CLINICAL GENETICS An International Journal of Genetics, Molecular and Personalized Medicine

Clin Genet 2016: 90: 324–333
Printed in Singapore. All rights reserved

CLINICAL GENETICS
doi: 10.1111/cge.12748

## Original Article

# Hereditary truncating mutations of DNA repair and other genes in *BRCA1*/*BRCA2*/*PALB2*-negatively tested breast cancer patients

Lhota F., Zemankova P., Kleiblova P., Soukupova J., Vocka M., Stranecky V., Janatova M., Hartmannova H., Hodanova K., Kmoch S., Kleibl Z. Hereditary truncating mutations of DNA repair and other genes in *BRCA1*/*BRCA2*/*PALB2*-negatively tested breast cancer patients.
Clin Genet 2016: 90: 324–333. © John Wiley & Sons A/S. Published by John Wiley & Sons Ltd, 2016

Hereditary breast cancer comprises a minor but clinically meaningful breast cancer (BC) subgroup. Mutations in the major BC-susceptibility genes are important prognostic and predictive markers; however, their carriers represent only 25% of high-risk BC patients. To further characterize variants influencing BC risk, we performed SOLiD sequencing of 581 genes in 325 BC patients (negatively tested in previous *BRCA1*/*BRCA2*/*PALB2* analyses). In 105 (32%) patients, we identified and confirmed 127 truncating variants (89 unique; nonsense, frameshift indels, and splice site), 19 patients harbored more than one truncation. Forty-six (36 unique) truncating variants in 25 DNA repair genes were found in 41 (12%) patients, including 16 variants in the Fanconi anemia (FA) genes. The most frequent variant in FA genes was c.1096_1099dupATTA in *FANCL* that also show a borderline association with increased BC risk in subsequent analysis of enlarged groups of BC patients and controls. Another 81 (53 unique) truncating variants were identified in 48 non-DNA repair genes in 74 patients (23%) including 16 patients carrying variants in genes coding proteins of estrogen metabolism/signaling. Our results highlight the importance of mutations in the FA genes' family, and indicate that estrogen metabolism genes may reveal a novel candidate genetic component for BC susceptibility.

### Conflict of interest

All authors declare no conflict of interest.

**F. Lhota[a,†], P. Zemankova[a,†], P. Kleiblova[a,b], J. Soukupova[a], M. Vocka[c], V. Stranecky[d], M. Janatova[a], H. Hartmannova[d], K. Hodanova[d], S. Kmoch[d] and Z. Kleibl[a]**

[a]Institute of Biochemistry and Experimental Oncology, First Faculty of Medicine, Charles University in Prague, Prague, Czech Republic, [b]Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University in Prague and General University Hospital in Prague, Prague, Czech Republic, [c]Department of Oncology, First Faculty of Medicine, Charles University in Prague and General University Hospital in Prague, Prague, Czech Republic, and [d]Institute of Inherited Metabolic Disorders, First Faculty of Medicine, Charles University in Prague and General University Hospital in Prague, Prague, Czech Republic

[†]These authors contributed equally to this work.

Key words: DNA repair – FANCL – genetic predisposition testing – hereditary breast cancer – high-throughput nucleotide sequencing – sequencing panel – SOLiD next-generation sequencing

Corresponding author: Zdenek Kleibl, MD, PhD, Institute of Biochemistry and Experimental Oncology, First Faculty of Medicine, Charles University in Prague, U Nemocnice 5, 128 53 Prague 2, Czech Republic.
Tel.: +420 22 496 5745;
fax.: +420 22 496 5732;
e-mail: zdekleje@lf1.cuni.cz

Received 12 December 2015, revised and accepted for publication 25 January 2016

Breast cancer (BC; OMIM#114480) emerges as a leading cause of cancer death in female population worldwide. Hereditary breast cancer (HBC) accounts approximately for 5–10% of cases. Clinical importance of HBC results from the high lifetime risk of BC development, increased risk of other associated cancers, early disease onset, and 50% probability of the mutant allele's transmission to the offspring (1). Hence, the identification of germline mutations that confer BC susceptibility is an important task of clinical oncogenetics with considerable clinical utility, including tailored healthcare focused on early cancer identification, preventive surgical strategies decreasing cancer risk, and specific therapeutic strategies (2). The most frequently mutated genes in HBC patients are *BRCA1* and *BRCA2*; however, mutations in these genes account for less than 25% of cases in HBC patients. Since the identification of major BC-susceptibility genes, numerous other predisposition genes have been identified. Their characterization has been strongly accelerated with the availability of next-generation sequencing (NGS) technologies (3). Mutational analyses of recently identified BC-susceptibility genes indicate that frequencies of their mutations are substantially lower than that in *BRCA1* and *BRCA2*, besides being highly variable among populations worldwide (4). However, mutations in these newly established BC-susceptibility genes could collectively epitomize another 25% of ascertained genetic risk in HBC patients and thus their analyses are gradually introduced into the clinical analyses (5).

A striking characteristic of the majority of known BC-susceptibility genes is the contribution of their protein products in DNA damage repair (6). On the other hand, the existence of known BC-susceptibility genes that code for proteins not directly involved in these processes (e.g. *PTEN*, *CDH1*, or *NF1*) indicates that non-DNA repair genes could also contribute to BC susceptibility (7).

Our previous gene-by-gene mutational analyses revealed that the most frequent mutations in Czech HBC patients are found in the *BRCA1* gene (8–10). Less frequently, we identified pathogenic variants in *BRCA2* or *PALB2* (11). In this study, we aimed to describe the presence of potentially pathogenic hereditary variants in other known BC-susceptibility genes using the targeted NGS and to identify further variants that may contribute to BC susceptibility in high-risk Czech BC patients.

## Materials and methods

Detailed methods are available in Supporting information methods.

### Patients and samples

The 325 successfully sequenced patients' samples were selected from a sample collection of high-risk Czech BC patients that fulfilled testing criteria described previously (8, 9, 11), were negatively tested for the presence of mutations in *BRCA1*/*BRCA2*/*PALB2*, and gave their informed consent approved by local ethical committee. As controls, we analyzed 105 samples obtained from Czech non-cancer elder females selected according to their age (>50 years; median age 71 years; ranged 54–95 years) from non-cancer controls described previously (12). The genotyping of the c.1096_1099dupATTA variant in *FANCL* was performed on additional sample sets of 337 high-risk BC patients, 673 sporadic BC patients and 686 non-cancer controls (13, 14) using high-resolution melting analysis and confirmed by Sanger sequencing (Fig. S1, Supporting information). Clinical and histopathological characteristics of analyzed high-risk BC patients are available in Tables S1 and S2.

### Sequencing gene panel

The targeted genes comprised two groups consisting of 141 DNA repair genes and 449 genes retrieved from Phenopedia database (15) using the disease term 'breast neoplasms' with at least two entries (assessed February 2012). Finally, 581 targeted genes (listed in Table S3) were sequenced successfully.

### Library construction, sequence capture and sequencing

Fragmented DNA was subjected to ligation of SOLiD sequencing adaptors and polymerase chain reaction (PCR)-based incorporation of bar codes, as described previously (16). The target DNA enrichment was performed by a custom SeqCap EZ Choice Library (Roche), and SOLiD sequencing primers were introduced by PCR. The final libraries were amplified by an emulsion PCR and sequenced on a SOLiD 4 System (Thermo Fisher, Waltham, MA, USA).

### Bioinformatics pipeline, variant filtration, and prioritization of missense variants

Sequencing reads were aligned to the human genome reference (GRCh37/hg19) using Novoalign (CS1.01.08). Picard was used for duplicate removal and SAMtools (0.1.8) for SAM-to-BAM conversion and calling of single nucleotide variants (SNVs) and small insertions and deletions (indels). Variant annotation was performed with ANNOVAR (17).

Variant filtration excluded off-target sequences and low confidence variants (sequence quality <150; sequencing coverage <10). We also excluded common variants with allelic frequencies in ESP6500 and 1000 Genomes databases >0.01. To reflect the population-specific variants and variants influencing cancer susceptibility, we excluded variants presented in no patient or in more than two controls.

To identify missense variants with a putative contribution to BC susceptibility, we performed a prioritization that considered five prediction algorithms (SIFT, PolyPhen-2, LRT, MutationTaster, and PhyloP) and two databases (ClinVar and HGMD) aggregating data about genotypes and corresponding clinical characteristics. Prioritized variants were considered those that were called by each prediction software as deleterious (or unknown) or considered as disease-associated in ClinVar and HGMD databases.

### Confirmation of truncating variants

All truncating variants were confirmed by conventional Sanger sequencing. The variants affecting a conservative splice site were analyzed from the blood-isolated patient's RNA, when available, using RT-PCR and sequencing as described previously (18). All primers are listed in Table S4.

### Statistical analysis

The differences among analyzed groups and subgroups were calculated by the chi-square test or Fischer exact test if the expected number of events was lower than six.

## Results

In the set of 325 patients' samples and 105 controls, we obtained 491,385 variants in exome and adjacent intronic sequences of 581 targeted genes. The mean sequencing coverage was 56.5 and 93% of the captured sequence was covered by >10 reads. Using the variant filtration, we identified 4540 rare variants in the final dataset representing 2647 unique variants of 496/581 targeted genes (85.4%). We found 144 truncating variants (either nonsense, frameshift indels, or splice site alterations), representing 89 unique variants, in 73/581 targeted genes (12.6%).

The set of 325 BC patients harbored 4053 rare variants (2647 unique) including 127 truncating variants (89 unique), 34 in-frame indels (22 unique), 2347 missense SNVs (1599 unique), and 1545 synonymous SNVs (937 unique). We primarily focused on the truncating variants that were identified in 105/325 (32.3%) BC patients (Fig. 1) and were all confirmed by Sanger sequencing. Nineteen patients carried more than one truncating variant (1 patient carried four, 1 patient carried three, and 17 patients carried two truncations), 86 patients carried one truncating variant. The group of truncating variants included 20 splicing variants (14 unique, each affecting one particular gene) flanking to intronic (±2 bp) sequences. Their impact on splicing was studied at the mRNA level (available from eight patients). Seven out of eight analyzed splicing variants showed frameshift (Figs S2 and S3). The prioritization analysis revealed 356 potentially pathogenic variants out of 1599 rare unique missense variants (22%).

### Hereditary variants in DNA repair genes

In 25 DNA repair genes, we identified 46/127 truncations (36/89 unique) in 41 (12.6%) BC patients (Table 1). The most frequent alterations affected genes that code for DNA double-strand break (DDSB)/interstrand crosslink (ICL) repair proteins. These included 16 patients carrying nine unique truncating variants in five Fanconi anemia (FA) genes (Fig. 1).

Another 19 patients harbored 19 unique variants affecting other genes coding for proteins involved in the DDSB repair pathways, including homologous recombination (HR; *ATM*, *EXO1*, *WRN*, *BLM*, *DCLRE1C*, *FAM175B/ABRO1*, *HELQ, NBN*, *RAD18*, *RAD50*, *RAD51D*, *CHEK2*, and *RFC1*) but also non-homologous end joining (NHEJ; *XRCC4*) repair. Finally, eight truncating variants, each in one patient, were identified in the genes that code for proteins involved in other DNA repair processes including single-strand DNA repair (*ATR, ATRIP*), nucleotide excision repair (NER; *ERCC2, ERCC6*), mismatch repair (MMR; *MSH5*), and direct removal of alkylated guanine (*MGMT*). Two patients carried truncating variants in more than one gene involved in different DNA repair pathways.

Altogether, 106 unique prioritized missense variants in 59 DNA repair genes were identified in 133 patients (34 of these variants, in 56 patients, were found in 15 genes in which at least 1 truncating variant was also detected; Table S5). The most frequent potentially pathogenic missense variants were found in *ATM* (12 variants in 17 patients) and *CHEK2* (4 variants in 13 patients). Among prioritized variants, we also identified pathogenic missense variants in *BRCA1* (c.115T>C; p.C39R), *TP53* (c.733G>A; p.G245S), and *CDH1* (c.1018A>G; p.T340A) in three young BC patients.

### Extended analysis of *FANCL* c.1096_1099dupATTA

The most frequent frameshift variant found in six BC patients and none NGS control was c.1096_1099dupATTA (p.T367Nfs*13) in *FANCL* (previously described in an FA patient belonging to the FA-L complementation group; OMIM#614083) (19). Because of the insufficient number of NGS controls, we first compared the frequency of this *FANCL* variant among our patients with data from the Exome Aggregation Consortium (ExAC) database (http://exac.broadinstitute.org; accessed May 2015) indicating an overrepresentation of this variant among our high-risk BC patients (Table 1). Therefore, we further analyzed another 337 high-risk BC Czech patients (329 females, 8 males; all *BRCA1*/*BRCA2*/*PALB2*-negative). Among these, we identified another three c.1096_1099dupATTA carriers with BC (all diagnosed before age of 38 years). Overall, the c.1096_1099dupATTA was identified in 9/662 high-risk BC individuals (1.3%).

To identify the carriers of c.1096_1099dupATTA in sporadic BC patients and other controls, we genotyped 673 unselected BC cases and 686 non-cancer controls (313 females and 373 males). This analysis revealed three carriers in each analyzed group, showing its frequency as 0.4% in both BC cases (3/693) and controls (3/791; including 105 NGS controls and 686 genotyped controls), respectively. Thus, the frequency of c.1096_1099dupATTA was significantly (Fisher exact test) overrepresented only among high-risk individuals ($p = 0.04$) but not in sporadic BC patients ($p = 0.9$). All 14 carriers among patients were females, while all three carriers in controls were males.

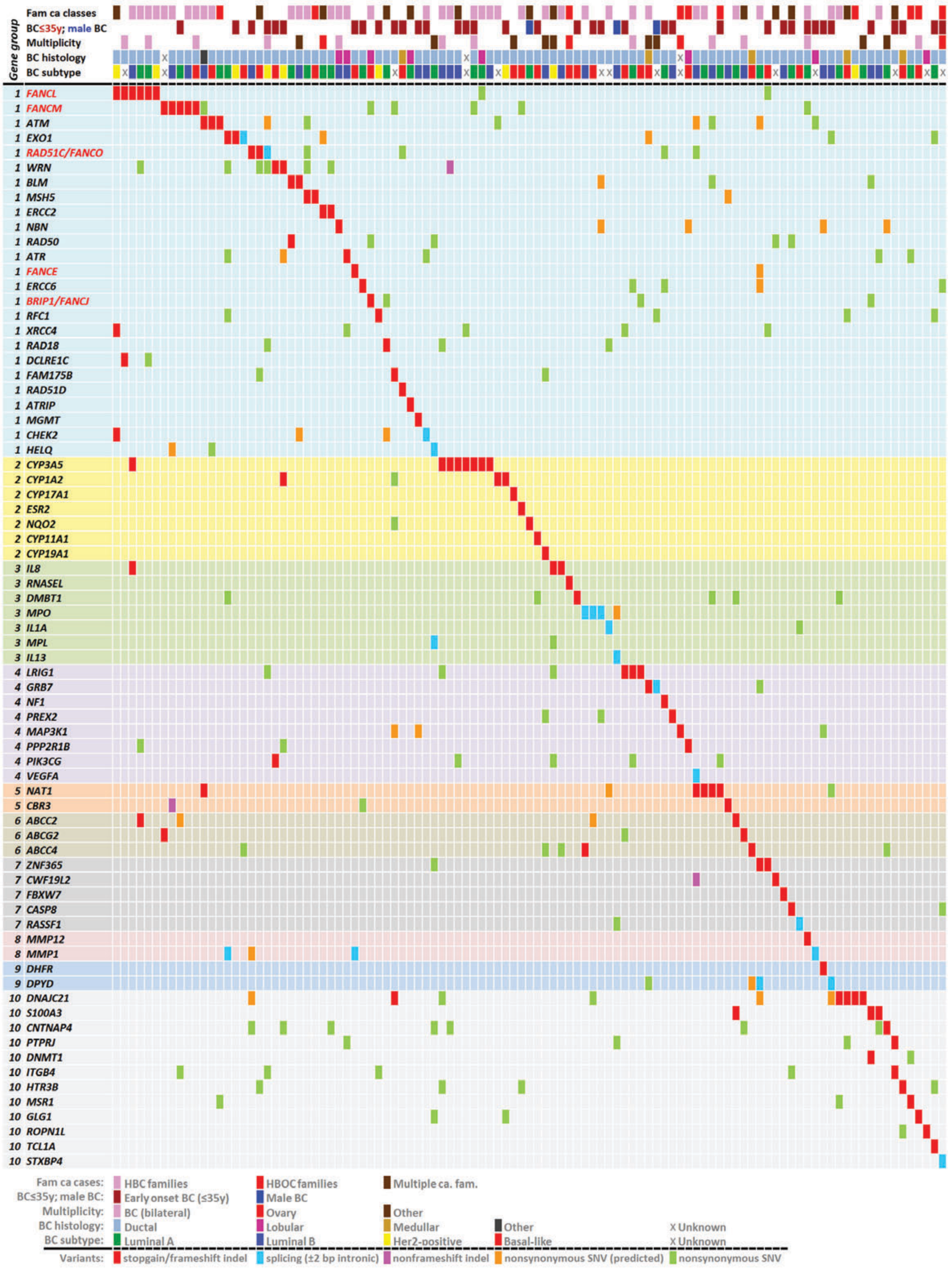


*Fig. 1.* Overview of variants in 73 genes (rows) affected by at least one truncating (nonsense, frameshift, or splicing) variant, that were identified in 105 BC patients (columns). Pathological characteristics of BC tumors (histology and subtypes) and selected clinical characteristics (BC in females at the age of <35 years or male BC, and the presence of familial cancer) are shown in five upper lines (color markings are displayed in Fig. 2; X denotes a missing information). The patients and genes are ordered according to the overall number of found variants, genes (with the Fanconia anemia gene members highlighted in red letters) are grouped by functional relationship of coded proteins (see note). Note: Genes in gene groups (1–10; number (*N*) patients with at least one truncating variant) were ascertained as follows: the genes coding proteins involved in DNA repair (1; *N* = 41); steroid hormones synthesis, turnover or signaling (2; *N* = 16); immune response (3; *N* = 11); membrane receptor signaling (4; *N* = 11); metabolism of xenobiotics (5; *N* = 6); membrane transport of molecules (6; *N* = 6); cell cycle/apoptosis regulation (7; *N* = 6); cell-to-cell communication (8; *N* = 4); nucleotide metabolism (9; *N* = 3); or other (unsorted) processes (10; *N* = 16). Color markings used for pathological and clinical characteristics (shown in legend) are identical to that presented in Fig. 2. Clinical and histopathological characteristics of truncating mutation carriers are shown in Table S6.

Table 1. List of 36 unique truncating variants (nonsense, frameshift indels, or splicing) that were found in 25 genes coding for proteins involved in DNA repair and DNA damage response identified in 41/325 BC patients (Pts) and 7/105 non-cancer controls (Ctrls)[a]

| Gene | HGVS coding | HGVS protein[b] | Classification | Rs number | HGMD/ ClinVar | Pts (*N*) | Ctrls (*N*) | ExAC (mut/all)[b, c] |
|---|---|---|---|---|---|---|---|---|
| *FANCL* | c.1096_1099dupATTA | p.T367Nfs*13 | Indel | | | 6 | 0 | 232/65648* |
| *FANCM* | c.1972C>T | p.R658* | Nonsense | | DM | 1 | 1 | 7/66502* |
| | c.3979_3980delCA | p.Q1327Vfs*16 | Indel | | | 1 | 0 | 0/66498* |
| | c.5101C>T | p.Q1701* | Nonsense | rs147021911 | DM? | 2 | 0 | 95/66562 |
| | c.5791C>T | p.R1931* | Nonsense | rs144567652 | DM | 1 | 0 | 63/66622 |
| *ATM* | c.3850delA | p.T1284Qfs*9 | Indel | | DM | 1 | 0 | n.r. |
| | c.7327C>T | p.R2443* | Nonsense | rs121434220 | DM/P | 2 | 0 | n.r. |
| *EXO1* | c.1522dupT | p.C508Lfs*7 | Indel | | | 1 | 0 | n.r. |
| | c.2358delG | p.L787Yfs*37 | Indel | | | 1 | 0 | n.r. |
| | c.2212-1G>C | p.V738_K743del | Splicing | rs4150000 | DM | 1 | 1 | 172/63478 |
| *CHEK2* | c.277delT | p.W93Gfs*17 | Indel | | | 1 | 0 | n.r. |
| | c.444+1G>A | p.R148Vfs*6 | Splicing | | DM | 2 | 0 | 11/66720* |
| *RAD51C* | c.502A>T | p.R168* | Nonsense | | | 2 | 0 | n.r. |
| | c.905-2_1delAG | p.L301Gfs*42 | Splicing | | DM | 1 | 0 | n.r. |
| *BLM* | c.1642C>T | p.Q548* | Nonsense | rs200389141 | DM | 2 | 0 | 21/66322* |
| *ERCC2* | c.230_231delTG | p.V77Afs*4 | Indel | | | 1 | 0 | n.r. |
| | c.1703_1704delTT | p.F568Yfs*2 | Indel | | DM | 1 | 2 | 11/65444* |
| *MSH5* | c.541C>T | p.R181* | Nonsense | rs147515280 | | 1 | 0 | 13/65882* |
| | c.1900C>T | p.R634* | Nonsense | | | 1 | 0 | n.r. |
| *WRN* | c.604A>T | p.K202* | Nonsense | | | 1 | 0 | n.r. |
| | c.4216C>T | p.R1406* | Nonsense | rs11574410 | | 1 | 0 | 87/65788 |
| *ATR* | c.5342T>A | p.L1781* | Nonsense | | | 1 | 0 | n.r. |
| *ATRIP* | c.827_828delAG | p.E276Gfs*2 | Indel | | | 1 | 0 | n.r. |
| *BRIP1* | c.2392C>T | p.R798* | Nonsense | rs137852986 | DM/P | 1 | 0 | 16/65688* |
| *DCLRE1C* | c.1903dupA | p.S635Kfs*6 | Indel | | | 1 | 0 | 27/66734* |
| *ERCC6* | c.3693C>G | p.Y1231* | Nonsense | | | 1 | 0 | n.r. |
| *FAM175B* | c.1084delC | p.Q362Kfs*19 | Indel | | | 1 | 0 | n.r. |
| *FANCE* | c.929dupC | p.V311Sfs*2 | Indel | | DM | 1 | 1 | n.r. |
| *HELQ* | c.2677-1G>A | p.Q348Pfs*17 | Splicing | rs200992133 | | 1 | 1 | 27/66528 |
| *MGMT* | c.207_210dupACGT | p.S70Yfs*5 | Indel | | | 1 | 0 | n.r. |
| *NBN* | c.657_661delACAAA | p.K219Nfs*16 | Indel | | | 1 | 1 | 21/65324 |
| *RAD18* | c.1430_1431insGCGG | p.T478Rfs*6 | Indel | | | 1 | 0 | n.r. |
| *RAD50* | c.1093C>T | p.R365* | Nonsense | | | 1 | 0 | n.r. |
| *RAD51D* | c.355_358deldelTGTA | p.C119Wfs*16 | Indel | | | 1 | 0 | n.r. |
| *RFC1* | c.2191delA | p.R731Gfs*7 | Indel | | | 1 | 0 | n.r. |
| *XRCC4* | c.25delC | p.H9Tfs*8 | Indel | | | 1 | 0 | 42/66632 |
| Total variants | | | | | | 46 | 7 | |

Variants listed in HGMD or ClinVar databases: DM, disease-causing (pathological) mutations; DM?, likely disease-causing (likely pathological) mutation; P, pathogenic. ExAC, Exome Aggregation Consortium; n.r., variant not reported in ExAC.

[a] The enhanced version of the table (including missense variant predicted as pathogenic, numbers of reference transcripts, and frequencies in ExAC, ESP6500, and 1000 genomes databases) is available as Table S5.

[b] ExAC allelic frequency in European non-Finnish population (mutated alleles/wt alleles).

[c] Asterisk (*) indicates significant differences ($p < 0.05$) between allelic frequencies in European non-Finnish population (ExAC) and in analyzed population of patients (Fisher exact test).

### Hereditary variants in non-DNA repair genes

The remaining 81/127 truncations (53/89 unique) in 48 non-DNA repair genes were identified in 74 (22.8%) BC patients (Table 2). We found variants in only non-DNA repair genes in 64 of them, while in 10 patients we also detected some truncating variants in DNA repair genes. To identify possible defects in pathways that may contribute to BC susceptibility, we sorted the affected genes into nine groups (Group 2–9 in Table 2 and Fig. 1) clustering functionally related proteins. Twelve genes (Group 10) comprised proteins with unrelated or unclear function. Sixteen carriers (5% of all patients) of eight different truncating variants have been identified in the ‘Group 2’ associating genes that code for proteins involved in steroid hormones metabolism or signaling.

Further, we detected 250 unique, prioritized, potentially pathogenic missense variants in 150 genes in 213 patients. The most frequent prioritized SNVs in non-DNA repair genes affected the *APC* gene (in eight patients).

