guidugli L, Wang X, Vallée MP, Monteiro AN, Tavtigian S, et al. A review of a multifactorial probability based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). *Hum Mutat.* (2012) 33:8–21. doi: 10.1002/humu.21627

11. Pauty J, Rodrigue A, Couturier A, Buisson R, Masson JY. Exploring the roles of PALB2 at the crossroads of DNA repair and cancer. *Biochem J.* (2014) 460:331–42. doi: 10.1042/BJ20140208

12. Zhang F, Fan Q, Ren K, Andreassen PR. PALB2 functionally connects the breast cancer susceptibility proteins BRCA1 and BRCA2. *Mol Cancer Res.* (2009) 7:1110–8. doi: 10.1158/1541-7786.MCR-09-0123

13. Zhang F, Ma J, Wu J, Ye L, Cai H, Xia B, et al. PALB2 links BRCA1 and BRCA2 in the DNA-damage response. *Curr Biol.* (2009) 19:524–9. doi: 10.1016/j.cub.2009.02.018

14. Ceccaldi R, Sarangi P, D'Andrea AD. The Fanconi anaemia pathway: new players and new functions. *Nat Rev Mol Cell Biol.* (2016) 17:337–49. doi: 10.1038/nrm.2016.48

15. Prakash R, Zhang Y, Feng W, Jasin M. Homologous recombination and human health: the roles of BRCA1, BRCA2, and associated proteins. *Cold Spring Harb Perspect Biol.* (2015) 7:a016600. doi: 10.1101/cshperspect.a016600

16. Xia B, Sheng Q, Nakanishi K, Ohashi A, Wu J, Christ N, et al. Control of BRCA2 cellular and clinical functions by a nuclear partner, PALB2. *Mol Cell* (2006) 22:719–29. doi: 10.1016/j.molcel.2006.05.022

17. Sy SM, Huen MS, Chen J. PALB2 is an integral component of the BRCA complex required for homologous recombination repair. *Proc Natl Acad Sci USA.* (2009) 106:7155–60. doi: 10.1073/pnas.0811159106

18. Anantha RW, Simhadri S, Foo TK, Miao S, Liu J, Shen Z, et al. Functional and mutational landscapes of BRCA1 for homology-directed repair and therapy resistance. *Elife* (2017) 6:e21350. doi: 10.7554/eLife.21350

19. Foo TK, Tischkowitz M, Simhadri S, Boshari T, Zayed N, Burke KA, et al. Compromised BRCA1-PALB2 interaction is associated with breast cancer risk. *Oncogene* (2017) 36:4161–70. doi: 10.1038/onc.2017.46

20. Sarkar M, Magliery TJ. Re-engineering a split-GFP reassembly screen to examine RING-domain interactions between BARD1 and BRCA1 mutants observed in cancer patients. *Mol Biosyst.* (2008) 4:599–605. doi: 10.1039/b802481b

21. Caleca L, Putignano AL, Colombo M, Congregati C, Sarkar M, Magliery TJ, et al. Characterization of an Italian founder mutation in the RING-finger domain of BRCA1. *PLoS ONE* (2014) 9:e86924. doi: 10.1371/journal.pone.0086924

22. Southey MC, Winship I, Nguyen-Dumont T. PALB2: research reaching to clinical outcomes for women with breast cancer. *Hered Cancer Clin Pract.* (2016) 14:9. doi: 10.1186/s13053-016-0049-2

23. Catucci I, Radice P, Milne RL, Couch FJ, Southey MC, Peterlongo P. The PALB2 p.Leu939Trp mutation is not associated with breast cancer risk. *Breast Cancer Res.* (2016) 18:111. doi: 10.1186/s13058-016-0762-9

24. Oliver AW, Swift S, Lord CJ, Ashworth A, Pearl LH. Structural basis for recruitment of BRCA2 by PALB2. *EMBO Rep.* (2009) 10:990–6. doi: 10.1038/embor.2009.126

25. Hussain H, Chong NF. Combined overlap extension PCR method for improved site directed mutagenesis. *Biomed Res Int.* (2016) 2016:8041532. doi: 10.1155/2016/8041532

26. Catucci I, Peterlongo P, Ciceri S, Colombo M, Pasquini G, Barile M, et al. PALB2 sequencing in Italian familial breast cancer cases reveals a high-risk mutation recurrent in the province of Bergamo. *Genet Med.* (2014) 16:688–94. doi: 10.1038/gim.2014.13

27. Catucci I, Casadei S, Ding YC, Volorio S, Ficarazzi F, Falanga A, et al. Haplotype analyses of the c.1027C>T and c.2167_2168delAT recurrent truncating mutations in the breast cancer-predisposing gene PALB2. *Breast Cancer Res Treat.* (2016) 160:121–9. doi: 10.1007/s10549-016-3981-y

28. Park JY, Singh TR, Nassar N, Zhang F, Freund M, Hanenberg H, et al. Breast cancer-associated missense mutants of the PALB2 WD40 domain, which directly binds RAD51C, RAD51 and BRCA2, disrupt DNA repair. *Oncogene* (2014) 33:4803–12. doi: 10.1038/onc.2013.421

Identification of Eight Spliceogenic Variants in BRCA2 Exon 16 by Minigene Assays

Eugenia Fraile-Bethencourt[1], Alberto Valenzuela-Palomo[1], Beatriz Díez-Gómez[1], Alberto Acedo[1,2†] and Eladio A. Velasco[1*]

[1] Splicing and Genetic Susceptibility to Cancer, Instituto de Biología y Genética Molecular, Consejo Superior de Investigaciones Científicas, Universidad de Valladolid, Valladolid, Spain, [2] Biome Makers Inc., San Francisco, CA, United States

Genetic testing of *BRCA1* and *BRCA2* identifies a large number of variants of uncertain clinical significance whose functional and clinical interpretations pose a challenge for genetic counseling. Interestingly, a relevant fraction of DNA variants can disrupt the splicing process in cancer susceptibility genes. We have tested more than 200 variants throughout 19 BRCA2 exons mostly by minigene assays, 54% of which displayed aberrant splicing, thus confirming the utility of this assay to check genetic variants in the absence of patient RNA. Our goal was to investigate *BRCA2* exon 16 with a view to characterizing spliceogenic variants recorded at the mutational databases. Seventy-two different BIC and UMD variants were analyzed with NNSplice and Human Splicing Finder, 12 of which were selected because they were predicted to disrupt essential splice motifs: canonical splice sites (ss; eight variants) and exonic/intronic splicing enhancers (four variants). These 12 candidate variants were introduced into the BRCA2 minigene with seven exons (14–20) by site-directed mutagenesis and then transfected into MCF-7 cells. Seven variants (six intronic and one missense) induced complete abnormal splicing patterns: c.7618-2A>T, c.7618-2A>G, c.7618-1G>C, c.7618-1G>A, c.7805G>C, c.7805+1G>A, and c.7805+3A>C, as well as a partial anomalous outcome by c.7802A>G. They generated at least 10 different transcripts: $\Delta 16p_{44}$ (alternative 3'ss 44-nt downstream; acceptor variants), $\Delta 16$ (exon 16-skipping; donor variants), $\Delta 16p_{55}$ (alternative 3'ss 55-nt downstream), $\Delta 16q_4$ (alternative 5'ss 4-nt upstream), $\Delta 16q_{100}$ (alternative 5'ss 4-nt upstream), $\blacktriangledown 16q^{20}$ (alternative 5'ss 20-nt downstream), as well as minor ($\Delta 16p_{93}$ and $\Delta 16,17p_{69}$) and uncharacterized transcripts of 893 and 954 nucleotides. Isoforms $\Delta 16p_{44}$, $\Delta 16$, $\Delta 16p_{55}$, $\Delta 16q_4$, $\Delta 16q_{100}$, and $\blacktriangledown 16q^{20}$ introduced premature termination codons which presumably inactivate BRCA2. According to the guidelines the American College of Medical Genetics and Genomics these eight variants could be classified as pathogenic or likely pathogenic whereas the Evidence-based Network for the Interpretation of Germline Mutant Alleles rules suggested seven class 4 and one class 3 variants. In conclusion, our study highlights the relevance of splicing functional assays by hybrid minigenes for the clinical classification of genetic variations. Hence, we provide new data about spliceogenic variants of BRCA2 exon 16 that are directly correlated with breast cancer susceptibility.

Keywords: breast cancer, BRCA2, DNA variants, splicing, hybrid minigenes

INTRODUCTION

Hereditary Breast and Ovarian Cancer (HBOC) represents 5–10% of all breast cancers. Nowadays, more than 25 HBOC susceptibility genes have been identified, most of them involved in DNA repair pathways (Nielsen et al., 2016). Deleterious variants of the most prevalent genes *BRCA1* (MIM# 113705) and *BRCA2* (MIM# 600185) confer up to 87% of risk to develop breast cancer by the age of 70 years (Petrucelli et al., 2013). Apart from specific founder deleterious mutations (Levy-Lahad et al., 1997; Infante et al., 2013), there have been described thousands of different BRCA1/2 variants at the mutation databases. According to Universal Mutation Database (UMD, http://www.umd.be; date last accessed 2017/06/16) 2,495 and 3,454 different variants have been detected in *BRCA1* and *BRCA2*, respectively, where a relevant fraction of them has been classified as variants of uncertain significance (VUS). These pose a challenge in clinical genetics since mutation carriers could benefit from preventive and prophylactic measures as well as new targeted therapies such as the Poly-ADP Ribose Polymerase Inhibitors (Ricks et al., 2015).

Standard approaches tend to classify DNA variants from the protein point of view. In this way, nonsense variants and frameshift insertions and deletions are automatically classified as pathogenic if they truncate critical protein domains [Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) class 5[1]]. However, upstream gene expression mechanisms, such as splicing, can be disrupted by DNA changes. In fact, splicing is a critical highly regulated process involved in many cell functions whose disruption has been directly related with disease, being common in cancer (Wang and Cooper, 2007; Douglas and Wood, 2011). Likewise, spliceogenic variants are more common than they are thought, and they are not restricted to the sequences of the canonical donor and acceptor sites since it has been suggested that up to 50% of exon variants could also affect splicing (López-Bigas et al., 2005). This can be explained by the wide range of splicing regulatory elements (SREs) that control this process, which include the conserved splice sites (5'ss and 3'ss), the branch point, polypyrimidine track, exonic/intronic splicing enhancers (ESEs/ISEs) and exonic/intronic splicing silencers (ESSs/ISSs) (Grodecká et al., 2017), as well as other regulatory components or the RNA secondary structure (Soemedi et al., 2017). Thus, all these factors cooperate with splicing factors and the spliceosome, to accurately remove introns (Will and Lührmann, 2011).

Interestingly, spliceogenic variants are often found in *BRCA2*. Our previous results showed that more than a half of tested *BRCA2* variants impaired splicing (Acedo et al., 2012, 2015; Fraile-Bethencourt et al., 2017). Moreover, the minigene technology was confirmed as a reliable tool to functionally assay potential splicing variants. Here, we aimed to check *BRCA2* exon 16 candidate variants to characterize the splicing effects using the pSAD-based minigene MGBR2_14-20, previously

employed to assay DNA variants of exons 17 and 18 (Fraile-Bethencourt et al., 2017). We have assayed 12 likely spliceogenic variants from HBOC patients reported in databases and selected after bioinformatics predictions. Wild-type (wt) and mutant minigenes assays showed that eight variants altered the splicing. Thus, we provide valuable information of spliceogenic *BRCA2* exon 16 variants that could be classified following ENIGMA and American College of Medical Genetics and Genomics (ACMG) guidelines (Richards et al., 2015).

MATERIALS AND METHODS

Ethical approval for this study was obtained from the Ethics Review Committee of the Hospital Universitario Río Hortega de Valladolid (6/11/2014).

Variant Collection and *In Silico* Analyses

BRCA2 introns 15 and 16 and exon 16 variants were collected from the BIC database[2] and the BRCA Share Database (UMD, date last accessed 2017/06/16; http://www.umd.be/BRCA2/) (Beroud et al., 2016). Variant descriptions were according to the *BRCA2* GenBank sequence NM000059.1 and the guidelines of the Human Genome Variation Society (HGVS[3]).

Wild-type and mutant sequences were analyzed with NNSPLICE[4] (Reese et al., 1997) and Human Splicing Finder version 3.0 (HSF[5]) (Desmet et al., 2009), which includes algorithms to detect splice sites, branch point, silencers, and enhancers (Fairbrother et al., 2002; Cartegni et al., 2003; Sironi et al., 2004; Wang et al., 2004; Yeo and Burge, 2004; Zhang and Chasin, 2004).

Minigene and Mutagenesis

MGBR2_ex14-20 was assembled as previously described (Fraile-Bethencourt et al., 2017). DNA variants and deletions were introduced by the QuikChange Lightning Kit (Agilent, Santa Clara, CA, United States). The wt minigene MGBR2_ex14-20 was used as template to generate 12 BIC/BRCA Share DNA variants and 4 microdeletions (**Table 1**). They were checked by SANGER sequencing at the Macrogen Spain facility (Macrogen, Madrid, Spain).

Transfection of Eukaryotic Cells

MCF-7 cells were plated ($\sim$2 $\times$ 10^5 cells/well) and grown to 90% confluency in 0.5 mL of medium (MEME, 10% fetal bovine serum, 2 mM glutamine, 1% non-essential amino acids, and 1% penicillin/streptomycin) in four-well plates (Nunc, Roskilde, Denmark). Transfections were made with 1 μg of minigene and 2 μL of low toxicity Lipofectamine (Life Technologies, Carlsbad, CA, United States) in Gibco™ Opti-Mem™ (Thermo Fisher Scientific, Waltham, MA, United States). Cells were incubated with 300 μg/mL of cycloheximide (Sigma-Aldrich, St. Louis,

---

[1]https://enigmaconsortium.org/library/general-documents/enigma-classification-criteria/

[2]https://research.nhgri.nih.gov/projects/bic/Member/index.shtml
[3]http://www.hgvs.org/mutnomen/
[4]http://www.fruitfly.org/seq_tools/splice.html
[5]http://www.umd.be/HSF3/

TABLE 1 | Mutagenesis primers of candidate splicing variants.

HGVS variants	Primers (5′ → 3′)
c.7618-2A>T	GTGTGTTTATTTTGTGTTGCTGTATACGTATGGCG
	CGCCATACGTATACAGCAACACAAAATAAACACAC
c.7618-2A>G	GTGTGTTTATTTTGTGTGGCTGTATACGTATGGCG
	CGCCATACGTATACAGCCACACAAAATAAACACAC
c.7618-1G>A	GTGTGTGTTTATTTTGTGTAACTGTATACGTATGGCGTTTC
	GAAACGCCATACGTATACAGTTACACAAAATAAACACACAC
c.7618-1G>C	GTGTGTGTTTATTTTGTGTACCTGTATACGTATGGCGTTTC
	GAAACGCCATACGTATACAGGTACACAAAATAAACACACAC
c.7625C>G	TTATTTTGTGTAGCTGTATAGGTATGGCGTTTCTAAACATT
	AATGTTTAGAAACGCCATACCTATACAGCTACACAAAATAA
c.7738C>T	ATGGACTGGAAAAGGAATATAGTTGGCTGATGGTGGATGG
	CCATCCACCATCAGCCAACTATATTCCTTTTCCAGTCCAT
c.7753G>A	AATACAGTTGGCTGATGGTAGATGGCTCATACCCTCCAAT
	ATTGGAGGGTATGAGCCATCTACCATCAGCCAACTGTATT
c.7772A>G	GGATGGCTCATACCCTCCAGTGATGGAAAGGCTGGAAAAG
	CTTTTCCAGCCTTTCCATCACTGGAGGGTATGAGCCATCC
c.7802A>G	GCTGGAAAAGAAGAATTTTGTAGGTACTCTATGCAAAAG
	CTTTTTGCATAGAGTACCTACAAAATTCTTCTTTTCCAGC
c.7805G>C	GGAAAAGAAGAATTTTATACGTACTCTATGCAAAAAGATT
	AATCTTTTTGCATAGAGTACGTATAAAATTCTTCTTTTCC
c.7805+1G>A	GGAAAAGAAGAATTTTATAGATACTCTATGCAAAAAGATTG
	CAATCTTTTTGCATAGAGTATCTATAAAATTCTTCTTTTCC
c.7805+3A>C	AAGAAGAATTTTATAGGTCCTCTATGCAAAAAGATTG
	CAATCTTTTTGCATAGAGGACCTATAAAATTCTTCTT

Microdeletions	Primers (5′ → 3′)
c.7620_7649del	TTGTGTGTGTTTATTTTGTGTAGCTAAAAATTAACAGCAAA
	AATGCAGAG
	CTCTGCATTTTGCTGTTAATTTTTAGCTACACAAAATAAAC
	ACACACAA
c.7645_7674del	GTATACGTATGGCGTTTCTAAACATTCTTTTCAGTTTCACAC
	TGAAGATT
	AATCTTCAGTGTGAAACTGAAAGAATGTTTAGAAACGCCA
	TACGTATAC
c.7748_7772del	ACTGGAAAAGGAATACAGTTGGCTGGAAAGGCTGGAAAAG
	AAGAATTTTA
	TAAAATTCTTCTTTTCCAGCCTTTCCAGCCAACTGTATTCCT
	TTTCCAGT
c.7773_7802del	ATGGTGGATGGCTCATACCCTCCAATAGGTACTCTATGCAA
	AAAGATTGT
	ACAATCTTTTTGCATAGAGTACCTATTGGAGGGTATGAGCC
	ATCCACCAT

MO, United States) for 4 h to inhibit nonsense-mediated decay (NMD). RNA was purified with the Genematrix Universal RNA Purification Kit (EURx, Gdansk, Poland) with on-column DNAse I digestion to degrade genomic DNA that could interfere with RT-PCR.

RT-PCR of Minigenes

Approximately 400 ng of RNA was retrotranscribed using RevertAid H Minus First Strand cDNA Synthesis Kit (Life Technologies, Carlsbad, CA, United States) and the gene-specific primer RTPSPL3-RV (5′-TGAGGAGTGAATTGGTCGAA-3′).

Samples were incubated at 42°C for 1 h, and reactions were inactivated at 70°C for 5 min. Then, 40 ng of cDNA was amplified in 50 μL reaction with pMAD_607FW (Patent P201231427, CSIC) and RTBR2_ex17RV2 (5′-GGCTTAGGCATCTATTAGCA-3′) or with RT_ex15FW (5′-CGAATTAAGAAGAAACAAAGG-3′) and pSAD_RT_RV (Patent P201231427, CSIC) using Platinum Taq DNA polymerase (Life Technologies, Carlsbad, CA, United States) (size of transcripts: 1018 and 1250 nt, respectively). Samples were denatured at 94°C for 2 min, followed by 35 cycles consisting of 94°C for 30 s, Td-2°C for 30 s, and 72°C (1 min/kb), and a final extension step at 72°C for 5 min. Sequencing reactions were performed by the sequencing facility of Macrogen Spain. Semi-quantitative fluorescent 26 cycles PCRs were done in triplicate with primers pMAD_607FW-FAM and RTBR2_ex17RV2 using Platinum Taq DNA polymerase (Life Technologies, Carlsbad, CA, United States). FAM-labeled products were run with Genescan LIZ-1200 as size standard (Life Technologies, Carlsbad, CA, United States) at the Macrogen facility and analyzed with the Peak Scanner software V1.0. Only peaks with heights ≥50 relative fluorescence unit (RFU) were considered. Mean peak areas of each transcript of three runs were used to quantify the relative abundance of each transcript.

RESULTS

Bioinformatics Analysis of Splicing Variants

Seventy-two variants were collected from the BIC and UMD databases. Among them, 35 had been previously classified as VUS and 34 as pathogenic or likely pathogenic. In order to select possible spliceogenic variants, they were analyzed by the splicing prediction software NNSplice and HSF (Supplementary Table S1). Selections were made following the next criteria: (a) ss creation or disruption; (b) branch point disruption; and (c) ESS creation (hnRNPA1). Curiously, NNSplice did not recognize exon 16 canonical 5′ss. In contrast, a very strong 100-nt upstream cryptic donor (NNSplice score: 0.99) was identified at position c.7706_7707. The MaxEnt results showed a weak canonical donor (4.68) and a strong cryptic donor (8.92) (**Table 2**).

Twelve variants were selected (**Figure 1**): six intronic (c.7618-2A>T, c.7618-2A>G, c.7618-1G>C, c.7618-1G>A, c.7805+1G>A, and c.7805+3A>C), five missense (c.7625C>G, c.7753G>A, c.7772A>G, c.7802A>G, and c.7805G>C), and one nonsense (c.7738C>T) variants. Four missense variants (c.7625C>G, c.7753G>A, c.7772A>G, and c.7802A>G) and c.7805+3A>C had been previously classified as VUS. Intronic variants (c.7618-2, -1 and c.7805+1, +3) and c.7805G > C disrupted the canonical ss, whereas variants c.7625C>G and c.7753G>A created new ss. DNA change c.7802A>G was selected because of its proximity to the canonical donor site and the presumable generation of an alternative "gt" donor site 4-nt upstream (underlined, TTT<u>GT</u>AGgtactc). Finally, bioinformatics results of c.7738C>T and c.7772A>G suggested the creation of one ESS (hnRNPA1) (**Table 2**).

TABLE 2 | Bioinformatics analysis of potential splicing variants of *BRCA2* exon 16.

DNA variants	NNSplice[1]	MaxEnt[1]	ESE finder[2]	ESE HSF[2]	hnRNPA1[2]
Ex16-Canonical ss	3′SS: 0.69	3′SS: 7.11			
	5′SS <0.4	5′SS: 4.68			
c.7618-2A>T	[−] 3′SS: 0.69 → <0.4	[−] 3′SS: 7.11 → −1.24			[−] (70.24)
c.7618-2A>G	[−] 3′SS: 0.69 → <0.4	[−] 3′SS: 7.11 → −0.83	[−] SRp55 (83.78)		[−] (70.24)
c.7618-1G>A	[−] 3′SS: 0.69 → <0.4	[−] 3′SS: 7.11 → −1.63	[−] SRp55 (77.06)		[−] (70.24)
c.7618-1G>C	[−] 3′SS: 0.69 → <0.4	[−] 3′SS: 7.11 → −0.94	[−] SRp55 (77.06)		[−] (70.24)
c.7625C>G (p.Thr2542Arg)	[+] 3′SS (<0.4 → 0.94); [+] 5′SS (<0.4 → 0.91)	[+] 5′SS (2.23 → 7.37)	[−] SRp55(89.80)	[−] 9G8 (62.28)	[+] (88.33)
c.7738C>T (p.Gln2580Ter)			[−] SC35 (81.01)		[+] (70.24)
c.7753G>A (p.Gly2585Arg)		[+] 5′SS (−1.33 → 3.06)			[+] (74.05)
c.7772A>G (p.Asn2591Ser)			[+] SF2/ASF (IgM) (74.85)	[−] 9G8 (60.94)	[+] (69.52)
c.7802A>G (p.Tyr2601Cys)			[−] SRp40 (83.89)		[−] (66.43)
c.7805G>C (p.Arg2602Thr)		[−] 5′SS: 4.68 → −2.79	[+] SRp55 (89.80)		[−] (69.05)
c.7805+1G>A		[−] 5′SS: 4.68 → −3.50	[−] SRp40 (83.89); [−] SC35 (82.36)		
c.7805+3A>C		[−] 5′SS: 4.68 → −3.85			

[+] and [−] symbols indicate creation or disruption of splicing motifs, respectively. Thresholds of the splicing programs: [1]Splice sites (ss): NNSPLICE (values 0–1): Cut-offs = 0.4 for both 5′- and 3′ ss (ss disruption <0.4, ss creation >0.4); MaxEnt: 3.0 for 5′ and 3′ ss (variation threshold ± 30% according to HSF). [2]Enhancers and silencers HSF: Human Splicing Finder matrices (default values, http://www.umd.be/HSF3/technicaltips.html): ESEfinder cut-offs (HSF scale, normalized to 0–100): SF2/ASF: 72.98/SF2/ASF (IgM-BRCA1): 70.51/SRp40: 78.08/SC35: 75.05/SRp55: 73.86; ESE motifs from HSF, cut-offs values (0–100): Tra2: 75.964/9G8: 59.245; hnRNP motifs: hnRNPA1: 65.476 (these values were considered the limits for the disruption or creation of a splicing regulatory element).

FIGURE 1 | Functional assays of spliceogenic candidate variants of *BRCA2* exon 16. (A) Nucleotide (c.7618_7805) and amino acid (p.2540_2602) sequences of *BRCA2* exon 16. Arrows indicate selected variants. (B) Agarose gel electrophoresis of RT-PCR products of the wt and mutant minigenes and the size standard 1 Kb Plus DNA Ladder at both sides of the gel. Amplification was made with primers pMAD607-FW and RTBR2_ex17RV2. Full-length transcript (V1-EX17) size: 1018 nt. Red arrows point to exon 16 skipping band (Δ16) (size: 830 nt).

Splicing Functional Assays of DNA Variants

The minigene MGBR2_ex14-20 had been already shown as a robust tool to assay possible spliceogenic variants contained in any of those exons and flanking introns (Fraile-Bethencourt et al., 2017). The wt construct produced a full-length transcript of the expected size (1806 nt), sequence, and structure (V1-BRCA2 exons 14-20-V2). To map the presence of putative splicing enhancers, a set of four overlapping exonic microdeletions were generated, which spanned 55-nt of the 5′- and 3′-ends (Fairbrother et al., 2004). This strategy had been previously shown to increase the accuracy of predictions of ESE disrupting variants (Acedo et al., 2015; Fraile-Bethencourt et al., 2017). None of the microdeletions induced splicing anomalies suggesting that this exon is not controlled by ESEs (data not shown). Consequently, ESE-disrupting variants, as unique selection criterion, were not chosen for subsequent functional tests (**Table 2**).