### Individual and disease characteristics in carriers of truncating variants

We found no significant differences in the characteristics of patients and tumors between the carriers of truncating

Table 2. List of 53 unique truncating variants (nonsense, frameshift indels, or splicing) in 48 non-DNA repair genes identified in 74/325 BC patients and 10/105 non-cancer controls[a]

| Gene | Gr | HGVS coding | HGVS protein[b] | Classification | Rs number | HGMD/ClinVar | Pts (*N*) | Ctrs (*N*) | ExAC (mut/all)[b], [c] |
|---|---|---|---|---|---|---|---|---|---|
| *CYP3A5* | 2 | c.92dupG | p.L32Tfs*3 | Indel | | | 7 | 1 | 732/66688 |
| | 2 | c.246dupG | p.A83Gfs*40 | Indel | | | 1 | 0 | 21/66728 |
| *CYP1A2* | 2 | c.816T>A | p.Y272* | Nonsense | rs140421378 | FTV | 3 | 0 | 16/65958* |
| *CYP11A1* | 2 | c.835delA | p.I279Yfs*10 | Indel | | DM | 1 | 0 | 2/66694* |
| *CYP17A1* | 2 | c.1072C>T | p.R358* | Nonsense | | DM | 1 | 0 | n.r. |
| *CYP19A1* | 2 | c.1058dupT | p.L353Ffs*10 | Indel | | | 1 | 0 | 1/60606* |
| *ESR2* | 2 | c.76G>T | p.E26* | Nonsense | | | 1 | 0 | 1/66734* |
| *NQO2* | 2 | c.628C>T | p.Q210* | Nonsense | | | 1 | 0 | 1/66386* |
| *IL8* | 3 | c.91G>T | p.E31* | Nonsense | rs188378669 | | 3 | 2 | 104/66426 |
| *DMBT1* | 3 | c.2227delC | p.Q743Rfs*4 | Indel | | | 1 | 0 | n.r. |
| *IL13* | 3 | c.174+2delT | p.(?) | Splicing | | | 1 | 0 | n.r. |
| *IL1A* | 3 | c.319+2T>C | p.(?) | Splicing | | | 1 | 0 | n.r. |
| *MPL* | 3 | c.79+2T>A | p.(?) | Splicing | rs146249964 | DM | 1 | 0 | 114/66230 |
| *MPO* | 3 | c.2031-2A>C | p.R677Wfs*73 | Splicing | rs35897051 | DM | 3 | 1 | 470/66434 |
| *RNASEL* | 3 | c.793G>T | p.E265* | Nonsense | rs74315364 | DM/P | 1 | 1 | 381/66212 |
| *LRIG1* | 4 | c.3149_3150delCG | p.A1050Gfs*17 | Indel | | | 3 | 0 | 102/66704* |
| *GRB7* | 4 | c.862C>T | p.Q288* | Nonsense | | | 1 | 0 | 2/48666* |
| | 4 | c.801+1G>C | p.(?) | Splicing | | | 1 | 0 | n.r. |
| *MAP3K1* | 4 | c.4151dupT | p.L1384Lfs*36 | Indel | | | 1 | 0 | n.r. |
| *NF1* | 4 | c.5690delG | p.G1897Vfs*28 | Indel | | | 1 | 0 | n.r. |
| *PIK3CG* | 4 | c.41_42delAG | p.E14Gfs*147 | Indel | | | 1 | 0 | 5/62474* |
| *PPP2R1B* | 4 | c.342_343delTG | p.V115Cfs*3 | Indel | | | 1 | 0 | 81/66082 |
| *PREX2* | 4 | c.3210_3213delAACA | p.D1072Vfs*17 | Indel | | | 1 | 0 | n.r. |
| *VEGFA* | 4 | c.1085+2T>C | p.(?) | Splicing | rs149528656 | | 1 | 0 | 15/66648* |
| *NAT1* | 5 | c.559C>T | p.R187* | Nonsense | rs5030839 | FP | 5 | 1 | 252/66632 |
| *CBR3* | 5 | c.533delA | p.D178Afs*46 | Indel | | | 1 | 0 | 102/66716 |
| *ABCC2* | 6 | c.3196C>T | p.R1066* | Nonsense | rs72558199 | DM/P | 2 | 0 | 35/66738* |
| *ABCC4* | 6 | c.2468dupA | p.N823Kfs*12 | Indel | | | 1 | 1 | 26/66634 |
| | 6 | c.1150C>T | p.R384* | Nonsense | | | 1 | 0 | n.r. |
| *ABCG2* | 6 | c.706C>T | p.R236* | Nonsense | rs140207606 | FP | 1 | 0 | 24/66634 |
| | 6 | c.736C>T | p.R246* | Nonsense | rs200190472 | FP/P | 1 | 0 | 5/66692* |
| *ZNF365* | 7 | c.1065G>A | p.W355* | Nonsense | rs142406094 | | 2 | 0 | 4/66740* |
| *CASP8* | 7 | c.106delG | p.E36Nfs*7 | Indel | | | 1 | 0 | n.r. |
| *CWF19L2* | 7 | c.1605delA | p.K535Nfs*4 | Indel | | | 1 | 0 | n.r. |
| *FBXW7* | 7 | c.310delC | p.H104Mfs*389 | Indel | | | 1 | 0 | n.r. |
| *RASSF1* | 7 | c.888+1G>A | p.V258Gfs*7 | Splicing | | | 1 | 0 | 1/64856* |
| *MMP1* | 8 | c.105+2T>C | pQ35Vfs*11 | Splicing | rs139018071 | FTV | 3 | 0 | 114/66230 |
| *MMP12* | 8 | c.327C>T | p.W109* | Nonsense | | | 1 | 0 | 0/65722* |
| *DPYD* | 9 | c.1905+1G>A | p.D581_N635del | Splicing | rs3918290 | DM | 2 | 1 | 389/66688 |
| *DHFR* | 9 | c.95delT | p.F32Sfs*7 | Indel | | | 1 | 0 | n.r. |
| *DNAJC21* | 10 | c.1503delA | p.K501Nfs*10 | Indel | | | 3 | 0 | 33/66560* |
| | 10 | c.1629delT | p.F543Lfs*4 | Indel | | | 2 | 0 | 5/11578* |
| *S100A3* | 10 | c.208delG | p.V70Wfs*83 | Indel | | | 3 | 1 | 291/66718 |
| *CNTNAP4* | 10 | c.3913G>T | p.E1305* | Nonsense | | | 1 | 0 | 3/56224* |
| *DNMT1* | 10 | c.1035dupC | p.K346Qfs*35 | Indel | | | 1 | 0 | n.r. |
| *GLG1* | 10 | c.3520C>T | p.R1174* | Nonsense | | | 1 | 0 | n.r. |
| *HTR3B* | 10 | c.871C>T | p.Q291* | Nonsense | | | 1 | 0 | n.r. |
| *ITGB4* | 10 | c.665delG | p.G222Efs*60 | Indel | | | 1 | 0 | n.r. |
| *MSR1* | 10 | c.569delT | p.L190Cfs*5 | Indel | | | 1 | 0 | n.r. |
| *PTPRJ* | 10 | c.1191T>A | p.Y397* | Nonsense | | | 1 | 0 | n.r. |
| *ROPN1L* | 10 | c.135T>A | p.Y45* | Nonsense | rs41280363 | | 1 | 1 | 126/66680 |
| *STXBP4* | 10 | c.181-1G>A | p.K60Vfs*28 | Splicing | | | 1 | 0 | 3/66388* |
| *TCL1A* | 10 | c.253C>T | p.R85* | Nonsense | | | 1 | 0 | n.r. |
| Total variants | | | | | | | 81 | 10 | |

Variants listed in HGMD or ClinVar databases: DM, disease-causing (pathological) mutations; DM?, likely disease-causing (likely pathological) mutation; P, pathogenic. ExAC, Exome Aggregation Consortium; n.r., variant not reported in ExAC.

[a] Genes are grouped (Gr 2–10) according to the functional relationship of coded proteins, as described in the legend of Fig. 1. The enhanced version of the table (including missense variant predicted as pathogenic, numbers of reference transcripts, and frequencies in ExAC, ESP6500, and 1000 genomes databases) is available as Table S5.

[b] ExAC allelic frequency in European non-Finnish population (mutated alleles/wt alleles).

[c] Asterisk (*) indicates Significant differences ($p < 0.05$) between allelic frequencies in European non-Finnish population (ExAC) and in analyzed population of patients (Fisher exact test).

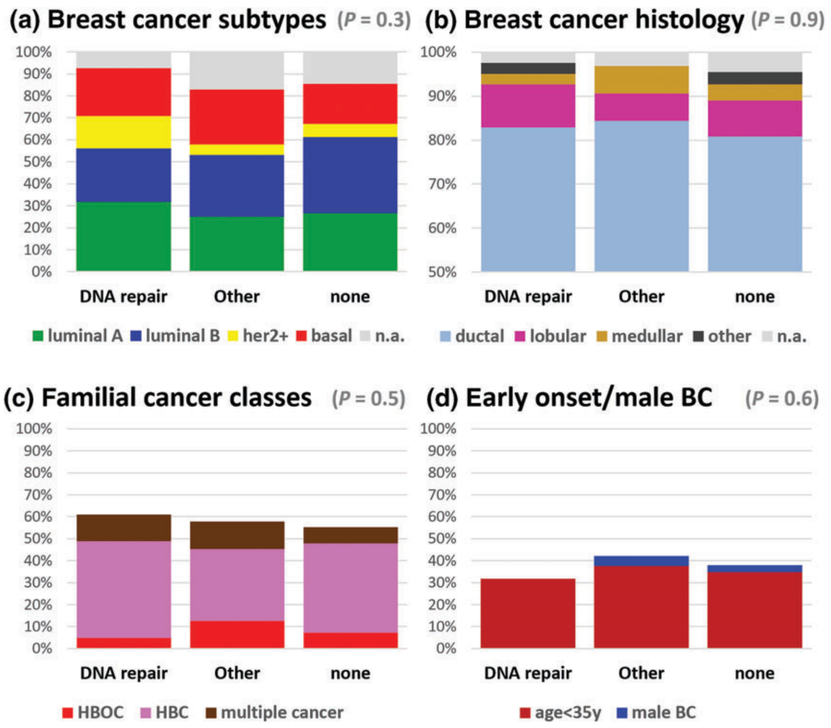


*Fig. 2.* Pathological characteristics of tumors and clinical characteristics of 325 analyzed BC patients grouped according to the presence of truncating variant in any DNA repair gene (DNA repair; 41 patients), variant in only other genes (other; 64 patients), and no truncating variant (none; 220 patients). The p-values (chi-square test) indicate insignificant differences in displayed characteristics among the analyzed subgroups.

variants in the DNA repair genes, carriers of truncating variants in other genes, and patients not carrying truncating variants (Fig. 2).

## Discussion

Panel NGS represents a reliable approach for the analysis of cancer susceptibility in clinical settings but also in identification of candidate genes in high-risk individuals. In contrast to exome or even genome NGS, it allows the identification of the carriers of pathogenic variants in a cost-effective manner, with flexibility in the selection of gene targets, sensitivity, and manageable bioinformatics load for routine practice (20). Our analysis revealed the presence of truncating variants in nearly one third of analyzed patients and 30 patients (9%) carried truncating variants in some of 15 genes (*ATM*, *ATR*, *BLM*, *BRIP1*, *ERCC2*, *FANCE*, *FANCL*, *FANCM*, *CHEK2*, *NBN*, *NF1*, *RAD50*, *RAD51C*, *RAD51D*, *WRN*) analyzed by currently clinically used NGS panels (5). Out of 73 genes with truncating variants, in 51 genes we found only a single truncation. This indicates that rare variants could be identified in a substantial proportion of high-risk individuals; however, their clinical interpretation and differentiation from incidental findings not associating with BC susceptibility would be difficult.

### Characterization of variants in DNA repair genes

The interesting result of our study is the high frequency of potentially pathogenic variants in five FA genes in 4.9% high-risk patients. FA genes code for DNA repair proteins contributing to genome stability maintenance by the ICL repair [reviewed in (21, 22)]. FA proteins form several protein–protein complexes (22). Hereditary bi-allelic mutations of FA genes are responsible for the development of FA characterized by congenital abnormalities, bone marrow failure, cellular hypersensitivity to DNA crosslinking agents and cancer susceptibility. The most frequent truncating variant was c.1096_1099dupATTA in *FANCL* that codes an ubiquitin ligase catalyzing the monoubiquitination of FANCI/FANCD2 (ID2) complex – the key step in FA pathway activation (23, 24). The c.1096_1099dupATTA variant was described by Ali et al. (19) in a patient that belonged to the FA-L complementation group. The mutated FANCL protein (p.T367Nfs*13) contains an aberrant chain of 12 amino acid residues that flanks to the PHD/RING finger domain catalyzing ubiquitin ligase activity. Ali et al. (19) performed its functional characterization revealing that the c.1096_1099dupATTA is a hypomorphic mutation resulting in the formation of altered protein with reduced binding to FA core complex and reduced FANCD2 monoubiquitination. Same variant was also identified by Akbari et al. (25) in a patient with familial esophageal squamous cell carcinoma. The results of our study, showing the overrepresentation of c.1096_1099dupATTA among high-risk BC patients, indicate that this variant may represent a novel BC-susceptibility allele. However, further studies, including segregation analyses providing information about the association of c.1096_1099dupATTA with cancer phenotype in affected families and analyses of the variant in other populations, will be necessary to evaluate its potential clinical utility. As four out of six c.1096_1099dupATTA carriers identified by our panel NGS carried also other truncating variants (Fig. 1), we could not rule out the possibility that c.1096_1099dupATTA could act as a rather modifying variant. The recurrent mutations affecting *FANCL* and *FANCM* at their C-termini indicate that truncating variants of FA genes located in far C-terminal regions may impair the FA pathway under specific and so far uncomprehended circumstances. Such phenomenon has been proposed also for the nonsense c.9976A>T (p.K3326*) variant in *BRCA2/FANCD1* truncating the last 93 amino acids. In contrast to the majority of *BRCA2* pathogenic variants, the p.K3326* has been recognized as only a modest BC-susceptibility allele (OR = 1.26) increasing a risk of other cancers (26).

In *FANCM*, coding a helicase contributing to the formation of the FA anchor complex, we identified four truncating variants in five patients. Truncations in *FANCM* were recently associated with susceptibility for triple-negative BC (27). In three patients (none of them triple-negative), we identified previously characterized nonsense or exon skipping mutations that were shown to increase BC risk (27, 28). The remaining two *FANCM* variants included the rare nonsense mutation c.1972C>T (p.R658*; in a luminal BC patient whose mother and her sister suffered from bilateral BC) and the novel mutation c.3979_3980delCA (p.Q1327Vfs*16; in a BC patient with multiple breast and colorectal cancer (CRC) cases in the family). The association between CRC and germinal *FANCM* mutation has recently been identified in CRC tumor samples obtained from two c.5791C>T (p.R1931*) carriers (29).

We have also identified three *RAD51C/FANCO* mutation carriers (0.9% of patients). The *RAD51C* was originally identified as OC-susceptibility gene (30); however, later data conferred also increased BC susceptibility (31). Recently, we described two other pathogenic *RAD51C* variants in two OC patients (13). These data indicate that mutations in *RAD51C* may affect ~1% of Czech high-risk BC or OC patients.

Finally, the carriers of variants in FA genes comprised also two basal-like BC patients carrying pathogenic variants in *FANCE* and *BRIP1/FANCJ*, respectively. Both variants were reported in association with esophageal cancer (25) and triple-negative BC (31), respectively.

We found rare truncating variants in several other genes associated with hereditary BC that code for DDSB repair proteins; however, we also identified several truncating variants in the genes coding proteins engaged in other DNA repair pathways. Among others, an interesting candidate is *EXO1*, which codes for exonuclease involved in numerous DNA repair pathways. Besides two indels, we identified and characterized the c.2212-1G>C splicing mutation resulting in six amino acids in-frame deletion (p.V738_K743del), involving the interaction of EXO1 with MSH2 during MMR (32). Contrary to our analysis, Wu et al. (33) characterized the identical variant as a frameshift in a patient with hereditary non-polyposis CRC. Moreover, we further identified also two rare *EXO1* prioritized missense variants [c.325G>A (p.E109K) and c.1105A>C (p.S369R)] in five patients. Clustering of mutations in *EXO1* and presence of mutations in other genes involved for example in NER (*ERCC2*, *ERCC6*) suggest that an impairment of these repair processes by hereditary alterations could increase BC susceptibility. The degree to which these variants may influence BC susceptibility remains to be investigated by further studies.

### Variants in non-DNA repair genes

The potentially deleterious hereditary variants were identified in 48/448 non-DNA repair genes, most frequently (in 16 BC patients) in the genes that code for the enzymes of steroid hormone metabolism and signaling. The group primarily included members of the cytochrome p450 superfamily contributing to the estrogen biosynthesis (CYP11A1, CYP17A1, CYP19A1) or catabolism (CYP3A5, CYP1A2) [reviewed in (34)]. Given that estrogens may affect BC etiology, variants in *CYP* genes may influence BC risk.

Variants in *CYP11A1* and *CYP17A1* identified in basal-like patients were previously described in patients suffering from severe congenital adrenal insufficiency (35) (OMIM#613743) and congenital adrenal hyperplasia (36) (OMIM#202110), respectively. Interestingly, Hopper et al. (37) reported p.R239* (c.775C>T) variant in *CYP17A1* in three *BRCA1/2*-negative young sisters with BC and hypothesized that this variant is responsible for dominantly inherited and possibly high-risk BC. Recently, Yang et al. (38) identified c.987delC (p.Y329*) variant in *CYP17A1* in a patient from an HBOC family. We found a novel variant, c.1058dupT (p.L353Ffs*10) in *CYP19A1*, in a patient with a BC and non-Hodgkin lymphoma duplicity whose mother developed bilateral BC. Mutations in similar positions cause aromatase deficiency (OMIM#613546).

Defects in estrogen-catabolizing enzymes suggest a mechanistically more obvious pathophysiological link to BC promotion. As estrogens are known substrates of CYP3A5 and CYP1A2 (34), 11 identified carriers of truncating variants in these genes could potentially have reduced estrogen clearance. We also found a nonsense variant in *NQO2* coding a quinone reductase eliminating estrogen quinones responsible for estrogen-initiated carcinogenesis (39) and one truncating variant in *ESR2* that codes for ERβ with anti-proliferative signaling (40). The high proportion of patients carrying constitutive truncating variants in steroid hormone metabolism genes supports the hypothesis of Hopper et al. (37) suggesting that cancer-causing mutations in these genes may represent a new pathophysiological mechanism linking genetic and environmental interactions in BC susceptibility.

The other functional groups associating the patients with truncating variants in functionally relevant genes were smaller. It is obvious that at least some variants in non-DNA repair genes have very limited (if any) impact on BC susceptibility and they rather represent incidental findings [e.g. mutations in *ABCC2* was identified in a patient with the Dubin–Johnson syndrome (41) or known *DPYD* mutation related to the fluoropyrimidines toxicity (42)]. Reporting of incidental findings is highly questionable and a matter of debate (43, 44).

### Disease and individual characteristics in carriers of truncating variants

Considering the patients and disease characteristics in the carriers of mutations in the major BC predisposing genes, the earlier age at BC diagnosis or more aggressive form of BC subtypes would be expected also in the carriers of mutations in other BC-susceptibility genes. In fact, we did not identify significant differences in clinical and histopathological characteristics between the carriers and non-carriers of truncating variants. This result did not change even when prioritized variants were added into the comparison (data not shown). Similar behavior was also documented recently in a large study of 1824 triple-negative BC patients analyzed by Couch et al. (31) for hereditary mutations in 17 genes, where significant differences in enrichment for family BC/OC history and tumor characteristics were identified only in the carriers of *BRCA1/2* mutations but not in carriers of non-*BRCA1/2* mutations. We suggest that some principal changes in the evaluation of clinical and histopathological characteristics will be required to assess the clinical importance of non-*BRCA1/2* BC-susceptibility genes. Since the frequencies of mutations in these genes are lower by order than that in *BRCA1/2*, an international and consortia effort will be required for such analyses.

## Conclusions

Our study identified truncating variants in 32% of patients, and 9% of patients were carriers of a truncating variant in the genes currently analyzed in clinical NGS panels for the cancer risk prediction. The most frequent truncating variants affected FA genes that, together with *BRCA1*, *BRCA2*, and *PALB2*, make this group the most important for cancer susceptibility in BC patients. Our results also show an overrepresentation of the *FANCL* variant c.1096_1099dupATTA in high-risk patients, indicating that this variant may represent a novel BC-susceptibility allele. Moreover, we identified potentially pathogenic variants in several rarely mutated DNA repair genes indicating that despite its low frequency, variants in these genes may influence the development of HBC in Czech patients. We believe that it is important to analyze such genes and in international co-operation to evaluate their contribution to the BC development because they may represent clinically valuable predictors of cancer risk in families of mutation carriers. Interestingly, in other analyzed genes, we found truncating variants in the genes coding the P450 enzymes of steroid hormones metabolism in 5% of BC patients. Therefore, this functional group may contribute to the explanation of so far undisclosed missing heritability in some high-risk BC patients. We are aware that exact role of both c.1096_1099dupATTA and *CYP* genes in BC susceptibility needs to be further clarified by independent and larger studies.

## Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web-site.

## Acknowledgements

The authors wish to thank all individuals who participated in this study. This study was supported by the grant of Ministry of Health of the Czech Republic NT14054 and by BIOCEV CZ.1.05/1.1.00/02.0109 and Charles University projects PRVOUK-P27/LF1/1, UK 260148/2015, and UNCE 204011.

## References

1. Robson M, Offit K. Clinical practice. Management of an inherited predisposition to breast cancer. N Engl J Med 2007: 357 (2): 154–162.
2. Couch FJ, Nathanson KL, Offit K. Two decades after BRCA: setting paradigms in personalized cancer care and prevention. Science 2014: 343 (6178): 1466–1470.
3. Rahman N. Realizing the promise of cancer predisposition genes. Nature 2014: 505 (7483): 302–308.
4. Karami F, Mehdipour P. A comprehensive focus on global spectrum of BRCA1 and BRCA2 mutations in breast cancer. Biomed Res Int 2013: 2013: 928562.
5. Easton DF, Pharoah PD, Antoniou AC et al. Gene-panel sequencing and the prediction of breast-cancer risk. N Engl J Med 2015: 372 (23): 2243–2257.
6. Kean S. Breast cancer. The 'other' breast cancer genes. Science 2014: 343 (6178): 1457–1459.
7. Ripperger T, Gadzicki D, Meindl A, Schlegelberger B. Breast cancer susceptibility: current knowledge and implications for genetic counselling. Eur J Hum Genet 2009: 17 (6): 722–731.
8. Pohlreich P, Stribrna J, Kleibl Z et al. Mutations of the BRCA1 gene in hereditary breast and ovarian cancer in the Czech Republic. Med Princ Pract 2003: 12 (1): 23–29.
9. Pohlreich P, Zikan M, Stribrna J et al. High proportion of recurrent germline mutations in the BRCA1 gene in breast and ovarian cancer patients from the Prague area. Breast Cancer Res 2005: 7 (5): R728–R736.
10. Ticha I, Kleibl Z, Stribrna J et al. Screening for genomic rearrangements in BRCA1 and BRCA2 genes in Czech high-risk breast/ovarian cancer patients: high proportion of population specific alterations in BRCA1 gene. Breast Cancer Res Treat 2010: 124 (2): 337–347.
11. Janatova M, Kleibl Z, Stribrna J et al. The PALB2 gene is a strong candidate for clinical testing in BRCA1- and BRCA2-negative hereditary breast cancer. Cancer Epidemiol Biomarkers Prev 2013: 22 (12): 2323–2332.
12. Kleibl Z, Havranek O, Hlavata I et al. The CHEK2 gene I157T mutation and other alterations in its proximity increase the risk of sporadic colorectal cancer in the Czech population. Eur J Cancer 2009: 45 (4): 618–624.
13. Janatova M, Soukupova J, Stribrna J et al. Mutation analysis of the RAD51C and RAD51D genes in high-risk ovarian cancer patients and families from the Czech Republic. PLoS One 2015: 10 (6): e0127711.
14. Mateju M, Stribrna J, Zikan M et al. Population-based study of BRCA1/2 mutations: family history based criteria identify minority of mutation carriers. Neoplasma 2010: 57 (3): 280–285.
15. Yu W, Clyne M, Khoury MJ, Gwinn M. Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. Bioinformatics 2010: 26 (1): 145–146.
16. Harakalova M, Mokry M, Hrdlickova B et al. Multiplexed array-based and in-solution genomic enrichment for flexible and cost-effective targeted next-generation sequencing. Nat Protoc 2011: 6 (12): 1870–1886.
17. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010: 38 (16): e164.

18. Sevcik J, Falk M, Macurek L et al. Expression of human BRCA1 Delta 17–19 alternative splicing variant with a truncated BRCT domain in MCF-7 cells results in impaired assembly of DNA repair complexes and aberrant DNA damage response. Cell Signal 2013: 25 (5): 1186–1193.
19. Ali AM, Kirby M, Jansen M et al. Identification and characterization of mutations in FANCL gene: a second case of Fanconi anemia belonging to FA-L complementation group. Hum Mutat 2009: 30 (7): E761–E770.
20. Schroeder C, Faust U, Sturm M et al. HBOC multi-gene panel testing: comparison of two sequencing centers. Breast Cancer Res Treat 2015: 152 (1): 129–136.
21. Kottemann MC, Smogorzewska A. Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. Nature 2013: 493 (7432): 356–363.
22. Walden H, Deans AJ. The Fanconi anemia DNA repair pathway: structural and functional insights into a complex disorder. Annu Rev Biophys 2014: 43 (1): 257–278.
23. Meetei AR, de Winter JP, Medhurst AL et al. A novel ubiquitin ligase is deficient in Fanconi anemia. Nat Genet 2003: 35 (2): 165–170.
24. Meetei AR, Yan Z, Wang W. FANCL replaces BRCA1 as the likely ubiquitin ligase responsible for FANCD2 monoubiquitination. Cell Cycle 2004: 3 (2): 179–181.
25. Akbari MR, Malekzadeh R, Lepage P et al. Mutations in Fanconi anemia genes and the risk of esophageal cancer. Hum Genet 2011: 129 (5): 573–582.
26. Delahaye-Sourdeix M, Anantharaman D, Timofeeva MN et al. A rare truncating BRCA2 variant and genetic susceptibility to upper aerodigestive tract cancer. J Natl Cancer Inst 2015: 107 (5): djv037.
27. Kiiski JI, Pelttari LM, Khan S et al. Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. Proc Natl Acad Sci U S A 2014: 111 (42): 15172–15177.
28. Peterlongo P, Catucci I, Colombo 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 (18): 5345–5355.
29. Smith CG, Naven M, Harris R et al. Exome resequencing identifies potential tumor-suppressor genes that predispose to colorectal cancer. Hum Mutat 2013: 34 (7): 1026–1034.
30. Loveday C, Turnbull C, Ruark E et al. Germline RAD51C mutations confer susceptibility to ovarian cancer. Nat Genet 2012: 44 (5): 475–476.
31. Couch FJ, Hart SN, Sharma P et al. Inherited mutations in 17 breast cancer susceptibility genes among a large triple-negative breast cancer cohort unselected for family history of breast cancer. J Clin Oncol 2015: 33 (4): 304–311.
32. Rasmussen LJ, Rasmussen M, Lee B et al. Identification of factors interacting with hMSH2 in the fetal liver utilizing the yeast two-hybrid system. *In vivo* interaction through the C-terminal domains of hEXO1 and hMSH2 and comparative expression analysis. Mutat Res 2000: 460 (1): 41–52.
33. Wu Y, Berends MJ, Post JG et al. Germline mutations of EXO1 gene in patients with hereditary nonpolyposis colorectal cancer (HNPCC) and atypical HNPCC forms. Gastroenterology 2001: 120 (7): 1580–1587.
34. Blackburn HL, Ellsworth DL, Shriver CD, Ellsworth RE. Role of cytochrome P450 genes in breast cancer etiology and treatment: effects on estrogen biosynthesis, metabolism, and response to endocrine therapy. Cancer Causes Control 2015: 26 (3): 319–332.
35. Hiort O, Holterhus PM, Werner R et al. Homozygous disruption of P450 side-chain cleavage (CYP11A1) is associated with prematurity, complete 46, XY sex reversal, and severe adrenal failure. J Clin Endocrinol Metab 2005: 90 (1): 538–541.
36. Hwang DY, Hung CC, Riepe FG et al. CYP17A1 intron mutation causing cryptic splicing in 17alpha-hydroxylase deficiency. PLoS One 2011: 6 (9): e25492.
37. Hopper JL, Hayes VM, Spurdle AB et al. A protein-truncating mutation in CYP17A1 in three sisters with early-onset breast cancer. Hum Mutat 2005: 26 (4): 298–302.
38. Yang X, Wu J, Lu J et al. Identification of a comprehensive spectrum of genetic factors for hereditary breast cancer in a Chinese population by next-generation sequencing. PLoS One 2015: 10 (4): e0125571.
39. Gaikwad NW, Yang L, Rogan EG, Cavalieri EL. Evidence for NQO2-mediated reduction of the carcinogenic estrogen ortho-quinones. Free Radic Biol Med 2009: 46 (2): 253–262.
40. Caiazza F, Ryan EJ, Doherty G, Winter DC, Sheahan K. Estrogen receptors and their implications in colorectal carcinogenesis. Front Oncol 2015: 5: 19.
41. Pacifico L, Carducci C, Poggiogalle E et al. Mutational analysis of ABCC2 gene in two siblings with neonatal-onset Dubin Johnson syndrome. Clin Genet 2010: 78 (6): 598–600.
42. Kleibl Z, Fidlerova J, Kleiblova P et al. Influence of dihydropyrimidine dehydrogenase gene (DPYD) coding sequence variants on the development of fluoropyrimidine-related toxicity in patients with high-grade toxicity and patients with excellent tolerance of fluoropyrimidine-based chemotherapy. Neoplasma 2009: 56 (4): 303–316.
43. Blackburn HL, Schroeder B, Turner C, Shriver CD, Ellsworth DL, Ellsworth RE. Management of incidental findings in the era of next-generation sequencing. Curr Genomics 2015: 16 (3): 159–174.
44. Christenhusz GM, Devriendt K, Dierickx K. Disclosing incidental findings in genetics contexts: a review of the empirical ethical research. Eur J Med Genet 2013: 56 (10): 529–540.

CLINICAL GENETICS An International Journal of Genetics, Molecular and Personalized Medicine

Clin Genet 2016: 90: 387–389
Printed in Singapore. All rights reserved

CLINICAL GENETICS
doi: 10.1111/cge.12842

## Correspondence

# RE: frameshift variant *FANCL**c.1096_1099dupATTA is not associated with high breast cancer risk

*To the Editor,*

In our recent publication, we noted a borderline association of the truncating variant c.1096_1099dupATTA in the *FANCL* gene with increased breast cancer (BC) risk in high-risk *BRCA1*/*BRCA2*/*PALB2*-negative BC patients (1). However, the subsequent analysis by Pfeifer et al. (2) genotyping the c.1096_1099dupATTA variant in 2370 samples from German and Macedonian BC patients and controls failed to confirm association of this variant with BC risk.

We agree with Pfeifer et al. that *FANCL**c.1096_1099dupATTA is unlikely a high-risk BC susceptibility allele with an immediate clinical utility. There are several lines of evidence that do not support strong involvement of this variant in BC susceptibility including: (i) the relative high frequency of this variant especially in European populations, (ii) functional characteristics demonstrating that cells expressing FANCL isoform coded by c.1096_1099dupATTA variant retain the residual FANCL functional capacity *in vitro* (3), and (iii) phenotypic characteristics that show only a mild Fanconi anemia (FA) complementation group L phenotype in the compound heterozygote carrying *FANCL**c.1096_1099dupATTA (alongside an another truncating *FANCL* variant) (3). Moreover, we identified a male c.1096_1099dupATTA homozygote (in controls) who had no signs of FA at his age of 57 years (Table 1).

We also agree that *FANCL**c.1096_1099dupATTA may confer a low (or lower) risk variant. We hypothesized that this variant may represent a modifying factor because its carriers were overrepresented only in a subgroup of high-risk BC patients in our study and also four out of six c.1096_1099dupATTA carriers analyzed by a panel next-gene sequencing (NGS) carried truncating variant(s) in other known or candidate cancer-susceptibility gene(s). After publication of our study reporting 15 carriers of c.1096_1099dupATTA in 2126 analyzed samples of Czech BC patients and controls, we identified another eight carriers using the CZECANCA multicancer panel NGS (4). The individual characteristics of c.1096_1099dupATTA carriers (Table 1) indicate a relatively low mean age at diagnose [47.2 years (range 28–76 years)] in 14 carriers with BC. We also recently identified three c.1096_1099dupATTA carries with ovarian cancer diagnosed at early age. Contrary to Pfeifer et al. who reported that only one out of 10 identified c.1096_1099dupATTA carriers had a family BC history, we have noticed a known familial history of BC (in a first or second degree relative) in nine out of 23 carriers (39%) and a familial history of some cancer in 15 carriers (65%). The c.1096_1099dupATTA variant was accompanied by another truncating variant(s) in nine out of 14 cancer patients analyzed by a panel NGS (Table 1). We suppose that these characteristics indicate that c.1096_1099dupATTA may (perhaps mildly) modify the breast (or other) cancer risk or cancer onset. However, further studies are required to estimate the risk of cancer development in c.1096_1099dupATTA carriers precisely. The segregation analyses and NGS analyses in families of carriers would be also required to evaluate the involvement of this hypomorphic variant in the risk of other cancer development or in modification of cancer onset.

Table 1. The clinical and genetic characteristics of c.1096_1099dupATTA carriers identified in the Czech population[a]

| Proband no. | Cancer diagnosis | Age at dg (*at analysis) | Family cancer history (date at diagnosis, or + death) | NGS panel | Truncating variant in cancer susceptibility or candidate genes |
|---|---|---|---|---|---|
| ***The carriers of c.1096_1099dupATTA in FANCL referred in Lhota et al. (1)*** | | | | | |
| 1249[b] | **BC** | 36 | **FM-BC**(+40); FF-Brain tumor(+73); **MM-BC**(+70). | SP | None. |
| 1252[b] | **BC** | 41 | F-RC(62); FM-RC(75); FB-PrC(65); **MM-BC.** | SP | *CHEK2*: c.444+1G>A (p.R148Vfs*6), c.277delT (p.W93Gfs*17); *XRCC4*: c.25delC (p.H9Tfs*8). |
| 748[b] | **BC** | 44 | **MS1-BC**(52); **MS2-BC**(58); **MS3-BC**(60). | SP | *ABCC2*: c.C3196T (p.R1066*). |
| C0211[b] | **BC&BC** | 42&43 | Unknown. | SP | *DCLRE1C*: c.1903dupA (p.S635Kfs*6). |
| 1316[b] | **BC** | 67 | **M-BC**(+50); **MS-BC**(80). | SP | *CYP3A5*: c.92dupG (p.L32Tfs*3); *IL8*: c.91G > T (p.E31*). |
| 1331[b] | **BC** | 76 | **M-BC**(75); **S-BC**(63). | SP | None. |
| 960[c] | **BC** | 33 | None. | CZ | None. |
| 1908[c] | **BC** | 38 | S-melanoma(48); FM-RC. | CZ | *MSR1*: c.877C>T (p.R293*). |
| 1782[c] | **BC** | 28 | None. | CZ | None. |
| A546[d] | **BC** | 47 | None. | n.d. | – |
| A626[d] | **BC** | 73 | Unknown. | n.d. | – |
| A032[d] | **BC** | 54 | None. | n.d. | – |
| K101[e] | None | *80 | B-myeloma(+65). | n.d. | – |
| C015 [e,f] | None | *57 | F-PrC. | n.d. | – |
| C308[e] | None | *50 | None. | n.d. | – |
| ***The carriers of c.1096_1099dupATTA in FANCL identified after publication of Lhota et al. (1) study*** | | | | | |
| 3100 | **BC** | 29 | None. | CZ | None. |
| 2946a16 | **BC** | 57 | **M-BC**(51); D-Hodgkin (24); F-LC(69); **FM-BC**(50). | CZ | None. |
| 2885a16 | None. | *44 | **M-BC**(69). | CZ | None. |
| 1846a15 | None | *37 | **MM-BC**(40)&**BC**(70); FB-CRC. | CZ | *DNAJC21*: c.1503delA (p.K501Nfs*10). |
| 3524a15 | None | *31 | M-CRC(48); MM-UBC; two cousins-BC. | CZ | None. |
| 868 | OC | 26 | None. | CZ | *BRCA2*: c.A9976T (p.K3326*). |
| 120 | OC | 39 | F-UBC(58). | CZ | *TSHR*: c.2102dupG (p.Q702Pfs*17). |
| 2864 | OC | 48 | F-LC(+74); FS2-GBC; **FS3-BC(+82)**; FB-LC(+74). | CZ | *RAD51C*: c.502A>T (p.R168*). |

B, brother; CZ, CZECANCA (CZEch CAncer paNel for Clinical Application) cancer panel (219 genes), described in (4); CRC, colorectal cancer; D, daughter; GBC, gallbladder cancer; FB, father's brother; FM, father's mother; LC, lung cancer; MM, mother's mother; MS, mother's sister; n.d., not done; PrC, prostate cancer; RC, renal cancer; S, sister; SP, SOLiD gene panel (581 genes) used in Lhota et al. (1); UBC, urinary bladder cancer.

[a] 20/23 carriers (except K101, C015, and C308) were females. All BC histologies were ductal BCs (except the second medullary BC in patient C0211).

[b] Six carriers identified by NGS analysis in 325 high-risk BC patients.

[c] Three carriers identified in additional sample set of 337 high-risk BC patients by genotyping (analyzed further by CZECANCA panel NGS).

[d] Three carriers identified in 673 sporadic BC patients.

[e] Three carriers (males) identified in 686 controls.

[f] Homozygote for the c.1096_1099dupATTA variant in the *FANCL* gene.

## Acknowledgements

This study was supported by the Ministry of Health of the Czech Republic, grant nr. 15-27695A and Charles University in Prague projects PRVOUK P27 and SVV UK 260256/2016

*P. Zemankova*[a]
*F. Lhota*[a,b]
*P. Kleiblova*[a,c]
*J. Soukupova*[a]
*M. Vocka*[d]
*M. Janatova*[a]
*Z. Kleibl*[a]
[a] *Institute of Biochemistry and Experimental Oncology,*
*First Faculty of Medicine, Charles University in Prague*,
*Prague, Czech Republic*
[b] *Centre for Medical Genetics and Reproductive Medicine*
*GENNET*,
*Prague, Czech Republic*
[c] *Institute of Biology and Medical Genetics,*
*First Faculty of Medicine, Charles University in Prague*
*and General University Hospital in Prague,*
*Prague, Czech Republic*
[d] *Department of Oncology,*
*First Faculty of Medicine, Charles University in Prague*
*and General University Hospital in Prague,*
*Prague, Czech Republic*
*e-mail:* zdekleje@lf1.cuni.cz

## References

1. Lhota F, Zemankova P, Kleiblova P et al. Hereditary truncating mutations of DNA repair and other genes in BRCA1/BRCA2/PALB2-negatively tested breast cancer patients. Clin Genet 2016: 90 (4): 324–333.
2. Pfeifer K, Schürmann P, Bogdanova N et al. Frameshift variant FANCL*c.1096_1099dupATTA is not associated with high breast cancer risk. Clin Genet 2016: 90 (4): 386–387.
3. Ali AM, Kirby M, Jansen M et al. Identification and characterization of mutations in FANCL gene: a second case of Fanconi anemia belonging to FA-L complementation group. Hum Mutat 2009: 30 (7): E761–E770.
4. Soukupová J, Zemánková P, Kleiblová P, Janatová M, Kleibl Z. [CZECANCA: CZEch CAncer paNel for Clinical Application- design and optimization of the targeted sequencing panel for the identification of cancer susceptibility in high-risk individuals from the Czech Republic]. Klin Onkol 2016: 29 (Suppl 1): S46–S54.

Correspondence: Zdenek Kleibl, MD, PhD,
Institute of Biochemistry and Experimental Oncology, First Faculty of
Medicine, Charles University in Prague, U Nemocnice 5, 128 53
Prague 2, Czech Republic.
Tel.: +420 224965745;
fax: +420 224965732;
e-mail: zdekleje@lf1.cuni.cz

RESEARCH ARTICLE

# Identification and Functional Testing of *ERCC2* Mutations in a Multi-national Cohort of Patients with Familial Breast- and Ovarian Cancer

**Andreas Rump[1,2,3]☯, Anna Benet-Pages[4]☯, Steffen Schubert[5]☯, Jan Dominik Kuhlmann[2,3,6,7]*, Ramūnas Janavičius[8,9], Eva Macháčková[10], Lenka Foretová[10], Zdenek Kleibl[11], Filip Lhota[11], Petra Zemankova[11], Elitza Betcheva-Krajcir[1,2,3], Luisa Mackenroth[1,2,3], Karl Hackmann[1,2,3,6], Janin Lehmann[5], Anke Nissen[4], Nataliya DiDonato[1,2,3], Romy Opitz[2,3,6,7], Holger Thiele[12], Karin Kast[2,3,6,7], Pauline Wimberger[2,3,6,7], Elke Holinski-Feder[4], Steffen Emmert[5,13], Evelin Schröck[1,2,3,6], Barbara Klink[1,2,3,6]**

**1** Institute for Clinical Genetics, Faculty of Medicine Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany, **2** German Cancer Consortium (DKTK), Dresden, Germany, **3** German Cancer Research Center (DKFZ), Heidelberg, Germany, **4** MGZ—Medical Genetics Center, Munich, Germany, **5** Clinic for Dermatology Venerology and Allergology, Göttingen, Germany, **6** National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany, **7** Department of Gynecology and Obstetrics, Medical Faculty and University Hospital Carl Gustav Carus, Technische Universität Dresden, Germany, **8** Vilnius University Hospital Santariskiu Clinics, Hematology, Oncology and Transfusion Medicine Center, Vilnius, Lithuania, **9** State Research Institute Innovative Medicine Center, Vilnius, Lithuania, **10** Masaryk Memorial Cancer Institute, Brno, Czech Republic, **11** Institute of Biochemistry and Experimental Oncology, First Faculty of Medicine, Charles University in Prague, Prague, Czech Republic, **12** Cologne Center for Genomics, Cologne, Germany, **13** Clinic of Dermatology, Rostock, Germany

☯ These authors contributed equally to this work.
* jan.kuhlmann@uniklinikum-dresden.de

OPEN ACCESS

**Citation:** Rump A, Benet-Pages A, Schubert S, Kuhlmann JD, Janavičius R, Macháčková E, et al. (2016) Identification and Functional Testing of *ERCC2* Mutations in a Multi-national Cohort of Patients with Familial Breast- and Ovarian Cancer. PLoS Genet 12(8): e1006248. doi:10.1371/journal.pgen.1006248

**Editor:** Charis Eng, Cleveland Clinic Genomic Medicine Institute, UNITED STATES

**Received:** June 8, 2016

**Accepted:** July 18, 2016

**Published:** August 9, 2016

**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Funding:** The Czech-Brno study was supported by MH CZ - DRO (MMCI, 00209805) and RECAMO, CZ.1.05/2.1.00/03.0101. The Czech-Prague study was supported by Ministry of Health of the Czech Republic, grant nr. 15-28830A and projects of the Charles University in Prague PRVOUK P27 and SVV UK 260256/2016. Functional analyses were funded by the Claudia von Schilling Foundation for breast cancer research (to SS and SE). The Lithuanian

## Abstract

The increasing application of gene panels for familial cancer susceptibility disorders will probably lead to an increased proposal of susceptibility gene candidates. Using *ERCC2* DNA repair gene as an example, we show that proof of a possible role in cancer susceptibility requires a detailed dissection and characterization of the underlying mutations for genes with diverse cellular functions (in this case mainly DNA repair and basic cellular transcription). In case of *ERCC2*, panel sequencing of 1345 index cases from 587 German, 405 Lithuanian and 353 Czech families with breast and ovarian cancer (BC/OC) predisposition revealed 25 mutations (3 frameshift, 2 splice-affecting, 20 missense), all absent or very rare in the ExAC database. While 16 mutations were unique, 9 mutations showed up repeatedly with population-specific appearance. Ten out of eleven mutations that were tested exemplarily in cell-based functional assays exert diminished excision repair efficiency and/or decreased transcriptional activation capability. In order to provide evidence for BC/OC predisposition, we performed familial segregation analyses and screened ethnically matching controls. However, unlike the recently published *RECQL* example, none of our recurrent *ERCC2* mutations showed convincing co-segregation with BC/OC or significant

study was supported by the Research Counsil of Lithuania (Project SEN18/2015). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

**Contributions:**

**Conceptualization:** AR ABP SS BK.
**Data curation:** AR ABP SS.
**Formal analysis:** AR ABP SS RJ ZK EM.
**Funding acquisition:** PW EHF SE ES BK.
**Investigation:** AR ABP SS RJ ZK EM.
**Methodology:** AR ABP SS RJ ZK EM BK.
**Project administration:** BK.
**Resources:** AR ABP SS JDK RJ LF ZK FL PZ EBK LM KH JL AN ND RO HT KK.
**Supervision:** AR.
**Validation:** AR ABP SS.
**Visualization:** AR ABP SS AN.
**Writing - original draft:** AR ABP SS JDK SE BK.
**Writing - review & editing:** AR JDK.

overrepresentation in the BC/OC cohort. Interestingly, we detected that some deleterious founder mutations had an unexpectedly high frequency of > 1% in the corresponding populations, suggesting that either homozygous carriers are not clinically recognized or homozygosity for these mutations is embryonically lethal. In conclusion, we provide a useful resource on the mutational landscape of ERCC2 mutations in hereditary BC/OC patients and, as our key finding, we demonstrate the complexity of correct interpretation for the discovery of "bonafide" breast cancer susceptibility genes.

## Author Summary

Approximately 5–10% of breast/ovarian cancer (BC/OC) cases have inherited an increased risk of developing this malignancy. However, mutations in the two major breast cancer susceptibility genes BRCA1 and BRCA2 explain only 15–20% of all familial BC/OC cases. With the emergence of the high throughput NGS-technology, the number of proposed novel candidate genes for breast cancer predisposition continuously increases. However, a "bonafide" proof of cancer susceptibility requires a detailed characterization of candidate mutations, which we addressed in the current study. Using the DNA repair gene *ERCC2* as an example, we performed a comprehensive multi-center approach, analyzing *ERCC2* mutations in 1000+ patients with hereditary BC/OC. We identified 25 potential candidate mutations for cancer breast cancer susceptibility, some of them affecting *ERCC2* functional activity in appropriate cell-culture based assays. However, a more dissected analysis showed no convincing co-segregation with BC/OC and there was no longer a significant overrepresentation in BC/OC when compared to regionally matched controls instead of the global ExAc variant data base, pointing to the relevance of founder-mutations. In conclusion, we provide a useful resource on the mutational landscape of *ERCC2* mutations in hereditary BC/OC patients and, as our key finding, we highlight the complexity of correct interpretation for the discovery of "bonafide" breast cancer susceptibility genes.

## Introduction

Since it became evident that only 15%-20% of the familial risk for BC/OC can be explained by mutations in the major breast cancer-susceptibility genes *BRCA1* and *BRCA2* [1], the search for additional BC/OC susceptibility loci has been pursued. In times of limited sequencing power this pursuit was based on carefully selected candidate genes which typically came from (i) cancer-associated syndromes (ii) linkage screens in large *BRCA1/2*-negative families and (iii) case–control association studies using single-nucleotide polymorphisms [2,3]. Since sequencing power is no longer an issue, the candidate approach is on its decline and about to be replaced by next generation sequencing (NGS) of large gene panels which, taken together, cover a total of more than 100 genes, only 21 of which have been associated with breast cancer so far [4]. This offers amazing opportunities for detection of novel susceptibility loci but also bears the danger of substantial misuse [4], because variants picked up by these panels are not clinically validated. Therefore, post-marketing data validation is absolutely essential [5]. Rare variants, however, need huge case-control datasets in order to reach the requested statistical significance of $P<0.0001$ [4]. Until such large datasets become available, variant validation needs to focus on mutations that are clearly deleterious on functional level but still frequent enough to be validated by a few thousand controls. Such recurrent yet harmful variants are best identified by screening various populations for founder mutations. In *NBN*, for example, a protein-truncating variant (c.657del5) has been identified in Eastern Europe, which is sufficiently common to allow its evaluation in a BC/OC case–control study [6]. Also the successful validation of deleterious Polish and Canadian founder mutations in *RECQL* [7] underlines the huge potential of multi-national BC/OC cohorts.

In this study we sequenced 1345 BC/OC cases from 3 different Central- and East European countries with multi-gene panels and identified recurrent founder mutations in *ERCC2*, which were functionally validated in cell-culture based assays. As essential component of transcription factor IIH, the ERCC2 protein is involved in basal cellular transcription [8] and nucleotide excision repair (NER) of DNA lesions [9]. The most known inherited disease associated with bi-allelic mutations in *ERCC2* is Xeroderma pigmentosum type D (XPD, OMIM 278730), a hereditary cancer-prone syndrome characterized by extreme skin photosensitivity and early development of multiple skin tumors [10]. Therefore, *ERCC2* is a plausible candidate gene for cancer susceptibility. On the other hand, bi-allelic mutations in *ERCC2* can also lead to syndromes without increased propensity to tumor development, namely Trichothiodystrophy 1 (TTD; OMIM 601675) and cerebrooculofacioskeletal syndrome (COFS2; OMIM 610756). This indicates that not all functionally relevant *ERCC2* mutations increase cancer susceptibility in their carriers.

## Results and Discussion

### Panel sequencing identifies a broad spectrum of rare variants as well as recurrent founder mutations in *ERCC2*

Within the entire set of 1345 BC/OC index cases, we have detected three different frame-shift (fs) mutations [p.(Val77fs), p.(Phe568fs) and p.(Ser746fs)], one splice-acceptor site mutation (c.1903-2A>G), one nucleotide exchange that activates a cryptic splice site (c.2150C>G) and 20 rare missense mutations (Table 1, Fig 1). Whereas 14 mutations were unique (2 fs, 1 splice-site, 11 missense), 11 mutations (1 fs, 1 splice-affecting, 9 missense) have been found in 43 independent families. The most frequent mutation was p.(Asp423Asn) identified in 8 carriers from Lithuania and one from the Czech Republic. The common polymorphisms p.(Lys751Gln) and p.(Asp312Asn) have each been encountered in approximately 64% of our cases; since these variants have been considered to be functionally irrelevant [11], we did not include them in our functional study. Among the 20 rare missense variants reported in Table 1, thirteen are predicted by various computer algorithms to be pathogenic (Table 1 and S4 Table). Further computational analysis of the conservation (PhyloP) and depletion (CADD) scores [12] for the mutated nucleotides strongly supported pathogenicity for these variants (S2 Fig). Mapping the mutated AA positions onto the ERCC2 protein structure revealed a widespread distribution pattern (Fig 1). Residues 13, 450, 461, 513, 536, 576, 592, 601, 611, 631, 678 cluster at the helicase motifs of the HD1 and HD2 catalytic domains and residues 166, 167, 188, 215, 280, 316, 423, 487, 722 locate at the TFIIH transcription factor complex binding domains (Arch, FES, and C-terminal). XPD-causing mutations located at the HD2 domain have been shown to inactivate helicase repair capability without disrupting protein structure. Mutations causing trichothiodystrophy (TTD, OMIM 601675), on the other hand, are located well away from the catalytic site of the enzyme and destabilize ERCC2 structure and TFIIH protein interactions [13–15]. We suggest that BC/OC relevant mutations might affect both—catalytic activity as well as protein stability.