Selected variants were introduced into the minigene and functionally assayed in MCF-7 cells. Agarose electrophoresis clearly showed that three of them (c.7805G>C, c.7805+1G>A, and c.7805+3A>C) induced aberrant splicing patterns (**Figure 1**). However, the high resolution and sensitivity of fluorescent capillary electrophoresis allowed us to identify a total of eight variants, including the three previous ones, that disrupted splicing: c.7618-2A>T, c.7618-2A>G, c.7618-1G>C, c.7618-1G>A, c.7802A>G, c.7805G>C, c.7805+1G>A, and c.7805+3A>C (**Figure 2**). Actually, this approach is able to detect rare transcripts with a relative abundance below 1% or can resolve transcripts that differ only in a few nucleotides (e.g., only 4-nt between the canonical and Δ16q4 isoforms). A total of at least 10 different aberrant transcripts were characterized

by fragment analysis and sequencing: Δ16p44 (44-nt deletion; alternate 3′ss 44-nt downstream), Δ16p55 (55-nt del; alternate 3′ss 55-nt downstream), Δ16 (exon 16 skipping), Δ16q4 (4-nt del; alternate 5′ss 4-nt upstream), Δ16q100 (100-nt del; alternate 5′ss 100-nt upstream), $\blacktriangledown$16q20 (20-nt insertion; alternate 5′ss 20-nt downstream), minor (Δ16p93 and Δ16,17p69), and uncharacterized transcripts of 893 and 954 nt (**Figure 2** and **Table 3**). On the one hand, fragments analysis and sequences revealed that 3′ss disrupting variants (positions −2 and −1) provoked the use of a cryptic acceptor 44-nt downstream (Δ16p44) within exon 16 (**Figure 2A**). Interestingly, this cryptic 3′ss was not recognized either by NNSplice or MaxtEnt. The loss of 44-nt at 5′ of exon 16 would suppose a frameshift deletion and a premature termination codon (PTC) (p.L2540Qfs*11).
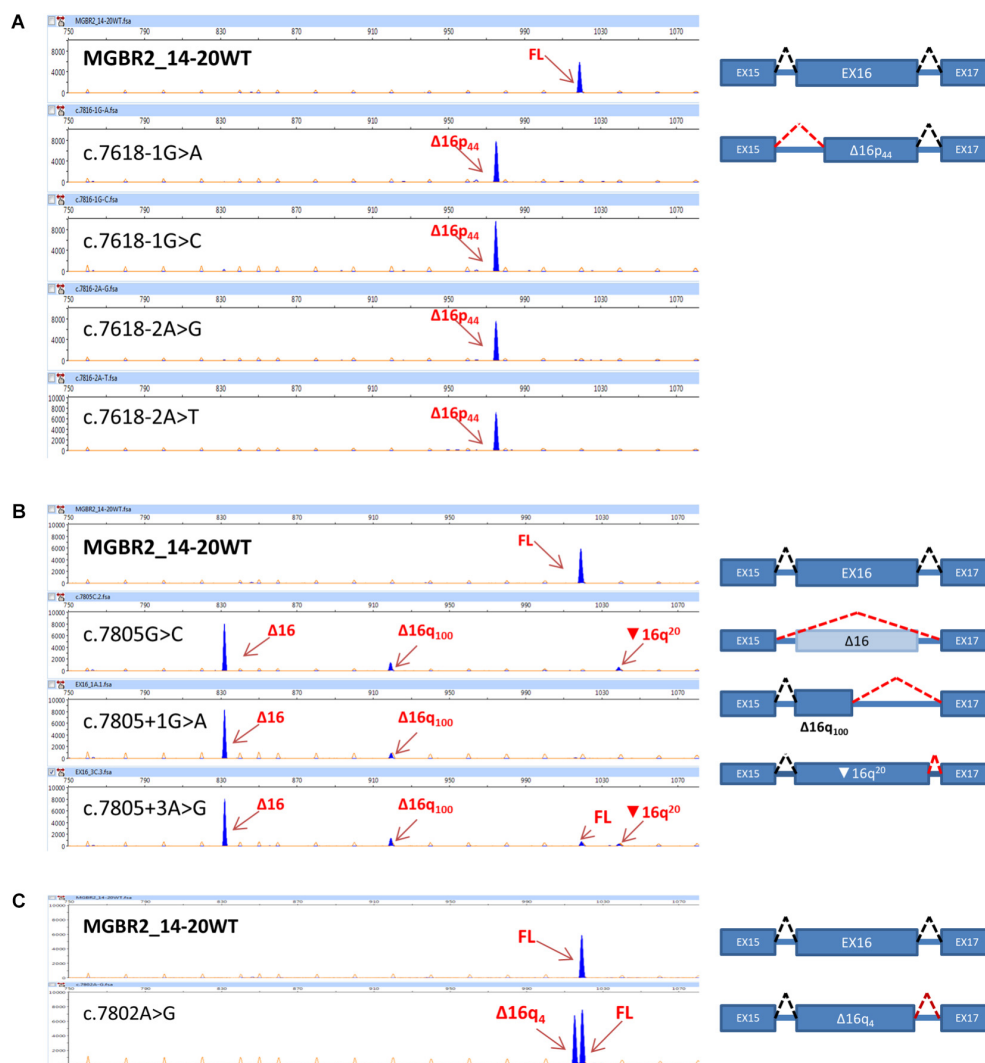
FIGURE 2 | Fluorescent capillary electrophoresis of transcripts from *BRCA2* exon 16 variants. On the left, screenshots of electropherograms are shown. cDNA was amplified with primers FAM-labeled pMAD_607FW and RTBR2_ex17RV2. Arrows indicate transcripts (blue peaks). Full-length transcript: 1018 nt. Size standard was Genescan LIZ 1200 (orange/faint peaks). Fragments were analyzed with the Peak Scanner software v1.0. Fragment sizes (bp) and relative fluorescent units are indicated on the *x*- and *y*-axes, respectively. On the right, diagrams of the splicing patterns are shown. Boxes represent exons, discontinued black lines represent canonical splicing, and discontinue red lines represent aberrant splicing. **(A)** Acceptor site variants. **(B)** Donor site variants. **(C)** Alternative donor variant.

TABLE 3 | Quantification of transcripts of spliceogenic variants of exon 16 by fluorescent capillary electrophoresis.

	Canonical	Δ16p$_{44}$	Δ16p$_{55}$	Δ16	Δ16q$_4$	Δ16q$_{100}$	▼16q^{20}	Others
MGBR2_14-20	100%							
DNA variants								
c.7618-2A>T		96.9 ± 0.4%	1.8 ± 0.3%	1.3 ± 0.03%				
c.7618-2A>G		97.2 ± 0.4%						2.8 ± 0.4% (954 nt)
c.7618-1G>A		91.5 ± 0.2%	4.7 ± 0.2%	0.7 ± 0.03%				1.7 ± 0.03% (Δ16p$_{93}$);
								1.4 ± 0.04% (Δ16,17p$_{69}$)
c.7618-1G>C		92.6 ± 0.3%	1.9 ± 0.2%	2.8 ± 0.05%				1.1 ± 0.05% (Δ16p$_{93}$)
								0.7 ± 0.03% (893 nt)
								0.9 ± 0.07% (Δ16,17p$_{69}$)
c.7625C>G	100%							
c.7738C>T	100%							
c.7753G>A	100%							
c.7772A>G	100%							
c.7802A>G	54.3 ± 0.3%				45.7 ± 0.3%			
c.7805G>C				77.6 ± 0.6%		14.4 ± 0.3%	6.5 ± 0.3%	1.5 ± 0.06% (Δ16,17p$_{69}$)
c.7805+1G>A				88.0 ± 0.2%		10.1 ± 0.1%		1.9 ± 0.02% (Δ16,17p$_{69}$)
c.7805+3A>C	7.6 ± 0.2%			75.3 ± 0.5%		13.3 ± 0.2%	3.8 ± 0.3%	

HGVS-RNA effect of transcripts: Δ16p$_{44}$, r.7618_7661del; Δ16q$_4$, r.7802_7805del; Δ16, r.7618_7805del; Δ16q$_{100}$, r.7706_7805del; Δ16p$_{55}$, r.7618_7672del; ▼16q^{20}, r.7805_7806ins7805+1_7805+20; Δ16p$_{93}$, r.7618_7710del; Δ16,17p$_{69}$, r.7618_7874del.

On the other hand, 5′ss variants (positions +1 and +3) produced exon 16 skipping (Δ16), which means a frameshift deletion through the loss of 188-nt from r.7618 to r.7805 (**Figure 2B**). Consequently, BRCA2 would be truncated with a PTC four codons downstream (p.L2540Gfs*4). Last exon nucleotide variant (c.7805G>C) induced the same outcome (Δ16) highlighting the importance of this position conservation (G in nearly 80% in all exons) in exon recognition (Zhang, 1998). Fragment analysis of variants c.7805G>C, c.7805+1G>A, and c.7805+3A>C also showed ∼14% of transcript Δ16q$_{100}$, which corresponds with the use of the previously mentioned cryptic 5′ss within exon 16 (NNSplice: 0.99; MaxEnt: 8.92), provoking r.7706_7805del (p.K2570Lfs*45) (**Table 3** and **Figure 2B**). Finally, missense variant c.7802A>G created a new 5′ss, which resulted in ∼45% of the aberrant transcript Δ16q$_4$ (**Figure 2C**). The loss of four nucleotides would introduce a PTC into the protein (p.Y2601Wfs*46). Thus, our results showed clearly how these eight variants disrupted splicing. Moreover, seven of them (c.7618-2A>T; c.7618-2A>G; c.7618-1G>C; c.7618-1G>A; c.7805G>C; c.7805+1G>A; c.7805+3A>C) generated more than ∼92% of frameshift transcripts.

DISCUSSION

Nowadays, with the advent of new generation sequencing technologies and, namely, cancer-gene panels (Slavin et al., 2015), thousands of variants are being described. However, their classifications as neutral or deleterious variants pose a challenge in Human Genetics. In fact, some deleterious variants can be missed because they are synonymous or intronic. Moreover, a significant fraction of *BRCA2* variants are considered VUS and require additional proofs to be reclassified, including functional tests. Here, we have shown that the minigene MGBR2_14-20 is a robust tool to functionally assay candidate spliceogenic variants of the *BRCA2* exon 16. Until now, we have comprehensively studied candidate splicing variants from 20 out of 27 *BRCA2* exons (Sanz et al., 2010; Acedo et al., 2012, 2015; Fraile-Bethencourt et al., 2017). Thus, we have found six intronic and two missense *BRCA2* variants which alter the splicing and could confer cancer risk.

BRCA2 exon 16 codifies from Leucine 2540 to Arginine 2602 (p.2540_2602). Interestingly, according to the International Agency for Research on Cancer (IARC[6]), this is a conserved region, since there is ∼22% of ultra-conserved aminoacids from human to sea urchin and ∼54% between mammals. Furthermore, this protein segment belongs to FANCD2- and DSS1-binding domains. Fanconi Anemia group D2 (FANCD2) protein binds to aminoacids from position p.2350 to p.2545 of BRCA2 and it has been suggested to have a role in the repair process (Hussain et al., 2004). DSS1 (Delete in Split hand/Split foot) protein, which binds to BRCA2 at positions p.2467_2957 (Marston et al., 1999), is an essential element of BRCA2 stability, since its loss supposes a dramatic decrease of BRCA2 levels (Li et al., 2006). Altogether, this highlights the value of exon 16 in BRCA2 function. Moreover, exon 16 skipping supposes a frame-shift deletion and the generation of a PTC (p.L2540Gfs*4), which would truncate the protein and subsequently loss the C-terminal region that would compromise BRCA2 function.

This study, based on minigene technology, provides detailed information about the impact on splicing of 12 *BRCA2* exon 16 variants. Aberrant splicing outcomes were found in eight of these variants, six intronic and two missense changes. Intriguingly, none of the aberrant transcripts described here was previously reported as natural alternative splicing events of the *BRCA2*

---

[6]http://agvgd.hci.utah.edu/BRCA2_Spur.html

gene (Fackenthal et al., 2016). Among them, seven (c.7618-2A>T, c.7618-2A>G, c.7618-1G>A, c.7618-1G>C, c.7805G>C, c.7805+1G>A, and c.7805+3A>C) provoked more than ∼92% of frameshift transcripts. Interestingly, previous studies of variant c.7618-1G>A in lymphoblastoid cells showed that 3′ss disruption induced transcripts Δ16p$_{44}$ and Δ16,17p$_{69}$ (Whiley et al., 2011). Here we found both transcripts, but also other minor ones: Δ16p$_{55}$, Δ16p$_{93}$, Δ16 (Table 3). Additionally, according to our data Δ16p$_{44}$ is the main transcript (∼91%) that other authors also identified but described as a minor transcript in agarose gels (Whiley et al., 2011). These differences could be due to: (i) the cell line; (ii) the use of cycloheximide to inhibit the NMD; (iii) the fact that we work with a single-mutant allele, avoiding the wt counterpart effect; and (iv) the high sensitivity of fluorescent capillary electrophoresis, which can detect rare transcripts versus agarose electrophoresis. In any case, both results show that c.7618-1G>A severely disrupted splicing. On the other hand, variant c.7805G>C was previously reported to result in Δ16 and Δ16q$_{100}$, with the total absence of the canonical transcript (Bonnet et al., 2008). This outcome matches our results (Table 3): Δ16 as the main transcript (∼78%), followed by Δ16q$_{100}$ (∼14%), and the lack of the full-length transcript. It is also worthy to mention that we detected other minor transcripts due to the high sensitivity of fluorescent capillary electrophoresis (▼16q$_{20}$ at ∼6.5% and Δ16,17p$_{69}$ at ∼1.5%) that otherwise could not be easily detected on agarose gels. In any case, the spliceogenic effects of variants c.7618-1G>A and c.7805G>C were supported by our data.

Variant c.7802A>G probably generated the most conflicting result since it triggered ∼54% of canonical transcript and ∼46% of Δ16q$_4$, so that its interpretation is more complex. The transcript Δ16q$_4$, caused by the use of a new 5′ss, generated a frameshift deletion and the protein truncation by a PTC 46 codons downstream (p.Y2601Wfs*46). However, it is still unclear if ∼54% of full-length transcript can preserve BRCA2 function, given that, for example, 20–30% of BRCA1 transcript is able to maintain BRCA1 activity (de la Hoya et al., 2016). It is also important to keep in mind that full-length transcript carries a missense variant (p.Y2601W) that, according to IARC alignment[7], Tyrosine 2601 is highly conserved from human to sea urchin, suggesting an important function in the protein. Moreover, PolyPhen-2 (Adzhubei et al., 2010) predicted that this aminoacid change is damaging with the maximum score (1.0). Curiously, c.7802A>G was reported a family with a significant history of primary cancers (colorectal, lymphoma, and breast cancers) which carried biallelic BRCA2 mutations (c.7802A>G and c.1845_1856delCT). However, patients did not present the typical FA phenotype, which suggested that p.Y2601W BRCA2 maintained at least enough BRCA2 activity to prevent early childhood FA features (Degrolard-Courcet et al., 2014). Nevertheless, this missense change remains classified as VUS in ClinVar[8].

On the other hand, variant c.7625C>G was previously computed to disrupt one SRp55 motif (Pettigrew et al., 2008),

although functional mapping by microdeletions indicated that exon 16 is likely not regulated by splicing enhancers. Nevertheless, this change was selected because it presumably created new strong 3′ and 5′ss as well, both with a NNSplice score >0.9 (Table 2). However, c.7625C>G only produced the full-length transcript without any splicing anomaly. The protein would even carry the missense variant p.T2542R. However, consistent with PolyPhen, this change might be considered as benign with a score of 0.0, which could be explained by the low conservation of the affected threonine. Anyway, further functional and association studies must be performed to interpret this variant. Other variant that resulted in a normal splicing pattern was the nonsense variant c.7738C>T (p.Q2580X), that *a priori* had been classified as pathogenic. In this case, the protein would be truncated at codon 2580 losing 839 aminoacids of the C-terminal where the DSS1-binding site, the DNA-binding domain, the RAD51C-binding site, and the cyclin-dependent kinase (CDK) phosphorylation site are located (Roy et al., 2012). Interestingly, this variant was found in an Italian non-Ashkenazi BRCA1 and BRCA2 double heterozygote family (Musolino et al., 2005).

According to the ACMG guidelines (Table 4; Richards et al., 2015), five variants (c.7618-2A>T, c.7618-2A>G, c.7618-1G>A, c.7618-1G>C, and c.7805+1G>A) can be classified as pathogenic as they match criteria PVS1 (very strong evidence of pathogenicity: null variant – nonsense, frameshift, canonical ±1 or 2 ss, initiation codon, single or multiexon deletion – in a gene where LOF is a known mechanism of disease), PS3 (strong evidence: well-established *in vitro* or *in vivo* functional studies supportive of a damaging effect on the gene or gene product), PM2 (moderate evidence: absent from controls in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium), PP3 (supporting evidence: multiple lines of computational evidence support a deleterious effect on the gene or gene product: conservation, evolutionary, splicing impact, etc.), and PP5 (reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation). On the other hand, variants c.7802A>G, c.7805G>C, and c.7805+3A>C were classified as likely pathogenic as they match criteria PS3, PM2, PP3, and PP5.

Similarly, following the ENIGMA rules for variant classification[9], all variants, except for c.7802A>G, should be reclassified as class 4 (likely pathogenic) because they are "considered extremely likely to alter splicing based on position" and are "predicted bioinformatically to alter the use of the native donor/acceptor site." Conversely, minigenes are not considered robust approaches to functionally test these variants yet ("... results from construct-based mRNA assays alone are not considered sufficiently robust to be used as evidence for variant classification ..."). However, this specific minigene with BRCA2 exons 14–20 was confirmed as a robust tool since it reproduced patient RNA results from eight variants (Fraile-Bethencourt et al., 2017), and also c.7618-1G>A and c.7805G>C of this study, so

---

[7]http://agvgd.hci.utah.edu/BRCA2_Spur.html

[8]https://www.ncbi.nlm.nih.gov/clinvar/variation/185651/#clinical-assertions

[9]https://enigmaconsortium.org/library/general-documents/enigma-classification-criteria/

TABLE 4 | Classification of variants according to the ENIGMA and ACMG rules.

DNA variants	Main aberrant transcripts[1]	HGVS RNA/protein effects	Previous classification[2]	ACMG classification[3]	Enigma classification[4]
c.7618-2A>T	$\Delta16p_{44}$	r.7618_7661del/p.L2540Qfs*11	5 – Causal	Pathogenic: PVS1, PS3, PM2, PP3, PP5	Class 4
c.7618-2A>G	$\Delta16p_{44}$	r.7618_7661del/p.L2540Qfs*11	5 – Causal	Pathogenic: PVS1, PS3, PM2, PP3, PP5	Class 4
c.7618-1G>A	$\Delta16p_{44}$	r.7618_7661del/p.L2540Qfs*11	5 – Causal	Pathogenic: PVS1, PS3, PM2, PP3, PP5	Class 4
c.7618-1G>C	$\Delta16p_{44}$	r.7618_7661del/p.L2540Qfs*11	5 – Causal	Pathogenic: PVS1, PS3, PM2, PP3, PP5	Class 4
c.7802A>G	$\Delta16q_4$	r.7802_7805del/p.Y2601Wfs*46	3 – VUS	Likely pathogenic: PS3, PM2, PP3, PP5	Class 3
c.7805G>C	$\Delta16$, $\Delta16q_{100}$	r.[7618_7805del; 7706_7805del]/ p.[L2540Gfs*4; K2570Lfs*45]	5 – Causal	Likely pathogenic: PS3, PM2, PP3, PP5	Class 4
c.7805+1G>A	$\Delta16$, $\Delta16q_{100}$	r.[7618_7805del; 7706_7805del]/ p.[L2540Gfs*4; K2570Lfs*45]	5 – Causal	Pathogenic: PVS1, PS3, PM2, PP3, PP5	Class 4
c.7805+3A>C	$\Delta16$, $\Delta16q_{100}$	r.[7618_7805del; 7706_7805del]/ p.[L2540Gfs*4; K2570Lfs*45]	3 – VUS	Likely pathogenic: PS3, PM2, PP3, PP5	Class 4

[1]*Transcripts were annotated according to previous reports of the ENIGMA consortium (Colombo et al., 2014).* Δ, *skipping or deletion;* ▼, *insertion; p: alternative acceptor site; q, alternative donor site; subscript number, number of deleted nt; superscript number, number of inserted nt. Thus,* $\Delta16$ *indicates exon 16 skipping and* $\Delta16p_{44}$ *indicates loss of 44 nt at the exon 16 acceptor site (or new acceptor site 44 nt upstream). All transcripts and their quantification data are described in **Table 2**.* [2]*Previous classifications according to the BRCA share (c.7618-2A>T, c.7618-2A>G, c.7618-1G>A, c.7618-1G>C, c.7802A>G, and c.7805G>C) and the BIC mutation databases (c.7805+1G>A and c.7805+3A>C).* [3]*ACMG criteria: PVS1, null variant (nonsense, frameshift, canonical ± 1 or 2 ss, etc.) in a gene where LOF is a known mechanism of disease; PS3, well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product; PM2, absent from controls in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium; PP3, multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.); PP5, reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation.* [4]*ENIGMA. Class 4: Nucleotide positions that are considered extremely likely to alter splicing and/or variants are predicted bioinformatically to alter the use of the native donor/acceptor site. Class 3: "In the absence of clinical evidence to assign an alternative classification, variant allele tested for mRNA aberrations . . . is found to produce mRNA transcript(s) predicted to encode intact full-length protein . . . ."*

that these seven class 4 variants could be even reclassified as class 5. Finally, c.7802A>G was classified as class 3 because it did not meet the above standards and induce a partial aberrant outcome with more than 50% of the canonical transcript.

In summary, we detected eight spliceogenic *BRCA2* exon 16 variants that should be classified as pathogenic or likely pathogenic according to the ACMG guidelines (**Table 4**). Moreover, they account for 22% of causal variants of exon 16 and 11% of all recorded variants of this exon at the mutation databases. Taken together this and our previous studies, we have tested 283 *BRCA1/2* variants under the splicing perspective, 154 of which induced anomalous patterns and 111 could be classified as pathogenic or likely pathogenic. These data remark the importance of variants of splicing regulatory sequences, which are often underestimated because most of them are placed in non-coding regions of the protein. Until now, genetic family-based studies have set up the impact of some variants on cancer risk. However, because of the exponential increment in the number of variants, their low frequencies and different nature, functional assays are strictly required. In this context, minigene technology constitutes a robust tool which can be used to functionally test spliceogenic candidate variants of any disease-gene without the interference of the counterpart wt allele. Certainly, pSAD-based minigenes represented valuable tools to functionally check variants of the SERPINA1 (severe alpha-1 antitrypsin deficiency) and CHD7 (Charge Syndrome) genes (Lara et al., 2014; Villate et al., 2018). RNA assays provide essential data for the initial characterization of VUS and improve the genetic counseling of hereditary diseases.

AUTHOR CONTRIBUTIONS

EF-B contributed to the bioinformatics analysis, minigene construction, manuscript writing, and performed most of the splicing functional assays. BD-G and AV-P participated in minigene construction, mutagenesis experiments, and functional assays. AA participated in minigene construction and functional mapping experiments. EV conceived the study and the experimental design, supervised all the experiments, and wrote the manuscript. All authors contributed to data interpretation, revisions of the manuscript, and approved the final version of the manuscript.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00188/full#supplementary-material

REFERENCES

Acedo, A., Hernández-Moro, C., Curiel-García, Á, Díez-Gómez, B., and Velasco, E. A. (2015). Functional classification of BRCA2 DNA variants by splicing assays in a large minigene with 9 exons. Hum. Mutat. 36, 210–221. doi: 10.1002/humu.22725

Acedo, A., Sanz, D. J., Durán, M., Infante, M., Pérez-Cabornero, L., Miner, C., et al. (2012). Comprehensive splicing functional analysis of DNA variants of the BRCA2 gene by hybrid minigenes. Breast Cancer Res. 14:R87. doi: 10.1186/bcr3202

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249. doi: 10.1038/nmeth0410-248

Beroud, C., Letovsky, S. I., Braastad, C. D., Caputo, S. M., Beaudoux, O., Bignon, Y. J., et al. (2016). BRCA share: a collection of clinical BRCA gene variants. Hum. Mutat. 37, 1318–1328. doi: 10.1002/humu.23113

Bonnet, C., Krieger, S., Vezain, M., Rousselin, A., Tournier, I., Martins, A., et al. (2008). Screening BRCA1 and BRCA2 unclassified variants for splicing mutations using reverse transcription PCR on patient RNA and an ex vivo assay based on a splicing reporter minigene. J. Med. Genet. 45, 438–446. doi: 10.1136/jmg.2007.056895

Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q., and Krainer, A. R. (2003). ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res. 31, 3568–3571. doi: 10.1093/nar/gkg616

Colombo, M., Blok, M. J., Whiley, P., Santamariña, M., Gutiérrez-Enríquez, S., Romero, A., et al. (2014). Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium. Hum. Mol. Genet. 23, 3666–3680. doi: 10.1093/hmg/ddu075

de la Hoya, M., Soukarieh, O., López-Perolio, I., Vega, A., Walker, L. C., van Ierland, Y., et al. (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, 2256–2268. doi: 10.1093/hmg/ddw094

Degrolard-Courcet, E., Sokolowska, J., Padeano, M. M., Guiu, S., Bronner, M., Chery, C., et al. (2014). Development of primary early-onset colorectal cancers due to biallelic mutations of the FANCD1/BRCA2 gene. Eur. J. Hum. Genet. 22, 979–987. doi: 10.1038/ejhg.2013.278

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 37:e67. doi: 10.1093/nar/gkp215

Douglas, A. G., and Wood, M. J. (2011). RNA splicing: disease and therapy. Brief Funct. Genomics 10, 151–164. doi: 10.1093/bfgp/elr020

Fackenthal, J. D., Yoshimatsu, T., Zhang, B., de Garibay, G. R., Colombo, M., De Vecchi, G., et al. (2016). Naturally occurring BRCA2 alternative mRNA splicing events in clinically relevant samples. J. Med. Genet. 53, 548–558. doi: 10.1136/jmedgenet-2015-103570

Fairbrother, W. G., Holste, D., Burge, C. B., and Sharp, P. A. (2004). Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol. 2:E268. doi: 10.1371/journal.pbio.0020268

Fairbrother, W. G., Yeh, R.-F., Sharp, P. A., and Burge, C. B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science 297, 1007–1013. doi: 10.1126/science.1073774

Fraile-Bethencourt, E., Díez-Gómez, B., Velásquez-Zapata, V., Acedo, A., Sanz, D. J., and Velasco, E. A. (2017). Functional classification of DNA variants by hybrid minigenes: Identification of 30 spliceogenic variants of BRCA2 exons 17 and 18. PLoS Genet. 13:e1006691. doi: 10.1371/journal.pgen.1006691

Grodecká, L., Buratti, E., and Freiberger, T. (2017). Mutations of pre-mRNA splicing regulatory elements: Are predictions moving forward to clinical diagnostics? Int. J. Mol. Sci. 18:E1668. doi: 10.3390/ijms18081668

Hussain, S., Wilson, J. B., Medhurst, A. L., Hejna, J., Witt, E., Ananth, S., et al. (2004). Direct interaction of FANCD2 with BRCA2 in DNA damage response pathways. Hum. Mol. Genet. 13, 1241–1248. doi: 10.1093/hmg/ddh135

Infante, M., Durán, M., Acedo, A., Sánchez-Tapia, E. M., Díez-Gómez, B., Barroso, A., et al. (2013). The highly prevalent BRCA2 mutation c.2808_2811del (3036delACAA) is located in a mutational hotspot and has multiple origins. Carcinogenesis 34, 2505–2511. doi: 10.1093/carcin/bgt272

Lara, B., Martínez, M. T., Blanco, I., Hernández-Moro, C., Velasco, E. A., Ferrarotti, I., et al. (2014). Severe alpha-1 antitrypsin deficiency in composite heterozygotes inheriting a new splicing mutation QOMadrid. Respir. Res. 15:125. doi: 10.1186/s12931-014-0125-y

Levy-Lahad, E., Catane, R., Eisenberg, S., Kaufman, B., Hornreich, G., Lishinsky, E., et al. (1997). Founder BRCA1 and BRCA2 mutations in Ashkenazi Jews in Israel: frequency and differential penetrance in ovarian cancer and in breast-ovarian cancer families. Am. J. Hum. Genet. 60, 1059–1067.

Li, J., Zou, C., Bai, Y., Wazer, D. E., Band, V., and Gao, Q. (2006). DSS1 is required for the stability of BRCA2. Oncogene 25, 1186–1194. doi: 10.1038/sj.onc.1209153

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G., and Guigó, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett. 579, 1900–1903. doi: 10.1016/j.febslet.2005.02.047

Marston, N. J., Richards, W. J., Hughes, D., Bertwistle, D., Marshall, C. J., and Ashworth, A. (1999). Interaction between the product of the breast cancer susceptibility gene BRCA2 and DSS1, a protein functionally conserved from yeast to mammals. Mol. Cell. Biol. 19, 4633–4642. doi: 10.1128/MCB.19.7.4633

Musolino, A., Naldi, N., Michiara, M., Bella, M. A., Zanelli, P., Bortesi, B., et al. (2005). A breast cancer patient from Italy with germline mutations in both the BRCA1 and BRCA2 genes. Breast Cancer Res. Treat. 91, 203–205. doi: 10.1007/s10549-004-7704-4

Nielsen, F. C., van Overeem Hansen, T., and Sørensen, C. S. (2016). Hereditary breast and ovarian cancer: new genes in confined pathways. Nat. Rev. Cancer 16, 599–612. doi: 10.1038/nrc.2016.72

Petrucelli, N., Daly, M. B., and Feldman, G. L. (2013). BRCA1 and BRCA2 Hereditary Breast and Ovarian Cancer. Seattle, WA: University of Washington.