### Functional testing identifies *ERCC2* mutations with deleterious effects on protein level

So far, 11 variants (9 recurrent founder mutations and 2 unique variants; Fig 2C) were tested in functional assays for nucleotide excision repair (NER) capability (Fig 2A) as well as

**Table 1. Mutations and rare variants in ERCC2 identified through panel sequencing of individuals with familial breast and/or ovarian cancer.** AA = amino acid; N = sample size; n.a. = not applicable; n.t. = not tested; CZ = Czech Republic, GE = Germany, LT = Lithuania. The cumulative assessment is based on the results of various effect prediction algorithms; details see S4 Table.

| Variant description | | | | | Predicted effect | Functional effect | | BC/OC cases | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Position | Exon | Nucleotide change | AA change | rs-ID | Cumulative assessment | Complementation of NER-deficient cells | Negative modulation of transcription | GE | CZ | LT | total | Tumor type |
| hg19 | (23) | NM_000400.3 | max = 760 aa | | | | | N = 587 | N = 353 | N = 405 | N = 1345 | |
| 19:45873459 | 2 | c.37C>T | p.(Pro13Ser) | - | pathogenic | n.t. | n.t. | 1 | 0 | 0 | 1 | BC |
| 19:45872203 | 4 | c.230_231delTG | p.(Val77Alafs) | - | n.a. | n.t. | n.t. | 0 | 1 | 0 | 1 | BC+OC |
| 19:45868194 | 7 | c.496C>T | p.(Arg166Cys) | - | pathogenic | n.t. | n.t. | 0 | 0 | 2 | 2 | BC |
| 19:45868191 | 7 | c.499G>C | p.(Glu167Gln) | rs367829012 | benign | n.t. | n.t. | 1 | 0 | 0 | 1 | BC |
| 19:45868127 | 7 | c.563G>C | p.(Gly188Ala) | - | benign | n.t. | n.t. | 1 | 0 | 0 | 1 | BC |
| 19:45867756 | 8 | c.644C>T | p.(Pro215Leu) | - | pathogenic | n.t. | n.t. | 0 | 0 | 1 | 1 | BC |
| 19:45867354 | 10 | c.839G>A | p.(Arg280His) | - | pathogenic | n.t. | n.t. | 0 | 0 | 1 | 1 | BC |
| 19:45867247 | 10 | c.946C>G | p.(Gln316Glu) | - | benign | n.t. | n.t. | 1 | 0 | 0 | 1 | BC |
| 19:45860928 | 13 | c.1267G>A | p.(Asp423Asn) | rs143710107 | benign | no | yes | 0 | 1 | 8 | 9 | 4xBC, 5xOC |
| 19:45860760 | 14 | c.1349G>A | p.(Arg450His) | rs146632315 | pathogenic | yes | no | 2 | 0 | 0 | 2 | BC |
| 19:45860626 | 15 | c.1381C>G | p.(Leu461Val) | rs121913016 | benign | yes | yes | 3 | 0 | 0 | 3 | 2xBC, 1xOC |
| 19:45860548 | 15 | c.1459C>T | p.(Arg487Trp) | rs562132292 | pathogenic | no | yes | 0 | 0 | 4 | 4 | 2xBC, 2xOC |
| 19:45858929 | 16 | c.1537G>T | p.(Asp513Tyr) | - | pathogenic | yes | yes | 1 | 0 | 0 | 1 | BC |
| 19:45858047 | 17 | c.1606G>A | p.(Val536Met) | rs142568756 | pathogenic | yes | yes | 2 | 0 | 0 | 2 | BC |
| 19:45856554 | 18 | c.1703_1704delTT | p.(Phe568fs) | - | pathogenic | no | no | 1 | 3 | 1 | 5 | BC |
| 19:45856532 | 18 | c.1726G>A | p.(Glu576Lys) | rs201165309 | pathogenic | n.t. | n.t. | 1 | 0 | 0 | 1 | BC |
| 19:45856397 | 19 | c.1775G>A | p.(Arg592His) | rs147224585 | pathogenic | yes | no | 1 | 0 | 7 | 8 | BC |
| 19:45856370 | 19 | c.1802G>A | p.(Arg601Gln) | rs140522180 | pathogenic | yes | yes | 2 | 1 | 0 | 3 | BC |
| 19:45856074 | 20 | c.1832T>C | p.(Val611Ala) | - | benign | n.t. | n.t. | 0 | 1 | 0 | 1 | BC |
| 19:45856015 | 20 | c.1891C>T | p.(Arg631Cys) | rs144511865 | pathogenic | no | no | 1 | 0 | 1 | 2 | 1xBC, 1xOC |
| 19:45855909 | IVS 20 | c.1903-2A>G | splice site | - | n.a. | n.t. | n.t. | 1 | 0 | 0 | 1 | BC+OC |
| 19:45855778 | 21 | c.2032G>C | p.(Val678Leu) | - | benign | n.t. | n.t. | 0 | 0 | 1 | 1 | BC |
| 19:45855507 | 22 | c.2150C>G | splice effect | rs144564120 | pathogenic | n.t. | n.t. | 3 | 0 | 0 | 3 | 2xBC, 1xOC |
| 19:45855492 | 22 | c.2165G>A | p.(Arg722Gln) | rs138569838 | pathogenic | n.t. | n.t. | 0 | 1 | 0 | 1 | BC |
| 19:45854932 | 23 | c.2238delA | p.(Ser746fs) | - | n.a. | yes | yes | 1 | 0 | 0 | 1 | OC |

doi:10.1371/journal.pgen.1006248.t001

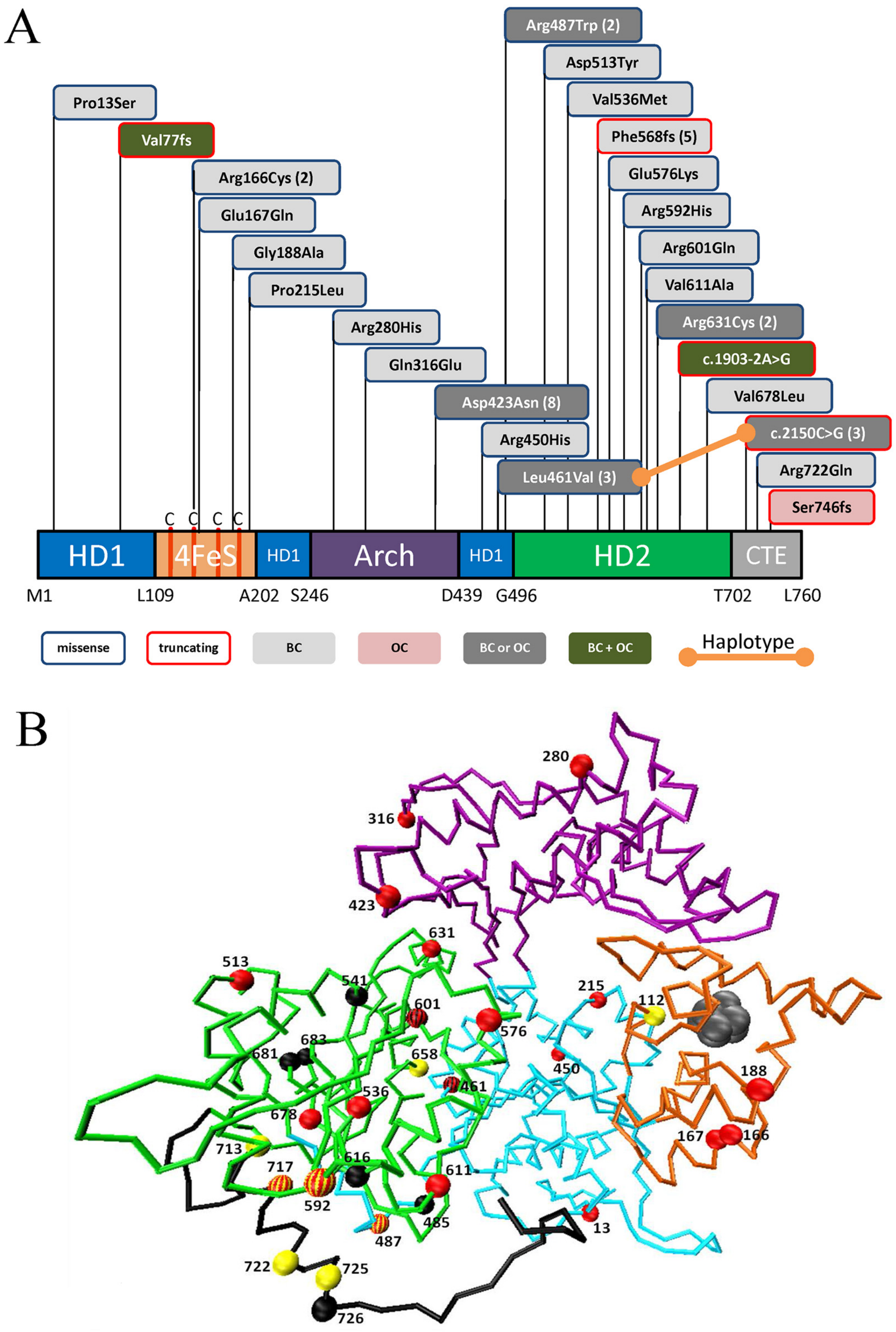


**Fig 1. Domain structure and modeling of the ERCC2 mutations.** (A) Mutations in the XPD/ERCC2 protein domains. The diagram shows the ERCC2 protein with the four XPD domains shown as HD1 (blue), HD2 (green), FeS (Orange) and Arch (purple). The human enzyme has a C-terminal (grey) extension (CTE) that probably forms an interaction surface with the p44 protein. Disease-relevant *ERCC2* mutation sites are indicated in boxes (blue or red frame: missense or truncating mutation, respectively; fillings: light-gray, cases with breast cancer

(BC); pink, case with ovarian cancer only (OC); dark-gray: cases with either breast- or ovarian cancer (BC or OC); dark-green, patients with both breast- and ovarian cancer (BC + OC)). Numbers in brackets indicate recurrent mutations. (B) Structural placement of mutations on a C-alpha trace model of human ERCC2. The residues targeted by HBOC-causing mutations are represented as space-filled red spheres. Xeroderma pigmentosum (XP) and trichothiodystrophy (TTD) disease causing mutations sites as reported in ClinVar are shown in yellow and black spheres. Missense variants at residue position 423, 461, 487, 568, 461 and 722 have been found in both BC/OC as well as XP (red-yellow spheres) and TTD (red-black spheres) patients.

doi:10.1371/journal.pgen.1006248.g001

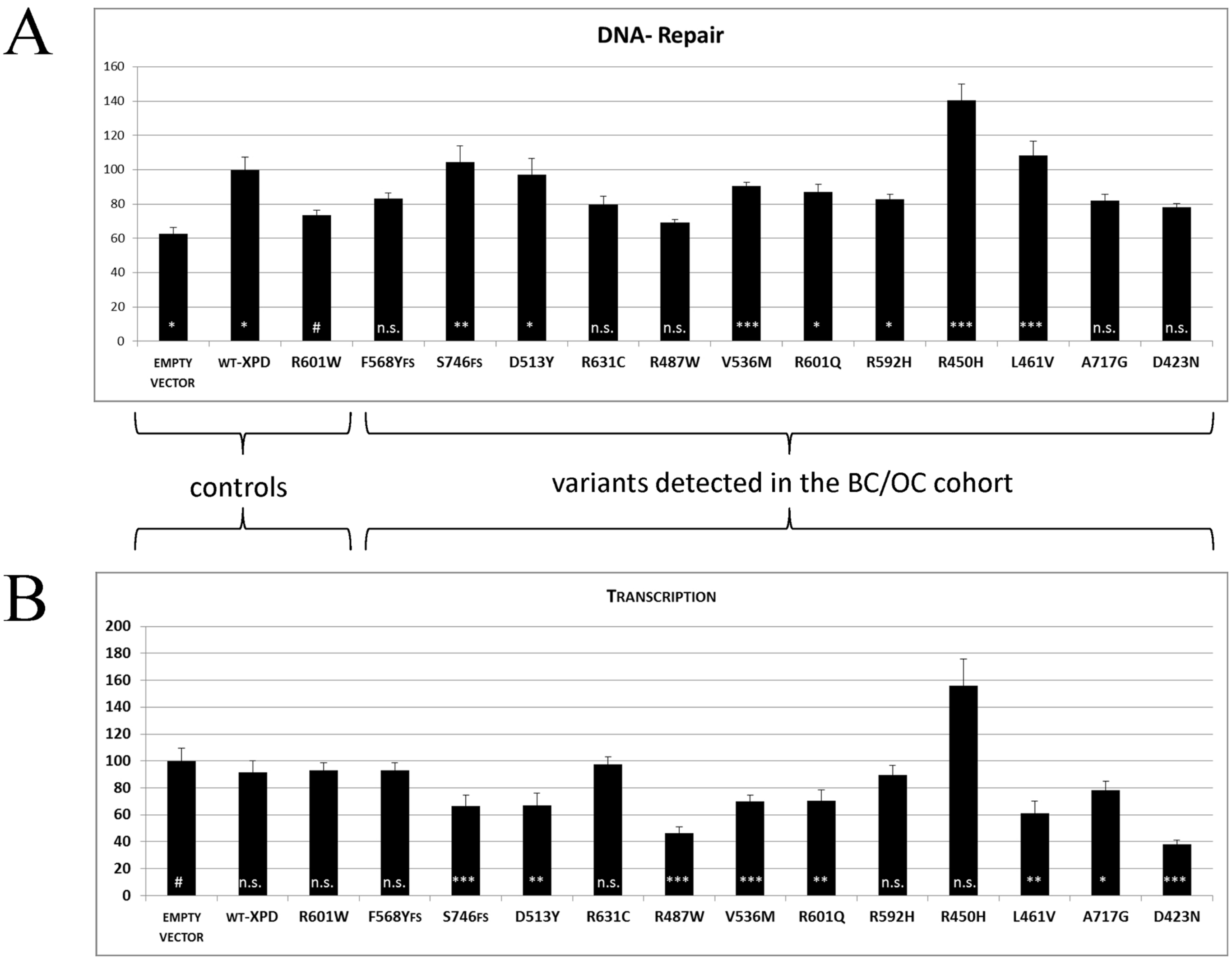
transcription (Fig 2B). Whereas six out of the 11 BC/OC-associated *ERCC2* variants tested in this study, have not yet been linked to any disease [AA positions 423, 450, 513, 536, 631, 746], five AA positions have already been found to be mutated in either TTD [AA 461 [16], 487 [17], 568 [18,19], 592 [20]] or XPD [AA 601 [21]] (Figs 1B and 2C). According to our functional assays, four ERCC2 protein variants [p.(Asp423Asn), p.(Arg487Trp), p.(Phe568Tyrfs) and p.(Arg631Cys)] failed to enhance functional NER of an UV-treated reporter gene plasmid indicating the impairment of ERCC2 repair capacity. The remaining seven tested variants retained some NER capability (Fig 2A). Concerning transcription, we detected a dominant negative influence of seven ERCC2 protein variants [p.(Asp423Asn), p.(Leu461Val), p.(Arg487Trp), p.(Asp513Tyr), p.(Val536Met), p.(Arg601Gln), p.(Ser746fs)] on reporter gene expression (Fig 2B) indicating transcription blocking. In summary, 10 of 11 mutations display diminished excision repair efficiency and/or decreased transcriptional activation capability, with p.(Asp423Asn) and p.(Arg487Trp) being the variants with the highest impact on protein function.

## The majority of the *ERCC2* mutations are founder mutations

The hallmarks of a founder mutation are recurrent appearance, population specificity and haplotype sharing. As to recurrent appearance, 11 out of 25 *ERCC2* mutations were seen at least twice in our BC/OC cohort (last column in Table 1). Among the 11 recurrent variants, 5 were identified exclusively in one of the three populations tested in this study (e.g. p.(Arg487Trp): 4x LT only) and another 5 were significantly overrepresented in one of the 3 populations (e.g. p.(Asp423Asn): 8x LT, 1x CZ, 0x GE). For two of the population-enriched recurrent founder mutations, we could also demonstrate haplotype sharing: (i) the mutation c.1381C>G (rs121913016) always co-occurred and co-segregated with mutation c.2150C>G (rs144564120), a haplotype which has been observed repeatedly in TTD/XPD patients [9,16,22]. (ii) In almost all cases (10/11) the frame-shift mutation c.1703_1704delTT co-occurred with the c.1758+32C>G polymorphism (rs238417). Furthermore, these two variants are only 84 nt apart from each other and all NGS-reads covering both variants showed these variants simultaneously, i.e. these variants are definitely localized in *cis* on the same DNA molecule.

## Even small region-specific control cohorts outnumber huge public variant databases

In the variant discovery phase of this project, the frequencies of *ERCC2* mutations found in the BC/OC cohort were compared to the corresponding frequencies in public databases provided by the NHLBI Exome Sequencing Project (ESP) and the Exome Aggregation Consortium (ExAC). As shown in Table 2, some intriguing mutations, like p.(Phe568fs) and p.(Asp423Asn), have very low frequencies according to ExAC, suggesting significant odds ratios (OR). As a first proof of principle measure, we performed segregation analysis. However, none of our recurrent *ERCC2* mutations showed convincing co-segregation with BC/OC (Fig 3). Moreover, as soon as a small number of population-specific control probands has been sequenced, it became clear that almost all founder mutations in the BC/OC cohort showed



C

| Mutation / characteristics | Asp 423 Asn | Arg 450 His | Leu 461 Val | Arg 487 Trp | Asp 513 Tyr | Val5 36M et | Phe 568 fs | Arg 592 His | Arg 601 Gln | Arg 631 Cys | Ser 746 fs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Codon linked to disease | TTD | --- | TTD | TTD | --- | --- | TTD | --- | XPD | --- | --- |
| In-silico effect prediction | Bgn | Pat | Bgn | Pat | Pat | Pat | fs | Pat | Pat | Pat | fs |
| Recurrent mutation | Yes | Yes | Yes | Yes | --- | Yes | Yes | Yes | Yes | Yes | --- |

**Fig 2. Nucleotide excision repair (NER) capacity and Transcriptional activity of breast cancer associated XPD/ERCC2 variants.** (A) Several XPD/ERCC2 variants cloned into an expression vector were analyzed regarding to complementation of *ERCC2*-defective XP6BE cells overexpressing the NER-deficient R601W XPD mutant [15] (normalization for overexpression artifacts). Black bars indicate the mean relative repair capacity (in %, WT-XPD was set to 100%) of an UV irradiated firefly luciferase reporter gene plasmid (UVC 1000 J/m$^2$) obtained by host cell reactivation (n>6 in triplicates). Red lines mark the range between DNA-repair levels of empty vector, i.e. residual repair activity of the cells, and

WT-XPD, i.e. 100% repair capacity. (B) Dominant modulation of firefly luciferase reporter gene expression (without irradiation) via overexpression of XPD/ERCC2 BC/OC-associated variants was estimated in the transcriptionally-proficient but repair-deficient XPD/ERCC2-defective XP6BE cells. Black bars indicate the mean relative reporter gene expression (in %, empty vector control was set to 100%), obtained by CMV-promotor driven basal transcription (n>6 in triplicates). Error bars indicate the standard error of the mean. Significance levels were calculated, after pairwise testing for normal distribution of the values, using appropriate statistical tests for comparison of two groups (T-Test or U-Test, # = reference group, *** = p<0.001, ** = p<0.01, * = p<0.05, n.s. = not significant). (C) Additional characteristics of the mutations tested for repair efficiency and transcriptional activity.

doi:10.1371/journal.pgen.1006248.g002

similar frequencies in the ethnically matching control cohorts. The only exception so far is the Lithuanian mutation p.(Arg487Trp), which was found 4 times in the Lithuanian BC/OC cohort and not (yet) in the corresponding control cohort (Table 2). With just above 100 individuals this cohort is way too small to be of any statistical relevance. Therefore, the acquisition of additional samples is mandatory. But even in this very early phase of variant (de-)validation it becomes evident that regionally matching control cohorts–as small as they may be–are superior to any huge global cohort. Since genotypic data allow to locate the geographic origin of a given individual within a few hundred kilometers [23], the term “regionally matching” should be defined as “less than ca. 300 km distance from the recruitment center”. As a consequence, regionally matching controls are even superior to population-specific controls, because populations do mix, especially in regions close to national borders. The p.Phe568fs mutation, for example, has been seen only once in a German BC/OC index case and never in the 1844 German controls. Based on population-specific data we would have been very excited about this finding. But the German case was recruited in Dresden, close to the Czech border, and in Prague, 118 km away, the same mutation has been found twice in a small control cohort of only 105 non-cancer females. This underlines the importance of regional controls and multinational studies for reliable variant validation.

## *ERCC2* mutations with tumorigenic relevance are probably located in very small and scattered areas of the protein

Due to its involvement in DNA repair and due to encoding a helicase like *RECQL* [7], *ERCC2* is a plausible gene candidate for familial cancer susceptibility. Bi-allelic mutations in *ERCC2*, however, can cause the cancer-prone disease XPD as well as the “non-cancer”-disease TTD [27] and there is no evident genotype-phenotype correlation [19]. The pathogenic p.(Arg112His) mutation, for example, has been identified in TTD patients as well as in a patient with major features of XPD [19]. Furthermore, impairment of DNA repair capacity is not correlated with tumor burden: the mutation p.(Phe568Tyrfs), for example, has been identified in non-cancer TTD patients twice, but not once in cancer-prone XPD patients, although this study (Fig 2) as well as a previous study [19] clearly show diminished repair capability of this frameshift variant. From these observations we have to conclude that a limited subset of mutations in *ERCC2* might predispose to cancer but these mutations are not likely to cluster in a defined area of the gene nor do they necessarily affect a specific sub-function of the ERCC2 protein. Therefore, cancer predisposing *ERCC2* mutations are very likely to be discovered only on the basis of familial co-segregation with cancer and overrepresentation in cancer cohorts vs. region-specific controls.

## The incidence of *ERCC2*-related diseases is not in line with the frequency of deleterious founder mutations in the corresponding populations

Although the founder mutations tested in this study may not predispose to BC/OC they still confer carrier status for the recessive disorders XPD (OMIM 278730), TTD (OMIM 601675)

**Table 2. *ERCC2* allele frequencies (%) in BC/OC patients and corresponding control cohorts.** The allele frequency is counted on the basis of sample size (in brackets) and number of observed cases (see Table 1) with hetero- and homozygosity.

| AA / nt change | CZ | CZ | LT | LT | GE | GE | ExAc |
|---|---|---|---|---|---|---|---|
| (N = 25) | BC/OC | Ctrl | BC/OC | Ctrl | BC/OC | Ctrl | vers. 0.2 |
| | [353][a] | [453][b] | [405] | [103] | [587][c] | [1844][d] | [variable][e] |
| Pro13Ser | 0 | 0 | 0 | 0 | 0.0851 | 0 | 0 |
| Val77Alafs | 0.1416 | 0 | 0 | 0 | 0 | 0 | 0 |
| Arg166Cys | 0 | 0 | 0.2469 | 0 | 0 | 0 | 0 |
| Glu167Gln | 0 | 0 | 0 | 0 | 0.0851 | 0 | 0.0033 |
| Gly188Ala | 0 | 0 | 0 | 0 | 0.0851 | 0 | 0 |
| Pro215Leu | 0 | 0 | 0.1234 | 0 | 0 | 0 | 0 |
| Arg280His | 0 | 0 | 0.1234 | 0 | 0 | 0 | 0.0072 |
| Gln316Glu | 0 | 0 | 0 | 0 | 0.0851 | 0 | 0.0152 |
| Asp423Asn | 0.1416 | 0.1104 | 0.9876 | 1.456 | 0 | 0.0542 | 0.0248 |
| Arg450His | 0 | 0 | 0 | 0 | 0.1704 | 0.0813 | 0.0214 |
| Leu461Val | 0 | 0 | 0 | 0 | 0.2553 | 0.1356 | 0.1345 |
| Arg487Trp | 0 | 0 | 0.4938 | 0 | 0 | 0 | 0.0034 |
| Asp513Tyr | 0 | 0 | 0 | 0 | 0.0851 | 0 | 0 |
| Val536Met | 0 | 0 | 0 | 0 | 0.1704 | 0 | 0.0231 |
| p.Phe568fs | 0.4249 | 0.4415 | 0.1234 | 0 | 0.0851 | 0 | 0.0093 |
| Glu576Lys | 0 | 0 | 0 | 0 | 0.0851 | 0.0542 | 0.0008 |
| Arg592His | 0 | 0 | 0.8642 | 0 | 0.0851 | 0 | 0.0332 |
| Arg601Gln | 0 | 0.1104 | 0 | 0 | 0.1704 | 0.0542 | 0.0175 |
| Val611Ala | 0.1416 | 0 | 0 | 0 | 0 | 0 | 0.0042 |
| Arg631Cys | 0 | 0 | 0.1234 | 0 | 0.0851 | 0 | 0.0025 |
| c.1903-2A>G | 0 | 0 | 0 | 0 | 0.0851 | 0 | 0 |
| Val678Leu | 0 | 0 | 0.1234 | 0 | 0 | 0 | 0 |
| c.2150C>G | 0 | 0 | 0 | 0 | 0.2553 | 0.0813 | 0.0349 |
| Arg722Gln | 0.1416 | 0 | 0 | 0 | 0 | 0 | 0.0067 |
| p.Ser746fs | 0 | 0 | 0 | 0 | 0.0851 | 0 | 0 |

BC/OC = index cases with breast- and/or Ovarian cancer; Crtl = healthy or non-cancer related individuals; CZ = Czech Republic, GE = Germany, LT = Lithuania; AA = Amino acid; nt = nucleotide; ExAC = Exome Aggregation Consortium, Cambridge, MA (URL: http://exac.broadinstitute.org) [accessed May 2015];

[a] 28 samples from Brno (TruSight-Cancer) + 325 samples from Prague [24,25] (custom panel with 581 genes);

[b] 105 female non-cancer samples from Prague [25,26] (custom panel with 581 genes) + 108 female non-cancer samples from Brno, sequenced in pools with the TruSight-Cancer panel + 240 non-cancer samples from Prague, sequenced in pools with the TruSight-Cancer panel;

[c] 271 samples from Dresden + 316 samples from Munich (MGZ), all sequenced with the TruSight-Cancer panel;

[d] 1629 individual exome samples from the Cologne Center for Genomics (CCG) + 79 individual non-BC/OC TruSight-One samples from Dresden + 136 individual non-BC/OC TruSight-Cancer samples from Dresden and Munich (MGZ);

[e] Since the exome data have been collected from various sources with various enrichment strategies, the sample size varies for each variant. Each allele frequency has been calculated with the corresponding sample size for that allele.

doi:10.1371/journal.pgen.1006248.t002

and COFS2 (OMIM 610756). Even the TTD-causing mutation p.(Phe568fs) alone has been detected in 7 of 806 samples from the Czech Republic (CZ), i.e. the frequency of heterozygous carriers of this mutation is approx. 0.86%. According to Hardy-Weinberg equilibrium model, this would result in a TTD incidence of 1/30.000. Based on combined data from the DNA repair diagnostic centers in France, West-Germany, Italy, the Netherlands and the United Kingdom the actual incidence for TTD is 1.2 per million [28]. Since it is reasonable to assume

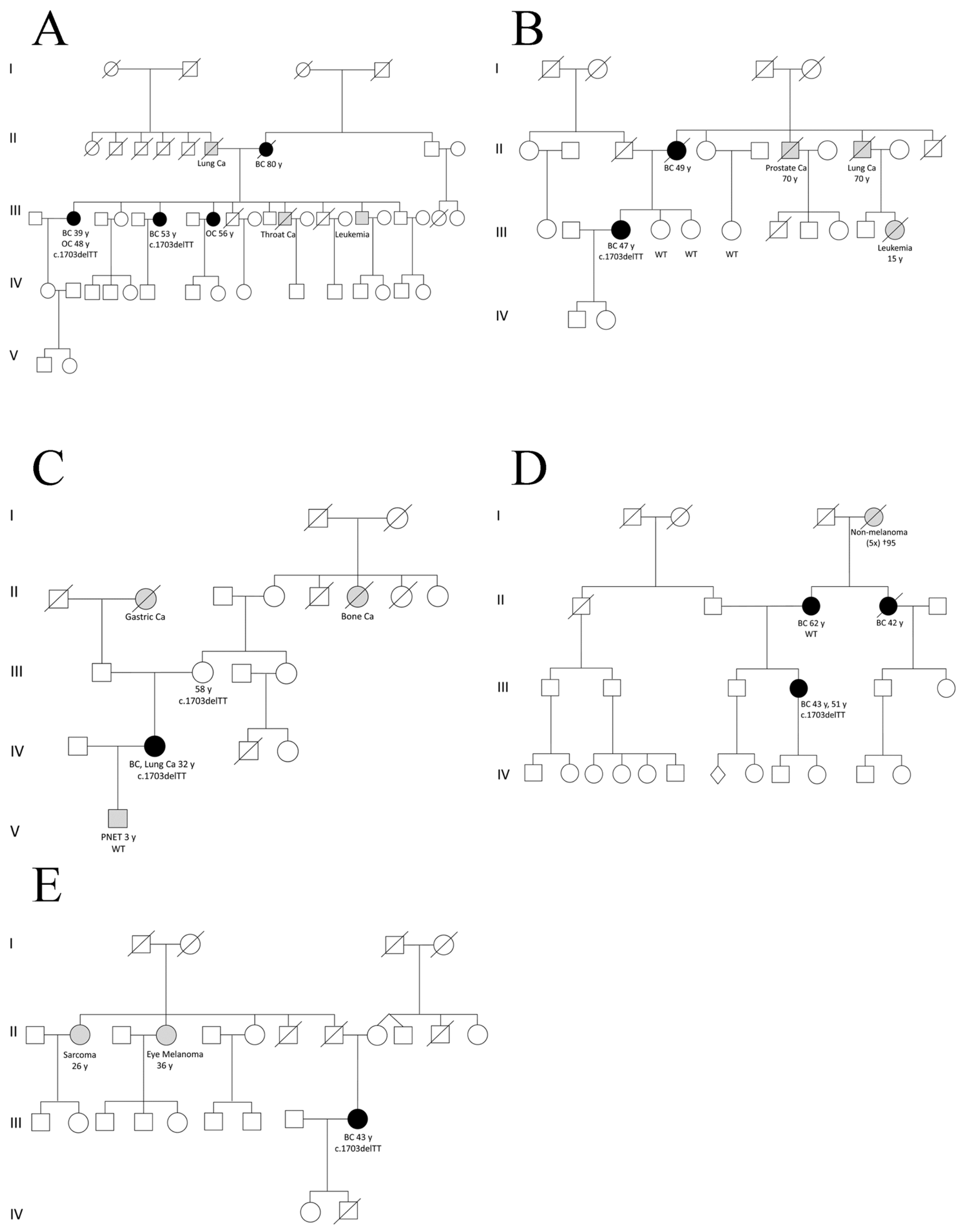


**Fig 3. *ERCC2* frameshift mutation c.1703_1704delTT (p.Phe568fs) in familial breast and ovarian cancer pedigrees.** Individuals with breast cancer (BC), ovarian cancer (OC) or both (BC, OC) are shown as circles filled in black. Individuals tested positive for the familial mutation are indicated in detail; those with WT (wild-type) have been tested negative. All affected individuals with BC or OC not tested for germline mutations in ERCC2 were either deceased or refused testing. (A) German, (B) Lithuanian and (C-E) Czech pedigrees.

doi:10.1371/journal.pgen.1006248.g003

that (i) a TTD incidence of 1/30.000 would not be missed by the clinical geneticists in CZ and (ii) the publications reporting p.(Phe568fs) as TTD-causing [9,19] are not wrong, there is one logical explanation for the discrepancy between allele frequency and disease incidence: homozygosity for p.Phe568fs is embryonic lethal. This is in-line with the observation that complete loss of ERCC2 activity is not compatible with life in homozygous knock-out mice [29] and it is also consistent with the observation that all XPD and TTD patients tested so far have residual ERCC2 activity [30]. Since an elevated TTD/XPD incidence has not been reported in Lithuania either, we can assume that homozygosity of the frequent Lithuanian founder mutation p.(Asp423Asn) (Table 2), which clearly displayed functional deficiency in our experiments (Fig 2), is embryonic lethal as well.

In conclusion, this multi-national study of *ERCC2* mutations in patients with familial BC/OC and regionally matching controls identified and functionally verified a broad spectrum of unique and recurrent *ERCC2* mutations. Although the frequent founder mutations are not very likely to predispose to BC/OC, some mutations, like p.(Val77Alafs), that are unique to the BC/OC cohort are worth to be considered in future large-scale association studies.

## Materials and Methods

### Ethics statement

Informed written consent was obtained from all patients and the study was approved by the Local Research Ethics Committee (EK 162072007).

### Subjects, families and pedigrees

We enrolled affected individuals from 587 German BC and BC/OC pedigrees with hereditary gynecological malignancies through a genetic counseling program at two centers (Dresden, Munich) from the "German Consortium for hereditary breast- and ovarian cancer" (GC-HBOC) and at the Medical Genetics Center (MGZ) in Munich. Additional 131 BC- and 136 BC/OC families were collected at the Vilnius University Hospital Santariskiu Klinikos in Vilnius, Lithuania and 28 BC/OC families were gathered in the Czech Republic at Brno. The Czech Prague subgroup involved 325 BC patients negatively tested for presence of pathogenic *BRCA1* and *BRCA2* variants [24] and 105 non-cancer controls analyzed as described recently [25,26], and additional 240 controls [26] sequenced in pools. The BC pedigrees fulfilled the criterion that at least three affected females with breast cancer but no ovarian cancers were present (breast cancer pedigrees). In the BC/OC pedigrees, at least one case of breast and one ovarian cancer had occurred. All individuals with variant *ERCC2* alleles were checked for mutations in 10 BC/OC core genes defined by GC-HBOC (*ATM*, *BRCA1*, *BRCA2*, *CDH1*, *CHEK2*, *NBN*, *PALB2*, *RAD51C*, *RAD51D* and *TP53*). Informed consent was obtained from all people participating in the study, and the experiments were approved by the ethics committees of the institutions contributing to this project.

### TruSight-Cancer panel sequencing

DNA was obtained from peripheral blood of all patients. For panel enrichment approximately 85 ng genomic DNA was required. We used the TruSight Cancer Illumina kit (Illumina), which targets the coding sequences of 94 genes associated with a predisposition towards cancer (S1 Table), following the manufacturer's instructions. Sequencing was carried out on an Illumina MiSeq instrument as 150 bp paired-end runs with V2 chemistry. Reads were aligned to the human reference genome (GRCh37/hg19) using BWA (v 0.7.8-r455) with standard parameters. Duplicate reads and reads that did not map unambiguously were removed. The percentage of reads overlapping targeted regions and coverage statistics of targeted regions were calculated using Shell scripts. Single-nucleotide variants and small insertions and deletions (INDELs) were called using SAMtools (v1.1). We used the following parameters: a maximum read depth of 10000 (parameter -d), a maximum per sample depth of 10000 for INDEL calling (parameter -L), adjustment of mapping quality (parameter -C) and recalculation of per-Base Alignment Quality (parameter -E). Additionally, we required putative SNVs to fulfill the following criteria: a minimum of 20% of reads showing the variant base and the variant base is indicated by reads coming from different strands. For INDELs we required that at least 15% of reads covering this position indicate the INDEL. Variant annotation was performed with snpEff (v 4.0e) and Alamut-Batch (v 1.3.1) based on the RefSeq database. Only variants (SNVs/small INDELs) in the coding region and the flanking intronic regions (±15 bp) were evaluated.

## Custom breast cancer panel sequencing

The data related to the ERCC2 gene in this study were retrieved from the custom-made gene panel sequencing analysis described recently [25]. Briefly, genomic DNA was obtained from a peripheral blood of 325 BC Czech patients from the Prague area that were negatively tested for a presence of pathogenic variants in the *BRCA1* or *BRCA2* gene previously [24]. The frequency of population-specific variants was assessed by a concurrent analysis of 105 control DNAs obtained from non-cancer individuals [26]. One μg of genomic DNA was used for library construction. The DNA was fragmented by ultrasonication and edited for SOLiD sequencing. Target DNA enrichment was performed by a custom solution-based sequence capture (SeqCap EZ Choice Library, Roche) according to the NimbleGenSeqCap EZ Library SR User's Guide (Version 4.2, Roche). Five hundred and ninety targeted genes include 141 genes that code for known proteins involved in DNA repair and DNA damage response pathways, and an additional set of genes retrieved from Phenopedia at HuGE Navigator16 web site associated with "breast neoplasms" (assessed February 2012). Captured libraries were sequenced on SOLiD4 system. Finally, exonic regions of 581 genes were captured successfully with sufficient coverage. Reads were aligned to the human reference genome (GRCh37/hg19) using Novoalign (CS 1.01.08) with standard parameters. Conversion of SAM to BAM format was performed with SAMtools (0.1.8). Single-nucleotide variants and small insertions and deletions (INDELs) were called using SAMtools (0.1.8). Variant annotation was performed with ANNOVAR [31]. For final evaluations, small INDELs, intronic variants flanking ± 2 bp to exon borders, and rare SNPs (presented in 1000 genome or exome sequencing (ESP) projects with frequency <1%) were considered.

## Sanger sequencing

Validation of *ERCC2* variants in probands and family members was performed by classical Sanger sequencing. Additional DNAs from 8 HBOC patients affected by malignant melanoma (5 cases) or presence of melanoma in other family members (3 cases) were analyzed for the complete *ERCC2* coding region. *ERCC2* exons were amplified with intronic primers (S2 Table) and sequenced using the ABI Prism Terminator Cycle Sequencing Ready Reaction Kit (Applied Biosystems). Genomic DNA (50 ng) containing 1x PCR Master Mix (Qiagen) and 0.25 μM of each forward and reverse primers in 15 μl reaction volume was subjected to PCR amplification for 25 cycles (30 sec at 95°C, 30 sec at 64°C and 30 sec at 72°C).

## Functional validation of *ERCC2* variants

**Variant cloning.** Wild type *ERCC2* cDNA was amplified from reverse transcribed mRNA isolated from fibroblasts derived from healthy donors (RevertAid H Minus First strand cDNA synthesis kit; Thermo scientific, Waltham, MA, USA) using forward (5'TTAGGTACCATGA AGCTCAACGTGGACG) and reverse (5' TTATCTAGATCAGAGCTGCTGAGCAATCT) primers and cloned into the pJET1.2/blunt vector (CloneJET PCR Cloning Kit; Life technologies, Waltham, MA, USA). These primers carry *Kpn*I and *Xba*I restriction sites to release *ERCC2* cDNA by double restriction enzyme digestion (Life technologies). The *ERCC2* cDNA was purified from agarose gels using the Wizard SV Gel and PCR Clean-Up System (Promega, Klaus, Austria) and cloned into the pcDNA3.1(+) mammalian expression vector (Life technologies) and subsequently transformed into DH5α *E.coli* cells. Colony PCR (using T7 and M13 primers) and Sanger sequencing of the entire gene was performed using the BigDye Terminator v3.1 Cycle Sequencing Kit (Life technologies, for primers see S3 Table).

For generation of the *ERCC2* variants, site directed mutagenesis was applied using Phusion High-Fidelity DNA Polymerase (Life technologies) and specific primer pairs in either the classical protocol (for variants Ser746FS and D513Y, Stratagene) or an optimized site-directed method (all other variants, for primers see S3 Table). For the latter, template (100 ng *ERCC2* in pcDNA3.1(+)) was first subjected to dam methylation using dam methyl transferase (NEB, Frankfurt a. M., Germany). Afterwards, a first PCR was conducted with the forward-primer using Phusion polymerase (Life technologies) in a 2-Step PCR protocol with 5 minutes of annealing and elongation at 72°C for 18 cycles. Then over-night enzyme digestion with *Dpn*I (Life technologies) was followed by ethanol precipitation. A second PCR using reverse primers (same conditions) was performed with this template and ethanol precipitated. The final reaction product was subject to transformation of DH5α *E.coli* cells. Positive clones were verified by Sanger sequencing as described above.

**Assay set-up.** The host cell reactivation (HCR) assay measures the amount of nucleotide excision repair (NER) in actively transcribed genes. This dual reporter gene assay deploys the turnover rate of firefly luciferase substrate as readout for the NER capacity of host cells transfected with the (UV-) damaged reporter gene plasmid encoding for firefly luciferase [32]. HCR can be used for DNA repair capacity assessment of NER deficient host cells transfected with DNA repair gene variants as well as for measuring in situ transcription using non-irradiated firefly luciferase reporter gene plasmids [33,34].

*ERCC2*-deficient XP6BE-SV-immortalized fibroblasts were a generous gift of K.H. Kraemer (NIH, Bethesda, MD, USA) and harbor two differently mutated *ERCC2* alleles [p.Arg683Trp and an in-frame deletion of amino acids (AA) 36–61] [9]. XP6BE cells were transfected using Attractene Transfection Reagent (Qiagen, Hilden, Germany) according to the manufacture's advice, with plasmids coding for firefly luciferase (100 ng), renilla-luciferase (50 ng) and an empty pcDNA3.1(+) vector or XPD-variants cloned into the pcDNA3.1(+) expression vector (100 ng) (for cloning see above). The plasmid coding for firefly luciferase was divided into two fractions prior to transfection. One fraction was irradiated with 1000 J/m2 of UVC light, a second fraction stayed untreated. The non-irradiated renilla-luciferase plasmid serves as an internal control for normalization of transfection efficacy.

After incubation of transfected XP6BE cells for one day (37°C, 5% $CO_2$), which allows sufficient repair of the UV-photoproducts and protein expression of the luciferases, cells were lysed and analyzed using Dual-Luciferase Reporter Assay System (Promega, Klaus, Austria). The luminescence measurements were performed in a white Glomax 96 microplate using the Glomax luminometer (Promega, Klaus, Austria).

The relative repair capacity is estimated using this formula:

$$repair\ (\%) = \frac{\text{mean (irradiated firefly/renilla per well)}}{\text{mean (unirradiated firefly/renilla per well)}}\ x\ 100$$

The repair capacity of XP6BE cells transfected with the wild type *ERCC2* cDNA containing expression vector was set to 100%.

Transcriptional activity was calculated as the amount of firefly luciferase expression from non-irradiated plasmids in XP6BE cells transfected either with wild type *ERCC2* or breast cancer associated *ERCC2* variants containing expression vectors relative to the amount of firefly luciferase expression in XP6BE cells transfected with the empty expression vector. The latter was set to 100%. Every experiment (NER capacity as well as transcription) was conducted at least six times in triplicates.

## Modeling of ERCC2 protein structure

**Structural modeling of the ERCC2 variants.** Homology modeling of the human ERCC2 protein was performed with SWISS-MODEL (ExPASy). The crystal structure of the ATP-dependent DNA helicase Ta0057 from *Thermoplasma acidophilum* (RCSB:4A15, UniProt: Q9HM14) was used as template structure for modeling. Predicted models for the residue changes of the detected missense mutations in *ERCC2* were displayed and analyzed using Visual Molecular Dynamics (VMD) (S1 Fig). The predicted models were superimposed onto the Ta0057 structure with the MulitSeq tool integrated in VMD.

**In-silico interpretation of missense variants.** The probability of effect of non-synonymous mutations in *ERCC2* was calculated by the amino acid (AA) substitution prediction methods SIFT, PolyPhen2, Provean, Mutation Taster, MAPP, and AGVD (S4 Table). Based on these data, a summarizing rating was assessed (last column in S4 Table and Table 1). Distribution of PhyloP and Grantham scores [35] for dbSNP, ClinVar and all variants identified in *ERCC2* were analyzed. Statistical probability scores of PhyloP and Grantham scores and analysis of distribution plots are provided (S4 Table and S2 Fig).

## AA conservation alignment

A multiple alignment of ERCC2 AA sequences was done according to HomoloGene (NCBI) in order to assess the AA conservation of the detected variants in 20 species with homologous proteins (S3 Fig).

## Supporting Information

**S1 Fig. ERCC2 domain structure and overlay of ERCC2 missense mutations Arg478Trp and Asp423Asn.** A) Schematic showing the domain structure and canonical motifs of human ERCC2. Helicase motor domains HD1 (blue) and HD2 (green) form the DNA ATP-binding interface. The FeS (orange) and the Arch (purple) domains are inserted into HD1. The boundaries of the FeS cluster binding domain are indicated by red spheres. The human enzyme C-terminal (grey) extension (CTE) is indicated in grey. Domain boundaries are indicated by residue numbers. B,C) 3D representation of the native (cyan) and mutant (pink) overlayed ERCC2 protein structures show a detailed structural environment of the wild-type (green), Arg487 and Asp423 residues in comparison to the Arg487Trp and Asp423Asn mutants (red). Surrounding amino acids (AAs) are indicated as licorice. (B) Note the significant changes in the AA constellations Arg424, Thr425 induced by the by the Asp423Asn replacement. (C) The Arg487Trp AA replacement introduces a tryptophan residue which protrudes beyond the protein surface and might destabilize the interactions with the surrounding AAs His700, Glu690 and Leu701 within the protein loop.
(TIF)

**S2 Fig. Distribution of PhyloP and CADD scores for 1000G, ClinVar and the mutations identified in this study in the ERCC2 gene.** A) Evolutionary conservations (PhyloP) and Combined Annotation Dependent Depletion (CADD) scores are represented for all non-synonymous ERCC2 variants found in BC/OC patients. Blue: Variants with no significant functional effect; Red: variants which showed a deleterious functional effect by no complementation of NER-deficient cells and/or negative modulation of transcription; Green: variants not tested. B) This analysis was further extended to analyze these combined scores for all non-synonymous variants reported in 1000G and ClinVar with no reported clinical significance (Class 1–3), or ClinVar reported pathogenic variants (Class 4–5) to visualize the probability for the ERCC2 variants which have not been functionally tested to be pathogenic or benign. Heat maps show the distribution and frequency for the combined PhyloP and CADD scores in 1000G and ClinVar. Red colors indicate a low frequency and green colors a high frequency. ERCC2 variants showing no functional pathogenic effect (circle), pathogenic variants with NER complementation failure and/or negative modulation of transcription (triangles), and variants not tested in our functional studies (black square) are represented. ERCC2 variants with deleterious functional effects show a better overlap with ClinVar pathogenic variants (Class 4–5) by their location mostly restricted to dark green and yellow as indicated. In contrast, location of variants shows within the dark red plot region when compared to 1000G and ClinVar (Class 1–3).Variants not included in our functional studies show a similar distribution pattern as functional deleterious variants which overlaps with ClinVar pathogenic variants (Class 4–5). In total, most of the ERCC2 variants are located in areas of high conservation and high deleteriousness. Statistical probability scores for these analyses are provided in S4 Table. PhyloP and CADD scores for 1000G and ClinVar variants were obtained from the annotation browser SNiPA [36].
(TIF)

**S3 Fig. ERCC2 amino acid (AA) sequence alignment.** Multiple sequence alignment of protein regions from various species surrounding the identified human ERCC2 missense variants (S4 Table). Affected residues are indicated in red letters. The dotted lines correspond to sequence gaps or sequence regions not yet available. Except Glu167, all affected residues showed strong conservation across vertebrates (Arg166, Gly188, Arg280, Gln316, Asp423, Leu461, Arg487, Val611, Val678, Ala717, Arg722) or even across all species (Pro13, Pro215, Arg450, D513, Val536, Glu576, Arg592, Arg601, Arg631). The AA variability at codon 167 is in line with the results of the effect prediction algorithms which predict the Glu167Gln replacement as benign (S4 Table). Accession number of the ERCC2protein sequences used for AA sequence comparison are as follows: Homo sapiens (NP_000391.1); Pan troglodytes (NP_001233519.1); Macaca mulatta (XP_002808245.1); Canis lupus (XP_541562.3); Bos taurus (NP_001096787.1); Mus musculus (NP_031975.2); Rattus norvegicus (NP_001166280.1); Xenopus tropicalis (NP_001008131.1); Danio rerio (NP_957220.1); Drosophila melanogaster (NP_726036.2); Anopheles gambiae (XP_311900.4); Caenorhabditis elegans (NP_497182.2); Saccharomyces cerevisiae (NP_011098.3); Kluyveromyces lactis (XP_452994.1); Eremothecium gossypii (NP_986780.1); Schizosaccharomyces pombe (NP_593025.1); Magnaporthe oryzae (XP_003716866.1); Neurospora crassa (XP_956536.2); Arabidopsis thaliana (NP_171818.1); and Oryza sativa (NP_001054627.1).
(TIF)

**S1 Table. Genes covered by the TruSight-Cancer gene panel.**
(DOCX)

**S2 Table. ERCC2/XPD primers for Sanger-validation of NGS derived mutations and analysis of familial segregation.** All sequences shown in 5' → 3' direction. XPD = alternate name of *ERCC2*.
(DOCX)

**S3 Table. Primer pairs used for PCR amplification, Sanger sequencing, and site directed mutagenesis of ERCC2 variants.** All primers (de-salted and deprotected) were synthesized by Sigma-Aldrich (Taufkirchen, Germany).
(DOCX)

**S4 Table. Effect prediction of ERCC2 missense variants.** The probability of effect of non-synonymous mutations in ERCC2 was predicted by the computer programs: SIFT, Sorting Invariant from Tolerated (Score under 0,05: not tolerated; Range 0–1); PolyPhen-2, Classification following PSIC scores (HumVar, "benign"- "possibly damaging"—"probably damaging ", Range: 0–1); Provean, Protein Variation Effect Analyze; MAPP, Multivariate Analysis of Protein Polymorphism; Align-GVGD, Scores (C0, C15, C25, C35, C45, C55, C65) from C0 (likely benign) to C65 (likely pathogenic); CADD, Combined Annotation Dependent Depletion [12]. Dlt = deleterious, PrD = probably damaging, PsD = possibly damaging, Bgn = benign, Ntr = neutral. Conservation was calculated with PhyloP (Score range from -14.1 to 6.4). Grantham [35] distance scores (Range 0–215). AA exchanges in gray background are located in cis and form a haplotype.
(DOCX)

## Acknowledgments

We thank all the families for providing samples for this study. We are deeply indebted to the German Cancer Aid for establishing the German Consortium for Hereditary Breast- and Ovarian Cancer (GC-HBOC).

## References

1. Couch FJ, Nathanson KL, Offit K (2014) Two decades after BRCA: setting paradigms in personalized cancer care and prevention. Science 343: 1466–1470. doi: 10.1126/science.1251827 PMID: 24675953
2. Turnbull C, Rahman N (2008) Genetic predisposition to breast cancer: past, present, and future. Annu Rev Genomics Hum Genet 9: 321–345. doi: 10.1146/annurev.genom.9.081307.164339 PMID: 18544032
3. Antoniou AC, Easton DF (2006) Models of genetic susceptibility to breast cancer. Oncogene 25: 5898–5905. PMID: 16998504
4. Easton DF, Pharoah PD, Antoniou AC, Tischkowitz M, Tavtigian SV, et al. (2015) Gene-panel sequencing and the prediction of breast-cancer risk. N Engl J Med 372: 2243–2257. doi: 10.1056/NEJMsr1501341 PMID: 26014596
5. Evans BJ, Burke W, Jarvik GP (2015) The FDA and genomic tests—getting regulation right. N Engl J Med 372: 2258–2264. doi: 10.1056/NEJMsr1501194 PMID: 26014592
6. Zhang G, Zeng Y, Liu Z, Wei W (2013) Significant association between Nijmegen breakage syndrome 1 657del5 polymorphism and breast cancer risk. Tumour Biol 34: 2753–2757. doi: 10.1007/s13277-013-0830-z PMID: 23765759
7. Cybulski C, Carrot-Zhang J, Kluzniak W, Rivera B, Kashyap A, et al. (2015) Germline RECQL mutations are associated with breast cancer susceptibility. Nat Genet.
8. Singh A, Compe E, Le May N, Egly JM (2015) TFIIH Subunit Alterations Causing Xeroderma Pigmentosum and Trichothiodystrophy Specifically Disturb Several Steps during Transcription. Am J Hum Genet 96: 194–207. doi: 10.1016/j.ajhg.2014.12.012 PMID: 25620205
9. Takayama K, Salazar EP, Lehmann A, Stefanini M, Thompson LH, et al. (1995) Defects in the DNA repair and transcription gene ERCC2 in the cancer-prone disorder xeroderma pigmentosum group D. Cancer Res 55: 5656–5663. PMID: 7585650

10. Lehmann J, Schubert S, Emmert S (2014) Xeroderma pigmentosum: diagnostic procedures, interdisciplinary patient care, and novel therapeutic approaches. J Dtsch Dermatol Ges 12: 867–872. doi: 10.1111/ddg.12419 PMID: 25262888

11. Pabalan N, Francisco-Pabalan O, Sung L, Jarjanazi H, Ozcelik H (2010) Meta-analysis of two ERCC2 (XPD) polymorphisms, Asp312Asn and Lys751Gln, in breast cancer. Breast Cancer Res Treat 124: 531–541. doi: 10.1007/s10549-010-0863-6 PMID: 20379847

12. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46: 310–315. doi: 10.1038/ng.2892 PMID: 24487276

13. Fan L, Fuss JO, Cheng QJ, Arvai AS, Hammel M, et al. (2008) XPD helicase structures and activities: insights into the cancer and aging phenotypes from XPD mutations. Cell 133: 789–800. doi: 10.1016/j.cell.2008.04.030 PMID: 18510924

14. Kuper J, Braun C, Elias A, Michels G, Sauer F, et al. (2014) In TFIIH, XPD helicase is exclusively devoted to DNA repair. PLoS Biol 12: e1001954. doi: 10.1371/journal.pbio.1001954 PMID: 25268380

15. Liu H, Rudolf J, Johnson KA, McMahon SA, Oke M, et al. (2008) Structure of the DNA repair helicase XPD. Cell 133: 801–812. doi: 10.1016/j.cell.2008.04.029 PMID: 18510925

16. Botta E, Nardo T, Broughton BC, Marinoni S, Lehmann AR, et al. (1998) Analysis of mutations in the XPD gene in Italian patients with trichothiodystrophy: site of mutation correlates with repair deficiency, but gene dosage appears to determine clinical severity. Am J Hum Genet 63: 1036–1048. PMID: 9758621

17. Viprakasit V, Gibbons RJ, Broughton BC, Tolmie JL, Brown D, et al. (2001) Mutations in the general transcription factor TFIIH result in beta-thalassaemia in individuals with trichothiodystrophy. Hum Mol Genet 10: 2797–2802. PMID: 11734544

18. Zhou X, Khan SG, Tamura D, Patronas NJ, Zein WM, et al. (2010) Brittle hair, developmental delay, neurologic abnormalities, and photosensitivity in a 4-year-old girl. J Am Acad Dermatol 63: 323–328. doi: 10.1016/j.jaad.2010.03.041 PMID: 20633800

19. Broughton BC, Berneburg M, Fawcett H, Taylor EM, Arlett CF, et al. (2001) Two individuals with features of both xeroderma pigmentosum and trichothiodystrophy highlight the complexity of the clinical outcomes of mutations in the XPD gene. Hum Mol Genet 10: 2539–2547. PMID: 11709541

20. Nishiwaki Y, Kobayashi N, Imoto K, Iwamoto TA, Yamamoto A, et al. (2004) Trichothiodystrophy fibroblasts are deficient in the repair of ultraviolet-induced cyclobutane pyrimidine dimers and (6–4)photoproducts. J Invest Dermatol 122: 526–532. PMID: 15009740

21. Cleaver JE, Thompson LH, Richardson AS, States JC (1999) A summary of mutations in the UV-sensitive disorders: xeroderma pigmentosum, Cockayne syndrome, and trichothiodystrophy. Hum Mutat 14: 9–22. PMID: 10447254

22. Taylor EM, Broughton BC, Botta E, Stefanini M, Sarasin A, et al. (1997) Xeroderma pigmentosum and trichothiodystrophy are associated with different mutations in the XPD (ERCC2) repair/transcription gene. Proc Natl Acad Sci U S A 94: 8658–8663. PMID: 9238033

23. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. (2008) Genes mirror geography within Europe. Nature 456: 98–101. doi: 10.1038/nature07331 PMID: 18758442

24. Pohlreich P, Zikan M, Stribrna J, Kleibl Z, Janatova M, et al. (2005) High proportion of recurrent germline mutations in the BRCA1 gene in breast and ovarian cancer patients from the Prague area. Breast Cancer Res 7: R728–736. PMID: 16168118

25. Lhota F, Zemankova P, Kleiblova P, Soukupova J, Vocka M, et al. (2016) Hereditary truncating mutations of DNA repair and other genes in BRCA1/BRCA2/PALB2-negatively tested breast cancer patients. Clin Genet.