Pettigrew, C. A., Wayte, N., Wronski, A., Lovelock, P. K., Spurdle, A. B., and Brown, M. A. (2008). Colocalisation of predicted exonic splicing enhancers in BRCA2 with reported sequence variants. Breast Cancer Res. Treat. 110, 227–234. doi: 10.1007/s10549-007-9714-5

Reese, M. G., Eeckman, F. H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. J. Comput. Biol. 4, 311–323. doi: 10.1089/cmb.1997.4.311

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., 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, 405–423. doi: 10.1038/gim.2015.30

Ricks, T. K., Chiu, H.-J., Ison, G., Kim, G., McKee, A. E., Kluetz, P., et al. (2015). Successes and challenges of PARP inhibitors in cancer therapy. Front. Oncol. 5:222. doi: 10.3389/fonc.2015.00222

Roy, R., Chun, J., and Powell, S. N. (2012). BRCA1 and BRCA2: different roles in a common pathway of genome protection. Nat. Rev. Cancer 12, 68–78. doi: 10.1038/nrc3181

Sanz, D. J., Acedo, A., Infante, M., Durán, M., Pérez-Cabornero, L., Esteban-Cardeñosa, E., et al. (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. doi: 10.1158/1078-0432.CCR-09-2564

Sironi, M., Menozzi, G., Riva, L., Cagliani, R., Comi, G. P., Bresolin, N., et al. (2004). Silencer elements as possible inhibitors of pseudoexon splicing. Nucleic Acids Res. 32, 1783–1791. doi: 10.1093/nar/gkh341

Slavin, T. P., Niell-Swiller, M., Solomon, I., Nehoray, B., Rybak, C., Blazer, K. R., et al. (2015). Clinical application of multigene panels: challenges of next-generation counseling and cancer risk management. *Front. Oncol.* 5:208. doi: 10.3389/fonc.2015.00208

Soemedi, R., Cygan, K. J., Rhine, C. L., Glidden, D. T., Taggart, A. J., Lin, C.-L., et al. (2017). The effects of structure on pre-mRNA processing and stability. *Methods* 125, 36–44. doi: 10.1016/j.ymeth.2017.06.001

Villate, O., Ibarluzea, N., Fraile-bethencourt, E., and Valenzuela, A. (2018). Functional analyses of a novel splice variant in the CHD7 gene, found by next generation sequencing, confirm its pathogenicity in a Spanish patient and diagnose him with CHARGE syndrome. 9, 26–31. doi: 10.3389/fgene.2018.00007

Wang, G.-S., and Cooper, T. A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* 8, 749–761. doi: 10.1038/nrg2164

Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M., and Burge, C. B. (2004). Systematic identification and analysis of exonic splicing silencers. *Cell* 119, 831–845. doi: 10.1016/j.cell.2004.11.010

Whiley, P. J., Guidugli, L., Walker, L. C., Healey, S., Thompson, B. A., Lakhani, S. R., et al. (2011). Splicing and multifactorial analysis of intronic *BRCA1* and *BRCA2* sequence variants identifies clinically significant splicing aberrations up to 12 nucleotides from the intron/exon boundary. *Hum. Mutat.* 32, 678–687. doi: 10.1002/humu.21495

Will, C. L., and Lührmann, R. (2011). Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* 3:a003707. doi: 10.1101/cshperspect.a003707

Yeo, G., and Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* 11, 377–394. doi: 10.1089/1066527041410418

Zhang, M. (1998). Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.* 7, 919–932. doi: 10.1093/hmg/7.5.919

Zhang, X. H., and Chasin, L. A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.* 18, 1241–1250. doi: 10.1101/gad.1195304

Computational Tools for Splicing Defect Prediction in Breast/Ovarian Cancer Genes: How Efficient Are They at Predicting RNA Alterations?

Alejandro Moles-Fernández[1], Laura Duran-Lozano[1], Gemma Montalban[1], Sandra Bonache[1], Irene López-Perolio[2], Mireia Menéndez[3,4,5], Marta Santamariña[6], Raquel Behar[2], Ana Blanco[6], Estela Carrasco[7], Adrià López-Fernández[7], Neda Stjepanovic[7,8], Judith Balmaña[7,8], Gabriel Capellá[3,4,5], Marta Pineda[3,4,5], Ana Vega[6], Conxi Lázaro[3,4,5], Miguel de la Hoya[2], Orland Diez[1,9†] and Sara Gutiérrez-Enríquez[1*†]*

[1] Oncogenetics Group, Vall d'Hebron Institute of Oncology, Barcelona, Spain, [2] Laboratorio de Oncología Molecular – Centro de Investigación Biomédica en Red de Cancer, Instituto de Investigación Sanitaria San Carlos, Hospital Clínico San Carlos, Madrid, Spain, [3] Hereditary Cancer Program, Catalan Institute of Oncology, Institut d'Investigació Biomédica de Bellvitge, Hospitalet de Llobregat, Barcelona, Spain, [4] Program in Molecular Mechanisms and Experimental Therapy in Oncology (Oncobell), Institut d'Investigació Biomédica de Bellvitge, Hospitalet de Llobregat, Barcelona, Spain, [5] Centro de Investigación Biomédica en Red de Cáncer, Madrid, Spain, [6] Grupo de Medicina Xenómica-USC, Fundación Pública Galega de Medicina Xenómica-SERGAS, CIBER de Enfermedades Raras, Instituto de Investigación Sanitaria, Santiago de Compostela, Spain, [7] High Risk and Cancer Prevention Group, Vall d'Hebron Institute of Oncology, Barcelona, Spain, [8] Medical Oncology Department, University Hospital Vall d'Hebron, Barcelona, Spain, [9] Area of Clinical and Molecular Genetics, University Hospital Vall d'Hebron, Barcelona, Spain

In silico tools for splicing defect prediction have a key role to assess the impact of variants of uncertain significance. Our aim was to evaluate the performance of a set of commonly used splicing *in silico* tools comparing the predictions against RNA *in vitro* results. This was done for natural splice sites of clinically relevant genes in hereditary breast/ovarian cancer (HBOC) and Lynch syndrome. A study divided into two stages was used to evaluate SSF-like, MaxEntScan, NNSplice, HSF, SPANR, and dbscSNV tools. A discovery dataset of 99 variants with unequivocal results of RNA *in vitro* studies, located in the 10 exonic and 20 intronic nucleotides adjacent to exon–intron boundaries of *BRCA1*, *BRCA2*, *MLH1*, *MSH2*, *MSH6*, *PMS2*, *ATM*, *BRIP1*, *CDH1*, *PALB2*, *PTEN*, *RAD51D*, *STK11,* and *TP53,* was collected from four Spanish cancer genetic laboratories. The best stand-alone predictors or combinations were validated with a set of 346 variants in the same genes with clear splicing outcomes reported in the literature. Sensitivity, specificity, accuracy, negative predictive value (NPV) and Mathews Coefficient Correlation (MCC) scores were used to measure the performance. The discovery stage showed that HSF and SSF-like were the most accurate for variants at the donor and acceptor region, respectively. The further combination analysis revealed that HSF, HSF+SSF-like or HSF+SSF-like+MES achieved a high performance for predicting the disruption of donor sites, and SSF-like or a sequential combination of MES and SSF-like for predicting disruption of acceptor sites. The performance confirmation of these last results with the validation dataset, indicated that the highest sensitivity, accuracy, and NPV (99.44%, 99.44%, and 96.88, respectively) were attained with HSF+SSF-like or HSF+SSF-like+MES for donor sites and SSF-like (92.63%, 92.65%, and 84.44, respectively) for acceptor sites.

We provide recommendations for combining algorithms to conduct *in silico* splicing analysis that achieved a high performance. The high NPV obtained allows to select the variants in which the study by *in vitro* RNA analysis is mandatory against those with a negligible probability of being spliceogenic. Our study also shows that the performance of each specific predictor varies depending on whether the natural splicing sites are donors or acceptors.

INTRODUCTION

The increasing use of massive parallel sequencing of customized multi-gene panels, for germline clinical testing of hereditary breast and ovarian cancer (HBOC) and Lynch syndrome, is leading to higher detection of genetic variants of unknown significance (VUS).

All exonic or intronic VUS can be potentially spliceogenic by disrupting the *cis* DNA sequences that define exons, introns, and regulatory sequences necessary for a correct RNA splicing process. Specifically, the *cis* DNA elements include: (i) exon–intron boundary core consensus nucleotides (GT at +1 and +2 of the 5′donor site and AG at -1 and -2 of the 3′acceptor site); (ii) intronic and exonic nucleotides adjacent to these invariable nucleotides that are also highly conserved and have been found to be critical for splice site selection: CAG/**GU**AAGU in donor sites and NY**AG**/G in acceptor sites; (iii) branch point and polypyrimidine tract sequence motifs, essential for the spliceosome complex formation; (iv) intronic and exonic sequences that act as splicing enhancers (ISE and ESE) or silencers (ISS and ESS), regulatory motifs that are usually bound by serine/arginine (SR)-rich proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs), respectively (Cartegni et al., 2002; Soukarieh et al., 2016; Abramowicz and Gos, 2018). A nucleotide change in any of these elements could lead to incorrect splice site recognition, creating new ones or activating the cryptic ones, resulting in aberrant transcripts and in non-functional proteins associated with disease such as hereditary cancer.

Interestingly, it has recently been described that hereditary cancer genes (including some HBOC and Lynch genes) are enriched for spliceogenic variants (Rhine et al., 2018). This finding highlights the importance of both the identification and the functional interpretation of variants causing RNA alterations in hereditary cancer genes. In HBOC syndrome and Lynch Syndrome, the clinical classification of VUS is essential since carriers of pathogenic variants may benefit from cancer prevention and risk-reducing strategies, make informed decisions about prophylactic surgery, and benefit from targeted treatments (Moreno et al., 2016). Conversely, carriers of non-pathogenic variants can be excluded from intensive follow-ups and avoid unnecessary risk-reducing surgery (Eccles et al., 2015).

To detect splice site alterations, *in vitro* splicing assays with patient's RNA or minigenes are widely used. However, testing all variants detected in the vicinity of exon–intron boundaries can be time consuming and expensive. In consequence, to select variants to be experimentally evaluated, a large number of prediction programs have been developed. These splicing computational tools are based on different premises. The most commonly used are based on Position Weight Matrix (PWM), in which each nucleotide on the splice site sequence is scored and ranked based on its frequency from its aligned consensus sequence (Shapiro and Senapathy, 1987; Desmet et al., 2009). Neural network programs use sets of sequences from databases to identify splicing sites (Reese et al., 1997). Tools based on Maximum Entropy Distribution models take into account the dependencies between nucleotide positions (Yeo and Burge, 2004). Approaches like SPANR (Xiong et al., 2015) use DNA and RNA sequence information and a machine learning method, to predict splicing alterations, enabling the identification of variants affecting *cis* and *trans* splicing factors. Another type of splicing tool has been developed using ensemble learning methods (adaptive boosting and random forest) taking advantage of individual computational tools (Jian et al., 2014a).

Several studies have analyzed the performance of these tools for genes related to cancer and other diseases and report discordant results without a consensus guideline recommending which programs should be used (Houdayer et al., 2008, 2012; Holla et al., 2009; Vreeswijk et al., 2009; Desmet et al., 2010; Théry et al., 2011; Colombo et al., 2013; Jian et al., 2014a; Tang et al., 2016) (**Table 1**). Here, we present an evaluation of the performance of commonly used splicing *in silico* tools, comparing their output with the experimental evidences obtained by RNA *in vitro* analysis of variants detected in HBOC and Lynch syndrome genes. In the first phase of the study, we assessed the accuracy of the splicing *in silico* tools with a dataset of RNA *in vitro* outcomes collected from four Spanish cancer genetic units. Subsequently, we validated the best algorithms obtained in the discovery phase, with findings obtained after RNA analysis extracted from different curated databases and reported literature.

MATERIALS AND METHODS

Variant Selection
Discovery Set
We restricted the study to variants located within the last 10 exonic and 20 first intronic nucleotides from the 5′ splice donor site, and the last 20 intronic and the first 10 exonic nucleotides from the 3′ splice acceptor site (−10 to +20 and −20 to +10, respectively). *BRCA1*, *BRCA2, MLH1, MSH2, MSH6,* and *PMS2* variants were selected from HBOC and Lynch

TABLE 1 | Publications evaluating in silico splicing site tools.

Reference	Number of variants	Source of the variants and in vitro data	Gene(s)	Region analyzed	Experimental design	Prediction tools evaluated	Accuracy of recommended tools	Consensus guideline
Houdayer et al., 2008	39	*Experimental evidence	RB1	±60 nucleotides from an AG/GT site	One evaluation stage	NNSplice, PWM, MES, ASSA, ESEfinder, RESCUE-ESE	NA	Not specifically provided
Holla et al., 2009	18	Experimental evidence	LDLR	Intronic: 5' until +5. 3' until −16	One evaluation stage	MES, NNSplice, NetGene2	NA	Not specifically provided
Vreeswijk et al., 2009	29	Experimental evidence	BRCA1/BRCA2	Intronic: 5' until +60. 3' until −20	One evaluation stage	NNSplice, NetGene2, PWM, ASSA, MES, HSF	NA	Not specifically provided
Desmet et al., 2010	623	UMD locus-specific databases, HGMD, and datasets from previous studies	Multiple	Not specifically stated	One evaluation stage	GENSCAN, GeneSplicer, HSF, MES, NNSplice, SplicePort, SplicePredictor, SpliceView, SROOGLE	Invariable position: HSF, MES, SpliceView and SROOGLE 100%. Intronic SS +3, +5 and last exonic position: MES 100%. Other SS intronic position: MES and SplicePort 5' 76/68% and 3' 77.27/77.27%	Invariable position: HSF, MES, SpliceView and SROOGLE. Intronic SS +3, +5 and last exonic position: MES. Other SS intronic positions: MES and SplicePort
Théry et al., 2011	53	Experimental evidence	BRCA1/BRCA2	Not specifically stated	One evaluation stage	PWM, GeneSplicer, NNSplice, MES, HSF	NA	Not specifically provided
Houdayer et al., 2012	272	Experimental evidence	BRCA1/BRCA2	Not specifically stated	One evaluation stage	NNSplice, SSF, MES, ESEfinder, RESCUE-ESE, HSF	Accuracy as AUC: MES: 0.956, SSF-like: 0.914	Sequential MES and SSF
Colombo et al., 2013	24	Experimental evidence	BRCA1/BRCA2	Not specifically stated	One evaluation stage	PWM, MES, NNSplice, GeneSplicer, HSF, NetGene2, SplicePredictor, ASSA	NA	HSF and ASSA
Jian et al., 2014b	2,959	HGMD, SpliceDisease database and DBASS. Negative variants from 1000 Genomes Phase 1	Multiple	5': from −3 to +8. 3': from −12 to +2	Evaluation of individual tools + new model construction + validation stage	SSF-like, MES, NNSplice, GeneSplicer, HSF, NetGene2, GENSCAN, SplicePredictor, **dbscSNV	SSF-like: 91.1% MES: 89.5%/dbscSNV: 93.3%	SSF-like, MES/dbscSNV
Tang et al., 2016	272	HGMD (damaging variants) and negative variants from 1000 Genomes Phase 1	Multiple	Intronic: 5' from +3 to +7. 3' from −3 to −9	One evaluation stage	HSF, MES, NNSplice, ASSP	Accuracy as AUC: MES: 0.878 ASSP: 0.881 HSF: 0.834	MES, ASSP, and HSF combination
Leman et al., 2018	395	Experimental evidence	Multiple	5': from −3 to +8. 3': from −12 to +2	Training + evaluation stage	HSF, MES, SSF-like, NNSplice, GS, SPICE (MES and SSF combination)	SPICE 95.6%	SPICE (Th_{Se} threshold with MES and SSF combination)

*Experimental evidence: experimental in vitro RNA results collected specifically for the study, derived from either patient blood cells or minigene assay. **dbscSNV: database containing the adaptive boosting and random forests scores. UMD, Universal Mutation Database; HGMD, the Human Gene Mutation Database; DBASS, Aberrant Splice Database; Splice NNSplice, Site Prediction by Neural Network; PWM, Position Weight Matrix; MES, MaxEntScan; ASSA, Automated Splice-Sites Analyses; HSF, Human Splice Finder; SSF, Splice Site Finder; GS, GeneSplicer; SROOGLE, splicing regulation online graphical engine; ASSP, Alternative Splice Site Predictor; NA, information not available in the paper; SS, splicing site; AUC, area under the curve; Th_{Se}, optimal sensitivity threshold.

syndrome patients routinely analyzed for diagnostic purposes. We also included *ATM*, *BRIP1*, *CDH1*, *PALB2*, *PTEN*, *RAD51D*, *STK11*, and *TP53* variants obtained in a research series of *BRCA1* and *BRCA2* negative HBOC patients. Genetic variants with unequivocal experimental evidences showing presence or absence of alterations in the mRNA, were collected from four different Spanish centers: Hospital Universitari Vall d'Hebron (HUVH), Barcelona; Hospital Clínico San Carlos (HCSC) Madrid; Fundación Pública Galega de Medicina Xenomica (FPGMX), Santiago de Compostela; Institut Català d'Oncologia (ICO), Hospital Duran i Reynals, Barcelona.

The variants included in the discovery set were analyzed *in vitro* in carriers and controls. RNA was isolated from whole blood leukocytes or short-term lymphocyte cultures, phytohaemagglutinin stimulated, and treated with and without puromycin. The contributing laboratories used diverse isolation protocols and/or cDNA synthesis strategies following ENIGMA recommendations (Colombo et al., 2014; Whiley et al., 2014). Briefly, the splicing products generated by reverse transcription-polymerase chain reaction (RT-PCR) assays were characterized using agarose gel or capillary electrophoresis in a QIAxcel instrument with QIAxcel DNA High Resolution Kit (QIAGEN) or an Agilent 2100 Bioanalyzer (Agilent), and Sanger sequencing. PCR primers were designed to amplify at least one whole exon 5′ and 3′ flanking the exon harboring the variant of interest. Primer sequences are available upon request.

The study was approved by the Institutional Review Board of each participating center. Patients received genetic counseling and written informed consent was obtained for further genetic and research studies.

Validation Set

At this stage, the predictors that presented the best performance alone or in combination, were applied to compare their predictions with the *in vitro* RNA results from the dataset obtained through literature and databases. We chose a collection of variants reported in INSIGHT, ClinVar and published works that were (i) located within the regions defined for the discovery set; (ii) identified in the set of cancer risk genes included above; (iii) experimentally confirmed as spliceogenic and non-spliceogenic in blood samples or with minigene assay at least by RT-PCR, agarose gel and Sanger Sequencing analysis; and (iv) not located at exonic splicing enhancer (ESE) regions with specific experimental evidence of causing splicing alteration.

In silico Splice Tools

A total of six splice-site prediction software programs were selected for this study. Two ensemble prediction scores constructed by Jian et al. (2014a) using adaptive boosting and random forests ensemble learning methods, were extracted from dbscSNV database[1]. Splicing-based Analysis of Variants (SPANR), a computational model of splicing derived from the application of "deep learning" computer algorithms (Xiong

et al., 2015) was ascertained by its own web site[2]. Splice Site Finder (SSF-like) (based on Shapiro and Senapathy, 1987), MaxEntScan (MES) (Yeo and Burge, 2004), Splice Site Prediction by Neural Network (NNPLICE) (Reese et al., 1997), and Human Splicing Finder (HSF) (Desmet et al., 2009) accessed through Alamut Visual 2.10 (Interactive Biosoftware). The GeneSplicer program is also included in the splicing module of Alamut, but it was excluded from the study since we noticed it had an exceedingly high missing scores (no estimation was obtained for 30% of the variants analyzed; data not shown), which had also been reported by Jian et al. (2014a). SPANR and dbscSNV do not analyze insertions and deletions and dbscSNV gives estimations for variants only located from −3 to +8 at 5′ and −12 to +2 at 3′ (**Supplementary Table 1**).

To interrogate the splicing prediction tools, we calculated the score variation caused by the variant in the donor site or acceptor site. To do that, we compared the score computed in the wild-type sequence (WT) to the score computed in the variant sequence (VAR) as:

$$\%scorevariation = (VARscore - WTscore)/WTscore*100$$

We calculated the % score variation for four out of the six tools (SSF-like, HSF, MES, and NNSPLICE), since dbscSNV and SPANR already provide a score change.

To consider a % score change as a positive prediction of a splicing motif disruption caused by the variant, which would lead to aberrant splicing, we adopted thresholds pre-established in the literature (**Supplementary Table 1**). When two programs were combined, a correct prediction of splicing alteration was considered if at least one of them scored above the threshold. When three, four, five, or six programs were combined, all tools but one had to score above the threshold to indicate splicing alteration.

Performance Assessment

In the discovery and validation phases, the experimental RNA results for each collected variant were annotated as positive splicing alteration when they unequivocally, verified by gel electrophoresis and Sanger sequencing, lead to: exon skipping, use of a new or cryptic splice site or altered alternative transcript profile. In contrast, a negative splicing alteration was annotated when the *in vitro* RNA result was exactly the same as that obtained in control samples.

For both stages, we calculated the overall accuracy (ratio of overall correct predictions to the total number of predictions), specificity (correct identification of non-spliceogenic variants; true negative rate), and sensitivity (correct identification of deleterious variants; true positive rate). The positive predictive values (PPV, proportion of positive predictions that were true positives), negative predictive values (NPV, proportion of negative predictions that were true negatives), false negative rates (FNR, proportion of false negative detection), and false positive rates (FPR, proportion of false positive detection) were also

---

[1]https://sites.google.com/site/jpopgen/dbNSFP

[2]http://tools.genes.toronto.edu/

FIGURE 1 | *In vitro* RNA results collected in the **discovery set.** Experimental data are displayed according to variation location. Positive splicing alterations include: exon skipping, use of a new or cryptic splice site or an altered-alternative transcript profile. Negative splicing alteration: *in vitro* RNA result was exactly the same as that obtained in control samples. Spl, splicing.

calculated. Matthews correlation coefficient (MCC) was used to provide a balanced comparison between *in silico* tools.

RESULTS

Discovery Set

A total of 99 variants with unequivocal RNA *in vitro* results were studied, located within positions −10 to +20 from the 5′ donor site, and within −20 to +10 from the 3′ acceptor site (**Supplementary Table 2**). Forty-four of the 99 variants generated a splice defect, with 11 and 9 disrupting the canonical GT or AG dinucleotides, respectively. The 24 remaining variants with aberrant splicing were located outside invariable GT or AG positions, with 15 variants altering the 5′ splice site and nine altering the 3′ splice site. Fifty-five variants did not yield an aberrant splicing, all located outside invariant dinucleotides. **Figure 1** displays the number of positive and negative splicing results relative to variant location.

Six *in silico* tools were used to interrogate the 99 variants, and their corresponding % score variation was obtained. These outputs were compared to the experimental RNA results. The respective thresholds pre-established in the literature were adopted for each program (**Supplementary Table 1**).

Supplementary Table 2 lists the % score variation obtained from each splicing tool used to assess the 99 variants, highlighting which scores were in agreement with the RNA analysis outcome. Of note, seven insertions or deletions were not computed by SPANR and dbscSNV, while estimations for 33 substitutions were not provided by dbscSNV.

Table 2 shows separately, for 5′ (52 variants), 3′ (47 variants), and both splice sites (global, 99 variants), the results of performance analysis for each one of the tools. The six predictors detected wild type (WT) splice sites in reference sequences for all the genes of interest.

On average, predictions for variants located in 5′ regions have higher accuracy (90.98%), sensitivity (90.44%) and specificity (91.28%) compared to those located in 3′ regions (83.74%,

84.52%, and 82.30%, respectively) (**Table 2**). The predictions computed by HSF (with a score change threshold of −2%) were the most accurate and sensitive for variants at donor site, while for variants at acceptor sites or affecting either acceptor or donor sites (global), SSF-like were the most accurate (with a score change threshold of −5%). MES program (with a score change threshold of −15%) showed 100% of sensitivity on all predictions, but its specificity did not reach 87% in any case. In contrast, SPANR program showed the highest values of specificity for predictions of variants at donor site or all variants affecting either at acceptor or donor splice sites, but the lowest values of sensitivity (**Table 2**).

Accordingly, the lowest false negative rates for 5′ splice site were reached by the HSF and MES predictors, while at 3′ splice sites, the SSF-like and MES predictors obtained the lowest false negative rates (**Table 2** and **Figure 2**). In contrast, SPANR predictor had the highest false negative and the lowest false positive rates in almost all cases (**Table 2** and **Figure 2**). Regarding the estimation of the proportion of negative predictions that were true negatives (NPV), HSF or MES and SSF-like or MES achieved the highest values (100%) for donor and acceptor sites, respectively (**Table 2**).

The accuracy of all possible predictor combinations was further assessed. For 5′ donor splice sites, predictions of HSF alone or HSF together with seven different combinations, SSF-like+SPANR and SSF-like+MES+SPANR reached a 98.08% of accuracy with the highest sensitivity for all the models (100%), obtaining 96.15% of specificity, 0.96 MCC and 100% of NPV (**Supplementary Table 3**). For 3′ splice sites, a sequential combination recommended by Houdayer et al. (2012) using MES as first-line analysis with a cut-off of 15% followed by SSF-like with a 5% threshold achieved the best performance, with a 100% of sensitivity, 96.55% of specificity, 97.87 % of accuracy, 0.96 MCC, and 100% of NPV (**Supplementary Table 4**). However, SSF-like alone and two more combinations including it also showed a 100% of NPV together with 100% sensitivity and high values of accuracy

TABLE 2 | Performance of the individual *in silico* tools in the discovery dataset.

	Sensitivity	Specificity	Accuracy	MCC	Positive Predictive Value	Negative Predictive Value	False Negative Rate	False Positive Rate	False Discovery Rate	False Omission Rate
Donor (5′)										
HSF	**100.000**	96.154	**98.077**	**0.962**	96.296	**100.000**	**0.000**	3.846	3.704	**0.000**
SSF-like	96.154	96.154	96.154	0.923	96.154	96.154	3.846	3.846	3.846	3.846
MES	**100.000**	84.615	92.308	0.856	86.667	**100.000**	**0.000**	15.385	13.333	**0.000**
dbscSNV	91.667	90.000	91.176	0.795	95.652	81.818	8.333	10.000	4.348	18.182
NNS	92.308	80.769	86.538	0.735	82.759	91.304	7.692	19.231	17.241	8.696
SPANR	62.500	**100.000**	81.633	0.677	**100.000**	73.529	37.500	0.000	**0.000**	26.471
Acceptor (3′)										
SSF-like	**100.000**	89.655	**93.617**	**0.877**	85.714	**100.000**	**0.000**	10.345	14.286	**0.000**
MES	**100.000**	86.207	91.489	0.839	81.818	**100.000**	**0.000**	13.793	18.182	**0.000**
dbscSNV	93.750	77.778	88.000	0.736	**88.235**	87.500	6.250	22.222	**11.765**	12.500
HSF	83.333	82.759	82.979	0.649	75.000	88.889	16.667	17.241	25.000	11.111
NNS	88.889	68.966	76.596	0.563	64.000	90.909	11.111	31.034	36.000	9.091
SPANR	41.176	88.460	69.760	0.343	70.000	69.697	58.824	11.538	30.000	30.303
Global (5′ and 3′)										
SSF-like	97.727	92.727	**94.949**	**0.900**	91.489	98.077	2.273	7.273	8.511	1.923
MES	**100.000**	85.455	91.919	0.850	84.615	**100.000**	**0.000**	14.545	15.385	**0.000**
HSF	93.182	89.091	90.909	0.818	87.234	94.231	6.818	10.909	12.766	5.769
dbscSNV	92.500	84.211	89.831	0.767	**92.500**	84.211	7.500	15.789	**7.500**	15.789
NNS	90.909	74.545	81.818	0.653	74.074	91.111	9.091	25.455	25.926	8.889
SPANR	53.659	**94.118**	76.087	0.533	88.000	71.642	46.341	5.882	12.000	28.358

Results of the performance evaluation is grouped by donor, acceptor or both splice sites. The best performance scores are highlighted in bold. False Discovery Rate represents the rate of false positives of the total of variants positively predicted and False Omission Rate represents the rate of false negatives of the total negative predicted variants. dbscSNV, database consulted for extracting the adaptive boosting and random forests scores.