26. Kleibl Z, Havranek O, Novotny J, Kleiblova P, Soucek P, et al. (2008) Analysis of CHEK2 FHA domain in Czech patients with sporadic breast cancer revealed distinct rare genetic alterations. Breast Cancer Res Treat 112: 159–164. PMID: 18058223

27. Berneburg M, Clingen PH, Harcourt SA, Lowe JE, Taylor EM, et al. (2000) The cancer-free phenotype in trichothiodystrophy is unrelated to its repair defect. Cancer Res 60: 431–438. PMID: 10667598

28. Kleijer WJ, Laugel V, Berneburg M, Nardo T, Fawcett H, et al. (2008) Incidence of DNA repair deficiency disorders in western Europe: Xeroderma pigmentosum, Cockayne syndrome and trichothiodystrophy. DNA Repair (Amst) 7: 744–750.

29. de Boer J, Donker I, de Wit J, Hoeijmakers JH, Weeda G (1998) Disruption of the mouse xeroderma pigmentosum group D DNA repair/basal transcription gene results in preimplantation lethality. Cancer Res 58: 89–94. PMID: 9426063

30. Lehmann AR (2001) The xeroderma pigmentosum group D (XPD) gene: one gene, two functions, three diseases. Genes Dev 15: 15–23. PMID: 11156600

31. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164. doi: 10.1093/nar/gkq603 PMID: 20601685

32. Protic-Sabljic M, Kraemer KH (1985) One pyrimidine dimer inactivates expression of a transfected gene in xeroderma pigmentosum cells. Proc Natl Acad Sci U S A 82: 6622–6626. PMID: 2995975

33. Schafer A, Schubert S, Gratchev A, Seebode C, Apel A, et al. (2013) Characterization of three XPG-defective patients identifies three missense mutations that impair repair and transcription. J Invest Dermatol 133: 1841–1849. doi: 10.1038/jid.2013.54 PMID: 23370536

34. Khobta A, Anderhub S, Kitsera N, Epe B (2010) Gene silencing induced by oxidative DNA base damage: association with local decrease of histone H4 acetylation in the promoter region. Nucleic Acids Res 38: 4285–4295. doi: 10.1093/nar/gkq170 PMID: 20338881

35. Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185: 862–864. PMID: 4843792

36. Arnold M, Raffler J, Pfeufer A, Suhre K, Kastenmuller G (2015) SNiPA: an interactive, genetic variant-centered annotation browser. Bioinformatics 31: 1334–1336. doi: 10.1093/bioinformatics/btu779 PMID: 25431330

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

# The c.657del5 variant in the *NBN* gene predisposes to pancreatic cancer

Marianna Borecka [a], Petra Zemankova [a], Filip Lhota [a], Jana Soukupova [a], Petra Kleiblova [a], Michal Vocka [b], Pavel Soucek [c], Ivana Ticha [d], Zdenek Kleibl [a], Marketa Janatova [a,*]

[a] Institute of Biochemistry and Experimental Oncology, First Faculty of Medicine, Charles University in Prague, Prague, Czech Republic
[b] Department of Oncology, First Faculty of Medicine, Charles University in Prague and General University Hospital in Prague, Prague, Czech Republic
[c] Department of Toxicogenomics, National Institute of Public Health, Prague, Czech Republic
[d] Department of Pathology, First Faculty of Medicine, Charles University in Prague and General University Hospital in Prague, Prague, Czech Republic

ARTICLE INFO

*Article history:*
Received 10 March 2016
Received in revised form 22 April 2016
Accepted 29 April 2016
Available online 2 May 2016

*Keywords:*
Pancreatic ductal adenocarcinoma
The *NBN* gene
Inherited risk

ABSTRACT

Pancreatic ductal adenocarcinoma (PDAC) is the sixth most frequent cancer type in the Czech Republic with a poor prognosis that could be improved by an early detection and subsequent surgical treatment combined with chemotherapy. Genetic factors play an important role in PDAC risk. We previously identified one PDAC patient harboring the Slavic founder deleterious mutation c.657del5 in the *NBN* gene, using a panel next-generation sequencing (NGS). A subsequent analysis of 241 unselected PDAC patients revealed other mutation carriers. The overall frequency of c.657del5 in unselected PDAC patients (5/241; 2.07%) significantly differed from that in non-cancer controls (2/915; 0.2%; P = 0.006). The result indicates that the *NBN* c.657del5 variant represents a novel PDAC-susceptibility allele increasing PDAC risk (OR = 9.7; 95% CI: 1.9 to 50.2). The increased risk of PDAC in follow-up recommendations for *NBN* mutation carriers should be considered if other studies also confirm an increased frequency of c.657del5 carriers in PDAC patients from other populations.

## 1. Introduction

Pancreatic ductal adenocarcinoma (PDAC) is the sixth most frequent cancer type (with an incidence of 19.6/100,000 persons in 2013) and the fifth most frequent cause of cancer death in the Czech Republic (www.svod.cz). The prognosis of PDAC is poor with a 5-year survival of 7% and a median survival of 6 months (Siegel et al., 2015). Early detection and subsequent surgical treatment combined with chemotherapy can improve the 5-year survival up to 40% (Nakao et al., 2006). While population screening is not rational due to the low PDAC incidence, the identification of high-risk individuals, who may benefit from the available screening methods, is desirable.

A genetic predisposition is the major endogenous risk factor of PDAC development, together with chronic pancreatitis and diabetes mellitus (Becker et al., 2014). It has been estimated that 5–10% of PDAC patients have a positive family PDAC history. The genetic basis of most familial PDAC cases has not been explained yet; however, several PDAC-susceptibility genes have been identified, including genes (*BRCA1*, *BRCA2*, *PALB2*, *MLH1*, *MSH2*, *MSH6*, *PMS2*, *STK11*, *APC*, *CDKN2A*) associated with hereditary cancer syndromes (reviewed in (Becker et al., 2014)). The protein products of numerous PDAC-susceptibility genes are directly involved in DNA repair and the DNA damage response. The most prevalent mutations have been identified in *BRCA2* (up to 6% of patients and increasing PDAC risk 3.5-fold (Couch et al., 2007)) and *PALB2* (3% of patients (Jones et al., 2009)). Their protein products share a common functional role in the DNA double-strand break (DDSB) repair. The *NBN* gene encodes nibrin, a protein participating in the formation of the multiprotein MRN (*MRE11-RAD50-NBN*) complex, an inevitable sensor of DNA damage in the DDSB repair (Carney et al., 1998). Biallelic *NBN* mutations predispose to the autosomal recessive Nijmegen-breakage syndrome (NBS) characterized by chromosomal instability and an increased risk of lymphoid malignancies and other cancers (Varon et al., 1998). Heterozygous *NBN* mutations predispose to breast cancer (BC) (Gorski et al., 2003), non-Hodgkin lymphoma (Steffen et al., 2006), and prostate cancer (Cybulski et al., 2013); however, their role in PDAC predisposition has not been studied yet. The most frequent pathogenic mutation in NBS patients and *NBN*-associated cancers is the recurrent Slavic founder mutation c.657del5 (c.657_661delACAAA) (Varon et al., 2000).

The next-generation sequencing (NGS) technology introduced analyses of large gene collections into genetic analyses in patients

*Abbreviation:* PDAC, pancreatic ductal adenocarcinoma; BRCA1, breast cancer gene 1; BRCA2, breast cancer gene 2; PALB2, partner and localizer of BRCA2; MLH1, homolog of *Escherichia coli* MutL 1; MSH2, homolog of *E. coli* MutS 2; MSH6, homolog of *E. coli* MutS 6; PMS2, postmeiotic segregation increased; STK11, serine/threonine protein kinase 11; APC, adenomatous polyposis coli gene; CDKN2A, cyclin-dependent kinase inhibitor 2A; DDSB, DNA double-strand break; NBN, nibrin; MRN, MRE11/RAD50/NBN complex; NBS, Nijmegen breakage syndrome; NGS, next-generation sequencing; FFPE, formalin-fixed paraffin-embedded; HRM, high resolution melting; FET, Fischer's exact test; OR, odds ratio; CI, confidence interval; BC, breast cancer; CHEK2, checkpoint kinase 2; PARP, poly(ADP-ribose) polymerase.

[*] Corresponding author at: Institute of Biochemistry and Experimental Oncology, First Faculty of Medicine, Charles University in Prague, U Nemocnice 5, 128 53 Prague 2, Czech Republic.

*E-mail address:* mjana@lf1.cuni.cz (M. Janatova).

http://dx.doi.org/10.1016/j.gene.2016.04.056
0378-1119/© 2016 Elsevier B.V. All rights reserved.

with cancer susceptibility. Among others, *NBN* is routinely analyzed in many cancer gene sequencing panels. Recently, we have performed a study of germline variants influencing the breast cancer susceptibility in high-risk breast cancer patients using the custom panel NGS (Lhota et al., 2016). We subsequently used the identical approach for the analysis of pancreatic cancer predisposition in a PDAC patient from multiple cancer family. We identified the c.657del5 germline mutation in the *NBN* gene in this patient. Therefore, we aimed to determine the frequency of c.657del5 in unselected Czech PDAC patients.

## 2. Materials and methods

### 2.1. Panel NGS analysis in a patient with pancreatic ductal adenocarcinoma

In order to identify possible germline pathogenic variant in PDAC-susceptibility genes, we performed custom panel NGS targeting 581 genes in a PDAC patient from multiple cancer family (Fig. 1). The NGS and bioinformatics analysis was performed as described previously (Lhota et al., 2016) and revealed germline c.657del5 *NBN* variant. The mutation was confirmed by Sanger sequencing from independent PCR amplified blood DNA sample. The presence of the c.657del5 *NBN* variant in deceased proband's sister with gastric cancer (Fig. 1) was analyzed in DNA isolated from FFPE tumor tissue using the Cobas DNA Sample Preparation Kit (Roche).

### 2.2. Patients with pancreatic ductal adenocarcinoma

We genotyped c.657del5 *NBN* variant in blood-isolated DNA samples from 241 unselected, histopathologically-verified PDAC patients, which included 152 samples from the National Institute of Public Health [median age at diagnosis: 63 years (ranged 40–82); 59 females] and 89 samples from the Department of Oncology, General University Hospital in Prague [median age at diagnosis: 64 years (ranged 38–84); 49 females]. Information about family history of cancer in c.657del5 carriers was gathered from medical records when available.

The control group included 915 non-cancer individuals and it had been described and genotyped previously. All patients and controls were of Slavic descent and of Czech origin. The study was approved by the local Ethical Committees and a written informed consent was obtained from all participants.

### 2.3. The NBN c.657del5 genotyping

The exon 6 of the *NBN* gene was analyzed by a high resolution melting (HRM; LightCycler 480; Roche) using HOT FirePol EvaGreen HRM Mix (Solis BioDyne). The primer sequences had been described previously (Mateju et al., 2012). The presence of c.657del5 was confirmed by sequencing.

### 2.4. Statistical analysis

The difference between groups was calculated using the Fisher exact test (FET).

## 3. Results

We analyzed a PDAC patient (diagnosed at 64 years) from multiple-cancer family and identified the c.657del5 *NBN* mutation using the panel NGS (Fig. 1). Except to this germline mutation, we found no other truncating variants in other known PDAC-susceptibility genes (*BRCA1*, *BRCA2*, *PALB2*, *MLH1*, *MSH2*, *MSH6*, *PMS2*, *STK11*, *APC*, *CDKN2A*). The presence of c.657del5 mutation was confirmed also in the proband's sister deceased from gastric cancer (Fig. 1).

In the subsequent analysis, we genotyped c.657del5 in other 241 unselected PDAC patients and found five mutation carriers among them (2.07%). Thus, the frequency of c.657del5 among PDAC patients was significantly higher than that in previously analyzed controls (2/915), suggesting that the carriers of c.657del5 have an increased risk of PDAC development (OR = 9.7; 95%CI: 1.9–50.2; $P_{FET}$ = 0.006). A PDAC family history was documented in none of the five c.657del5 carriers from 241 unselected PDAC patients; however, one patient had family cancer history (a sister with gastric cancer), and another female patient suffered from a duplicity of BC (at 46 years) and PDAC

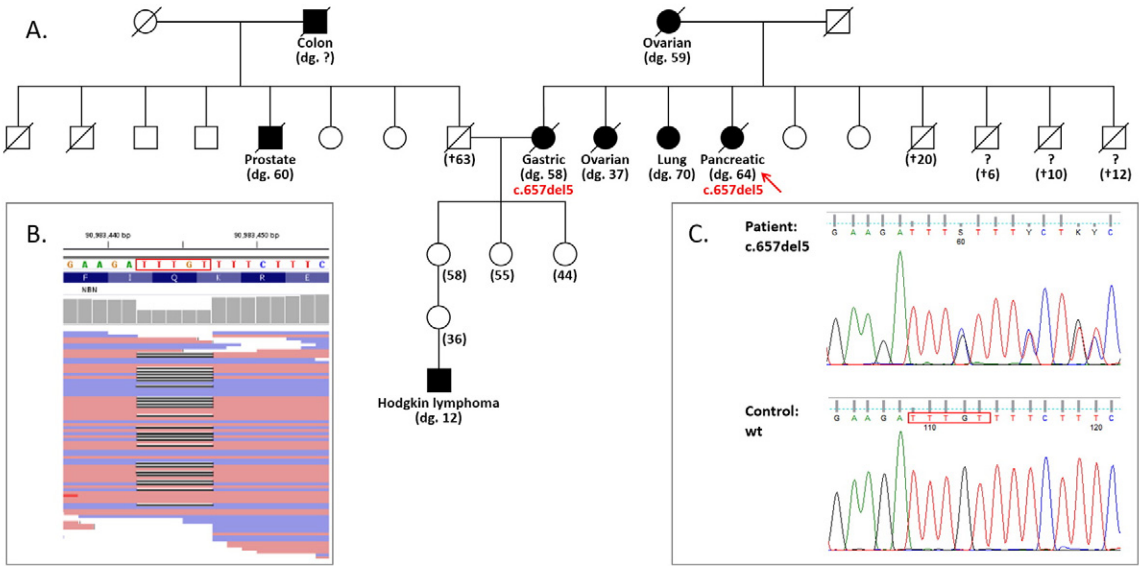


**Fig. 1.** Pedigree (A) of the multiple cancer family showing the proband with PDAC (indicated by an arrow) and her sister, both carrying c.657del5. DNA samples from other relatives were not available for genotyping. The ages of cancer diagnoses (dg.) or cessation (†) are indicated in the pedigree. The deletion of five nucleotides (TTTGT from reverse strand) is highlighted by a red frame in NGS analysis (B), confirmed by Sanger sequencing (C).

(at 64 years). The mean age at diagnosis for the c.657del5 carriers was 65.8 years (range 59–73).

## 4. Discussion

The highest frequency of *NBN* mutation carriers (up to 3.7% of patients) was found in BC patients from Central and Eastern Europe (Gorski et al., 2003). Recent meta-analysis indicated that c.657del5 is a moderate BC (OR = 2.51; 95%CI: 1.68–3.73) and lymphoma (OR = 2.93; 95%CI: 1.62–5.29) susceptibility allele, and that it also strongly increases the risk of prostate cancer (OR = 5.87; 95%CI: 2.51–13.75) (Gao et al., 2013). The association of the hereditary *NBN* mutations with BC susceptibility led to the inclusion of *NBN* into multigene cancer panel NGS analyses in high-risk individuals (Couch et al., 2015). Two studies have reported the results of hereditary mutation analysis performed by multi-gene panel testing in PDAC patients. While no truncating *NBN* mutation was identified in two previous studies of 290 and 638 patients, respectively (Grant et al., 2015; Roberts et al., 2016), Hu et al. found one c.657del5 carrier in 96 patients (also carrying the *CHEK2* mutation) (Hu et al., 2016). Recently, Lener et al. performed analysis of 10 prevalent founder mutations in *BRCA1*, *CHEK2*, *PALB2* and *NBN* (incl. c.657del5) in 383 pancreatic cancer patients and detected eight carriers of c.657del5 (2.09%), indicating the increased risk of pancreatic cancer in c.657del5 carriers (OR = 3.8; 95%CI: 1.68–8.60) in Poland (Lener et al., 2016).

The high frequencies of c.657del5 identified in PDAC patients in our and Lener et al. studies indicate that *NBN* is another DNA repair gene involved in PDAC-susceptibility. In comparison with our current study identifying 2.07% of c.657del5 carriers in unselected PDAC patients, earlier analyses found considerably lower frequencies of the mutation in Czech unselected BC (0.3%), colorectal cancer (0.3%), and lymphoma patients (0.8%) (Lhota et al., 2016; Pardini et al., 2009; Soucek et al., 2003). Our results and Lener et al. study (Lener et al., 2016) suggest that c.657del5 may be a novel PDAC-susceptibility allele significantly increasing the risk of PDAC development [combined OR calculated from this and Lener et al. studies comprising 624 pancreatic cancer patients (13 carriers of c.657del5) and 4915 controls (24 carriers) is 4.33; 95%CI 2.2–8.56; $p < 0.001$]. However, further studies in larger populations together with segregation analyses will be necessary to confirm our observation. They also may help specify the PDAC-associated risk more precisely, which is required for clinical management of the carriers and evaluation of c.657del5 as a putative predictive biomarker for therapy using DNA cross-linking agents or PARP inhibitors in carriers with PDAC (Schroder-Heurich et al., 2014).

Only the first patient identified in our preliminary NGS analysis had an indicative family cancer history (Fig. 1) and c.657del5 co-segregated with cancer diagnoses in the family. Its presence in the proband's sister with gastric cancer indicates that the carriers of c.657del5 may develop a broader spectrum of cancers. One in five carriers from the unselected PDAC group had a sister with gastric cancer (unfortunately, no DNA from this patient was available). The other mutation carriers displayed no family cancer history, just like the c.657del5 mutation carrier in the aforementioned report by Hu et al. (Hu et al., 2016). The similar mean age at PDAC diagnosis in carriers and non-carriers in our analysis (65.8 and 63.5 years, respectively) suggests that the c.657del5 mutation is not associated with an earlier disease onset.

In conclusion, our study suggests a novel role of the c.657del5 mutation in PDAC susceptibility. Future analyses of *NBN* in multi-gene cancer panels will help identify the hereditary pathogenic *NBN* mutations throughout the entire gene and enable a more accurate estimation of *NBN*-associated cancer risks.

## Conflict of interest

None.

## Acknowledgements

We would like to thank all patients who agreed to participate in this study, Marketa Dostalova and Marie Epsteinova for their excellent technical help. This study was supported by grants: IGA MZCR NT-14006-3/2013 and NT-14054-3/2013, PRVOUK-P27/LF1/1, and SVV-UK 260256/2016.

## References

Becker, A.E., Hernandez, Y.G., Frucht, H., Lucas, A.L., 2014. Pancreatic ductal adenocarcinoma: risk factors, screening, and early detection. World J. Gastroenterol. 20, 11182–11198.

Carney, J.P., Maser, R.S., Olivares, H., Davis, E.M., Le Beau, M., Yates 3rd, J.R., Hays, L., Morgan, W.F., Petrini, J.H., 1998. The hMre11/hRad50 protein complex and Nijmegen breakage syndrome: linkage of double-strand break repair to the cellular DNA damage response. Cell 93, 477–486.

Couch, F.J., Johnson, M.R., Rabe, K.G., Brune, K., de Andrade, M., Goggins, M., Rothenmund, H., Gallinger, S., Klein, A., Petersen, G.M., Hruban, R.H., 2007. The prevalence of BRCA2 mutations in familial pancreatic cancer. Cancer Epidemiol. Biomark. Prev. 16, 342–346.

Couch, F.J., Hart, S.N., Sharma, P., Toland, A.E., Wang, X., Miron, P., Olson, J.E., Godwin, A.K., Pankratz, V.S., Olswold, C., Slettedahl, S., Hallberg, E., Guidugli, L., Davila, J.I., Beckmann, M.W., Janni, W., Rack, B., Ekici, A.B., Slamon, D.J., Konstantopoulou, I., Fostira, F., Vratimos, A., Fountzilas, G., Pelttari, L.M., Tapper, W.J., Durcan, L., Cross, S.S., Pilarski, R., Shapiro, C.L., Klemp, J., Yao, S., Garber, J., Cox, A., Brauch, H., Ambrosone, C., Nevanlinna, H., Yannoukakos, D., Slager, S.L., Vachon, C.M., Eccles, D.M., Fasching, P.A., 2015. Inherited mutations in 17 breast cancer susceptibility genes among a large triple-negative breast cancer cohort unselected for family history of breast cancer. J. Clin. Oncol. 33, 304–311.

Cybulski, C., Wokolorczyk, D., Kluzniak, W., Jakubowska, A., Górski, B., Gronwald, J., Huzarski, T., Kashyap, A., Byrski, T., Dębniak, T., Gołąb, A., Gliniewicz, B., Sikorski, A., Switała, J., Borkowski, T., Borkowski, A., Antczak, A., Wojnar, L., Przybyła, J., Sosnowski, M., Małkiewicz, B., Zdrojowy, R., Sikorska-Radek, P., Matych, J., Wilkosz, J., Różański, W., Kiś, J., Bar, K., Bryniarski, P., Paradysz, A., Jersak, K., Niemirowicz, J., Słupski, P., Jarzemski, P., Skrzypczyk, M., Dobruch, J., Domagała, P., SA, N., Lubiński, J., 2013. Polish hereditary prostate cancer consortium. An inherited NBN mutation is associated with poor prognosis prostate cancer. Br. J. Cancer 108, 461–468.

Gao, P., Ma, N., Li, M., Tian, Q.B., Liu, D.W., 2013. Functional variants in NBS1 and cancer risk: evidence from a meta-analysis of 60 publications with 111 individual studies. Mutagenesis 28, 683–697.

Gorski, B., Debniak, T., Masojc, B., Mierzejewski, M., Medrek, K., Cybulski, C., Jakubowska, A., Kurzawski, G., Chosia, M., Scott, R., Lubiński, J., 2003. Germline 657del5 mutation in the NBS1 gene in breast cancer patients. Int. J. Cancer 106, 379–381.

Grant, R.C., Selander, I., Connor, A.A., Selvarajah, S., Borgida, A., Briollais, L., Petersen, G.M., Lerner-Ellis, J., Holter, S., Gallinger, S., 2015. Prevalence of germline mutations in cancer predisposition genes in patients with pancreatic cancer. Gastroenterology 148, 556–564.

Hu, C., Hart, S.N., Bamlet, W.R., Moore, R.M., Nandakumar, K., Eckloff, B.W., Lee, Y.K., Petersen, G.M., McWilliams, R.R., Couch, F.J., 2016. Prevalence of pathogenic mutations in cancer predisposition genes among pancreatic cancer patients. Cancer Epidemiol. Biomark. Prev. 25, 207–211.

Jones, S., Hruban, R.H., Kamiyama, M., Borges, M., Zhang, X., Parsons, D.W., Lin, J.C., Palmisano, E., Brune, K., Jaffee, E.M., Iacobuzio-Donahue, C.A., Maitra, A., Parmigiani, G., Kern, S.E., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., Eshleman, J.R., Goggins, M., Klein, A.P., 2009. Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene. Science 324, 217.

Lener, M.R., Scott, R.J., Kluźniak, W., Baszuk, P., Cybulski, C., Wiechowska-Kozłowska, A., Huzarski, T., Byrski, T., Kładny, J., Pietrzak, S., Soluch, A., Jakubowska, A., Lubiński, J., 2016. Do founder mutations characteristic of some cancer sites also predispose to pancreatic cancer? Int. J. Cancer http://dx.doi.org/10.1002/ijc.30116.

Lhota, F., Zemankova, P., Kleiblova, P., Soukupova, J., Vocka, M., Stranecky, V., Janatova, M., Hartmannova, H., Hodanova, K., Kmoch, S., Kleibl, Z., 2016. Hereditary truncating mutations of DNA repair and other genes in BRCA1/BRCA2/PALB2-negatively tested breast cancer patients. Clin. Genet. http://dx.doi.org/10.1111/cge.12748.

Mateju, M., Kleiblova, P., Kleibl, Z., Janatova, M., Soukupova, J., Ticha, I., Novotny, J., Pohlreich, P., 2012. Germline mutations 657del5 and 643C > T (R215W) in NBN are not likely to be associated with increased risk of breast cancer in Czech women. Breast Cancer Res. Treat. 133, 809–811.

Nakao, A., Fujii, T., Sugimoto, H., Kanazumi, N., Nomoto, S., Kodera, Y., Inoue, S., Takeda, S., 2006. Oncological problems in pancreatic cancer surgery. World J. Gastroenterol. 12, 4466–4472.

Pardini, B., Naccarati, A., Polakova, V., Smerhovsky, Z., Hlavata, I., Soucek, P., Novotny, J., Vodickova, L., Tomanova, V., Landi, S., Vodicka, P., 2009. NBN 657del5 heterozygous mutations and colorectal cancer risk in the Czech Republic. Mutat. Res. 666, 64–67.

Roberts, N.J., Norris, A.L., Petersen, G.M., Bondy, M.L., Brand, R., Gallinger, S., Kurtz, R.C., Olson, S.H., Rustgi, A.K., Schwartz, A.G., Stoffel, E., Syngal, S., Zogopoulos, G., Ali, S.Z., Axilbund, J., Chaffee, K.G., Chen, Y.C., Cote, M.L., Childs, E.J., Douville, C., Goes, F.S., Herman, J.M., Iacobuzio-Donahue, C., Kramer, M., Makohon-Moore, A., McCombie, R.W., McMahon, K.W., Niknafs, N., Parla, J., Pirooznia, M., Potash, J.B., Rhim, A.D., Smith, A.L., Wang, Y., Wolfgang, C.L., Wood, L.D., Zandi, P.P., Goggins, M., Karchin, R., Eshleman, J.R., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., Hruban, R.H., Klein, A.P., 2016. Whole genome sequencing defines the genetic heterogeneity of familial pancreatic cancer. Cancer Discov. 6, 166–175.

Schroder-Heurich, B., Bogdanova, N., Wieland, B., Xie, X., Noskowicz, M., Park-Simon, T.W., Hillemanns, P., Christiansen, H., Dörk, T., 2014. Functional deficiency of NBN, the Nijmegen breakage syndrome protein, in a p.R215W mutant breast cancer cell line. BMC Cancer 14, 434.

Siegel, R.L., Miller, K.D., Jemal, A., 2015. Cancer statistics, 2015. CA Cancer J. Clin. 65, 5–29.

Soucek, P., Gut, I., Trneny, M., Skovlund, E., Grenaker Alnaes, G., Kristensen, T., Børresen-Dale, A.L., Kristensen, V.N., 2003. Multiplex single-tube screening for mutations in the Nijmegen breakage syndrome (NBS1) gene in Hodgkin's and non-Hodgkin's lymphoma patients of Slavic origin. Eur. J. Hum. Genet. 11, 416–419.

Steffen, J., Maneva, G., Poplawska, L., Varon, R., Mioduszewska, O., Sperling, K., 2006. Increased risk of gastrointestinal lymphoma in carriers of the 657del5 NBS1 gene mutation. Int. J. Cancer 119, 2970–2973.

Varon, R., Vissinga, C., Platzer, M., Cerosaletti, K.M., Chrzanowska, K.H., Saar, K., Beckmann, G., Seemanova, E., Cooper, P.R., Nowak, N.J., Stumm, M., Weemaes, C.M., Gatti, R.A., Wilson, R.K., Digweed, M., Rosenthal, A., Sperling, K., Concannon, P., Reis, A., 1998. Nibrin, a novel DNA double-strand break repair protein, is mutated in Nijmegen breakage syndrome. Cell 93, 467–476.

Varon, R., Seemanova, E., Chrzanowska, K., Hnateyko, O., Piekutowska-Abramczuk, D., Krajewska-Walasek, M., Sykut-Cegielska, J., Sperling, K., Reis, A., 2000. Clinical ascertainment of Nijmegen breakage syndrome (NBS) and prevalence of the major mutation, 657del5, in three Slav populations. Eur. J. Hum. Genet. 8, 900–902.

# CZECANCA: CZEch CAncer paNel for Clinical Application – návrh a příprava cíleného sekvenačního panelu pro identifikaci nádorové predispozice u rizikových osob v České republice

# CZECANCA: CZEch CAncer paNel for Clinical Application – Design and Optimization of the Targeted Sequencing Panel for the Identification of Cancer Susceptibility in High-risk Individuals from the Czech Republic

Soukupová J.[1], Zemánková P.[1], Kleiblová P.[1,2], Janatová M.[1], Kleibl Z.[1]

[1] Ústav biochemie a experimentální onkologie, 1. LF UK v Praze
[2] Ústav biologie a lékařské genetiky, 1. LF UK a VFN v Praze

## Souhrn

Dědičná nádorová onemocnění tvoří malou, ale klinicky významnou část onkologických onemocnění, v České republice se jedná ročně o několik tisíc osob. Identifikace kauzální mutace v nádorových predispozičních genech má u těchto nemocných zásadní prognostický a v některých případech i prediktivní význam. Mimo to je podmínkou cílené preventivní péče o asymptomatické nosiče mutací v rodinách se zvýšeným rizikem vzniku nádorového onemocnění. Do současné doby bylo charakterizováno více než 150 nádorových predispozičních genů. Mutace většiny z nich se vyskytují vzácně, s výraznou populační specifičností a jejich klinická interpretace je často obtížná. Diagnostiku raritních variant technicky zjednodušují postupy využívající sekvenování nové generace, které umožňují vyšetření rozsáhlých sad genů. Za účelem racionalizace diagnostiky hereditárních nádorových syndromů v České republice jsme navrhli sekvenační panel „CZECANCA“, který cílí na vyšetření 219 genů asociovaných s dědičnými nádorovými onemocněními. Panel obsahuje přes 50 klinicky významných genů vysokého a středního rizika, zbývající geny tvoří málo prozkoumané a kandidátní predispoziční geny, jejichž vrozené mutace mají nejasnou klinickou interpretaci. Společně s návrhem panelu byl optimalizován postup vlastního sekvenování a bioinformatického zpracování sekvenačních dat pro tvorbu jednotné databáze genotypů analyzovaných vzorků. Cílem projektu je nabídnout použití sekvenačního panelu včetně optimalizovaného postupu sekvenování nové generace diagnostickým laboratořím v České republice a zajistit sdílení genotypů a klinických údajů o vyšetřovaných pacientech ve společné databázi za účelem zlepšení možnosti klinické interpretace vzácných mutací u vysoce rizikových osob.

## Klíčová slova

analýza genetické predispozice – dědičné nádorové syndromy – masivní paralelní sekvenování – databáze genetických informací – panelové sekvenování – cílené sekvenování – sekvenování nové generace (NGS)

Práce byla realizována za podpory Interní grantové agentury MZ ČR (IGA MZ ČR) pod grantovým číslem NT14054, Agentury pro zdravotnický výzkum ČR (AZV ČR) pod čísly NV15-28830A a NV15-27695A a Ligy proti rakovině Praha.

This work was supported by Czech Ministry of Health grants No. NT14054, NV15-28830A, NV15-27695A and The League Against Cancer Prague.

Autoři deklarují, že v souvislosti s předmětem studie nemají žádné komerční zájmy.

The authors declare they have no potential conflicts of interest concerning drugs, products, or services used in the study.

Redakční rada potvrzuje, že rukopis práce splnil ICMJE kritéria pro publikace zasílané do biomedicínských časopisů.

The Editorial Board declares that the manuscript met the ICMJE recommendation for biomedical papers.

doc. MUDr. Zdeněk Kleibl, Ph.D.
Ústav biochemie a experimentální onkologie
1. LF UK v Praze
U Nemocnice 5
128 53 Praha 2
e-mail: zdekleje@lf1.cuni.cz

Obdrženo/Submitted: 2. 10. 2015
Přijato/Accepted: 13. 10. 2015

http://dx.doi.org/10.14735/amko2016S46

## Summary

Individuals with hereditary cancer syndromes form a minor but clinically important subgroup of oncology patients, comprising several thousand cases in the Czech Republic annually. In these patients, the identification of pathogenic mutations in cancer susceptibility genes has an important predictive and, in some cases, prognostic value. It also enables rational preventive strategies in asymptomatic carriers from affected families. More than 150 cancer susceptibility genes have been described so far; however, mutations in most of them are very rare, occurring with substantial population variability, and hence their clinical interpretation is very complicated. Diagnostics of mutations in cancer susceptibility genes have benefited from the broad availability of next-generation sequencing analyses using targeted gene panels. In order to rationalize the diagnostics of hereditary cancer syndromes in the Czech Republic, we have prepared the sequence capture panel "CZECANCA", targeting 219 cancer susceptibility genes. Besides more than 50 clinically important high- and moderate-penetrance susceptibility genes, the panel also targets less common candidate genes with uncertain clinical relevance. Alongside the panel design, we have optimized the analytical and bioinformatics pipeline, which will facilitate establishing a collective nationwide database of genotypes and clinical data from the analyzed individuals. The key objective of this project is to provide diagnostic laboratories in the Czech Republic with a reliable procedure and collective database improving the clinical utility of next-generation sequencing analyses in high-risk patients, which would help improve the interpretation of rare or population-specific variants in cancer susceptibility genes.

## Key words

genetic predisposition testing – hereditary cancer syndromes – high-throughput nucleotide sequencing – genetic information databases – panel sequencing – sequence capture – next-generation sequencing (NGS)

## Úvod

Nemocní s dědičnými nádorovými onemocněními zaujímají malou (obvykle mezi 5 a 10 % všech případů), ale klinicky významnou část onkologických pacientů. S ohledem na celkovou incidenci onkologických onemocnění v ČR se tak jedná o několik tisíc vysoce rizikových pacientů ročně. Plošné testování na přítomnost nádorových predispozičních variant u všech onkologicky nemocných je v současnosti ekonomicky neúnosné, proto je vyšetření nádorové predispozice omezeno na vybrané skupiny pacientů na základě charakteristických znaků, které jsou rozvedeny u jednotlivých diagnóz v tomto supplementu. Hlavním rysem dědičných nádorových onemocnění je zvýšené (a často velmi vysoké) riziko vzniku nádorového onemocnění u nosičů patogenních mutací v postižených rodinách. Z tohoto důvodu je identifikace příčinných mutací v nádorových predispozičních genech u rizikových osob předpokladem účinné strategie léčebné péče, která může zahrnovat širokou škálu terapeutických a preventivních modalit snižujících výskyt či zlepšujících prognózu nádorových onemocnění.

Diagnostika dědičných nádorových onemocnění je jedním ze základních cílů současné onkogenetiky a jejích výstupů do klinické praxe. Nejprostudovanější jsou geny, jejichž mutace způsobují hereditární nádorové syndromy s vysokým rizikem vzniku onkologického onemocnění (např. *TP53* u Li-Fraumeni syndromu, *BRCA1* a *BRCA2* u syndromu hereditárního karcinomu prsu a ovarií či *APC* u familiární adenomatózní polypózy). Jejich genetickým podkladem jsou převážně monoalelické patogenní mutace ve vysoce penetrantních genech. Mutace v těchto genech však objasňují malou část geneticky podmíněných častých nádorových onemocnění [1]. Zbytek případů připadá na mutace v desítkách až stovkách dalších genů, z nichž jen některé jsou dnes dobře charakterizovány, a případy hereditárního onemocnění na předpokládaném podkladě polygenní dědičnosti [2]. V porovnání s mutacemi v hlavních predispozičních genech se patogenní varianty v těchto dalších predispozičních genech vyznačují nižší penetrancí [3,4], nádorový tropizmus u postižení konkrétního genu je méně vyhraněný [5], vyskytují se s významně nižší populační frekvencí a tato frekvence je mnohdy významně proměnlivá v jednotlivých populacích a etnikách [6]. Identifikace nosičů patogenních variant mimo oblast „klasických", vysoce penetrantních genů je tak tradičními genetickými postupy analyzujícími jednotlivé geny velmi nákladná a zdlouhavá.

Nástup moderních a výkonných molekulárně biologických technologií posledních let vede k významnému zrychlení identifikace nových predispozičních genů a genetických variant. V současné době jsou známy stovky genů, jejichž dědičné mutace prokazatelně či pravděpodobně zvyšují riziko vzniku nádorových onemocnění. Skutečnou revoluci do preklinického výzkumu, ale i klinické diagnostiky, přinesl nástup sekvenování nové generace (next-generation sequencing – NGS). Flexibilita této metody spočívá v masivním sekvenačním paralelizmu, kde v rámci jednoho sekvenačního běhu je možné analyzovat statisíce až miliony templátových molekul DNA. V rámci konkrétních genetických aplikací je možné tento paralelizmus využít pro sekvenování unikátních DNA templátů, v rámci např. celého genomu, nebo sekvenování vybraných úseků DNA u mnoha různých probandů, jako je tomu u panelového NGS. Pomocí NGS je možné v poměrně krátké době najednou identifikovat genetické varianty ve stovkách genů u desítek probandů s ekonomickými náklady nesrovnatelně nižšími, než by tomu bylo při analýze jednotlivých genů klasickými postupy molekulární biologie zahrnujícími prescreening mutací (DGGE/HA/ /DHPLC/HRMA/RFLP/PTT) s následnou charakterizací patogenní varianty Sangerovým sekvenováním [7]. Ve srovnání s klasickými analýzami však NGS vyžaduje specifické přístrojové vybavení a laboratorní přístupy, významné zvýšení nároků na bioanalytické zpracování a vysoké nároky na klinické hodnocení identifikovaných genetických variant.

V rámci předchozí studie zahrnující NGS a cílené na panel 581 genů jsme v souboru 325 *BRCA1/BRCA2/PALB2* negativních pacientek s karcinomem prsu nalezli 127 variant způsobujících zkrácení proteinového produktu v některém ze 73 genů u téměř třetiny vyšetřovaných pacientek [8]. Tato analýza, stejně jako další publikované výstupy NGS, prokázala vysoký výkon, spolehlivost a robustnost cíleného NGS [9]. Proto jsme se rozhodli připravit panel genů – CZECANCA (CZEch CAncer paNel for Clinical Application), který by umožnil komplexní, rentabilní a rychlou analýzu germinálních mutací v hlavních predispozičních genech, ale i kandidátních genech asociovaných se zvýšeným rizikem vzniku nejčastějších solidních nádorů v populaci vysoce rizikových pacientů v ČR.

## Koncepce projektu CZECANCA

Sekvenační projekt CZECANCA předpokládá použití panelu CZECANCA pro cílené obohacení (sequence capture) sekvenovaných oblastí DNA zahrnujících především kódující exony a intron-exonové přechody genů, jejichž hereditární varianty byly asociovány se zvýšeným rizikem vzniku nádorových onemocnění u jejich nosičů. Protože se s výjimkou vysoce penetrantních genů jedná obvykle o velmi málo frekventní varianty s předpokládanou výraznou populačně-specifickou variabilitou, je jejich správné klinické zhodnocení a klinická interpretace často velmi obtížná [10]. Zapojení řady klinických pracovišť využívajících jednotný technologický přístup založený na využití panelu CZECANCA umožní získání reprezentativního počtu genotypů u vysoce rizikových osob s různými nádorovými syndromy. Z důvodů minimalizace variability (a tím vznikajících technických chyb ve společné databázi) budou hrubá sekvenační data analyzována na našem pracovišti jednotnou bioinformatickou procedurou (pipeline). Tato data spolu se základními charakteristikami (fenotypem) sekvenovaných osob budou ukládána do jednotné databáze přístupné všem zúčastněným laboratořím. Sdílená databáze nebude obsahovat žádné údaje o vyšetřovaných osobách umožňující jejich identifikaci.

Projekt CZECANCA tak není omezen pouze na sekvenační panel, ale reprezentuje komplexní řešení analýzy nádorové predispozice za účelem zvýšení efektivity klinické interpretace variant nejasného významu a variant genů s nejasným rizikem (schéma 1). O výstupech z projektu CZECANCA bude pravidelně informována odborná veřejnost tak, aby složení sekvenačního panelu i postupy vyšetření odpovídaly aktuálnímu stavu vědeckých poznatků onkogenomiky, NGS a klinických požadavků. Aktuální informace budou dostupné na stránce www.czecanca.cz.

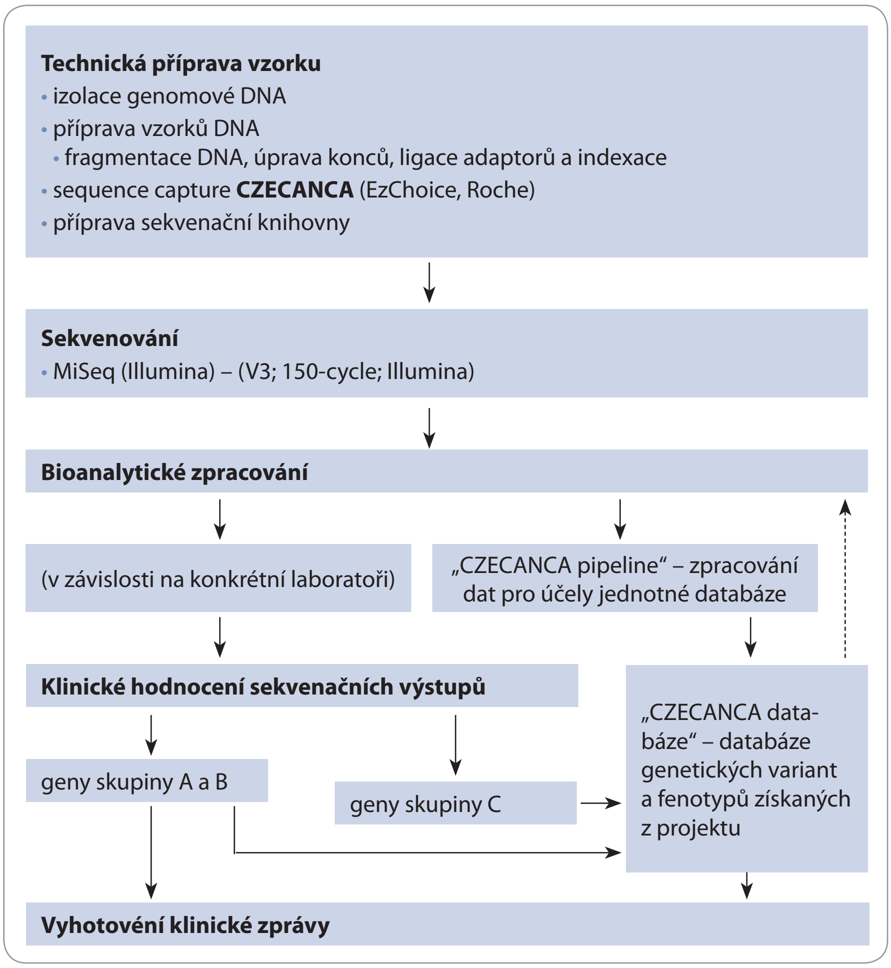

Technická příprava vzorku
• izolace genomové DNA
• příprava vzorků DNA
• fragmentace DNA, úprava konců, ligace adaptorů a indexace
• sequence capture CZECANCA (EzChoice, Roche)
• příprava sekvenační knihovny
Sekvenování
• MiSeq (Illumina) – (V3; 150-cycle; Illumina)
Bioanalytické zpracování
(v závislosti na konkrétní laboratoři)
„CZECANCA pipeline" – zpracování dat pro účely jednotné databáze
Klinické hodnocení sekvenačních výstupů
geny skupiny A a B
geny skupiny C
„CZECANCA databáze" – databáze genetických variant a fenotypů získaných z projektu
Vyhotovení klinické zprávy

**Schéma 1. Schematické znázornění postupů sekvenování a hodnocení sekvenačních výstupů v projektu CZECANCA (bližší vysvětlení v textu).**

Poznámka k výpočtu sekvenačního výstupu: Při velikosti cílové sekvence panelu CZECANCA (~ 600 kb), cílovém sekvenačním pokrytí 200krát, je pro analýzu jednoho vzorku DNA unikátního pacienta zapotřebí kapacity 120 Mb. S uvažovanou (dolní) mezí sekvenační kapacity chemie V3 (150-cycles), která činí ~ 4 Gb, lze teoreticky analyzovat vzorky 33 unikátních pacientů.

## Charakterizace panelu CZECANCA

Sekvenační panel CZECANCA je konstruován na bázi technologie SeqCap EZ choice (Nimblegen/Roche). Výběr genů zohledňoval četnost různých onkologických diagnóz v ČR, aktuální stav informací o genetické podstatě hereditárních nádorových syndromů, předpoklady pro identifikaci dalších kandidátních genů, ale i technické možnosti pro účinný a spolehlivý způsob cíleného obohacení pro účely panelového NGS. Z technických důvodů, které směřovaly k omezení oblastí genomu s výskytem neunikátních sekvencí (pseudogenů a repetitivních sekvencí), byly z návrhu stávající verze panelu CZECANCA vědomě vynechány některé známé predispoziční geny (např. *DIS3L2, DMBT1, PMS2, SBDS, SDHA, SDHC, SDHD*) nebo jejich neunikátní části (*CHEK2, NF1*). V současné podobě (verze 1.0) obsahuje panel

CZECANCA sondy cílící na kódující sekvence 219 genů (628 169 b).

Cílené geny jsou, z hlediska klinické významnosti pro účely diagnostiky nádorových predispozičních genů, rozděleny do tří skupin:

### Geny skupiny A

- **Prokázaná jasná (nezpochybnitelná) asociace s nádorovou hereditou.**
- **Hlavní predispoziční geny.**
- **Známé a významně zvýšené relativní riziko pro nosiče (RR > 5).**
- Klinicky relevantní alterace genů se referují v plném rozsahu klinickému genetikovi (tzn. patogenní mutace a VUS třídy 3–5 dle IARC).
- Konsenzuální klinická doporučení pro sledování nosičů mutací jsou jasně definována.
- Vyšetření příbuzných nosičů mutací se provádí z důvodu predikce.

### Geny skupiny B

- **Prokázána jasná asociace s nádorovou hereditou (evidentní na základě několika publikovaných studií).**
- **Geny/alely s (předpokládanou) střední penetrancí.**
- **Předpokládané významně zvýšené relativní riziko pro nosiče (RR 2–5).**
- **Asociace s nádorovou hereditou je evidentní, ale RR není přesně stanoveno (asociace s nádorovými hereditami je zjevná, ale není dostatek studií/nosičů mutací pro korektní stanovení RR).**
- Klinicky relevantní alterace genů se referují v plném rozsahu klinickému genetikovi (tzn. patogenní mutace a VUS třídy 4 nebo 5 dle IARC).
- Klinická doporučení pro sledování nosičů mutací nejsou jasně definována.
- Nosiči mutací jsou požádáni o spolupráci při vyšetření příbuzných pro stanovení segregace.
- V rodinách je provedena segregační analýza identifikované varianty.
- Interpretace výsledků prediktivního testování (pokud je prováděno) je následující:
  a) Pozitivně prediktivně testované jedince lze zařadit do preventivních sledovacích programů definovaných pro daný hereditární syndrom (pokud takový program existuje; je nutno vést v patrnosti případné modifikace těchto sledovacích programů). Preventivní chirurgické zákroky nejsou pouze na základě nosičství patogenní mutace v těchto genech indikovány, ale jsou ke zvážení při indikativní rodinné anamnéze.
  b) Negativně prediktivně testované jedince dále sledovat, zatím jen dle empirického rizika plynoucího z osobní a rodinné anamnézy.

### Geny skupiny C

- **Nejasná, avšak předpokládaná asociace s nádorovými hereditami. Informace o nádorové predispozici přinášejí pouze ojedinělé studie nebo preklinická data.**
- **Informace o klinickém významu genu pro nádorovou predispozici není známa, avšak produkt genu je zapojen v signální dráze, ve které poruchy v jiných genech (kódujících kooperující proteiny) prokazatelně souvisejí s nádorovou predispozicí.**
- **Klinická doporučení pro sledování nosičů mutací neexistují.**
- Alterace genů se nereferují a slouží pro vyhodnocení podílu variant sledovaných genů na vzniku nádorové predispozice u nemocných v ČR.
- Po vyhodnocení kolektovaných údajů mohou být nosiči kandidátních patogenních mutací požádáni o spolupráci při vyšetření příbuzných pro stanovení segregace v případě, že u probanda s indikativní rodinnou anamnézou nebyly nalezeny pravděpodobné patogenní mutace ve skupině A a B.

V tab. 1 jsou uvedeny základní charakteristiky genů ze skupiny A a B, které tvoří geny, jejichž patogenní mutace se prokazatelně podílejí na zvýšeném riziku vzniku nádorů u nosičů. Skupiny A a B odlišuje především existence klinických doporučení pro péči o nosiče mutací (ve skupině A).

Kromě genů ze skupiny A a B obsahuje panel CZECANCA i skupinu C, která zahrnuje geny, jejichž asociace s nádorovými onemocněními je mnohem méně známá (geny jsou vedeny v abecedním pořadí): *AIP, ALK, APEX1, ATMIN, ATR, ATRIP, AURKA, AXIN1, BABAM1, BRAP, BRCC3, BRE, BUB1B, C11ORF30, C19ORF40, CASP8, CCND1, CDC73, CDKN1B, CDKN1C, CEBPA, CEP57, CLSPN, CSNK1D, CSNK1E, CWF19L2, CYLD, DCLRE1C, DDB2, DHFR, DICER1, DMC1, DNAJC21, DPYD, EGFR, EPHX1, ERCC1, ERCC4, ERCC5, ERCC6, ESR1, ESR2, EXO1, EXT1, EXT2, EYA2, EZH2, FAM175A, FAM175B, FAN1, FANCA, FANCB, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, FBXW7, GADD45A, GATA2, GPC3, GRB7, HELQ, HNF1A, HOXB13, HRAS, HUS1, CHEK1, KAT5, KCNJ5, LIG1, LIG3, LIG4, LMO1, LRIG1, MAX, MCPH1, MDC1, MDM2, MDM4, MGMT, MMP8, MPL, MRE11A, MSH3, MSH5, MSR1, MUS81, NAT1, NCAM1, NELFB, NFKBIZ, NHEJ1, NSD1, OGG1, PARP1, PCNA, PHB, PHOX2B, PIK3CG, PLA2G2A, PMS1, POLB, PPM1D, PREX2, PRF1, PRKDC, PTTG2, RAD1, RAD17, RAD18, RAD23B, RAD50, RAD51, RAD51AP1, RAD51B, RAD52, RAD54B, RAD54L, RAD9A, RBBP8, RECQL5, RFC1, RFC2, RFC4, RHBDF2, RNF146, RNF168, RNF8, RPA1, RUNX1, SDHAF2, SETBP1, SETX, SHPRH, SMARCA4, SMARCE1, TCL1A, TELO2, TERF2, TERT, TLR2, TLR4, TMEM127, TOPBP1, TP53BP1, TSHR, UBE2A, UBE2B, UBE2I, UBE2V2, UBE4B, UIMC1, XPA, XPC, XRCC1, XRCC2, XRCC3, XRCC4, XRCC5, XRCC6, ZNF350, ZNF365*. Vyšetření genů ze skupiny C je nezbytné pro získání informace, zda jejich patogenní mutace mohou vysvětlovat zvýšenou četnost vzniku nádorových onemocnění u jejich nosičů v ČR. Tento význam bude analyzován při dosažení dostatečného počtu vyšetřených osob v databázi projektu.