FIGURE 2 | False negative and false positive rates for individual splicing prediction tools in the **discovery set**. dbscSNV, database consulted for extracting the adaptive boosting and random forests scores.
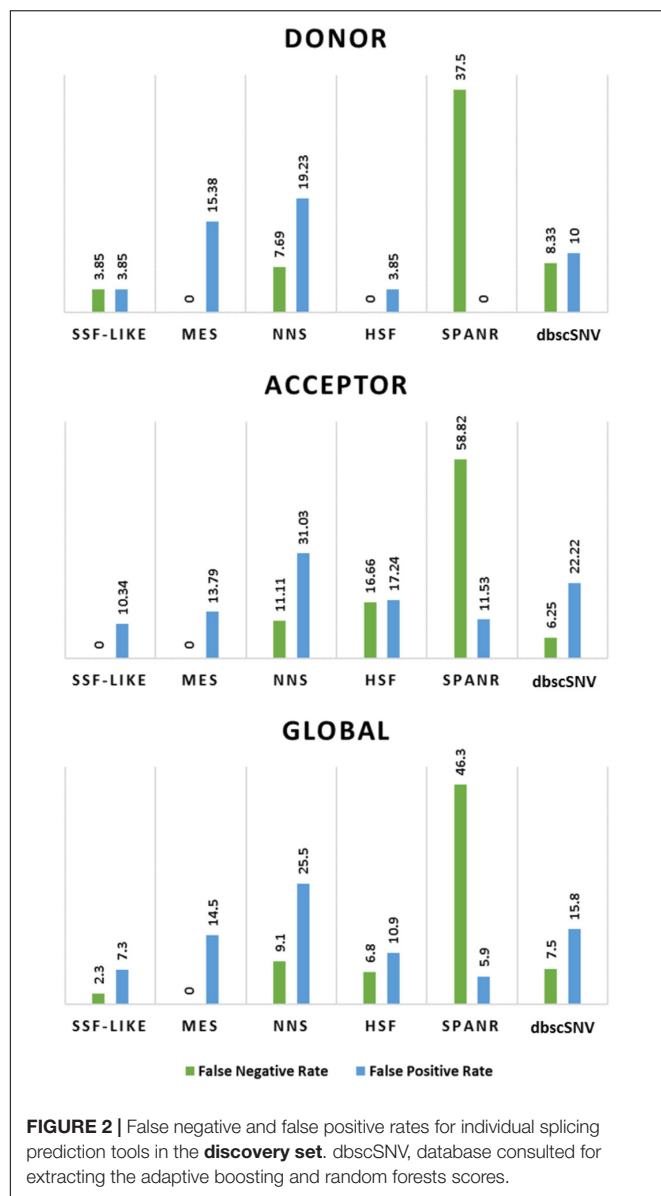
(for predictions at acceptor site, **Supplementary Table 4**). Considering the tool combinations for predicting disruption caused by variants located in any of the two splice sites (global), MES and SSF-like sequential combination achieved the best accuracy with a 96.97% and 0.94 of MCC, followed for two combinations, including SSF-like and MES, which showed 100% sensitivity and 100% of NPV (**Supplementary Table 5**).

Validation Set

In order to validate the predictors with the best performance obtained in the discovery set, we analyzed a dataset of 346 variants with RNA *in vitro* results published or detailed in free available databases. At donor region, 210 variants were included, 177 showing *in vitro* splicing alterations (65 at intronic GT positions) and 33 showing no splicing effects (all outside intronic

GT) (**Figure 3** and **Supplementary Table 6**). One hundred thirty-six variants were located at the acceptor region, 95 showing splicing alterations (67 of them at intronic AG positions), and 41 with absence of alterations (40 of them outside intronic AG) (**Figure 3** and **Supplementary Table 7**). Only SSF-like and SPANR were able to identify all WT splice sites in reference sequences for all the genes of interest.

We selected for validation, the HSF stand-alone and the combinations HSF+SSF-like and HSF+SSF-like+MES for 5′donor sites (**Supplementary Table 3**), and the SSF-like alone and the sequential MES and SSF combination for 3′acceptor sites (**Supplementary Table 4**), considering sensitivity, accuracy, MCC and NPV scores. We excluded the combinations including SPANR or dbscSNV since they do not provide predictions on insertions and deletions.

Overall, the *in silico* predictions in the validation dataset were more accurate for variants with effects on donor splice sites than acceptor sites (**Table 3** and **Figure 4**). These findings were in agreement with those results obtained with the discovery set (**Table 2**).

The data analysis indicated that for 5′ donor sites the best combinations, with 98.57% accuracy, 99.44% of sensitivity and 96.88% of NPV, are HSF+SSF-like or HSF+SSF-like+MES (**Table 3**) with very slight differences in performance, between the estimations of splicing effects for all variants (including variants placed at invariable dinucleotides) and for the group of variants located outside the two invariable nucleotides. For acceptor sites, the sequential combination of MES and SSF-like (Houdayer et al., 2012) and SSF-like stand-alone reached a performance with the same score of accuracy, 92.65%, but SSF-like showed a highest NPV (**Table 3**). Unlike the donor site, the accuracy of these predictors decreased (to 85.29%) when the variants analyzed did not include those at the two nucleotide invariables (AG) of the 3′ acceptor splice site (**Table 3**). For predictions of variants outside these dinucleotides, the rate of false negatives showed by SSF-like is slightly lower than those rates of MES and SSF-like sequential combination (25% versus 28.57%, respectively, **Table 3**).

DISCUSSION

The use of massive parallel sequencing in clinical diagnostics is leading to a significant increase in data and the detection of a high number of variants of uncertain significance (VUS) with potential effect on splicing which need interpretation. Therefore, prediction of the effect of DNA sequence variations on splicing using *in silico* tools has become a common approach. Several studies have been published on the performance and reliability of *in silico* predictions of the splicing impact of variants (Jian et al., 2014b). **Table 1** details the results obtained in these studies and shows that the recommendations provided about the most appropriate to be used are not concordant. However, the studies that give clear recommendations, always include one of the HSF, SSF, or MES programs, alternatively.

We have evaluated the reliability of *in silico* splicing effect predictions of six programs (MES, HSF, SSF-like, SPANR, NNSplice, and dbscSNV) comparing their scores with splicing

FIGURE 3 | *In vitro* RNA results collected in the **validation set.** Experimental data are displayed according to variation location. Variants located at 0 position are those that affect the invariable dinucleotide positions (GT or AG) plus other contiguous nucleotides. Positive splicing alterations include: exon skipping, use of a new or cryptic splice site or an altered-alternative transcript profile. Negative splicing alteration: *in vitro* RNA result was exactly the same as that obtained in control samples. Spl, splicing.

in vitro analysis outcomes of variants identified in hereditary cancer related genes. We elaborated the study in two stages, discovery and validation, to identify the best predictors or the best combination for their application in routine clinical testing, taking into account the percentages reached for sensitivity, specificity, accuracy and NPV as well as the score of Mathews Coefficient Correlation (MCC).

In the discovery stage, significant performance differences were appreciated among individual tools (**Table 2**). For global, as well as for 5′, and 3′ splice sites, low accuracies of SPANR and NNSplice contrasted with the high performance achieved by SSF, MES, and HSF, while dbscSNV demonstrated an intermediate accuracy.

At the second stage of our study, we validated the combinations of HSF with SSF-like or HSF+SSF-like+MES as the highest performance for splicing aberrations at donor sites, and SSF-like stand-alone at acceptor sites (**Table 3**). All these results are in agreement with the trend observed in the previous published results, where HSF or SSF or MES outperformed other methods (**Table 1**). Of note, besides high accuracy and sensitivity, these validated tools, combined or as stand-alone, also had high NPV. This is relevant in a clinical setting, since it allows to separate the variants with an extremely low or non-existent probability of being abnormally spliceogenic from those variants in which *in vitro* RNA studies are of interest, with the consequent saving of resources in the laboratory.

All of the three predictors are available through Alamut Visual 2.10 (Interactive Biosoftware, Rouen), allowing a high throughput analysis, which is essential in a massive parallel sequencing annotation pipeline. Yet, in the newest version of Alamut Visual (2.11) the HSF predictor is not included in its splicing module, it is freely available at Human Splice Finder website[3] or through VarAFT software[4], which allows the annotation of a large batch of variants. MES program is also freely

accessible via web[5,6], although caution should be taken when obtaining predictions via Alamut or via web, since differences have been reported (Tang et al., 2016). SSF-like tool is currently only accessible through Alamut, yet it has been recently published a free program named Splicing Prediction in Consensus Elements (SPiCE[7]) that combines predictions from SSF-like and MES (Leman et al., 2018). On the other hand, SPANR and dbscSNV are free and could be easily implemented in a pipeline (Xiong et al., 2015; Liu et al., 2017), but these tools are not able to interpret splicing alterations caused by insertion or deletions (6.36% of validation set variants), which represents a limitation for their use compared to the other tools.

Non-canonical GC-AG and AT-AC sequences at the splice site invariant positions occur in 0.56 and 0.09% of the splice site pairs, respectively (Abramowicz and Gos, 2018). In the list of the genes that we analyzed, only six splice sites vary from the canonical splice site GT-AG: *ATM* exon 50 donor site (GC), *BRCA2* exon 17 donor site (GC), *MUTYH* exon 14 donor site (GC), *PALB2* exon 12 donor site (GC), *STK11* exon 2 donor site (AT) and exon 3 acceptor site (AC). In our validation dataset, we only had variants at atypical *BRCA2* exon 17 donor site (GC), and among the studied tools, only SSF-like and SPANR were able to identify these atypical splicing sites and made a prediction for variants located nearby. As the performance of SSF-like is better than SPANR, we suggest the use of SSF-like to analyze these non-canonical splicing sites.

The tools analyzed in this article have only been interrogated to predict alteration at donor and acceptor splice sites. However, alterations in RNA may be produced by variant effects on other factors in *cis* (branch points, polypyrimidine tract, intronic and exonic splicing silencers and enhancers) or create new splice sites or activate cryptic ones. At the stage of validation, the rate of false negative predictions is significantly higher for acceptor sites

---

[3]http://www.umd.be/HSF3/

[4]https://varaft.eu/

[5]http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html

[6]http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html

[7]https://sourceforge.net/projects/spicev2-1/

TABLE 3 | Performance with the validation dataset of the best *in silico* tools previously selected from the results at discovery stage.

	Sensitivity	Specificity	Accuracy	MCC	Positive predictive Value	Negative Predictive Value	False Negative Rate	False Positive Rate	False Discovery Rate	False Omission Rate
Donor										
HSF										
All variants	96.045	90.909	95.238	0.831	98.266	81.081	3.955	9.091	1.734	18.919
Without invariable dinucleotides	94.643	90.909	93.793	0.830	97.248	83.333	5.357	9.091	2.752	16.667
HSF+SSF-like										
All variants	**99.435**	**93.939**	**98.571**	**0.946**	**98.876**	**96.875**	**0.565**	**6.061**	**1.124**	**3.125**
Without invariable dinucleotides	**99.107**	**93.939**	**97.931**	**0.941**	**98.230**	**96.875**	**0.893**	**6.061**	**1.770**	**3.125**
HSF+SSF-like+MES										
All variants	**99.435**	**93.939**	**98.571**	**0.946**	**98.876**	**96.875**	**0.565**	**6.061**	**1.124**	**3.125**
Without invariable dinucleotides	**99.107**	**93.939**	**97.931**	**0.941**	**98.230**	**96.875**	**0.893**	**6.061**	**1.770**	**3.125**
Acceptor										
MES and SSF-like sequential										
All variants	91.579	**95.122**	92.647	0.837	97.753	82.979	8.421	**4.878**	**2.247**	17.021
Without invariable dinucleotides	71.429	**95.000**	85.294	0.699	90.909	82.609	28.571	5.000	9.091	17.391
SSF-like										
All variants	**92.632**	92.683	**92.647**	0.832	96.703	**84.444**	**7.368**	7.317	3.297	**15.556**
Without invariable dinucleotides	**75.000**	92.500	**85.294**	0.695	87.500	**84.091**	**25.000**	7.500	12.500	**15.909**

The best performance scores are highlighted in bold. The atypical BRCA2 exon 17 native donor site (GC) was not estimated by HSF nor MES, and we have considered it as a failed prediction of the two tools for variants affecting this exon regardless of the in vitro splicing effect of the variant. False Discovery Rate represents the rate of false positives of the total of variants positively predicted and False Omission Rate represents the rate of false negatives of the total negative predicted variants.
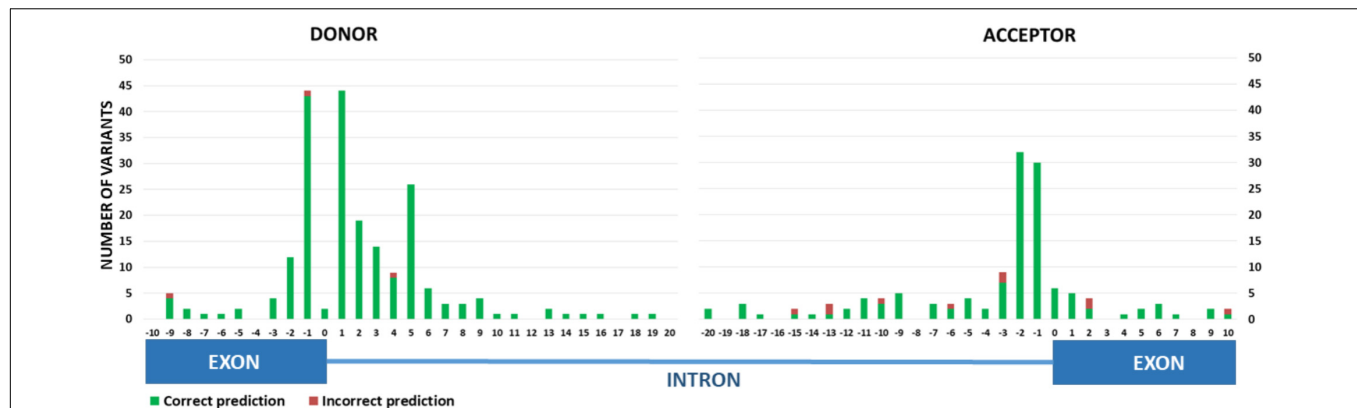
FIGURE 4 | Prediction performance of HSF+SSF-like for donor sites and SSF-like for acceptor sites with variants collected in the **validation set**. Correct prediction: *in silico* and *in vitro* results are concordant. Incorrect prediction: *in silico* and *in vitro* results are discordant.

than for donor sites (**Table 3**). This difference may be due to the greater complexity of the sequence adjacent to the 3′, with the presence of the branch point and the polypyrimidine tract. Therefore, variants located in these two last elements could alter RNA and not be detected as changes in the scores of the splicing sites computed by the predictors. For example, the variant c.1066-6T>G at *ATM* (included in the validation set), which is not predicted correctly by MES and SSF-like sequential combination (**Supplementary Table 7**), alters the polypyrimidine tract causing an aberrant transcript (Dörk et al., 2001).

Likewise, the *BRCA2* exonic variant c.467A>G, located nine nucleotides upstream from the 5′ donor site, causes the loss of these last nine nucleotides, while the HSF and SSF-like predicts that their scores for the native donor splice site of 88.9 and 84.5, respectively, are not changed by the variant, which it is misinterpreted as a false negative (**Supplementary Table 6**). Using some of the tools analyzed in our study to identify enhanced cryptic sites or creation of new splice sites, the variant is predicted to cause a new donor site at nine nucleotides from 5′, in concordance with *in vitro* results: SSF-like indicates a new donor site with a score of 96.9 against 84.5 of the natural splice site, MES 11.1 vs. 9.5 and HSF 98.2 vs. 88.9.

Furthermore, variants located in the exonic regions collected in our study could affect enhancer elements (ESEs) leading to an exon skipping, but they would not be correctly predicted by the analyzed tools. Although variants with specific experimental evidence of suffering this type of alteration were not included in our study, most articles consulted do not explicitly describe or exhaustively exclude the effect of ESEs. As an example, the *BRCA1* c.557C>A altering splicing variant gathered at validation set is not predicted to affect native acceptor site by SSF-like, but specific tools to predict splicing defect caused by regulatory sequence disruption indicates an ESE disturbance: ESRseq score of −1.567 (Ke et al., 2011) and HEXplorer ΔHZ$_{EI}$ = −30.24 (Erkelenz et al., 2014).

Computational tools or programs able to perform predictions on the disruption of all *cis* DNA elements would cover the whole landscape of aberrant RNA splicing yielded by spliceogenic VUS. Theoretically, SPANR is able to detect exon skipping caused by all

of the elements above mentioned, although our study indicated that this program has a low performance for at least to predict correctly alterations of donor and acceptor sites (**Table 2**). The HSF predictor accessed via its website[8], also predicts the impact of genetic variations on branch point elements and has been improved for the identification of natural non-canonical splice sites (Oetting et al., 2018). The breast cancer genes PRIORS probabilities program[9], gives MES estimations of disruption of natural splice sites and also computes the creation of new donor and acceptor splice sites using NNSplice, yet only for *BRCA1* and *BRCA2* genes (Vallée et al., 2016). However, the accuracy and performance of SPANR, HSF, and PRIORS predictions of variants placed in elements other than natural splice sites has not yet been evaluated.

To our knowledge, our study is the only that evaluates the accuracy of different tools separately for donor and acceptor sites, resulting in different recommendations for each one with high performance (**Table 1**).

One limitation of our study is the use of splicing *in silico* tools through a non-free commercial program, Alamut Visual 2.10, with the uncertainty of whether the predictions obtained through Visual Alamut are the same as those estimated directly by the tools in their respective free access websites. We have confirmed that HSF via web (see footnote 8; data not shown) and MES via SPICE (see footnote 7; **Supplementary Table 8**), at least for native splice sites, provide the same estimations than those provided by Alamut Visual 2.10. However, SSF-like predictions obtained through Alamut Visual 2.10 slightly differ from the predictions ascertained through SPICE (**Supplementary Table 8**). Therefore and considering our findings, we recommend as a free pipeline to use HSF accessed via web and MES via SPICE for donor and acceptor site predictions, respectively.

Another limitation is the higher number of variants causing splicing defects compared to the number of variants causing no

---

splicing alteration in our validation dataset. This bias is due to a tendency to report only variants that cause splicing defects. Some studies, in order to avoid this bias, have included common single nucleotide polymorphisms (SNPs) from control dataset, assuming that they do not cause alterations (**Table 1**). Likewise, reports of RNA *in vitro* effects of variants in the two invariable dinucleotides GT-AG are overrepresented, while those located further from splice junctions are less frequently analyzed.

CONCLUSION

In conclusion, to perform *in silico* analysis of VUS potentially affecting natural splice sites in hereditary cancer genes, we recommend the use of the HSF+SSF-like combination (with Δ-2% and Δ-5% as thresholds, respectively) for donor sites and SSF-like (Δ-5%) stand-alone for acceptor sites. These tools have shown in the validation stage a high sensitivity and especially a high NPV. Although the *in vitro* study of RNA remains the gold standard to evaluate the process of splicing, and it is not recommended to use these predictions as the sole source of evidence to make clinical assertions (Richards et al., 2015), our results indicate that these combined tools can be used to filter out VUS with a very low probability of altering splicing without losing true spliceogenic variants that will need deeper experimental validation. Complementing the analysis using specific predictors to identify variants that could affect elements other than splice sites (such as branch points or ESEs), may be useful for the screening of the whole RNA defect landscape. Lastly, it is worth stating that (i) the aim of this work was not to classify variants but to provide an *in silico* algorithm with the highest performance to predict an altered *in vitro* splicing regardless of whether the variants are benign or pathogenic; and (ii) the detection of splicing defect does not automatically denote the pathogenicity of the variant for which a comprehensive qualitative and quantitative RNA analysis is warranted as highlighted in ENIGMA[10] or ACGM guidelines (Richards et al., 2015) for variant classification.

AUTHOR CONTRIBUTIONS

AM-F, LD-L, SG-E, and OD: conception or design of the work. AM-F, LD-L, GM, SB, IL-P, MM, MS, RB, AB, EC, AL-F, NS, and

---

[10] https://enigmaconsortium.org/

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00366/full#supplementary-material

REFERENCES

Abramowicz, A., and Gos, M. (2018). Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J. Appl. Genet.* 59, 253–268. doi: 10.1007/s13353-018-0444-7

Cartegni, L., Chew, S. L., and Krainer, A. R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.* 3, 285–298. doi: 10.1038/nrg775

Colombo, M., Blok, M. J., Whiley, P., Santamariña, M., Gutiérrez-Enríquez, S., Romero, A., et al. (2014). Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the

ENIGMA consortium. *Hum. Mol. Genet.* 23, 3666–3680. doi: 10.1093/hmg/ddu075

Colombo, M., de Vecchi, G., Caleca, L., Foglia, C., Ripamonti, C. B., Ficarazzi, F., et al. (2013). Comparative in vitro and in silico analyses of variants in splicing regions of BRCA1 and BRCA2 genes and characterization of novel pathogenic mutations. *PLoS One* 8:e57173. doi: 10.1371/journal.pone.0057173

Desmet, F. O., Hamroun, D., Collod-Béroud, G., Claustres, M., and Béroud, C. (2010). "Bioinformatics identification of splice site signals and prediction of mutation effects," in *Research Advances in Nucleic Acids Research*, ed. R. M. Mohan (Kerala: Global Research Network), 1–16.

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* 37, 1–14. doi: 10.1093/nar/gkp215

Dörk, T., Bendix, R., Bremer, M., Rades, D., Klöpper, K., Bremer, M., et al. (2001). Spectrum of ATM gene mutations in a hospital-based series of unselected breast cancer patients. *Cancer Res.* 61, 7608–7615.

Eccles, E. B., Mitchell, G., Monteiro, A. N. A., Schmutzler, R., Couch, F. J., Spurdle, A. B., et al. (2015). BRCA1 and BRCA2 genetic testing-pitfalls and recommendations for managing variants of uncertain clinical significance. *Ann. Oncol.* 26, 2057–2065. doi: 10.1093/annonc/mdv278

Erkelenz, S., Theiss, S., Otte, M., Widera, M., Peter, J. O., and Schaal, H. (2014). Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. *Nucleic Acids Res.* 42, 10681–10697. doi: 10.1093/nar/gku736

Holla, ØL., Nakken, S., Mattingsdal, M., Ranheim, T., Berge, K. E., Defesche, J. C., et al. (2009). Effects of intronic mutations in the LDLR gene on pre-mRNA splicing: comparison of wet-lab and bioinformatics analyses. *Mol. Genet. Metab.* 96, 245–252. doi: 10.1016/j.ymgme.2008.12.014

Houdayer, C., Caux-Moncoutier, V., Krieger, S., Barrois, M., Bonnet, F., Bourdon, V., et al. (2012). Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. *Hum. Mutat.* 33, 1228–1238. doi: 10.1002/humu.22101

Houdayer, C., Dehainault, C., Mattler, C., Michaux, D., Caux-Moncoutier, V., Pagès-Berhouet, S., et al. (2008). Evaluation of in silico splice tools for decision-making in molecular diagnosis. *Hum. Mutat.* 29, 975–982. doi: 10.1002/humu.20765

Jian, X., Boerwinkle, E., and Liu, X. (2014a). In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.* 42, 13534–13544. doi: 10.1093/nar/gku1206

Jian, X., Boerwinkle, E., and Liu, X. (2014b). In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet. Med.* 16, 497–503. doi: 10.1038/gim.2013.176

Ke, S., Shang, S., Kalachikov, S. M., Morozova, I., Yu, L., Russo, J. J., et al. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* 21, 1360–1374. doi: 10.1101/gr.119628.110

Leman, R., Gaildrat, P., Gac, G. L., Ka, C., Fichou, Y., Audrezet, M., et al. (2018). Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico / in vitro studies: an international collaborative effort. *Nucleic Acids Res.* doi: 10.1093/nar/gky372 [Epub ahead of print].

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2017). dbNSFP v3.0: a one-stop database of functional predictions and annotations for human non-synonymous and splice site SNVs. *Hum. Mutat.* 37, 235–241. doi: 10.1002/humu.22932

Moreno, L., Linossi, C., Esteban, I., Gadea, N., Carrasco, E., Bonache, S., et al. (2016). Germline BRCA testing is moving from cancer risk assessment to a predictive biomarker for targeting cancer therapeutics. *Clin. Transl. Oncol.* 18, 981–987. doi: 10.1007/s12094-015-1470-0

Oetting, W. S., Béroud, C., Brenner, S. E., Greenblatt, M. S., Karchin, R., and Mooney, S. D. (2018). Methods and tools for assessing the impact of genetic variations. *Hum. Mutat.* 39, 454–458. doi: 10.1002/humu.23393

Reese, M. G., Eeckman, F. H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in genie. *J. Comput. Biol.* 4, 311–323. doi: 10.1089/cmb.1997.4.311

Rhine, C. L., Cygan, K. J., Soemedi, R., Maguire, S., Murray, M. F., Monaghan, S. F., et al. (2018). Hereditary cancer genes are highly susceptible to splicing mutations. *PLoS Genet.* 14:e1007231. doi: 10.1371/journal.pgen.1007231

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., and Gastier-Foster, J. (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. doi: 10.1038/gim.2015.30

Shapiro, M. B., and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* 15, 7155–7174. doi: 10.1093/nar/15.17.7155

Soukarieh, O., Gaildrat, P., Hamieh, M., Drouet, A., Baert-Desurmont, S., Frébourg, T., et al. (2016). Exonic splicing mutations are more prevalent than currently estimated and can be predicted by using in silico tools. *PLoS Genet.* 12:e1005756. doi: 10.1371/journal.pgen.1005756

Tang, R., Prosser, D. O., and Love, D. R. (2016). Evaluation of bioinformatic programmes for the analysis of variants within splice site consensus regions. *Adv. Bioinformatics* 2016:5614058. doi: 10.1155/2016/5614058

Théry, J. C., Krieger, S., Gaildrat, P., Révillion, F., Buisine, M. P., Killian, A., 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. doi: 10.1038/ejhg.2011.100

Vallée, M. P., Di Sera, T. L., Nix, D. A., Paquette, A. M., Parsons, M. T., Bell, R., et al. (2016). Adding in silico assessment of potential splice aberration to the integrated evaluation of BRCA gene unclassified variants. *Hum. Mutat.* 37, 627–639. doi: 10.1002/humu.22973

Vreeswijk, M. P., Kraan, J. N., Van Der Klift, H. M., Vink, G. R., Cornelisse, C. J., Wijnen, J. T., et al. (2009). Intronic variants in BRCA1 and BRCA2 that affect RNA splicing can be reliably selected by splice-site prediction programs. *Hum. Mutat.* 30, 107–114. doi: 10.1002/humu.20811

Whiley, P. J., de la Hoya, M., Thomassen, M., Becker, A., Brandao, R., Pedersen, I. S., et al. (2014). Comparison of mRNA splicing assay protocols across multiple laboratories: recommendations for best practice in standardized clinical testing. *Clin. Chem.* 60, 341–352. doi: 10.1001/jamasurg.2014.1086

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K. C., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347:1254806. doi: 10.1126/science.1254806

Yeo, G., and Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* 11, 377–394. doi: 10.1089/1066527041410418

Quantitative Analysis of *BRCA1* and *BRCA2* Germline Splicing Variants Using a Novel RNA-Massively Parallel Sequencing Assay

Suzette Farber-Katz[1], Vickie Hsuan[1], Sitao Wu[2], Tyler Landrith[1], Huy Vuong[2], Dong Xu[2], Bing Li[2], Jayne Hoo[3], Stephanie Lam[3], Sarah Nashed[4], Deborah Toppmeyer[4], Phillip Gray[3], Ginger Haynes[1], Hsiao-Mei Lu[2], Aaron Elliott[3], Brigette Tippin Davis[3] and Rachid Karam[1]*

[1] Translational Genomics Laboratory, Ambry Genetics, Aliso Viejo, CA, United States, [2] Department of Bioinformatics, Ambry Genetics, Aliso Viejo, CA, United States, [3] Department of Research and Development, Ambry Genetics, Aliso Viejo, CA, United States, [4] Rutgers Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick, NJ, United States

Clinical genetic testing for hereditary breast and ovarian cancer (HBOC) is becoming widespread. However, the interpretation of variants of unknown significance (VUS) in HBOC genes, such as the clinically actionable genes *BRCA1* and *BRCA2*, remain a challenge. Among the variants that are frequently classified as VUS are those with unclear effects on splicing. In order to address this issue we developed a high-throughput RNA-massively parallel sequencing assay—*CloneSeq*—capable to perform quantitative and qualitative analysis of transcripts in cell lines and HBOC patients. This assay is based on cloning of RT-PCR products followed by massive parallel sequencing of the cloned transcripts. To validate this assay we compared it to the RNA splicing assays recommended by members of the ENIGMA (Evidence-based Network for the Interpretation of Germline Mutant Alleles) consortium. This comparison was performed using well-characterized lymphoblastoid cell lines (LCLs) generated from carriers of the *BRCA1* or *BRCA2* germline variants that have been previously described to be associated with splicing defects. CloneSeq was able to replicate the ENIGMA results, in addition to providing quantitative characterization of *BRCA1* and *BRCA2* germline splicing alterations in a high-throughput fashion. Furthermore, CloneSeq was used to analyze blood samples obtained from carriers of *BRCA1* or *BRCA2* germline sequence variants, including the novel uncharacterized alteration *BRCA1* c.5152+5G>T, which was identified in a HBOC family. CloneSeq provided a high-resolution picture of all the transcripts induced by *BRCA1* c.5152+5G>T, indicating it results in significant levels of exon skipping. This analysis proved to be important for the classification of *BRCA1* c.5152+5G>T as a clinically actionable likely pathogenic variant. Reclassifications such as these are fundamental in order to offer preventive measures, targeted treatment, and pre-symptomatic screening to the correct individuals.

Keywords: genetic testing, hereditary breast and ovarian cancer, HBOC, *BRCA1*, *BRCA2*, RNA, Splicing, NGS

INTRODUCTION

Correct interpretation of genomic sequence variants, and subsequent classification of variants as benign or pathogenic, is of utmost importance to patient management, especially in clinically actionable genes such as the breast and ovarian cancer susceptibility genes *BRCA1* and *BRCA2* (OMIM 113705 and 600185, respectively). Variant interpretation is based on multiple lines of evidence (1), including molecular and functional analysis, highlighting the urgent need to develop and implement high-throughput functional assays for variant classification (2).

Genomic sequence variants in *BRCA1* and *BRCA2* have the potential to alter normal splicing of these genes (3). In fact, many alterations in *BRCA1* and *BRCA2* have been shown to be clinically significant by RNA studies and multifactorial likelihood analyses that combine bioinformatics, pathologic, and clinical data (4–6). These variants include those that affect splicing by abolishing or weakening the canonical splice sites at intron-exon boundaries, by creating a novel or activating a cryptic splice site, or by disrupting enhancer or silencer splicing regulatory sequences (7).

Recommendations for mRNA analysis best practice in clinical testing were published by the Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) (8), a consortium established in 2009 with the purpose of sharing data, methods, and resources to facilitate the classification of sequence variants in hereditary breast and ovarian cancer (HBOC) genes (9). ENIGMA recommends the use of RT-PCR and digital or capillary electrophoresis to detect abnormal transcripts based on the length of the product observed, followed by cloning and Sanger sequencing to characterize the sequence of these transcripts (8). However, the consortium notes that evaluation of splicing results for variant carriers can be complicated by the detection of normal alternatively spliced transcripts (8). Both *BRCA1* and *BRCA2* genes are known to undergo alternative splicing, and clinical and functional data indicate that alternatively spliced transcripts may retain function (10, 11). Two fundamental issues in determining the functional significance of normal and abnormal spliced transcripts are whether a transcript is out-of-frame, and therefore predicted to be targeted to degradation by the nonsense-mediated RNA decay (NMD) pathway (12), and the level at which these transcripts are expressed (13). Therefore, a combination of qualitative and quantitative analysis is needed to provide proper characterization of splice variations, and to establish the clinical significance of these specific alterations.

With these in mind, we developed CloneSeq, a high-throughput RNA-based massively parallel sequencing (MPS) technique designed to perform quantitative and qualitative characterization of splicing alterations in a time-frame necessary for clinical testing. Here we describe this technique and perform a comparison of CloneSeq with the techniques recommended by the ENIGMA consortium (8). We performed this comparison using four well-characterized lymphoblastoid cell lines (LCLs) generated from carriers of the *BRCA1* or *BRCA2* germline variants *BRCA1* c.5467+5G>T, *BRCA1* c.135-1G>T, *BRCA2* c.8632+1G>A, or *BRCA2* c.9501+3A>T. These variants have been previously described and are known to be associated with

splicing defects (8). Additionally, we used a similar strategy to analyze blood samples obtained from carriers of *BRCA1* or *BRCA2* germline sequence variants (**Figure 1A**), including the uncharacterized alteration *BRCA1* c.5152+5G>T identified in a novel HBOC family.

MATERIAL AND METHODS

Samples

This study was approved and carried out in accordance with the recommendations of the Western Institutional Review Board (WIRB). All subjects gave written informed consent in accordance with the Declaration of Helsinki. Blood from normal healthy controls or patients participating in the Ambry Genetics Family Studies program was drawn in PAXgene Blood RNA Tubes and stored according to the manufacturer's recommendations (PreAnalytiX, Hombrechtikon, Switzerland). RNA was extracted using the PAXgene Blood RNA Kit according to the recommended protocol (PreAnalytiX). Informed consent was obtained from all participants. Breast RNA was purchased from Amsbio (Lake Forest, CA, USA) and BioChain (Newark, CA, USA). RNA quality was determined using the RIN number calculated by the TapeStation 2200 with RNA ScreenTape or High Sensitivity RNA ScreenTape (Agilent, Santa Clara, CA, USA).

Cell Lines

Lymphoblastoid cell lines (LCLs) were obtained from the Kathleen Cuningham Consortium for Research into Familial Breast Cancer (kConFab, Melbourne, Australia) from 4 carriers of *BRCA1* or *BRCA2* variants and 2 controls. Genotypes were verified by Sanger sequencing. LCLs were maintained according to the recommendations of kConFab. Inhibition of nonsense-mediated decay (NMD) was performed using puromycin (300 μg/ml) or cycloheximide (100 μg/ml) for 4 h, as previously described (8, 14).

RNA Analysis

cDNA was generated using the SuperScript IV First-Strand Synthesis System (Thermo Fisher Scientific, Chino, CA, USA). PCR was performed using either Platinum SuperFi PCR Master Mix (Thermo Fisher Scientific) or HotStarTaq Master Mix (Qiagen, Valencia, CA, USA) as previously described (8).

PCR products were analyzed using digital electrophoresis with D1000 ScreenTape and Reagents on the TapeStation 2200 (Agilent). Capillary electrophoresis (CE) was performed on an ABI 3730xl using MapMarker1000 as a standard (BioVentures, Murfreesboro, TN, USA). Primers were tagged at the 5′ end with FAM or HEX for detection by CE. CE analysis was performed with GeneMapper software (Thermo Fisher Scientific).

PCR products were cloned into pGEM-T Easy and transformed into bacteria according to the manufacturer's recommended protocol (Promega, Fitchburg, WI, USA). Individual white colonies were picked, amplified by rolling-circle replication, and Sanger sequenced by Genewiz (La Jolla, CA, USA).

For CloneSeq, cDNA, PCR, and cloning were performed as described above. All colonies on a plate were scraped and
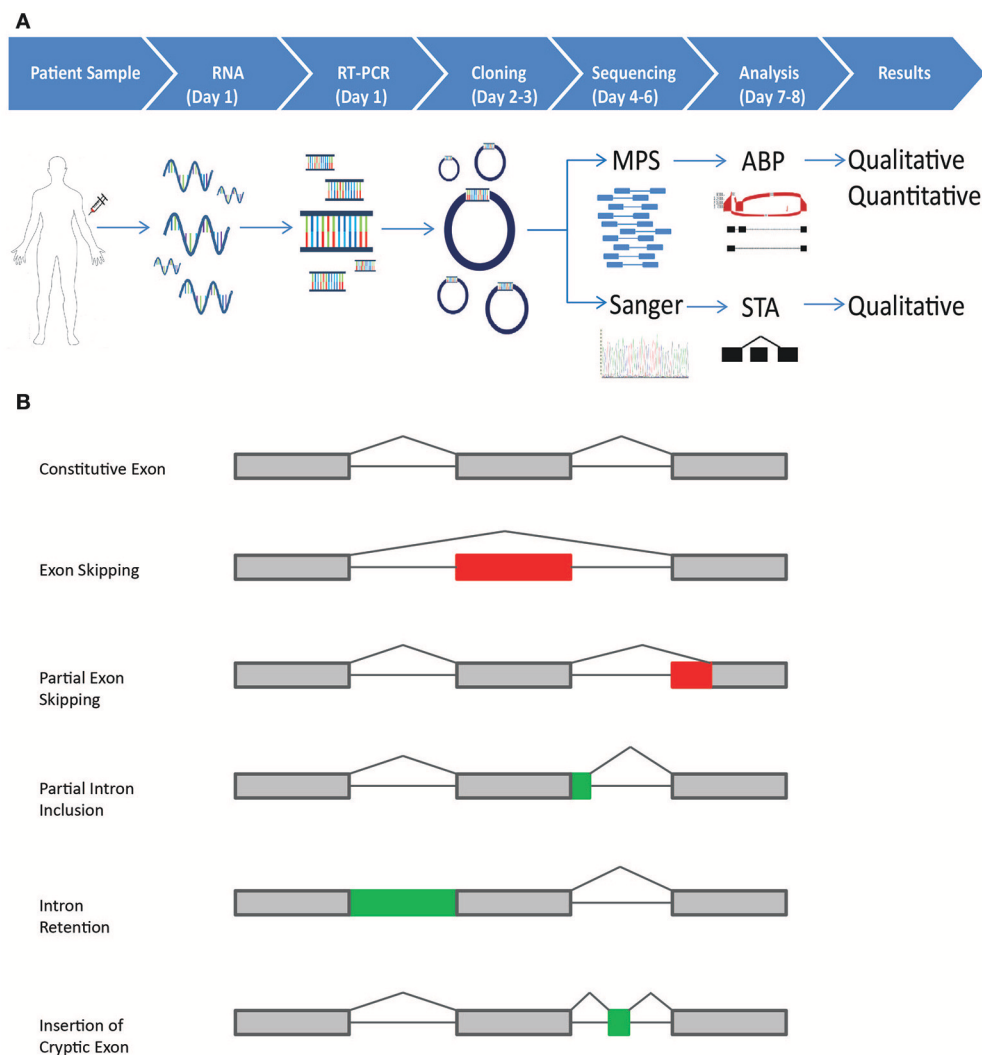
FIGURE 1 | Schematic representation of the CloneSeq protocol and splicing events detected by the bioinformatics pipeline. **(A)** Blood from normal healthy controls and patients participating in the Ambry Genetics Family Studies program was drawn in PAXgene Blood RNA Tubes. RT-PCR was performed following ENIGMA recommendations. RT-PCR products were cloned into pGEM-T Easy and transformed into bacteria. For CloneSeq, all colonies on a plate were scraped and suspended in PBS. Plasmids were extracted, CloneSeq libraries were constructed, and Massively Parallel Sequencing (MPS) was performed, which generated 2 × 250 paired-end reads. The mapped reads are then analyzed by the customized Ambry Bioinformatic Pipeline (ABP) software to generate qualitative and quantitative data for splicing events, including exon skipping, alternative 5′ donor site, alternative 3′ acceptor site, and intron retention. We confirmed CloneSeq results by comparing the data with ENIGMA-recommended assays, in which several individual positive colonies were picked, amplified by rolling-circle amplification, and Sanger sequenced. Single-transcript alignment (STA) was performed to characterize the transcripts' sequences. **(B)** The five types of alternative splicing events, as described by Diederichs et al. that can be detected by the ABP: (1) exon skipping; (2) partial exon skipping (as a result of the usage of alternative exonic donor or acceptor site); (3) partial intron inclusion (as a result of the usage of alternative intronic donor or acceptor site);(4) intron retention;(5) insertion of cryptic exons.

suspended in PBS. Plasmids were extracted with the GeneJET Plasmid Miniprep kit (Thermo Fisher Scientific). CloneSeq libraries were constructed according to the protocol outlined by KAPA Biosystems (Wilmington, MA, USA) using the Hyper Prep kit. Briefly, DNA was sheared to an average size of 250–350 bp using sonication (Covaris, Woburn, MA, USA). DNA fragment ends were repaired and phosphorylated. An "A" base was added to the 3′ end of the blunted fragments, followed by ligation of single-indexed adapters via T-A mediated ligation. The size and concentration of the DNA library were determined using the TapeStation 2200. Massively Parallel Sequencing (MPS) was performed on an Illumina MiSeq, which generated 2 × 250 paired-end reads. Sequencing reads were aligned to the hg19 reference genome and analyzed using Ambry's Bioinformatic Pipeline (see below).

For whole transcriptome RNA-Seq, globin mRNA and ribosomal RNA were depleted using the Globin-Zero Gold rRNA removal kit (Illumina, San Diego, CA, USA). After depletion, RNA was fragmented and single-indexed cDNA libraries were generated using an RNA Hyper Prep kit

(KAPA Biosystems). Quality control was performed using the TapeStation 2200. Libraries were checked for average fragment size, concentration, and the presence of spurious peaks such as adapter dimers. Concentration was confirmed using a Qubit Fluorometer (ThermoFisher). Libraries were sequenced to a depth of 1×10^8 paired end reads (2×150 bp) per sample on the Illumina NextSeq platform. Sequencing reads were aligned to the hg19 reference genome and analyzed using Ambry's Bioinformatic Pipeline for alternative splicing events and differentially expressed genes.

Bioinformatics Analysis

Paired-end RNA-seq reads (2×250 bp) and Sanger sequencing reads ($\sim$1,100 bp) were first aligned to the hg19 human reference genome. For Sanger reads, GMAP aligner (version 2016-04-04) was used with default parameters to perform single transcript alignment (STA) of very long reads. For CloneSeq reads, STAR aligner v2.5.2a was used with default parameters except the "outSAMtype" parameter was set to "BAM SortedByCoordinate." The mapped reads were then analyzed by our customized Ambry Bioinformatics Pipeline (ABP) software to detect splicing events such as exon skipping, alternative 5′ donor site, alternative 3′ acceptor site, and intron retention (15). These events are detected by the pipeline, based on the alignments against the reference genome (**Figure 1B**): (1) exon skipping, if there is no reads align to one exon or several consecutive exons; (2) partial exon skipping, if there is no read alignment in one end of an exon; (3) partial intron inclusion, if there is alignment in one end of an intron; (4) intron retention, if there is alignment in a whole intron; (5) cryptic exon, if there is alignment in the middle of an intron and no alignment in the rest of the intron. Schematic representations of these splicing events are illustrated in **Figure 1B**. To quantify splicing events, we calculated the percentage of alternative splicing event against a given transcript/isoform: percent of alternative splicing event = (number of reads supporting alternative splicing event)/(number of all reads in the region covering alternative splicing event). To filter out noise caused by sequencing and alignment errors, or due to the expression of ultra-rare isoforms, the splicing events with "number of reads supporting alternative splicing event" <20, or "number of all reads in the region covering splicing event" <50, or "percent of splicing event" <2.5% were filtered out. HGVS nomenclature values were approximate for intron retention and alternative splicing site events due to differences in alignments based on NGS reads.

RESULTS

Quantitative and Qualitative RNA Analysis of the Variant *BRCA1* c.5467+5G>T

The variant *BRCA1* c.5467+5G>T, which impairs the native donor splice site of *BRCA1* exon 23, has been described to result in skipping of exon 23 (Δ23) (8). This variant is currently classified as VUS by ENIGMA (class 3). LCLs were obtained from carriers of the variant and 2 controls, and reverse-transcriptase PCR (RT-PCR) was performed using the conditions

recommended by Whiley et al. including the use of the same primers, reverse transcriptase, and NMD inhibitors puromycin (puro) and cycloheximide (CHX) (8). Skipping of exon 23 (Δ23) was clearly detected by digital electrophoresis of the RT-PCR products in the *BRCA1* c.5467+5G>T carrier's LCL, but not in two control LCLs treated with NMD inhibitors (**Figure 2A**). RT-PCR products were cloned and CloneSeq was performed on these samples for sequence characterization and quantification. The sequence and absolute number of reads observed in the carrier and control cell lines are shown using Sashimi plots (**Figure 2B**). Sashimi plots provide a quantitative visualization of aligned MPS reads that enables quantitative comparison of exon usage across samples (16). A total of 4,018 reads supporting exon 23 skipping (r.5407_5467del61) were detected in the *BRCA1* c.5467+5G>T LCL, whereas none was detected in the control LCL (**Figure 2B**). Abnormal transcripts levels were then measured as a "percent spliced in index" (PSI) (**Figure 2C**). PSI demonstrates the ratio between reads including or excluding exons, indicating how efficiently sequences of interest are spliced into transcripts (17). This analysis indicated that $\sim$25% of transcripts expressed by the *BRCA1* c.5467+5G>T LCL contains skipping of exon 23, whereas skipping of exon 23 was not detected in negative control LCLs (**Figure 2C**).

To validate the CloneSeq results we performed, in the same set of samples, the mRNA splicing assays recommended by the members of the ENIGMA consortium (8), including capillary electrophoresis (**Figure 2D**), and Sanger sequencing of subcloned transcripts (**Figure 2E**). Capillary electrophoresis clearly detected Δ23 in the *BRCA1* c.5467+5G>T LCL, in addition to the full-length WT transcript (**Figure 2D**). Sanger sequencing also detected Δ23 exclusively in the *BRCA1* c.5467+5G>T carrier's LCL (**Figure 2E**).

CloneSeq Characterization of the Pathogenic Alteration *BRCA1* c.135-1G>T

The variant *BRCA1* c.135-1G>T, which impairs the native acceptor site of *BRCA1* exon 5, has been associated with multiple splicing isoforms (8, 18), including an abundant transcript with skipping of exon 5 (Δ5). This variant is currently classified as pathogenic by ENIGMA (class 5). RT-PCR for the *BRCA1* c.135-1G>T carrier's LCL and control LCLs was performed following ENIGMA recommendations and analyzed by digital electrophoresis, which detected a band consistent in size with Δ5 (**Figure 3A**). RT-PCR products were cloned and CloneSeq performed, and the sequence and absolute number of reads observed in the carrier and control cell lines are shown using Sashimi plots (**Figure 3B**). A total of 10,902 reads supporting exon 5 skipping (r.135_212del78) were detected in the *BRCA1* c.135-1G>T LCL, whereas only 584 reads were detected in the control LCL (**Figure 3B**). Quantification of splicing events indicated that the *BRCA1* c.135-1G>T LCL expresses $\sim$50% of transcripts with skipping of exon 5, whereas LCL negative controls have negligible levels of Δ5 (**Figure 3C**). Individual colonies were selected for transcript confirmation by Sanger sequencing, and the Δ5 transcript was the most
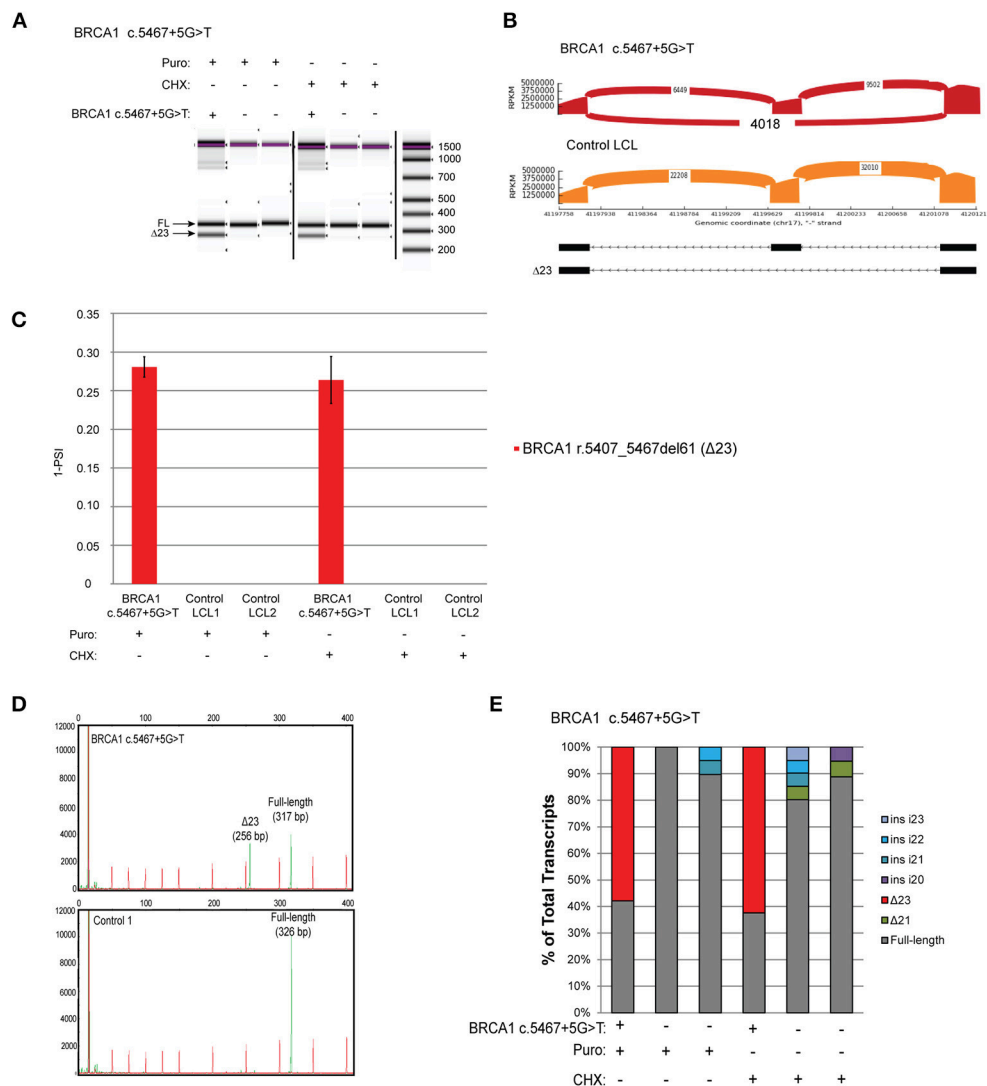
FIGURE 2 | Quantitative and Qualitative RNA Analysis of the variant *BRCA1* c.5467+5G>T **(A)** Digital electrophoresis analysis of the RT-PCR performed on the *BRCA1* c.5467+5G>T carrier LCL and control LCLs treated with puro or CHX. **(B)** Sashimi plots of CloneSeq performed in the *BRCA1* c.5467+5G>T carrier LCL and control LCL. **(C)** Relative quantification of CloneSeq results shown as "percent spliced in index" (PSI). **(D)** Capillary electrophoresis analysis of the RT-PCR performed in the *BRCA1* c.5467+5G>T carrier LCL and control LCL. **(E)** RT-PCR products were cloned and individual colonies were selected for Sanger sequencing. Median relative frequency of each detected transcript is graphed (*n* = 3 biological replicates).

abundant abnormal transcripts in the *BRCA1* c.135-1G>T LCL (**Figure 3D**).

CloneSeq Characterization of the Pathogenic Alteration *BRCA2* c.8632+1G>A

BRCA2 c.8632+1G>A, located at intron 20 of *BRCA2*, was shown to result in skipping of exon 20 (Δ20) (8), and is currently classified by ENIGMA as a pathogenic alteration (class 5). RT-PCR of the *BRCA2* c.8632+1G>A carrier's LCL and control LCLs was performed following ENIGMA recommendations. Digital electrophoresis analysis of the RT-PCR products detected several

alternative transcripts, in addition to the full-length mRNA (**Figure 4A**). RT-PCR products were cloned and CloneSeq was performed. The sequence and absolute number of reads observed are shown in Sashimi plots for the carrier and control cell lines demonstrating a total of 3,002 reads supporting exon 20 skipping (r.8488_8632del145) in the *BRCA2* c.8632+1G>A LCL, whereas only 25 reads supporting Δ20 were detected in the control LCL (**Figure 4B**). Quantification of splicing events indicated that ~20% of the transcripts expressed by the *BRCA2* c.8632+1G>A LCL have skipping of exon 20, whereas LCL negative controls have negligible levels of Δ20 (**Figure 4C**). Several alternative spliced transcripts, detected both in the *BRCA2* c.8632+1G>A LCL and controls, were also identified
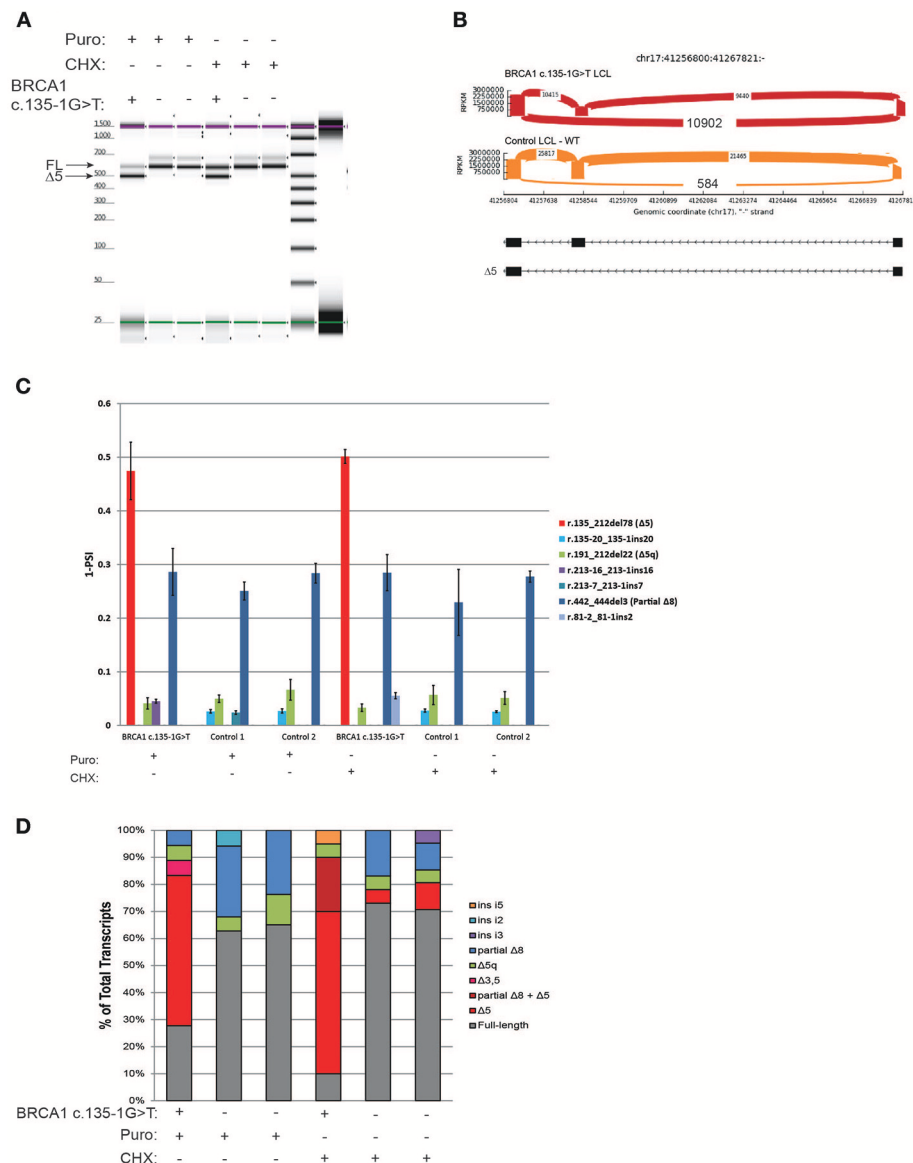
FIGURE 3 | CloneSeq characterization of the pathogenic alteration *BRCA1* c.135-1G>T **(A)** Digital electrophoresis analysis of the RT-PCR performed on the *BRCA1* c.135-1G>T carrier LCL and control LCLs treated with puro or CHX. **(B)** Sashimi plots of CloneSeq performed in the *BRCA1* c.135-1G>T carrier LCL and control LCL. **(C)** Relative quantification of CloneSeq results shown as "percent spliced in index" (PSI). **(D)** RT-PCR products were cloned and individual colonies were selected for Sanger sequencing. Median relative frequency of each detected transcript is graphed (*n* = 3 biological replicates).