## Sekvenování s panelem CZECANCA

Primární optimalizace sekvenování s panelem CZECANCA probíhala za využití sekvenátoru MiSeq (Illumina), avšak principiálně lze obohacenou knihovnu pravděpodobně použít na libovolné sekvenační platformě současné generace. Použití v současnosti nejrozšířenějšího přístroje firmy Illumina zjednodušuje následné bioinformatické analýzy a snižuje variabilitu výskytu technických sekvenačních artefaktů.

Vstupním materiálem pro přípravu knihovny obohacené o cílové úseky genomové DNA z panelu CZECANCA je fragmentovaná DNA. Fragmentaci lze

**Tab. 1. Přehled základních charakteristik 54 genů zařazených do skupiny A nebo B s charakterizací funkcí kódovaných proteinů a asociací nádorových onemocnění v příslušných lokalizacích spojených s nosičstvím dědičných patogenních variant.**

| Skupina | Gen | OMIM | Oficiální název | Základní funkce proteinu | Prs | Ovarium | Děloha | Kolon a rektum | Žaludek | Slinivka | Mozek | Kůže (a névy) | Ledvina | Prostata | Endokrinní tkáně | Sarkomy | Leukemie/lymfomy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | *APC* | 611731 | adenomatous polyposis coli | IC signalizace: Wnt | | | | ŽM | | | ŽM | | ŽM | | ŽM | | |
| B | *ATM* | 607585 | ataxia-telangiectasia mutated gene | reparace DNA | Ž | | | ŽM | | ŽM | | | | | | | ŽM |
| A | *BAP1* | 603089 | BRCA1 associated protein-1 | reparace DNA | Ž? | | | | | | | ŽM | ŽM | | | | |
| B | *BARD1* | 601593 | BRCA1 associated RING domain 1 | reparace DNA | Ž | | | | | | | | | | | | |
| B | *BLM* | 604610 | Bloom syndrome; *RECQL3* | reparace DNA | Ž | | | ŽM | | | | | | | | | ŽM |
| A | *BMPR1A* | 601299 | bone morphogenetic protein receptor, type IA | IC signalizace: TGF-β | | | | ŽM | ŽM | | | | | | | | |
| A | *BRCA1* | 113705 | breast cancer 1 | reparace DNA | ŽM | Ž | Ž | | | ŽM | | | | M | | | |
| A | *BRCA2* | 600185 | breast cancer 2 | reparace DNA | ŽM | Ž | Ž | | | ŽM | ŽM | ŽM | ŽM | M | | | ŽM |
| B | *BRIP1* | 605882 | BRCA1 interacting protein C-terminal helicase 1 | reparace DNA | Ž | Ž | | | | | | | | | | | |
| A | *CDH1* | 192090 | cadherin 1, type 1, E-cadherin (epithelial) | IC signalizace: Wnt | Ž | | Ž | | ŽM | | | | | M | | | |
| A | *CDK4* | 123829 | cyclin-dependent kinase 4 | runěčný cyklus | | | | | | | | ŽM | | | | | |
| A | *CDKN2A* | 600160 | cyclin-dependent kinase inhibitor 2A; *p16(INK4A)* | runěčný cyklus | | | | | | ŽM | | ŽM | | | | | |
| A | *EPCAM* | 185535 | epithelial cellular adhesion molecule | mezibuněčná signalizace | | | | ŽM | | | | | | | | | |
| B | *ERCC2* | 126340 | excision repair cross-complementation group 1 | reparace DNA | | | | | | | | ŽM | | | | | |
| B | *ERCC3* | 133510 | excision repair cross-complementation group 3 | reparace DNA | | | | | | | | ŽM | | | | | |
| B | *FANCC* | 613899 | Fanconi anemia, complementation group C | reparace DNA | Ž | | | | | | | | | | | | ŽM |
| B | *FANCM* | 609644 | Fanconi anemia, complementation group M; *(FAAP250)* | reparace DNA | Ž | | | | | | | | | | | | |
| A | *FH* | 136850 | fumarate hydratase | metabolizmus živin | | | | | | | | | ŽM | | | | |
| A | *FLCN* | 607273 | folliculin | IC signalizace: ? | | | | | | | | | ŽM | | | | |

**Tab. 1 – pokračování. Přehled základních charakteristik 54 genů zařazených do skupiny A nebo B s charakterizací funkcí kódovaných proteinů a asociací nádorových onemocnění v příslušných lokalizacích spojených s nosičstvím dědičných patogenních variant.**

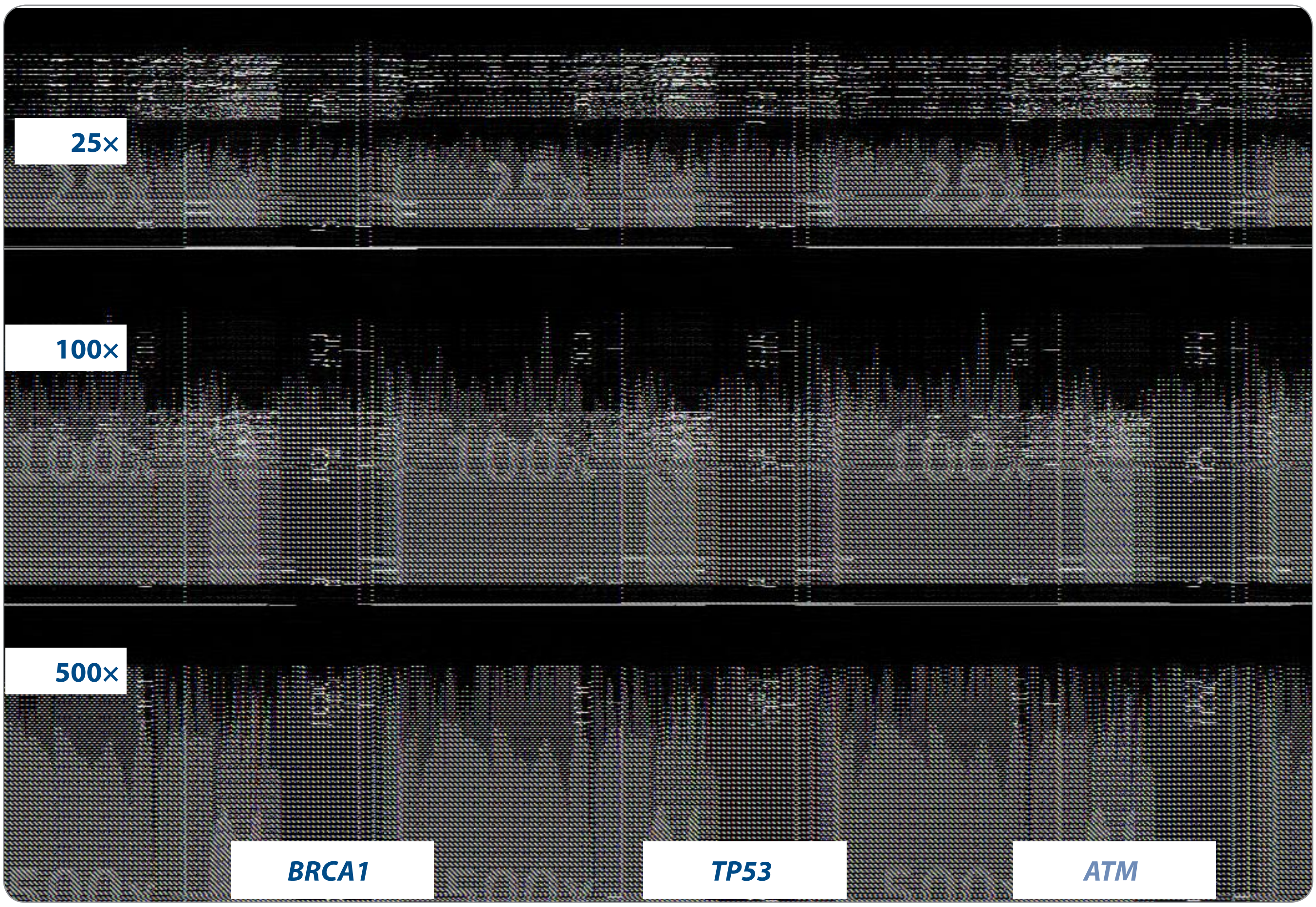
| Skupina | Gen | OMIM | Oficiální název | Základní funkce proteinu | Prs | Ovarium | Děloha | Kolon a rektum | Žaludek | Slinivka | Mozek | Kůže (a névy) | Ledvina | Prostata | Endokrinní tkáně | Sarkomy | Leukemie/lymfomy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| B | *CHEK2** | 604373 | checkpoint kinase 2 | reparace DNA | ŽM | Ž | Ž | ŽM | | ŽM | | | | M | ŽM | | |
| A | *KIT* | 164920 | v-KIT viral oncogene homolog | IC signalizace: RTK | | | | ŽM | ŽM | | | | | | | | ŽM |
| A | *MEN1* | 613733 | multiple endocrine neoplasia type I | reparace DNA/ /regulace GE | | | | | | ŽM | | | | | ŽM | | |
| A | *MET* | 164860 | MET protooncogene | IC signalizace: RTK | | | | | | | | | ŽM | | | | |
| A | *MLH1* | 120436 | mutL homolog 1 | reparace DNA | Ž | Ž | Ž | ŽM | | | | | | | | | |
| A | *MLH3* | 604395 | mutL homolog 3 | reparace DNA | | | Ž | ŽM | | | | | | | | | |
| A | *MSH2* | 609309 | mutS homolog 2 | reparace DNA | | Ž | Ž | ŽM | | | | | | | | | |
| A | *MSH6* | 600678 | mutS homolog 6 | reparace DNA | | | Ž | ŽM | | | | | | | | | |
| A | *MUTYH* | 604933 | mutY homolog | rreparace DNA | | | | ŽM | | | | | | | | | |
| B | *NBN* | 602667 | nibrin | reparace DNA | Ž | | | | | | | ŽM | | | | | ŽM |
| A | *NF1** | 613113 | neurofibromin 1 | IC signalizace: Ras | | | | | | | ŽM | | | | ŽM | ŽM | |
| A | *NF2* | 607379 | neurofibromin 2 | cytoskelet | | | | | | | ŽM | | | | | | |
| B | *PALB2* | 610355 | partner and localizer of BRCA2; *FANCN* | reparace DNA | Ž | | | | | ŽM | | | | | | | ŽM |
| B | *POLD1* | 174761 | polymerase (DNA directed), β | reparace DNA | | | | ŽM | | | | | | | | | |
| B | *POLE* | 174762 | polymerase (DNA directed), ε | reparace DNA | | | | ŽM | | | | | | | | | |
| B | *PRKAR1A* | 188830 | protein kinase, cAMP-dependent, regulatory, type Iα | IC signalizace: GPCR | | | | | | | | | | | ŽM | | |
| A | *PTEN* | 602954 | phosphatase and tensin homolog | IC signalizace: Akt | Ž | | Ž | ŽM | | | ŽM | ŽM | ŽM | | ŽM | | |
| A | *PTCH1* | 601309 | patched, drosophila, homolog of, 1 | IC signalizace: Hedgehog | | | | | | | | ŽM | | | | | |
| B | *RAD51C* | 602774 | RAD51 paralog C; *FANCO* | reparace DNA | Ž | Ž | | | | | | | | | | | |
| B | *RAD51D* | 602954 | RAD51 paralog D; *RAD51L3* | reparace DNA | Ž | Ž | | | | | | | | | | | |
| A | *RB1* | 614041 | retinoblastoma 1 | buněčný cyklus | | | | | | | ŽM | | | | | | |
| B | *RECQL* | 600537 | RecQ helicase-like | reparace DNA | Ž | | | | | | | | | | | | |
| B | *RECQL4* | 603780 | RecQ helicase-like 4 | reparace DNA | Ž | | | | | | | | | | | | |
| A | *RET* | 164761 | rearranged during transfection protooncogene | IC signalizace: RTK | | | | | | | | | | | ŽM | | |

**Tab. 1 – pokračování. Přehled základních charakteristik 54 genů zařazených do skupiny A nebo B s charakterizací funkcí kódovaných proteinů a asociací nádorových onemocnění v příslušných lokalizacích spojených s nosičstvím dědičných patogenních variant.**

| Skupina | Gen | OMIM | Oficiální název | Základní funkce proteinu | Prs | Ovarium | Děloha | Kolon a rektum | Žaludek | Slinivka | Mozek | Kůže (a névy) | Ledvina | Prostata | Endokrinní tkáně | Sarkomy | Leukemie/lymfomy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | *SDHB* | 185470 | succinate dehydrogenase complex, subunit b | metabolizmus živin | | | | | ŽM | | | | ŽM | | ŽM | ŽM | |
| B | *SLX4* | 613278 | S. cerevisiae, homolog of SLX4 (*FANCP*) | reparace DNA | Ž | | | | | | | | | | | | |
| A | *SMAD4* | 600993 | SMA- and MAD-related protein 4 | IC signalizace: TGF-β/TF | | | | ŽM | | | | | | | | | |
| A | *SMARCB1* | 601607 | SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin, subfamily b, member 1 | regulace GE | | | | | | | ŽM | | | | | | |
| A | *STK11* | 602216 | serine/threonine kinase 11 | IC signalizace: metabolizmus/ /růst buněk | Ž | Ž | Ž | ŽM | ŽM | ŽM | | | | | | | |
| B | *SUFU* | 607035 | suppressor of fused, drosophila, homolog of | IC signalizace: Hedgehog | | | | | | | ŽM | ŽM | | | | | |
| A | *TP53* | 191170 | tumor protein p53 | reparace DNA | Ž | | | | | | ŽM | | ŽM | | ŽM | ŽM | |
| A | *TSC1* | 191100 | tuberous sclerosis-1 | IC signalizace: Akt | | | | | | | ŽM | | ŽM | | | ŽM | |
| A | *TSC2* | 191092 | tuberous sclerosis-2 | IC signalizace: Akt | | | | | | | ŽM | | ŽM | | | ŽM | |
| A | *VHL* | 608537 | von Hippel-Lindau | IC signalizace: hypoxie | | | | | | ŽM | | | | | ŽM | | |
| B | *WRN* | 604611 | Werner syndrome, RecQ helicase-like | reparace DNA | | | | | | | ŽM | | | | | ŽM | |
| A | *WT1* | 607102 | Wilms tumor, type 1 | IC signalizace: regulace GE | | | | | | | | | ŽM | | | | |

Některé z exonů genů označených * byly z návrhu panelu vynechány z důvodu vysokého výskytu pseudogenů.
Ž – ženy, M – muži, IC – intracelulární, RTK – receptorová tyrozinkináza, TF – transkripční faktor, GE – genová exprese

provádět ultrazvukem (např. systém Covaris) nebo enzymaticky (např. KAPA HyperPlus, Roche) s DNA o doporučeném vstupním množství 0,3–1 μg. Po úpravě fragmentů DNA (end-repair, A-tailing, ligace adaptorů) provádíme selekci fragmentů o vhodné délce, které pak značíme v průběhu LM-PCR (ligation-mediated PCR) indexy specifickými pro každý vzorek individuální DNA. Takto označené vzorky pak můžeme proporcionálně spojit. Počet spojených vzorků (= společně analyzovaných pacientů) závisí na velikosti panelu, požadovaném pokrytí (= počet čtení každého nukleotidu) a kapacitě sekvenátoru. Za použití panelu CZECANCA lze při cíleném sekvenačním pokrytí 200krát vyšetřit na systému MiSeq (Illumina) bezpečně 30 pacientů v jednom sekvenačním běhu. Fragmenty DNA všech analyzovaných vzorků jsou následně společně hybridizovány se sondami panelu CZECANCA – probíhá obohacování knihovny. Po ukončení hybridizace jsou biotinylované sondy s navázanými fragmenty DNA vychytány magnetic-



**Obr. 1. Homogenita pokrytí u třech vybraných genů z CZECANCA panelu (*BRCA1, TP53* a *ATM*) při různé cílené hloubce sekvenačního pokrytí (coverage: 25×, 100×, 500×; oranžová linka).**

V grafech jsou pomocí skriptu Boudalyzer znázorněny pokrytí všech jednotlivých bází v oblasti všech kódujících exonů zobrazených genů.

kými kuličkami na základě vazby biotin-streptavidin. Vychytané fragmenty DNA jsou následně amplifikovány a po přečištění je obohacená knihovna připravena k sekvenování. Pro sekvenování na MiSeq používáme sekvenační chemii V3 (150-cycle, Illumina).

Zásadní důraz byl kladen na homogenní sekvenační pokrytí (počet čtení jednotlivých nukleotidů sekvenovaného úseku DNA) jednotlivých genů a robustní reprodukovatelnost umožňující minimalizovat sekvenační chyby mezi jednotlivými analýzami a mezi laboratořemi (obr. 1). Otázka sekvenačního pokrytí je dlouhodobě významně diskutované téma. V současné době je za hodnověrné považováno pokrytí 35–50krát pro jednonukleotidové záměny a malé delece/inzerce napříč genomem [11].

## Bioinformatické zpracování pro účely jednotné databáze (CZECANCA pipeline)

Bioinformatické zpracování sekvenačních dat je založeno na protokolu vypracovaném pracovníky Ústavu dědičných a metabolických poruch (Mgr. Viktor Stránecký, Ph.D. a doc. Ing. Stanislav Kmoch, CSc.) upraveném v Ústavu biochemie a experimentální onkologie (Mgr. Petra Zemánková) 1. LF UK v Praze.

Bioinformatické zpracování pro účely jednotné databáze předpokládá sdílení sekvenačních dat cestou BaseSpace (https://basespace.illumina.com). Soubory jednotlivých vyšetřovaných osob jsou kódovány pracovištěm sdílícím sekvenační data, které jako jediný subjekt má přístup k identifikaci svého konkrétního vzorku. Čtení v podobě fastq souboru jsou namapována pomocí alignovacího softwaru na lidský genom, zároveň vzniká SAM (Sequence Alignment Map) soubor. Pomocí aplikace Picard tools je převeden na BAM soubor, což je binární verze předchozího SAM. K tomuto kroku patří také odstranění duplikátů pomocí stejné aplikace. V této části se také provádí tzv. realignment, který nám umožňuje GATK (The Genome Analysis Toolkit, https://www.broadinstitute.org/gatk/). Soubory BAM slouží pro zobrazení čtení v příslušném prohlížeči, např. Integrative Genomics Viewer (IGV). Dalším důležitým krokem v procesu zpracování dat je převod na VCF (Variant Call File), k čemuž slouží GATK. V tomto souboru se nacházejí varianty nalezené u příslušného pacienta. Tento výstup je zpracován pomocí anotačního softwaru, např. ANNOVAR (http://annovar.openbioinformatics.org/en/latest/), kde se každé variantě přiřadí její biologická funkce a záznamy, jako je přítomnost varianty v databázích ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/) a HGMD (http://www.hgmd.cf.ac.uk/ac/index.php), frekvence varianty v mezinárodních sekvenačních projektech 1 000 genomů (http://www.1000genomes.org/) nebo ESP6500 (https://esp.gs.washington.edu/drupal/) nebo EXAC (http://exac.broadinstitute.org/).

## Implementace projektu CZECANCA

Projekt CZECANCA je připraven po technické stránce a v posledním období bylo na základě optimalizovaného protokolu na našem pracovišti analyzováno přes 200 indikovaných osob. V současné době je finalizována podoba společné databáze genotypů a fenotypových charakteristik sekvenovaných pacientů. Údaje o pacientech by měly zahrnovat data, která se vztahují k onkologické diagnóze (věk diagnózy, histologii, imunohistochemická vyšetření, stupeň diferenciace a rozsah onemocnění), osobní anamnéze (věk, pohlaví) a onkologické rodinné anamnéze (reprezentované ideálně rodokmenem).

Vytvoření společné databáze předpokládá rovněž získání genotypů zdravé populace z vyšetření reprezentativního počtu vzorků kontrolního souboru osob bez onkologické diagnózy. Tento důležitý požadavek pro identifikaci populačně specifických genetických variant bez souvislosti s nádorovými onemocněními bude nezbytné řešit formou specifických grantových projektů.

Pro správnou interpretaci získaných sekvenačních dat, zejména v případě nově identifikovaných či kandidátních predispozičních genů, a tím následně pro správnou péči o vyšetřované jedince, je nezbytná úzká spolupráce mezi vyšetřující laboratoří, ambulantním genetikem a ošetřujícím onkologem/gynekologem. Neexistence této funkční spolupráce vede ke ztrátě řady cenných informací a ve svém důsledku může vést k poškození testovaného probanda, resp. dalších členů jeho rodiny. Na svém významu tak ještě více nabývá kvalitně odebraná rodinná anamnéza včetně informací o zdravých příbuzných a zejména pak její pravidelná aktualizace ošetřujícím lékařem.

Domníváme se, že společné úsilí zainteresovaných pracovišť je racionální cestou k dosažení cíle, kterým je zlepšení klinické diagnostiky dědičných nádorových onemocnění, které by mělo přinést zlepšení péče o nosiče mutací v nádorových predispozičních genech.

## Poděkování

Zásadní podíl na provedených bioinformatických analýzách a navržených postupech pro jednotné zpracování dat má Mgr. Viktor Stránecký, Ph.D. a jeho kolegové z Ústavu dědičných metabolických poruch 1. LF UK a VFN v Praze, kterým velmi děkujeme za jejich pomoc a podporu. Za přípravu a kritické poznámky k této práci a složení CZECANCA panelu děkujeme především doc. MUDr. Lence Foretové, Ph.D. a RNDr. Evě Macháčkové, Ph.D. z Oddělení epidemiologie a genetiky nádorů, MOÚ v Brně a řadě dalších kolegů za přínosné poznámky a doporučení.

## Literatura

**1.** Foretova L, Petrakova K, Palacova M et al. Genetic testing and prevention of hereditary cancer at the MMCI – over 10 years of experience. Klin Onkol 2010; 23(6): 388–400.

**2.** Rahman N. Realizing the promise of cancer predisposition genes. Nature 2014; 505(7483): 302–308. doi: 10.1038/nature12981.

**3.** Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009; 458(7239): 719–724. doi: 10.1038/nature07943.

**4.** Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. Nat Genet 2008; 40(1): 17–22.

**5.** Saam J, Arnell C, Theisen A et al. Patients tested at a laboratory for hereditary cancer syndromes show an overlap for multiple syndromes in their personal and familial cancer histories. Oncology 2015; 89(5): 288–293. doi: 10.1159/000437307.

**6.** Kleibl Z, Novotny J, Bezdickova D et al. The CHEK2 c.1100delC germline mutation rarely contributes to breast cancer development in the Czech Republic. Breast Cancer Res Treat 2005; 90(2): 165–167.

**7.** Schroeder C, Faust U, Sturm M et al. HBOC multi-gene panel testing: comparison of two sequencing centers. Breast Cancer Res Treat 2015; 152(1): 129–136. doi: 10.1007/s10549-015-3429-9.

**8.** Lhota F, Stranecky V, Boudova P et al. Targeted next-gen sequencing in high-risk BRCA1- and BRCA2-negative breast cancer patients. Curr Oncol 2014; 21(2): e376.

**9.** Kluska A, Balabas A, Paziewska A et al. New recurrent BRCA1/2 mutations in Polish patients with familial breast/ovarian cancer detected by next generation sequencing. BMC Med Genomics 2015; 8: 19. doi: 10.1186/s12920-015-0092-2.

**10.** Cybulski C, Lubiński J, Wokołorczyk D et al. Mutations predisposing to breast cancer in 12 candidate genes in breast cancer patients from Poland. Clin Genet 2014; 88(4): 366–370. doi: 10.1111/cge.12524.

**11.** Sims D, Sudbery I, Ilott NE et al. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 2014; 15(2): 121–132. doi: 10.1038/nrg3642.

RESEARCH ARTICLE

# Validation of CZECANCA (CZEch CAncer paNel for Clinical Application) for targeted NGS-based analysis of hereditary cancer syndromes

**Jana Soukupova[1]*, Petra Zemankova[1], Klara Lhotova[1], Marketa Janatova[1], Marianna Borecka[1], Lenka Stolarova[1], Filip Lhota[1,2], Lenka Foretova[3], Eva Machackova[3], Viktor Stranecky[4], Spiros Tavandzis[5], Petra Kleiblova[1,6], Michal Vocka[7], Hana Hartmannova[4], Katerina Hodanova[4], Stanislav Kmoch[4], Zdenek Kleibl[1]***

**1** Institute of Biochemistry and Experimental Oncology, First Faculty of Medicine, Charles University, Prague, Czech Republic, **2** Centre for Medical Genetics and Reproductive Medicine, Gennet, Prague, Czech Republic, **3** Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, Czech Republic, **4** Research Unit for Rare Diseases, Department of Paediatrics and Adolescent Medicine, First Faculty of Medicine, Charles University and General University Hospital in Prague, Prague, Czech Republic, **5** Department of Medical Genetics, AGEL Laboratories, AGEL Research and Training Institute, Novy Jicin, Czech Republic, **6** Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University and General University Hospital in Prague, Prague, Czech Republic, **7** Department of Oncology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Prague, Czech Republic

* zdekleje@lf1.cuni.cz (ZK); jana.soukupova@lf1.cuni.cz (JS)

## Abstract

### Background

Carriers of mutations in hereditary cancer predisposition genes represent a small but clinically important subgroup of oncology patients. The identification of causal germline mutations determines follow-up management, treatment options and genetic counselling in patients' families. Targeted next-generation sequencing-based analyses using cancer-specific panels in high-risk individuals have been rapidly adopted by diagnostic laboratories. While the use of diagnosis-specific panels is straightforward in typical cases, individuals with unusual phenotypes from families with overlapping criteria require multiple panel testing. Moreover, narrow gene panels are limited by our currently incomplete knowledge about possible genetic dispositions.

### Methods

We have designed a multi-gene panel called CZECANCA (CZEch CAncer paNel for Clinical Application) for a sequencing analysis of 219 cancer-susceptibility and candidate predisposition genes associated with frequent hereditary cancers.

### Results

The bioanalytical and bioinformatics pipeline was validated on a set of internal and commercially available DNA controls showing high coverage uniformity, sensitivity, specificity and

OPEN ACCESS

**Citation:** Soukupova J, Zemankova P, Lhotova K, Janatova M, Borecka M, Stolarova L, et