(**Figure 4C**). Individual colonies were selected for transcript confirmation by Sanger sequencing, which identified the most abundant abnormal transcript Δ20 in the *BRCA2* c.8632+1G>A LCL, in addition to confirming other minor alternative isoforms detected by CloneSeq (**Figure 4D**).

Characterization of the Variant *BRCA2* c.9501+3A>T in LCLs and Blood Samples

BRCA2 c.9501+3A>T is located in the native donor site of intron 25. This variant was reported to result in low levels of skipping of exon 25 (Δ25), and it is currently classified as

benign (class 1) by ENIGMA (8). RT-PCR was performed on the carrier's LCL and on the control cells. Digital electrophoresis identified a minor band consistent with the size of Δ25 in the *BRCA2* c.9501+3A>T LCL that was not detected in negative controls (**Figure 5A**). RT-PCR products were cloned, and individual colonies were selected for Sanger sequencing, which detected Δ25 (r.9257_9501del245) in the *BRCA2* c.9501+3A>T LCL, in addition to low levels of other alternatively spliced transcripts in controls (**Figure 5B**). To more accurately quantify the relative abundance of Δ25, we performed CloneSeq. The assay detected a total of 2,883 reads supporting exon 25
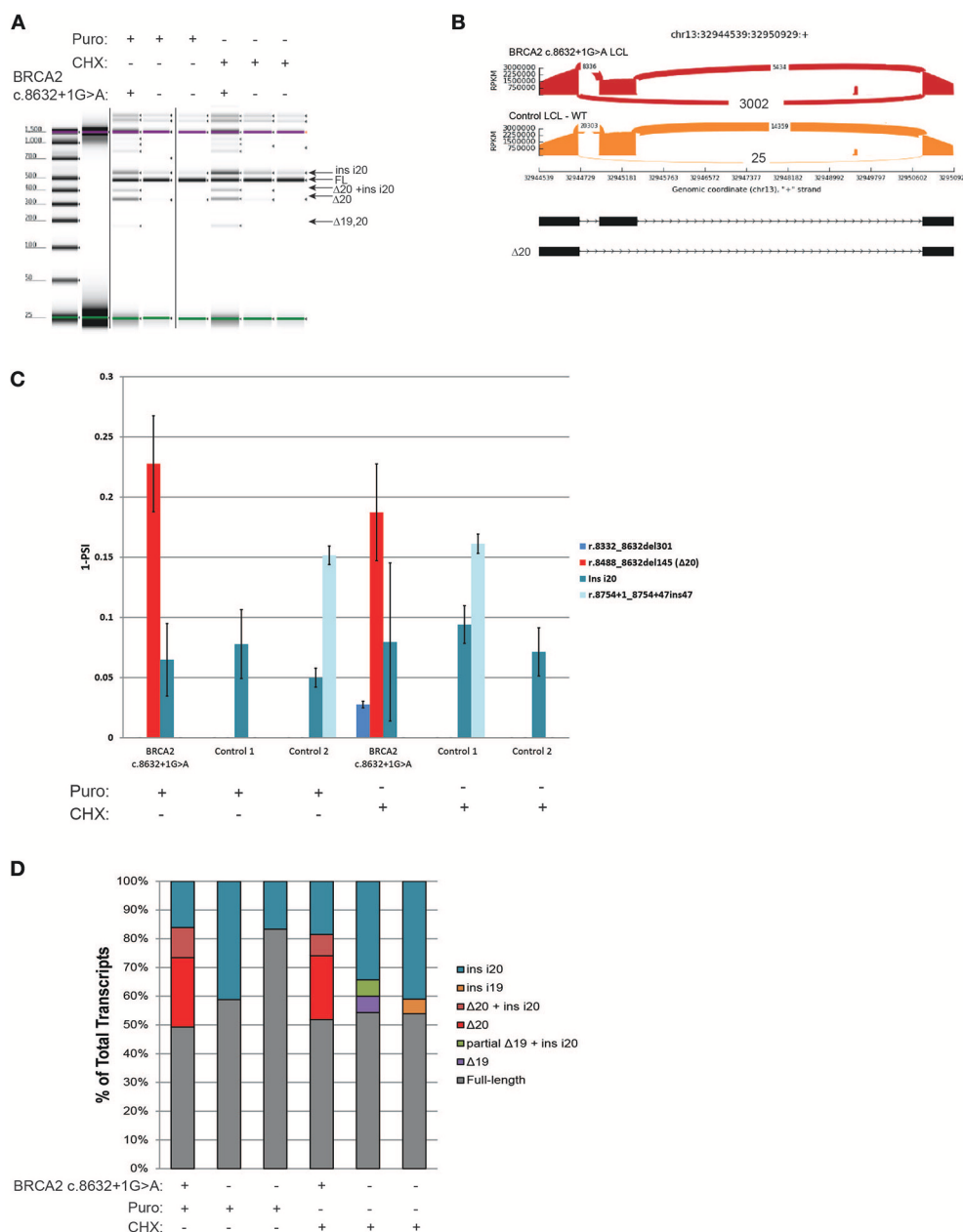
FIGURE 4 | CloneSeq characterization of the pathogenic alteration *BRCA2* c.8632+1G>A **(A)** Digital electrophoresis analysis of the RT-PCR performed on the *BRCA2* c.8632+1G>A carrier LCL and control LCLs treated with puro or CHX. **(B)** Sashimi plots of CloneSeq performed in the *BRCA2* c.8632+1G>A carrier LCL and control LCL. **(C)** Relative quantification of CloneSeq results shown as "percent spliced in index" (PSI). **(D)** RT-PCR products were cloned and individual colonies were selected for Sanger sequencing. Median relative frequency of each detected transcript is graphed (*n* = 3 biological replicates).

skipping in the *BRCA2* c.9501+3A>T LCL, whereas no reads supporting skipping of this exon were detected in the control LCL (**Figure 5C**). The reads supporting Δ25 were ~10% of the total splicing events detected in the *BRCA2* c.9501+3A>T LCL (**Figure 5D**).

Subsequently, we compared CloneSeq LCL results with an analysis of RNA isolated from the blood cells of carriers

of the *BRCA2* c.9501+3A>T alteration. RT-PCR performed on RNA from individuals that are heterozygous for *BRCA2* c.9501+3A>T (proband and mother) and negative controls (father, LCL-, normal breast RNA, and normal blood controls) detected Δ25 only in the positive samples (**Figure 5E**). CloneSeq was performed to quantify Δ25 in these samples, which detected 9,279 and 4,066 splicing events supporting Δ25 in the proband
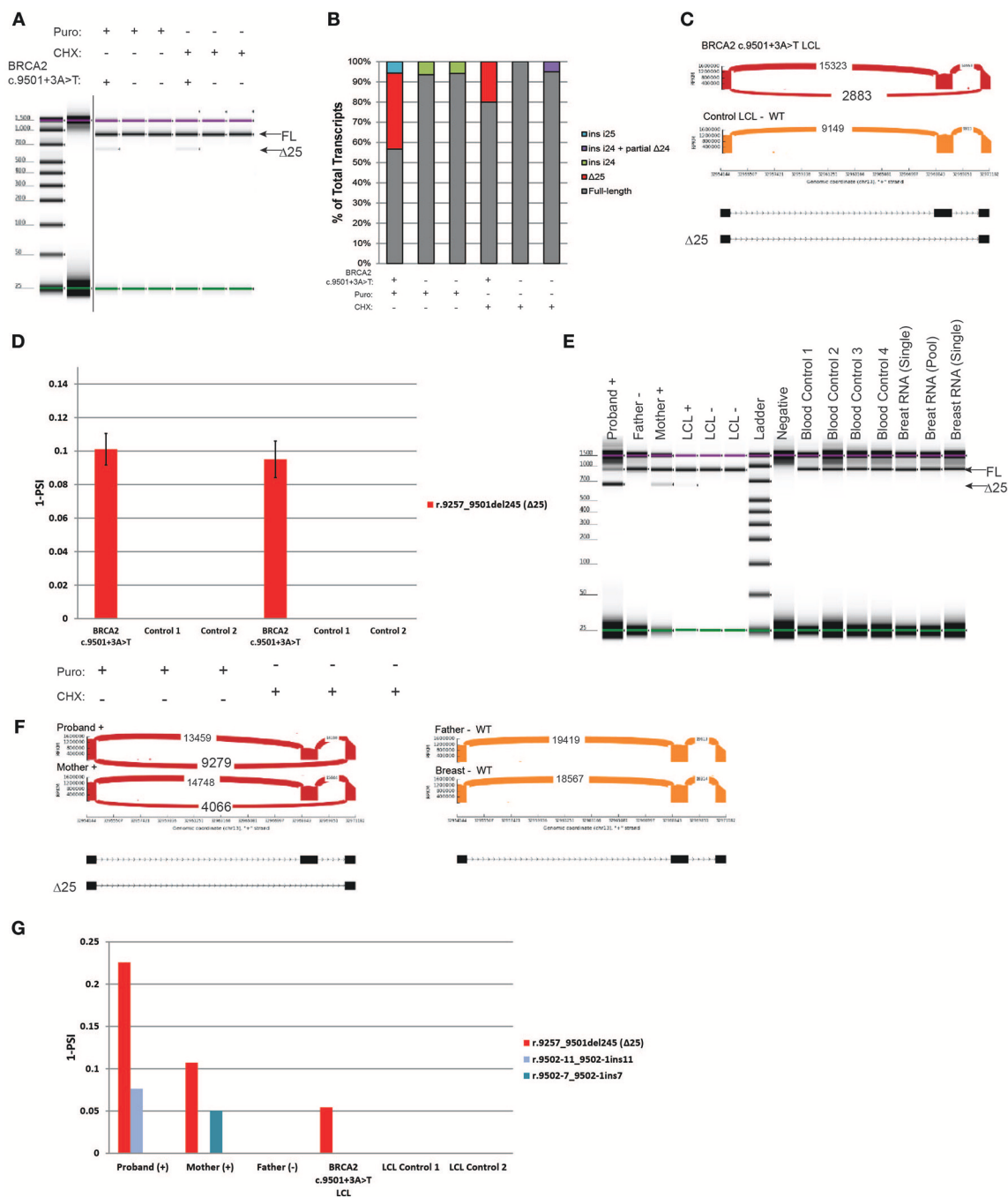
FIGURE 5 | Characterization of the variant *BRCA2* c.9501+3A>T in LCLs and patient samples **(A)** Digital electrophoresis analysis of the RT-PCR performed on the *BRCA2* c.9501+3A>T carrier LCL and control LCLs treated with puro or CHX. **(B)** RT-PCR products were cloned and individual colonies were selected for Sanger sequencing. Median relative frequency of each detected transcript is graphed (*n* = 3 biological replicates). **(C)** Sashimi plots of CloneSeq performed in the *BRCA2* c.9501+3A>T carrier LCL and control LCL. **(D)** Relative quantification (PSI) of CloneSeq results obtained from the *BRCA2* c.9501+3A>T carrier LCL and control LCL. **(E)** Digital electrophoresis analysis of the RT-PCR performed on RNA obtained from the blood of *BRCA2* c.9501+3A>T carriers and control samples. **(F)** Sashimi plots of CloneSeq performed on RNA obtained from the blood of the *BRCA2* c.9501+3A>T carriers (proband and mother) and control individuals negative for the alteration (father and control breast tissue). **(G)** Relative quantification (PSI) of CloneSeq results obtained from *BRCA2* c.9501+3A>T carriers' blood, *BRCA2* c.9501+3A>T carrier LCL (LCL+), and negative controls (Father's blood and control LCLs).

and mother's samples respectively, while none were detected in the negative controls (**Figure 5F**). Interestingly, the percentage of Δ25 transcripts varies among different carriers. For the proband,

Δ25 represents ~20% of splicing events, while it accounted for ~10% in the mother's sample (**Figure 5G**), ~5% in the LCL+ *BRCA2* c.9501+3A>T without NMD inhibition (**Figure 5G**),

and ∼10% of splicing events when the LCL+ is treated with inhibitors (**Figure 5D**).

Characterization of a Novel Variant, *BRCA1* c.5152+5G>T

We next analyzed a novel uncharacterized VUS, *BRCA1* c.5152+5G>T, identified in a HBOC family (**Figure 6A**). This rare variant is located in the donor splice site of intron 18 at a highly conserved position, and was predicted by several splicing *in silico* programs to abolish the splice site (data not shown). We obtained blood samples from patients that are heterozygous for *BRCA1* c.5152+5G>T (proband and father) as well as samples from negative individuals (proband's mother and sister). To characterize the variant we performed whole transcriptome sequencing (WTS) on the proband and control blood samples. WTS detected 14 reads supporting skipping of exon 18 (Δ18, r.5075_5152del78) and 12 reads supporting the WT transcript in the proband (**Figure 6B**, top). In the control blood sample, only WT reads ($n = 36$) were detected (**Figure 6B**, bottom). Primers were designed in the flanking exons and RT-PCR was performed. Digital electrophoresis analysis of RT-PCR products identified a band corresponding to Δ18, exclusively in samples from the heterozygous carriers (**Figure 6C**). Sanger sequencing of subcloned transcripts was then performed, which confirmed Δ18 sequence (**Figure 6D**), indicating it results in in-frame skipping of the important BRCT functional domain of BRCA1 (19). Finally, using CloneSeq, we detected and quantified the Δ18 transcript in heterozygous individuals, which was undetectable in non-carriers (**Figure 6E**). CloneSeq detected 11,824 reads supporting Δ18 (r. c.5075_5152del78) and 13,268 reads supporting WT transcript in the proband (**Figure 6E**, left top). In the proband's father, CloneSeq detected 7,412 reads supporting Δ18, and 6,229 reads supporting WT transcript (**Figure 6E**, left bottom). Quantitatively, we found heterozygous individuals to express ∼40% of Δ18 transcripts (**Figure 6F**). Of note, analysis of the LCL harboring the pathogenic alteration *BRCA1* c.5152+1G>T, affecting the same donor splice site as *BRCA1* c.5152+5G>T, also led to similar expression of the abnormal transcript Δ18 (**Figures 6C,F**). Altogether, these data were used to reclassify *BRCA1* c.5152+5G>T from VUS to likely pathogenic, and therefore a clinically actionable alteration.

DISCUSSION

Genetic testing for HBOC is becoming increasingly widespread in the era of precision medicine (20, 21). The implementation of next-generation sequencing has resulted in an explosion of genetic data, and germline variants with unknown function are regularly detected by clinical diagnostic laboratories (22). In particular, VUS in clinically actionable genes, such as the HBOC susceptibility genes *BRCA1* and *BRCA2*, pose a quandary to medical providers and patients (23–25). A specific challenge is the large percentage of VUS in the *BRCA1* and *BRCA2* genes predicted to affect splicing by *in silico* tools, but lack RNA evidence (26). In part, this is due to the scarcity of high-throughput assays designed to interrogate the impact of

variants on splicing. The American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) guidelines for the interpretation of germline sequence variants recommends the use of multiple types of evidence for classifying variants identified by DNA genetic testing, such as functional evidence, allele frequency data, computational and *in silico* predictions, and phenotype/family history (1). Therefore, RNA testing is critical to perform a comprehensive interpretation of sequence variants predicted to affect splicing. With this in mind, there are several RNA assays that have been used to characterize splicing alterations, each assay possessing its own advantages and drawbacks. These include the use of hybrid minigenes for the characterization of candidate splicing variants (27, 28). Given that a hybrid minigene assay analyzes the splicing outcome of a single allele, it is a great toll to evaluate allele-specific expression, i.e., the demonstration that the variant allele produces highly expressed abnormal transcripts predicted to induce NMD or to disrupt clinically important residues, an important step for the classification of splicing variants as pathogenic (10, 11). A caveat of this technique is the requirement of constructing and using artificial vectors (which are not available to most commercial diagnostic laboratories), and the need to test cell lines instead of samples derived from the patient being evaluated. The assays recommended by the ENIGMA consortium to characterize *BRCA1* and *BRCA2* transcripts include capillary electrophoresis and Sanger sequencing of subcloned transcripts, (8, 29). These assays are accessible to most laboratories, can be performed in patient samples, do not require the use of expression vectors, and can perform qualitative characterization of splicing variants; however, these assays only provide semi-quantitative data and lack the throughput required to analyze a large amount of variants (30, 31). Alternatively, there are quantitative approaches, including real-time and digital PCR, that provide robust and reliable quantitative data (29, 32), but cannot perform qualitative analysis (i.e., these assays are unable to reveal the precise transcript sequence). Here we describe a novel RNA MPS method, CloneSeq, and demonstrated that this technique is capable of performing reliable high-throughput quantitative and qualitative analysis of splicing variants, a necessary feature to obtain evidence for the large number of alterations predicted to affect splicing in HBOC genes.

Using CloneSeq coupled to our custom ABP bioinformatics analysis we were able to detect all major splicing aberrations described by the ENIGMA consortium in four well-characterized LCLs (8), with the advantage of obtaining absolute and relative quantification of the expressed transcripts (**Figures 2–5**). CloneSeq proficiently detected abnormal transcripts, as well as less abundant alternative splicing events, as it analyzed thousands of cloned reads. We validated the CloneSeq results using digital and capillary electrophoresis and Sanger sequencing of subcloned transcripts. Digital and capillary electrophoresis identified the abundant abnormally spliced transcripts detected by CloneSeq, but these methods were incapable of providing proper quantification of transcript levels due to their semi-quantitative nature. Additionally, sequencing was needed to
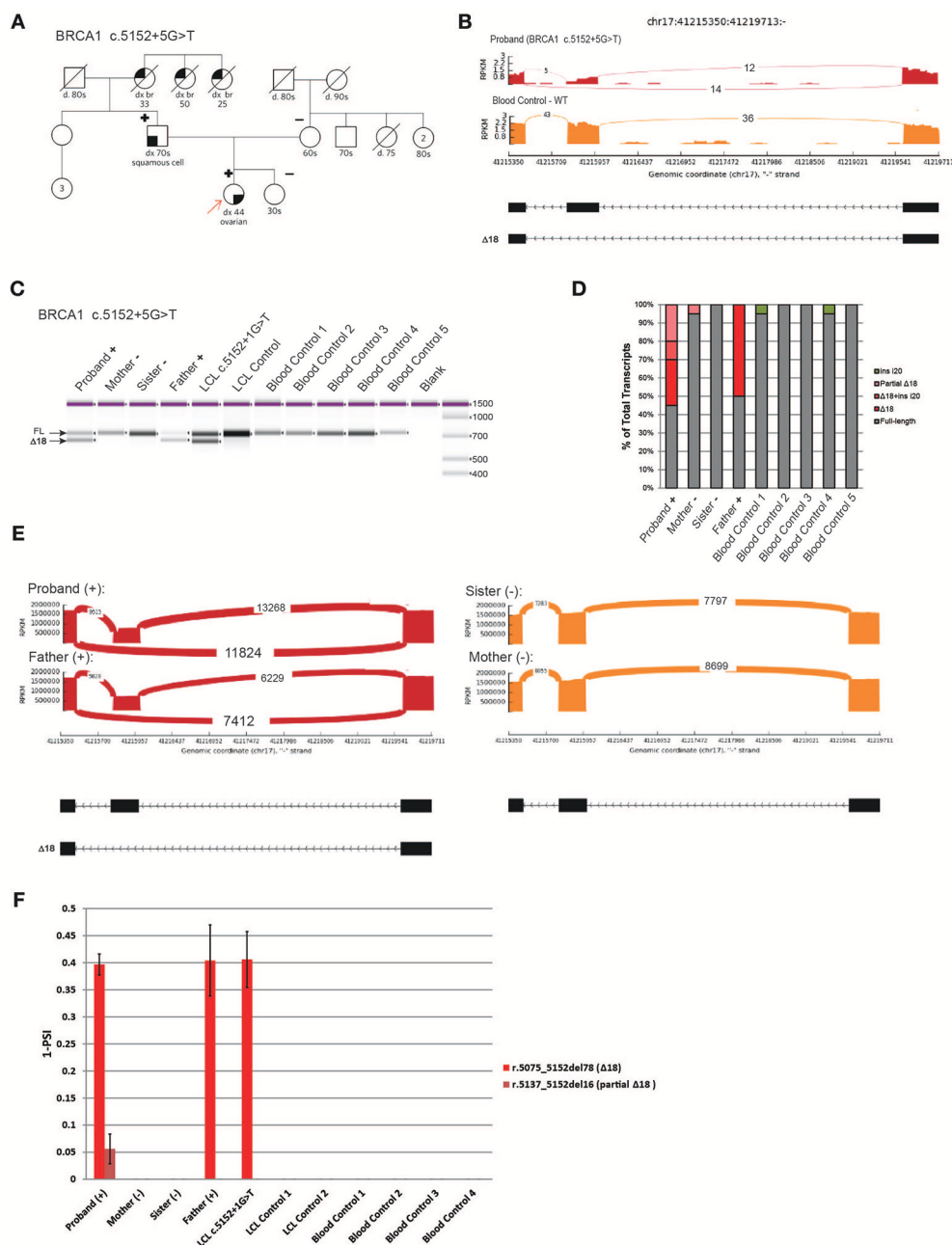
FIGURE 6 | Characterization of a novel variant, *BRCA1* c.5152+5G>T **(A)** Pedigree of HBOC family carrying the variant *BRCA1* c.5152+5G>T. **(B)** Sashimi plots of whole transcriptome sequencing performed on RNA obtained from the blood of a *BRCA1* c.5152+5G>T carrier and control sample. **(C)** Digital electrophoresis analysis of the RT-PCR performed in RNA obtained from blood of *BRCA1* c.5152+5G>T carriers (proband and father), *BRCA1* c.5152+1G>T carrier LCL, and control samples negative for the alteration (RNA obtained from the mother and sister's blood, negative LCL, negative unrelated blood controls). **(D)** RT-PCR products were cloned and individual colonies were selected for Sanger sequencing. Median relative frequency of each detected transcript is graphed (n = 3 biological replicates). **(E)** Sashimi plots of CloneSeq performed in RNA obtained from blood of the *BRCA1* c.5152+5G>T carriers (proband and father) and control individuals negative for the alteration (sister and mother). **(F)** Relative quantification (PSI) of CloneSeq results obtained from *BRCA1* c.5152+5G>T carriers' blood (proband and father), *BRCA1* c.5152+1G>T carrier LCL, and several controls negative for the alteration (mother and sister, control LCL, and unrelated blood controls).

confidently identify the exact splicing event detected by digital and capillary electrophoresis. Both CloneSeq and Sanger sequencing were able to precisely determine the sequence of the transcripts, however, CloneSeq performed sequencing

of tens of thousands of transcripts (average number of *BRCA1* or *BRCA2* mapped reads *per* sample tested was 24,803). In comparison, Sanger sequencing is limited to low-throughput sequencing of colonies, each containing a

single subcloned transcript (median=56 clones sequenced *per* LCL). As examples of its high analytical sensitivity, CloneSeq was able to detect the splicing isoforms induced by the variants *BRCA1* c.135-1G>T and *BRCA2* c.8632+1G>A, and rare alternative spliced isoforms previously reported by ENIGMA (**Figures 3, 4**). Ultimately, CloneSeq's targeted ultra-deep *locus* sequencing of *BRCA1* or *BRCA2* proved to be a fundamental feature for the bioinformatics qualitative and quantitative characterization of splicing alterations in these genes.

In addition to LCLs, we analyzed RNA extracted from blood samples obtained from variant carriers and controls. For the previously characterized *BRCA2* c.9501+3A>T benign variant (class 1), we were able to compare LCL RNA data with RNA data from the blood of heterozygous *BRCA2* c.9501+3A>T carriers. Similar to Sanger sequencing, CloneSeq detected the major abnormal splicing event associated with this variant, skipping of exon 25, both in LCLs and in carriers' blood RNA. Because quantification of abnormally and alternatively spliced transcripts is fundamental to predict pathogenicity (10, 11, 29), we quantified the impact of this benign variant on splicing levels. The percentage of skipped exon 25 identified in different *BRCA2* c.9501+3A>T carriers ranged from ∼20 to ∼10% of total splicing events, suggesting that an alteration resulting in less than ∼20% of abnormal splicing is clinically benign. However, it is important to note, this is an indirect measurement of each allele's expression, since we were unable to perform allele-specific expression in the individuals we tested due to the lack of informative variants in the coding sequence of *BRCA2* in the respective samples. In order to mitigate this caveat, we ran a series of normal blood and tissue controls to identify and differentiate any physiologic alternatively spliced isoform from abnormal transcripts identified in the variant carriers (**Figures 5E–G**). The CloneSeq results are also in agreement with a minigene single-allele analysis of this alteration that reported ∼13% of Δ25 is induced by the variant *BRCA2* c.9501+3A>T (27). By quantifying the impact that variants with benign clinical behavior have on splicing, CloneSeq could be used in the future to identify a splicing threshold that must be reached by abnormal transcripts in order to classify a *BRCA1* or *BRCA2* alteration as pathogenic.

Lastly, we analyzed blood samples obtained from a HBOC family carrying a novel uncharacterized VUS, *BRCA1* c.5152+5G>T. To characterize the VUS we performed whole transcriptome sequencing in the proband and control blood samples. WTS detected 14 reads supporting skipping of exon 18, and 12 reads supporting the WT transcript in the proband. Using CloneSeq, we were able to detect in the proband 11,824 reads supporting skipping of exon 18 and 13,268 reads supporting the WT transcript. Comparatively, the number of reads detected by WTS vs. CloneSeq highlights the higher analytical sensitivity of the later. This supports the notion that WTS can provide biased results due to low detection yields and other technical limitations (33, 34). On the other hand, CloneSeq provided the sufficient sequencing depth necessary for transcript characterization and quantification, which proved to be indispensable to reclassify the VUS *BRCA1* c.5152+5G>T as a clinically actionable likely pathogenic alteration.

Massively parallel sequencing is revolutionizing cancer genetics by enabling the detection and characterization of sequence variants at unprecedented scale and speed. For example, depending on the technology and protocol used, the number of individuals tested *per* variant, and the number of controls tested, CloneSeq can concomitantly perform analysis of multiple variants in a single MPS run (up to 96 samples). From the initial RNA extraction steps to the final bioinformatics analysis, the protocol described here can analyze multiple samples in less than 10 days (**Figure 1A**). Even though NGS technologies have evolved quickly over the past decade, leading to a substantial decrease in the cost per megabase (30), the cost of NGS assays may still pose challenges to laboratories with low throughput (31). However, implementation of automated steps and the development of innovative sequencing technologies could reduce the cost and time-frame of CloneSeq even further in the near future. Besides laboratory costs, one important issue to consider is the impact RNA genetic testing has on variant classification. One of the caveats of DNA MPS multi-gene testing is the high rate of VUS results (35). Since RNA genetic testing has the potential to reduce VUS rates, future research should investigate the broader impact these tests have on the overall clinical management of patients identified with germline variants in *BRCA1* and *BRCA2*.

CONCLUSION

CloneSeq is an alternative to the current splicing assays recommended by the ENIGMA consortium. Due to its high-throughput format, quantitative and qualitative abilities, and high analytical sensitivity, CloneSeq has the potential to improve the interpretation of splicing sequence variants detected by HBOC clinical genetic tests. The enhanced classification of these germline variants as either disease-causing or neutral is fundamental to offering preventive pre-symptomatic measures or targeted treatment to the correct individuals.

AUTHOR CONTRIBUTIONS

FUNDING

ACKNOWLEDGMENTS

REFERENCES

1. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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* (2015) 17:405–24. doi: 10.1038/gim.2015.30

2. Pesaran T, Karam R, Huether R, Li S, Farber-Katz S, Chamberlin A, et al. Beyond DNA: an integrated and functional approach for classifying germline variants in breast cancer genes. *Int J Breast Cancer* (2016) 2016:2469523. doi: 10.1155/2016/2469523

3. Nielsen FC, van Overeem Hansen T, Sorensen CS. Hereditary breast and ovarian cancer: new genes in confined pathways. *Nat Rev Cancer* (2016) 16:599–612. doi: 10.1038/nrc.2016.72

4. Farrugia DJ, Agarwal MK, Pankratz VS, Deffenbaugh AM, Pruss D, Frye C, et al. Functional assays for classification of BRCA2 variants of uncertain significance. *Cancer Res.* (2008) 68:3523–31. doi: 10.1158/0008-5472.CAN-07-1587

5. Walker LC, Whiley PJ, Couch FJ, Farrugia DJ, Healey S, Eccles DM, et al. Detection of splicing aberrations caused by BRCA1 and BRCA2 sequence variants encoding missense substitutions: implications for prediction of pathogenicity. *Hum Mutat.* (2010) 31:E1484–1505. doi: 10.1002/humu.21267

6. Thomassen M, Blanco A, Montagna M, Hansen TV, Pedersen IS, Gutierrez-Enriquez S, et al. Characterization of BRCA1 and BRCA2 splicing variants: a collaborative report by ENIGMA consortium members. *Breast Cancer Res Treat.* (2012) 132:1009–23. doi: 10.1007/s10549-011-1674-0

7. Caminsky N, Mucaki EJ, Rogan PK. Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. *F1000Res* (2014) 3:282. doi: 10.12688/f1000research.5654.1

8. Whiley PJ, de la Hoya M, Thomassen M, Becker A, Brandao R, Pedersen IS, et al. Comparison of mRNA splicing assay protocols across multiple laboratories: recommendations for best practice in standardized clinical testing. *Clin Chem.* (2014) 60:341–52. doi: 10.1373/clinchem.2013.210658

9. Spurdle AB, Healey S, Devereau A, Hogervorst FB, Monteiro AN, Nathanson KL, et al. ENIGMA–evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Hum Mutat.* (2012) 33:2–7. doi: 10.1002/humu.21628

10. Colombo M, Blok MJ, Whiley P, Santamarina M, Gutierrez-Enriquez S, Romero A, et al. Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium. *Hum Mol Genet.* (2014) 23:3666–80. doi: 10.1093/hmg/ddu075

11. Fackenthal JD, Yoshimatsu T, Zhang B, de Garibay GR, Colombo M, De Vecchi G, et al. Naturally occurring BRCA2 alternative mRNA splicing events in clinically relevant samples. *J Med Genet.* (2016) 53:548–58. doi: 10.1136/jmedgenet-2015-103570

12. Karam R, Wengrod J, Gardner LB, Wilkinson MF. Regulation of nonsense-mediated mRNA decay: implications for physiology and disease. *Biochim Biophys Acta* (2013) 1829:624–33. doi: 10.1016/j.bbagrm.2013.03.002

13. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* (2008) 456:470–6. doi: 10.1038/nature07509

14. Karam R, Carvalho J, Bruno I, Graziadio C, Senz J, Huntsman D, et al. The NMD mRNA surveillance pathway downregulates aberrant E-cadherin transcripts in gastric cancer cells and in CDH1 mutation carriers. *Oncogene* (2008) 27:4255–60. doi: 10.1038/onc.2008.62

15. Diederichs S, Bartsch L, Berkmann JC, Frose K, Heitmann J, Hoppe C, et al. The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. *EMBO Mol Med.* (2016) 8:442–57. doi: 10.15252/emmm.201506055

16. Katz Y, Wang ET, Silterra J, Schwartz S, Wong B, Thorvaldsdottir H, et al. Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics* (2015) 31:2400–2. doi: 10.1093/bioinformatics/btv034

17. Schafer S, Miao K, Benson CC, Heinig M, Cook SA, Hubner N. Alternative splicing signatures in RNA-seq data: percent spliced in (PSI). *Curr Protoc Hum Genet.* (2015) 87:11–4. doi: 10.1002/0471142905.hg1116s87

18. Tesoriero AA, Wong EM, Jenkins MA, Hopper JL, Brown MA, Chenevix-Trench G, et al. Molecular characterization and cancer risk associated with BRCA1 and BRCA2 splice site variants identified in multiple-case breast cancer families. *Hum Mutat.* (2005) 26:495. doi: 10.1002/humu.9379

19. Clark SL, Rodriguez AM, Snyder RR, Hankins GD, Boehning D. Structure-function of the tumor suppressor BRCA1. *Comput Struct Biotechnol J.* (2012) 1:e201204005. doi: 10.5936/csbj.201204005

20. Evans DG, Wisely J, Clancy T, Lalloo F, Wilson M, Johnson R, et al. Longer term effects of the Angelina Jolie effect: increased risk-reducing mastectomy rates in BRCA carriers and other high-risk women. *Breast Cancer Res.* (2015) 17:143. doi: 10.1186/s13058-015-0650-8

21. Afghahi A, Kurian AW. The changing landscape of genetic testing for inherited breast cancer predisposition. *Curr Treat Options Oncol.* (2017) 18:27. doi: 10.1007/s11864-017-0468-y

22. LaDuca H, Stuenkel AJ, Dolinsky JS, Keiles S, Tandy S, Pesaran T, et al. Utilization of multigene panels in hereditary cancer predisposition testing: analysis of more than 2,000 patients. *Genet Med.* (2014) 16:830–7. doi: 10.1038/gim.2014.40

23. Pal T, Cragun D, Lewis C, Doty A, Rodriguez M, Radford C, et al. A statewide survey of practitioners to assess knowledge and clinical practices regarding hereditary breast and ovarian cancer. *Genet Test Mol Biomarkers* (2013) 17:367–75. doi: 10.1089/gtmb.2012.0381

24. Garcia C, Lyon L, Littell RD, Powell CB. Comparison of risk management strategies between women testing positive for a BRCA variant of unknown significance and women with known BRCA deleterious mutations. *Genet Med.* (2014) 16:896–902. doi: 10.1038/gim.2014.48

25. Kurian AW, Li Y, Hamilton AS, Ward KC, Hawley ST, Morrow M, et al. Gaps in incorporating germline genetic testing into treatment decision-making for early-stage breast cancer. *J Clin Oncol.* (2017) 35:2232–9. doi: 10.1200/JCO.2016.71.6480

26. Houdayer C, Caux-Moncoutier V, Krieger S, Barrois M, Bonnet F, Bourdon V, et al. Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined *in silico/in vitro* studies on BRCA1 and BRCA2 variants. *Hum Mutat.* (2012) 33:1228–38. doi: 10.1002/humu.22101

27. Acedo A, Hernandez-Moro C, Curiel-Garcia A, Diez-Gomez B, Velasco EA. Functional classification of BRCA2 DNA variants by splicing assays in a large minigene with 9 exons. *Hum Mutat.* (2015) 36:210–21. doi: 10.1002/humu.22725

28. Fraile-Bethencourt E, Diez-Gomez B, Velasquez-Zapata V, Acedo A, Sanz DJ, Velasco EA. Functional classification of DNA variants by hybrid minigenes: identification of 30 spliceogenic variants of BRCA2 exons 17 and 18. *PLoS Genet.* (2017) 13:e1006691. doi: 10.1371/journal.pgen.1006691

29. de la Hoya M, Soukarieh O, Lopez-Perolio I, Vega A, Walker LC, van Ierland Y, et al. 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.* (2016) 25:2256–68. doi: 10.1093/hmg/ddw094

30. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet.* (2016) 17:333–51. doi: 10.1038/nrg.2016.49

31. Caspar SM, Dubacher N, Kopps AM, Meienberg J, Henggeler C, Matyas G. Clinical sequencing: from raw data to diagnosis with lifetime value. *Clin Genet.* (2018) 93:508–19. doi: 10.1111/cge.13190

32. Colombo M, Lopez-Perolio I, Meeks HD, Caleca L, Parsons MT, Li H, et al. The BRCA2 c.68-7T > A variant is not pathogenic: a model for clinical calibration of spliceogenicity. *Hum Mutat.* (2018) 39:729–41. doi: 10.1002/humu.23411

33. Costa V, Aprile M, Esposito R, Ciccodicola A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur J Hum Genet.* (2013) 21:134–42. doi: 10.1038/ejhg. 2012.129

34. Park E, Pan Z, Zhang Z, Lin L, Xing Y. The expanding landscape of alternative splicing variation in human populations. *Am J Hum Genet.* (2018) 102:11–26. doi: 10.1016/j.ajhg.2017.11.002

35. Bunnell AE, Garby CA, Pearson EJ, Walker SA, Panos LE, Blum JL. The clinical utility of next generation sequencing results in a community-based hereditary cancer risk program. *J Genet Couns.* (2017) 26:105–12. doi: 10.1007/s10897-016-9985-2

Check for updates

Investigation of Experimental Factors That Underlie *BRCA1/2* mRNA Isoform Expression Variation: Recommendations for Utilizing Targeted RNA Sequencing to Evaluate Potential Spliceogenic Variants

Vanessa L. Lattimore[1]*, John F. Pearson[2,3], Margaret J. Currie[1], Amanda B. Spurdle[4], kConFab Investigators[5], Bridget A. Robinson[1,6] and Logan C. Walker[1]

[1] *Mackenzie Cancer Research Group, Department of Pathology and Biomedical Science, University of Otago, Christchurch, New Zealand,* [2] *Biostatistics and Computational Biology Unit, University of Otago, Christchurch, New Zealand,* [3] *Department of Pathology and Biomedical Science, University of Otago, Christchurch, New Zealand,* [4] *Genetics and Computational Biology Division, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia,* [5] *Peter MacCallum Cancer Centre, Melbourne, VIC, Australia,* [6] *Canterbury Regional Cancer and Haematology Service, Canterbury District Health Board, Christchurch Hospital, Christchurch, New Zealand*

PCR-based RNA splicing assays are commonly used in diagnostic and research settings to assess the potential effects of variants of uncertain clinical significance in *BRCA1* and *BRCA2*. The Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) consortium completed a multicentre investigation to evaluate differences in assay design and the integrity of published data, raising a number of methodological questions associated with cell culture conditions and PCR-based protocols. We utilized targeted RNA-seq to re-assess *BRCA1* and *BRCA2* mRNA isoform expression patterns in lymphoblastoid cell lines (LCLs) previously used in the multicentre ENIGMA study. Capture of the targeted cDNA sequences was carried out using 34 *BRCA1* and 28 *BRCA2* oligonucleotides from the Illumina Truseq Targeted RNA Expression platform. Our results show that targeted RNA-seq analysis of LCLs overcomes many of the methodology limitations associated with PCR-based assays leading us to make the following observations and recommendations: (1) technical replicates ($n > 2$) of variant carriers to capture methodology induced variability associated with RNA-seq assays, (2) LCLs can undergo multiple freeze/thaw cycles and can be cultured up to 2 weeks without noticeably influencing isoform expression levels, (3) nonsense-mediated decay inhibitors are essential prior to splicing assays for comprehensive mRNA isoform detection, (4) quantitative assessment of exon:exon junction levels across *BRCA1* and *BRCA2* can help distinguish between normal and aberrant isoform expression patterns. Experimentally derived recommendations from this study will facilitate the application of targeted RNA-seq platforms for the quantitation of *BRCA1* and *BRCA2* mRNA aberrations associated with sequence variants of uncertain clinical significance.

Keywords: BRCA1, BRCA2, mRNA, quantitative, splicing, next-generation sequencing, targeted next-generation sequencing, mRNA isoforms

INTRODUCTION

At least 20% of hereditary breast and ovarian cancer cases contain germline pathogenic variants in breast cancer susceptibility genes *BRCA1* (MIM #113705) or *BRCA2* (MIM #600185) (1). Functioning as tumor suppressor genes, *BRCA1* and *BRCA2* repair single and double-stranded breaks in DNA, a process which can be compromised when variants that disrupt pre-mRNA splicing to create aberrant splice isoforms are present (2–4). These variants may directly disrupt splice sites or splicing regulatory regions, such as exonic splicing enhancers or exonic splicing silencers (5). Resulting splicing aberrations, such as major deletion/retention events and frame shifts, can lead to loss of function through the introduction of premature termination codons, leading to non-functional isoforms that are generally destroyed by nonsense-mediated decay (NMD), or *via* the production of truncated proteins (6). In addition, variants located at splicing regulatory regions, such as exonic splicing enhancers, have been shown to significantly alter the abundance of natural *BRCA1/2* isoforms (7, 8). We and others have recently employed next-generation sequencing technologies to explore the expression of mRNA isoforms in *BRCA1* and *BRCA2* variant carriers (9–11). A better understanding of expression level changes that reflect normal variation in *BRCA1/2* splicing patterns between individuals would improve our understanding of isoform regulation for identifying variability that is likely to be of clinical relevance.

In-depth qualitative data published for *BRCA1* and *BRCA2* allows for normally expressed mRNA isoforms to be distinguished more easily from aberrant isoforms (12, 13). The Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) consortium developed a 5-tier classification system, which uses mRNA splicing information to help interpret the pathogenicity of possible spliceogenic variants (14). Splicing assays for *BRCA1* and *BRCA2* mRNA isoforms have historically employed a PCR-based qualitative (or semi-quantitative) approach, with only very recent work expanding into a quantitative analysis. To assess key elements for splicing assay design and the integrity of published splicing data, a multicentre quality control investigation was conducted by the ENIGMA Splicing Working Group (15). This study highlighted the need to standardize splicing protocols between laboratories after raising a number of methodological issues associated with current PCR-based protocols, including (1) primer design that encompasses only a subset of the *BRCA1* and *BRCA2* exons, (2) non-standardized use of NMD inhibitors, (3) isoforms infrequently confirmed by sequencing, and (4) a qualitative or semi-quantitative approach to assess mRNA expression patterns, as opposed to quantitative assessment (14, 15). Although this study demonstrated variation in analytical sensitivity between samples, and the same sample between participating laboratories, the impact of underlying experimental factors remained unclear.

Targeted RNA-seq technologies potentially address many of the difficulties currently associated with PCR-based assays. For example, RNA-seq platforms enable detection of mRNA isoforms both qualitatively and quantitatively (16, 17), thus producing comprehensive transcript profiles across the entire gene(s). Assessment of the analytical sensitivity of targeted RNA-seq to both qualitatively and quantitatively measure *BRCA1/2* isoform expression in relation to experimental factors would provide a deeper understanding of the sources of mRNA isoform variation in these genes, but has yet to be evaluated.

In this study, we carried out targeted RNA-seq to assess *BRCA1/2* mRNA isoform expression patterns in lymphoblastoid cell lines (LCLs) previously utilized by the ENIGMA-led multi-centre study (15). We describe a comprehensive assessment of naturally and aberrantly occurring *BRCA1* and *BRCA2* mRNA isoforms in relation to experimental factors. Our results show that quantitation of relative levels of naturally occurring transcripts is not significantly impacted by key elements of cell-storage and culture protocols, with the possible exception of NMD-inhibition. We provide recommendations for future use of targeted RNA-seq for the analysis of variants that may disrupt RNA splicing.

MATERIALS AND METHODS

This study was approved by the Southern Health and Disability Ethics Committee (12/STH/44).

Samples

27 LCLs derived from 17 *BRCA1* or *BRCA2* rare variant carriers, and 10 healthy controls (Figure S1 and Table S1 in Supplementary Material) were obtained from Kathleen Cuningham Consortium for Research into Familial Breast Cancer. Eighteen of the cell lines (sample IDs LCL1–8 and LCL18–27) were previously used in a multi-center methods splicing analysis study coordinated through the ENIGMA consortium (15).

Variants included in this study are referred to by the recommended Human Genome Variation Society[1] nomenclature, including use of A in the ATG translation initiation codon to start the nucleotide numbering for *BRCA*1 (GenBank accession—NM_007294.3) and *BRCA2* (GenBank accession—NM_000059.3) (18).

Cell lines were cultured in RPMI 1640 media, with fetal calf serum (10%) and penicillin–streptomycin (1%), while incubated at 37°C in a 5% CO_2 atmosphere. RNA was isolated from cycloheximide treated (4 h 100 µg/mL) and untreated cells using the Qiagen RNeasy Mini Kit. RNA was reverse transcribed into cDNA using Superscript III First Strand Synthesis System (Invitrogen), according to the manufacturer's instructions.

RNA-seq Analysis

Targeted RNA-seq was undertaken on all cycloheximide treated and untreated LCLs using the Illumina Truseq Targeted RNA Expression kit. The targeted sequencing assay was custom designed in Illumina's design studio using 34 *BRCA1* and 28 *BRCA2* oligonucleotides chosen from a database of validated predesigned probes (Tables S2 and S3 and Figures S2 and S3 in Supplementary Material). Capture of the targeted cDNA sequences was performed according to the manufacturer's specifications (Figure S4 in Supplementary Material). Sequence analysis was carried out using the Illumina MiSeq platform.

---

[1]http://www.hgvs.org (Accessed: January 15, 2016).

Whole transcriptome sequencing analysis was also carried out for one control (LCL24). Using these data, *BRCA1* and *BRCA2* mRNA isoform splicing expression patterns were also obtained. RNA from LCL24 was sequenced with and without cycloheximide treatment on a HiSeq2000 using the Truseq® Stranded mRNA kit (Illumina).

Read Mapping and Processing

Targeted RNA-seq and whole transcriptome read mapping was undertaken as previously described (19). Briefly, the Homo_sapiens.GRCh37.72 reference genome was downloaded from Ensembl[2] and the chromosomes arranged into lexicographic order prior to mapping. Sequence reads were mapped using the two pass approach of the Spliced Transcripts Alignment to a Reference (STAR) aligner using the default settings unless specified otherwise (20). Maximum intron length was set to 100,000 nucleotides to accommodate for splice junctions that span the length of each gene. Detected splice junctions for each sample were extracted from STAR's SJ.out file for further analysis. Sample-specific read counts are listed in Table S4 in Supplementary Material.

Targeted RNA-seq Data Analysis: Normalization

Raw read counts of all alternative splicing events were normalized to measure the relative number of individual spliced exon/exon junctions in each sample. This was achieved by calculating the read depth of the full-length transcript (**Figure 1**). To normalize

read depth, the total read count between two adjacent exons (exons 2 and 3 in *BRCA1* and *BRCA2*, respectively) was used as the reference junction (RJ) (hereon referred to as "reference junction") to calculate the relative expression of the full-length and alternative mRNA transcripts for both *BRCA1* and *BRCA2*, respectively. To achieve this, the sum of all non-overlapping exon 2–3 alternative splicing events was deducted from the RJ for each gene independently. This leaves a proportion of reads that solely represent the full-length transcript (**Figure 1B**i). To determine the relative proportion of each splice junction, the total number of reads for each sample was first calculated. This consisted of the sum of the reads encompassing the alternative splicing events together and the previously calculated full-length transcript (using RJ exons 2–3, see above) (**Figure 1B**ii). The relative proportion of the individual isoforms in each sample is determined by dividing its respective read count by the total number of reads for that sample (**Figure 1B**iii). These expression values were incorporated into a comparative expression analysis, using the mean and SE (95%) of each isoform across the controls. This approach does not account for the possibility that some of the detected alternative splicing events may occur concurrently. Relative expression ranges were calculated based on the criteria that at least two control samples expressed the transcript with more than 10 reads, each sample was represented by more than 10,000 reads per gene, and each sample expressed at least two minor transcripts for the studied gene.

Alternative events were excluded from these calculations if they had questionable probe efficiency (such as was observed for *BRCA1* Δ9–10) or were common NAG events (21). These are
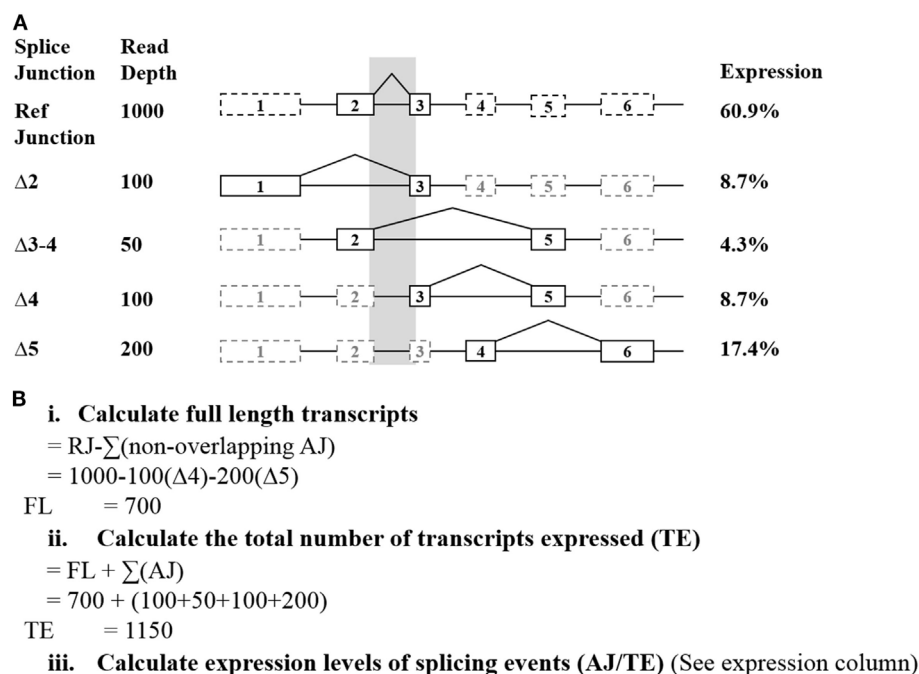
FIGURE 1 | Exemplar of how the relative expression of each splice junction was calculated from targeted RNA-seq data. **(A)** The shaded region indicates the alternative splicing events excluded from the full-length calculations. Solid lines indicate the exons directly involved with an exon skipping event.
(B) Calculations used to determine the relative expression of each detected junction. Abbreviations: AJ, alternate junction; RJ, reference junction.

shown to commonly co-occur with the other events detected, and so are not deducted as separate alternative splicing events when calculating the full-length transcript. The resulting proportions were compared using complementary log–log confidence intervals (22).

Quantitative PCR (qPCR) Validation of Splicing Events

The expression levels observed for the *BRCA1* exon 10–11, Δ10, and Δ9–10 junctions were assessed using Taqman qPCR assays and Roche LightCycler® 480 platform. Primers were designed to encompass each targeted junction type, with the corresponding probe spanning the junction (Table S5 in Supplementary Material). All qPCR assays were carried out in triplicate.

RESULTS

Identification of BRCA1 and BRCA2 Isoforms From Targeted RNA-seq Data

Using targeted RNA-seq, we detected 40 *BRCA1* and 17 *BRCA2* alternate isoforms in LCLs from non-variant carrier controls cultured with or without a NMD inhibitor (Tables S6–S7 in Supplementary Material). These include 25/63 of the *BRCA1* isoforms identified by Colombo et al. (12) and/or Davy et al. (10) (Table S6 in Supplementary Material), in addition to 5/22 *BRCA2* isoforms identified by Fackenthal et al. (13) (Table S7 in Supplementary Material). In addition to these, six naturally occurring *BRCA1* isoforms previously reported by Colombo et al. (12) were detected in variant carriers. Targeted RNA-seq also detected 13 *BRCA1* and 11 *BRCA2* isoforms that have not been reported previously in healthy controls. The novel transcripts identified here increases the total number of *BRCA1* and *BRCA2* splicing events observed in control samples to 70 and 34, respectively.

Of the previously reported naturally occurring isoforms not detected in *BRCA1* and *BRCA2* using targeted RNA-seq (12, 13), the majority (28/32 in *BRCA1* and 13/17 in *BRCA2*) were due to target probe placement (restricted to the options predesigned by Illumina), while the remainder (11/32 *BRCA1* and 4/17 *BRCA2*) were presumably not expressed at levels that were detectable in our cell lines (Tables S6 and S7 in Supplementary Material). From our observations, and those reported in previous publications (12), all *BRCA1* mRNA exons have been shown to be spliced out in at least one naturally occurring alternative mRNA transcript. By contrast, six *BRCA2* exons (8, 14, 21, and 24–27) were not found to be involved in an exon skipping event in this study or by Fackenthal et al. (13). Nineteen alternative transcripts (11 *BRCA1* and eight *BRCA2)* were detected solely in the samples treated with a NMD inhibitor (Tables S6–S9 in Supplementary Material). Of the 19, 14 are out of frame.

To assess the targeted RNA-seq method for evaluating transcript profiles in rare variant carriers (LCL1–8), we compared our data with those previously reported from the PCR-based ENIGMA multicentre study (15) (Table S10 in Supplementary Material). In this study, a total of 37 *BRCA1* and 11 *BRCA2* alternative splicing events were identified by targeted RNA-seq in addition to those detected by reverse transcriptase-PCR (RT-PCR) from

the multicentre study (Table S10 in Supplementary Material). We found that 33/37 *BRCA1* and all 11 *BRCA2* of these events fell outside the region targeted by the RT-PCR assays. By comparison, 26 *BRCA1* and 12 *BRCA2* splicing events were exclusively detected in these samples in the multicentre study. However, 21 of these events were not detected by the Truseq Targeted RNA Expression platform due to the absence of probes targeting those regions, while 11 were due to the events involving multiple separate regions, which are not detectable together using this platform. Of the remaining six events not detected by Targeted RNA-seq (*BRCA1*, Δ5–6, Δ9, Δ9–11, Δ9–12, Δ11–12, Δ22), three (Δ9, Δ22, and Δ9–11) were respectively identified by three laboratories, which always included the two laboratories utilizing capillary electrophoresis for detection. Further to this, *BRCA1* Δ5–6 was identified solely by laboratories utilizing capillary electrophoresis, which was the most sensitive detection method used in the multi-center study (15).

Targeted RNA-seq identified another five splicing events described by Whiley et al. (15) that were solely present in four out of the eight LCLs each carrying a known spliceogenic rare variant (Table S10 in Supplementary Material, *BRCA1* c.671−2A>G − Δ9−11 and Δ10−11; *BRCA1* c.5467+5G>C − Δ23; *BRCA2* c.8632+1G>A − Δ19−20; *BRCA2* c.9501+3A>T − Δ25). In contrast to the multicentre study, targeted RNA-seq was unable to detect the Δ5 event associated with the pathogenic variant *BRCA1* c.135−1G>T, likely as a result of a low read count (Table S10 in Supplementary Material). These data further highlight the complexity associated with detection and interpretation of *BRCA1* and *BRCA2* splicing patterns when assays are designed across the whole transcript.

Quantitative Assessment of BRCA1 and BRCA2 Transcripts

To derive quantitative information from the targeted RNA-seq data, we separately calculated the relative expression range for *BRCA1* ($n = 25$) and *BRCA2* ($n = 14$) transcripts from the study LCLs that did not contain known splice disrupting variants in the respective gene assayed. The number of alternative splicing events detectable across multiple LCLs was double in cycloheximide treated cell lines compared to that found in non-treated cells (**Figures 2A,B**), thus the following quantitative data corresponds to treated cells only. A correlation was observed between the number of detected alternative events and the total read count per sample for *BRCA1* ($R^2 = 0.68$) and *BRCA2* ($R^2 = 0.69$) (Figure S5 in Supplementary Material). The questionable probe efficiency observed for alternative event *BRCA1* Δ9–10 using targeted RNA-seq was confirmed with qPCR to be overinflated (Table S11 in Supplementary Material) and so was excluded from the normalization calculations.

The full-length transcripts were found to be the most highly expressed mRNA isoforms for both *BRCA1* and *BRCA2* when comparing the relative levels of all mRNA isoforms expressed for each gene, while they also had the greatest expression variability (**Figure 2**). No *BRCA1* splicing events were expressed above 20% of the total number of detected transcripts, whereas the expression ranges of *BRCA2* Δ9–10, Δ12, and ▼20 exceeded this level.

FIGURE 2 | Relative expression of *BRCA1* and *BRCA2* mRNA isoforms in rare variant samples compared to controls. **(A)** Natural expression ranges of mRNA splice isoforms calculated from lymphoblastoid cell lines (LCLs) not containing any known spliceogenic variants in *BRCA1* **(A)** and *BRCA2* **(C)**. Colored symbols overlaid indicate the relative mRNA isoform expression in LCLs containing known *BRCA1* **(B)** or *BRCA2* **(D)** splice disrupting variants. Only mRNA splice isoforms that were detected by more than 10 reads in at least two controls were included. Mean and upper and lower limits shown for each isoform [SE (95%)].

Despite a high mRNA expression variability detected for natural isoforms, results still highlighted notable isoform expression differences between variant carrier and control LCLs not measured by PCR-based methods in the ENIGMA multi-center study (15). The most significant difference was for Δ10 in LCL5 (*BRCA1* c.[594−2A>C; 641A>G]), which expressed an 8.8-fold increase compared to controls. The isoforms Δ15 and ▼21 were also upregulated in this sample (2.8- and 5.0-fold increase, respectively). Expression differences were also found

for ▼25 (4.5-fold increase) for LCL8 (*BRCA2* c.8632+1G>A); ▼20 (2.3-fold increase) and ▼25 (2.4-fold increase) for LCL1 (*BRCA2* c.426-12_426-_8delGTTTT); and ▼21 and Δ11 (1.3-fold increase) for LCL4 (*BRCA1* c.671−2A>G) (**Figures 2B,D**).

To compare targeted RNA-seq and qPCR-derived expression levels, the relative expression of *BRCA1* Δ9–10, Δ10, and the exon 10–11 junction was calculated using a RJ (*BRCA1* exons 2–3) in LCL5 (*BRCA1* c.594−2A>C) and compared to controls. Consistent with recent findings (9), both targeted RNA-seq and

qPCR assays measured a significant increase in expression levels of Δ10 (28.688-fold$_{RNA-seq}$ – 16.262-fold$_{qPCR}$), similar levels of Δ9–10 (0.869-fold$_{RNA-seq}$ – 1.150-fold$_{qPCR}$), and reduced 10–11 junction levels (0.718-fold$_{RNA-seq}$ – 0.427-fold$_{qPCR}$) in LCL5 compared to the controls (Table S11 in Supplementary Material).

The Effect of LCL Storage and Culture Conditions on BRCA1 and BRCA2 mRNA Isoform Expression

A key observation from the ENIGMA multicentre study was the variability in *BRCA1* and *BRCA2* isoform detection between laboratories using different cell processing and assay protocols (15). However, it was unclear whether these inter-laboratory differences are due to untested aspects of the laboratory protocol.

To explore this possibility, we assessed the effect of cell culture and storage conditions on *BRCA1* and *BRCA2* isoform expression in RNA extracted at six time points from LCL7 with fortnightly freeze/thaw cycles. RNA was sequenced using targeted RNA-seq with technical triplicates (**Figure 3**).

Detection of the more prominently expressed alternative splicing events (for example, *BRCA1* Δ9–10 and Δ1Aq) was more consistent across time points and between technical replicates than it was for the minor events (**Figures 3** and **4**). Testing for significant isoform expression differences by time point using the linear model found no consistent effect for *BRCA1* and *BRCA2* in either NMD inhibitor treated or untreated samples (Figures S6–S9 in Supplementary Material). Together these results suggest that expression variability does exist at the intra-laboratory level and this variability is greatest for mRNA

Full length *BRCA1* junctions	1			2			3			4			5			6		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1-2	989	4	1113	881	2079	1528	8	1154	803	724	828	566	1370	1576	1194	1282	1612	475
2-3	2890	1138	626	1662	142	817	1684	775	947	1377	1432	2230	2300	1234	1604	2027	1185	2240
5-6	1453	467	1511	181	310	865	6	222	825	2428	588	951	2453	164	1125	1316	992	1766
6-7	1561	2442	1239	535	257	313	586	23	600	50	10	402	671	184	1261	908	984	1301
7-8	1528	1061	500	879	573	653	3019	1	1329	443	433		281	1228	550	645	1502	289
8-9	1830	102	264	521	1124	1	1730	2	295	787	1790	2	1836	668	695	661	869	1627
9-10	1284	1209	668	509	410	585	694	129	1592	969	5	1121	393	124	1357	587	183	1232
10-11	4075	11571	14653	19210	21415	17745	18730	25292	15401	6722	26440	14778	9609	19835	10851	6870	11033	8287
11-12	2947	2212	1437	376	2517	404	377	53	1105	2702	14	3311	2476	1099	2147	3973	1683	2152
12-13	3057	5108	2893	1648	610	2778	1421	2252	3181	4066	1608	1378	4127	1464	3054	2739	3168	4445
13-14				210			2		48								104	
14-15	1913	828	1616	1673	1565	567	2275	438	694	1715	670	1917	1461	453	1627	1888	1410	1258
15-16	1641	441	1366	486	768	745	8	522	917	2252	1002	944	1194	391	648	1040	1075	430
17-18	1707	2088	484	1254	638	423	6	405	1099	1213	1142	1859	1469	145	460	1567	855	954
18-19	665	208		976	298	246	437		512	1426	1142	152	1182	550	319	710	572	350
19-20	1416	1173	1544	593	410	438	481	190	264	2073	475	1004	1307	403	631	1121	1468	785
20-21	1107	611	1074	488	527	1428	5	1508	1344	1611	9	1272	954	413	2608	2003	1539	1815
21-22	743	2639	1030	1021	559	989	947	28	902	2735	16	1344	1321	1749	1549	1713	1530	1497
22-23	5210	3912	2520	2149	1210	4466	3402	3582	3620	1950	24	2584	2969	1662	3694	6046	4086	6142
23-24	351		5		235	1	809	1	185	10	2	790			1		279	321
Alternative splicing events																		
Δ9-10	580	551	1863	3677	1927	3327	2738	2167	2219	2261	1882	2464	825	4720	1612	779	1681	1443
Δ1Aq	1126	1594	1838	794	2263	973	5	406	1574	6	10	5	1424	906	1577	909	1184	404
Δ8p (3nt)	889	232	8	113			5		478	1348	2	319	2		737	468	44	394
Δ13p 3nt	405				1					971	2			626	301	60	536	1
Δ11q (3309nt)	174			10		470	2	539				2	2	233	135	140	40	
Δ10	116	407	382	1		213		266				2	73	1	132		125	
▼1aA	30		3	110	40								111		19	62	142	
▼1	18					19	48	44	19	135			12			21	22	70
Δ14p (3nt)					100		2		1	1				115		92	22	63
Δ22				42			2			1			1			215		
Δ17															103			
Δ2, Δ3p (7nt)	2	2			1	2	3			2				1	1	1	1	
Ins 8p (8nt), Δ7 (8nt)	4			1	1				1	1				3	1	0	2	1
▼1, Δ2-9, Δ10p		2	5			1		2						1		1	0	1
▼1															2			
1Aq Δ2-10														1				
1B Δ2-10													1					
▼2																		1

Detection level key

	10,000
	5,000
	1,000
	10

FIGURE 3 | *BRCA1* mRNA isoforms detected at six time points in an lymphoblastoid cell line (sample #7, Table S1 in Supplementary Material) treated with an nonsense-mediated decay inhibitor. A freeze–thaw process was undertaken after time points two and four. Three technical replicates are listed under each time point.

Full length BRCA2	1-1	1-2	1-3	2-1	2-2	2-3	3-1	3-2	3-3	4-1	4-2	4-3	5-1	5-2	5-3	6-1	6-2	6-3
2-3	2053	2795	21	2311	1164	3516	9	1429	3431	1364	1820	1150	227	1269	1585	1409	2376	2935
4-5	570	918	712	3477	4056	829	1856	1108	635	2038	431	1178	832	2846	281	824	1236	1559
5-6							2											
6-7	4001	1581	2378	2313	2993	3553	4145	3332	3546	1921	4569	3392	4008	5772	6287	3386	3369	3009
8-9	2542	2236	3412	1167	2116	2011	1910	1674	2115	1110	97	2669	3513	1505	1145	2690	1478	1830
9-10	3199	4599	3238	3463	688	1946	6	2496	4649	2092	4859	2421	5349	4933	5296	5462	2170	2908
10-11	4715	7375	2083	8935	8197	4977	6762	4038	3308	8491	9472	6569	1461	9132	5074	4596	5329	6088
11-12	997	1577	1793	297	975	1486	6	1489	1318	2678	2048	2228	2816	1881	1356	2498	4317	1870
12-13	1434	1437	2204	346	553	758	441	2364	1512	1464	10	3495	986	788	705	886	950	818
13-14	1331	2029	780	2946	1725	1623	9	1843	2195	1837	17	3346	1013	1543	679	1125	2670	2159
14-15	821	421	796	117	1956	352	88	245	557	826	2	394	621	161	338	507	420	302
15-16	1461	2006	5901	1650	1629	901	2990	1437	388	1357	334	1448	1461	817	1593	1725	1875	1670
16-17	1804	884	2816	2495	5	3107	2473	1136	2965	2296	5	12	2281	2241	1816	2517	1344	2911
17-18	246	563	5	748	1300	275	3	2		2		2	2	4	1	144	239	437
20-21	2113	1247	1735	1273	2453	1930	839	3116	817	1479	1335	2754	2318	494	2284	1796	1435	2115
21-22	3582	1427	2146	968	4779	1282	2558	4867	1377	2775	15	1737	3927	1237	2940	2076	1175	1715
22-23	1122	4	1598	1171		812	16	6080	2988	607	4150	1086	1780	2084	1826	1964	1928	1204
24-25	4846	2572	1677	2503	3774	2617	9852	269	2540	5158	2749	2774	3902	1729	2857	2772	3212	2247
25-26	2934	3865	6059	2914	688	6020	6026	2472	5231	2187	8065	2613	3054	931	3088	3038	3886	3317
Alternative splicing events																		
Δ12	1	776		593	77	944	3		2		5	3	3			168	341	
▼21 (21 nt)	92	267	633	308	864	135		589	234	309			222	98		222	60	251
▼25a		377				386									287	40	103	226
Δ11								25			720				1			1
▼25b		284				322									220	33	76	186
▼24a	118	757			3		3			1				2		113		
▼24b													261					116
▼24c													261					116
Δ12p (3nt)	1	4	5	2	5		7	1		6		5	8	4	335	7	7	9
Δ9-10								162					210					
Δ4	13																	

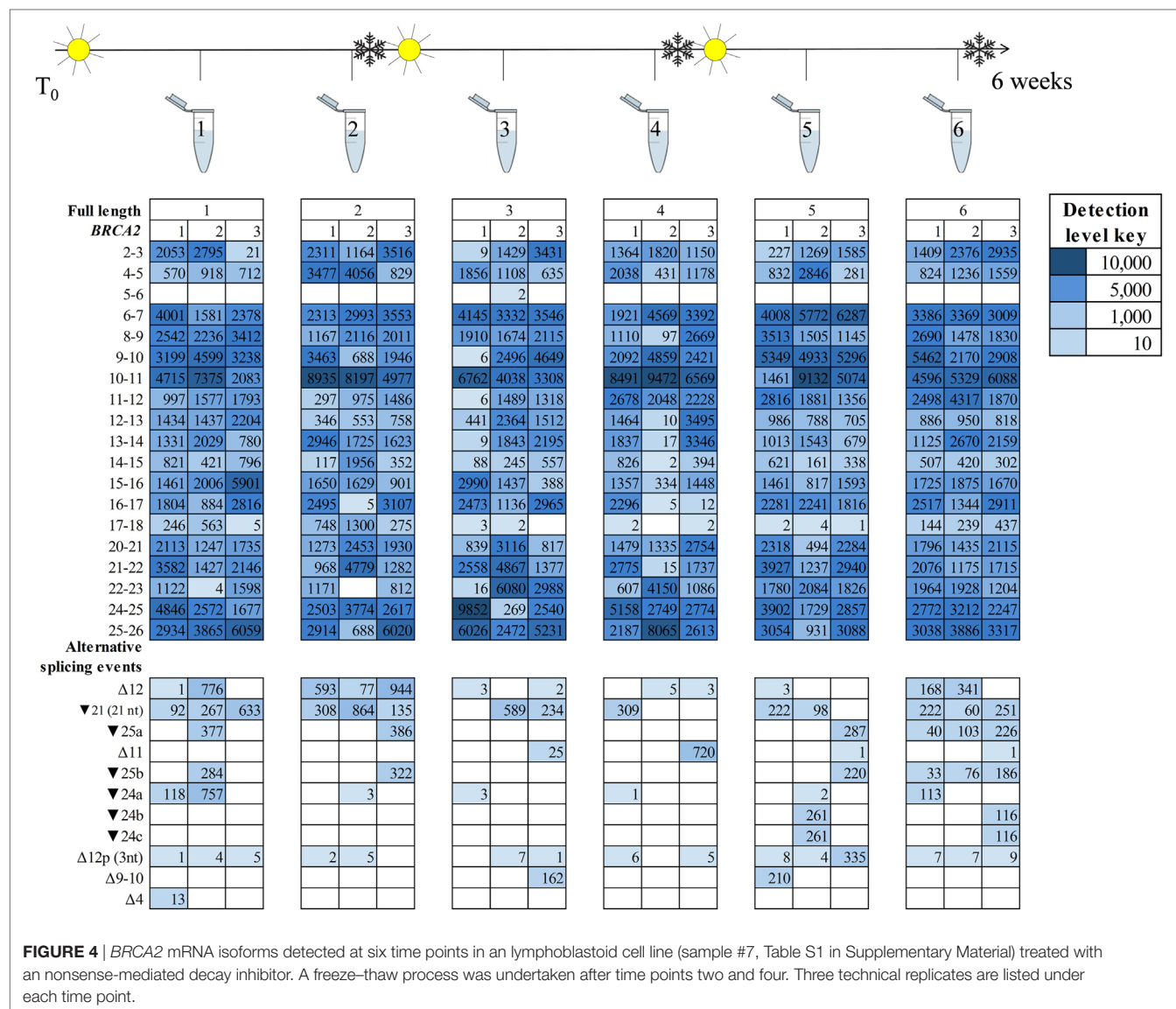Detection level key: 10,000 / 5,000 / 1,000 / 10

FIGURE 4 | BRCA2 mRNA isoforms detected at six time points in an lymphoblastoid cell line (sample #7, Table S1 in Supplementary Material) treated with an nonsense-mediated decay inhibitor. A freeze–thaw process was undertaken after time points two and four. Three technical replicates are listed under each time point.

isoforms detected at low levels, such as the samples not treated with NMD inhibitors (**Figures 3** and **4**; Figures S10 and S11 in Supplementary Material). However, there was no evidence for systematic effects relating to the number of freeze/thaw storage cycles or whether the cells have been analyzed after 1 or 2 weeks growth.

DISCUSSION

BRCA1 and BRCA2 mRNA splicing assays are often carried out in a diagnostic and research setting to assess the effects of variants of uncertain clinical significance. To date, PCR-based mRNA splicing assays have been the method of choice to assess mRNA transcripts qualitatively. While genetic variation has been suggested to induce abnormal isoform expression changes (23), such aberrations are not easy to detect using non-quantitative or semi-quantitative techniques. Here, we have utilized a targeted RNA-seq approach to qualitatively and quantitatively assess the

expression profile of BRCA1 and BRCA2 mRNA isoforms in LCLs previously assayed in a multicentre study (15). This whole-gene approach is much more comprehensive than traditional PCR-based splicing assays, as evidenced by the detection of 55 mRNA isoforms (30 known and 25 novel) using only a small fraction of the samples used for the reported catalog of 80 BRCA1/2 isoforms (12, 13). Several transcripts were not detectible due to limitations with the capture design and/or reduced sensitivity, however, lowly expressed BRCA1 and BRCA2 isoforms that have not been identified by previous studies were able to be detected using this platform (12, 13). Additional splicing events may be present, but at very low levels, requiring a much higher sequencing depth for detection. Further work is required to determine if such transcripts are clinically important, and so establish whether their detection is required for an understanding of breast cancer risk.

The targeted RNA-seq platform utilized in this work was able to overcome previously reported limitations of PCR-based assays

(14, 15) by providing multiple exon coverage across *BRCA1* or *BRCA2* for each assay, sequence confirmation of splicing event, and quantitative assessment of isoform expression patterns. Moreover, our study demonstrated that the use of NMD inhibitors is important for obtaining detectable levels of full length and alternative splicing events using the Illumina Truseq Targeted RNA Expression platform.

We show the utility of targeted RNA-seq to quantitatively identify previously reported upregulated splicing events, such as Δ10 (*BRCA1* c.[594−2A>C; 641A>G]) and Δ11 (*BRCA1* c.671−2A>G) (**Figure 2**), while the expression levels of many of the other isoforms in the variant carriers were within the range seen in controls. Our study also identified higher levels of Δ15 and ▼21 for LCL4 (*BRCA1*c.[594−2A>C; 641A>G]) than expected in controls, which have not previously been reported, likely because these small differences are not easily observable with semi-quantitative technologies. However, these changes are not associated with pathogenicity in *BRCA1* c.[594−2A>C; 641A>G] carriers (9), and any association between the variant and each splicing event remains unclear. It is possible that future research with additional control samples may show that these small expression changes are within the natural expression range. Interestingly, the relative expression ranges observed for all alternative events appeared to be more tightly regulated in *BRCA1* than *BRCA2*. While they do not appear to overlap any important domains, three *BRCA2* mRNA isoforms were expressed at greater levels than those associated with *BRCA1* (**Figure 2**). These differences suggest that greater variability in expression for some *BRCA2* isoforms is tolerated in LCLs, however, further research is required to established *BRCA1* and *BRCA2* isoform expression patterns in cancer specific tissue, such as normal breast and ovarian epithelia.

Splicing data from this study, in addition to those from recent reports (10–13), show that every exon in *BRCA1* and 20/27 exons for *BRCA2* are skipped in at least one natural isoform. This highlights how quantitative assessment of aberrant splicing would be very beneficial in these highly variable genes as it would provide a more comprehensive detection method of the splicing changes present. It also suggests that seven *BRCA2* exons are likely to be highly conserved, so any changes involving these exons are likely to be detrimental in the cell.

To identify technical factors that also contribute *BRCA1* and *BRCA2* mRNA expression differences between samples, we assessed the effect of cell culture and storage conditions on *BRCA1* and *BRCA2* isoform expression across the six experimental time points. Our results showed variability in the isoforms expressed at any given time, irrespective of the number of storage events and culture time. Moreover, variability is greater when RNA-seq assays generate relatively low number of sequence reads (**Figures 3** and **4**; Table S12 in Supplementary Material).

Targeted RNA-seq platform utilize short fragmented library reads to detect mRNA splicing events. Such platforms are, therefore, limited in their ability to determine whether multiple events occur on the same transcript. Recently, we carried out a study using the MinION™ (Oxford Nanopore Technologies, Oxford, UK) long read sequencer and obtained whole transcript information for *BRCA1* in a normal sample, which showed evidence of co-occurring splice events in *BRCA1* (24). Further work involving long read sequencing of *BRCA1* and *BRCA2* would help to further distinguish transcript exon structure regarding all deletion and retention events for variant carriers. Results from such research will also accurately define out-of-frame transcripts which are prone to NMD.

Here, we utilize targeted RNA-seq technology to provide a comprehensive review and quantitative assessment of naturally occurring mRNA isoforms in *BRCA1* and *BRCA2*. While qualitative analysis alone has been assumed to be sufficient for identifying aberrant events, the more quantitative RNA-seq offers improvements to the accuracy and capabilities of PCR-based assays, overcoming many of the detection limitations presented by the earlier technologies. These results lead us to make the following recommendations: (1) technical replicates ($n > 2$) of the variant carrier are necessary to capture methodology induced variability associated with RNA-seq assays, (2) LCLs can undergo multiple freeze/thaw cycles and can be cultured up to 2 weeks without noticeably influencing isoform expression levels, (3) NMD inhibitors are essential prior to splicing assays for comprehensive mRNA isoform detection, (4) quantitative assessment of exon:exon junction levels across *BRCA1* and *BRCA2* can help distinguish between normal and aberrant isoform expression patterns. While advances in probe design, and possibly the uptake of long read sequencers, are essential to allow detection of all expressed mRNA isoforms using this platform, the decreasing costs of RNA-seq technology, alongside an increasing understanding of bioinformatics capabilities, will likely increase the progression away from PCR-based assessment of gene expression, as evidenced by recent work by Davy et al. (10). In addition, the advanced capabilities of RNA-seq promise to aid in evaluating the clinical significance of variants in *BRCA1* and *BRCA2*, but further exploration is required to determine whether these variants are influencing expression. Quantitative assessment of *BRCA1/2* isoforms in a greater number of control samples using other RNA-seq platforms will further improve our understanding of "normal" expression and provide an invaluable reference for establishing the occurrence of aberrant splicing when assessing genetic variants. Furthermore, careful assay design will be crucial for obtaining data across the gene, thus enabling an interpretation of splicing changes for the entire transcript as opposed to selected regions.

AUTHOR CONTRIBUTIONS

Conception or design of the work—VL, LW, JP, BR, MC, and AS. Resources—kI and LW. Data collection and drafting the article—VL and LW. Data analysis and interpretation—VL, JP, and LW. Critical revision of the article and final approval of the version to be published—all authors.

ACKNOWLEDGMENTS

REFERENCES

1. Antoniou AC, Cunningham AP, Peto J, Evans DG, Lalloo F, Narod SA, et al. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *Br J Cancer* (2008) 98(8):1457–66. doi:10.1038/sj.bjc.6604305

2. Sanz DJ, Acedo A, Infante M, Duran M, Perez-Cabornero L, Esteban-Cardenosa E, et al. A high proportion of DNA variants of *BRCA1* and *BRCA2* is associated with aberrant splicing in breast/ovarian cancer patients. *Clin Cancer Res* (2010) 16(6):1957–67. doi:10.1158/1078-0432.CCR-09-2564

3. Whiley PJ, Guidugli L, Walker LC, Healey S, Thompson BA, Lakhani SR, et al. Splicing and multifactorial analysis of intronic *BRCA1* and *BRCA2* sequence variants identifies clinically significant splicing aberrations up to 12 nucleotides from the intron/exon boundary. *Hum Mutat* (2011) 32(6):678–87. doi:10.1002/humu.21495

4. Ahlborn LB, Dandanell M, Steffensen AY, Jonson L, Nielsen FC, Hansen TV. Splicing analysis of 14 BRCA1 missense variants classifies nine variants as pathogenic. *Breast Cancer Res Treat* (2015) 150(2):289–98. doi:10.1007/s10549-015-3313-7

5. Gaildrat P, Krieger S, Thery JC, Killian A, Rousselin A, Berthet P, et al. The BRCA1 c.5434C->G (p.Pro1812Ala) variant induces a deleterious exon 23 skipping by affecting exonic splicing regulatory elements. *J Med Genet* (2010) 47(6):398–403. doi:10.1136/jmg.2009.074047

6. Jung H, Lee D, Lee J, Park D, Kim YJ, Park WY, et al. Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat Genet* (2015) 47(11):1242–8. doi:10.1038/ng.3414

7. Walker LC, Whiley PJ, Couch FJ, Farrugia DJ, Healey S, Eccles DM, et al. Detection of splicing aberrations caused by *BRCA1* and *BRCA2* sequence variants encoding missense substitutions: implications for prediction of pathogenicity. *Hum Mutat* (2010) 31(6):E1484–505. doi:10.1002/humu.21267

8. Tammaro C, Raponi M, Wilson DI, Baralle D. BRCA1 EXON 11, a CERES (composite regulatory element of splicing) element involved in splice regulation. *Int J Mol Sci* (2014) 15(7):13045–59. doi:10.3390/ijms150713045

9. de la Hoya M, Soukarieh O, Lopez-Peroliom IEA, Vega A, Walker LC, van Ierland Y, et al. 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* (2016) 25(11):2256–68. doi:10.1093/hmg/ddw094

10. Davy G, Rousselin A, Goardon N, Castera L, Harter V, Legros A, et al. Detecting splicing patterns in genes involved in hereditary breast and ovarian cancer. *Eur J Hum Genet* (2017) 25(10):1147–54. doi:10.1038/ejhg.2017.116

11. Hojny J, Zemankova P, Lhota F, Sevcik J, Stranecky V, Hartmannova H, et al. Multiplex PCR and NGS-based identification of mRNA splicing variants: analysis of BRCA1 splicing pattern as a model. *Gene* (2017) 637:41–9. doi:10.1016/j.gene.2017.09.025

12. Colombo M, Blok MJ, Whiley P, Santamarina M, Gutierrez-Enriquez S, Romero A, et al. Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium. *Hum Mol Genet* (2014) 23(14):3666–80. doi:10.1093/hmg/ddu075

13. Fackenthal JD, Yoshimatsu T, Zhang B, de Garibay GR, Colombo M, De Vecchi G, et al. Naturally occurring *BRCA2* alternative mRNA splicing events in clinically relevant samples. *J Med Genet* (2016) 53(8):548–58. doi:10.1136/jmedgenet-2015-103570

14. Walker LC, Whiley PJ, Houdayer C, Hansen TV, Vega A, Santamarina M, et al. Evaluation of a 5-tier scheme proposed for classification of sequence variants using bioinformatic and splicing assay data: inter-reviewer variability and promotion of minimum reporting guidelines. *Hum Mutat* (2013) 34(10):1424–31. doi:10.1002/humu.22388

15. Whiley P, de la Hoya M, Thomassen M, Becker A, Brandao R, Pedersen IS, et al. Comparison of mRNA splicing assay protocols across multiple laboratories: recommendations for best practice in standardized clinical testing. *Clin Chem* (2014) 60(2):341–52. doi:10.1373/clinchem.2013.210658

16. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* (2009) 10:57–63. doi:10.1038/nrg2484

17. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* (2010) 28(5):511–5. doi:10.1038/nbt.1621

18. den Dunnen JT, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat* (2000) 15(1):7–12. doi:10.1002/(SICI)1098-1004(200001)15:1<7::AID-HUMU4>3.0.CO;2-N

19. Lattimore V. *Genetic Variation and mRNA Splicing in Familial Breast Cancer Genes*. University of Otago, Christchurch (2017).

20. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* (2013) 29(1):15–21. doi:10.1093/bioinformatics/bts635

21. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, et al. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat Genet* (2004) 36(12):1255–7. doi:10.1038/ng1469

22. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Stat Sci* (2001) 16(2):101–17. doi:10.1214/ss/1009213285

23. Thery JC, Krieger S, Gaildrat P, Revillion F, Buisine MP, Killian A, et al. Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. *Eur J Hum Genet* (2011) 19(10):1052–8. doi:10.1038/ejhg.2011.100

24. de Jong LC, Cree S, Lattimore V, Wiggins GAR, Spurdle AB; kConFab Investigators, et al. Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events. *Breast Cancer Res* (2017) 19(1):127. doi:10.1186/s13058-017-0919-1

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership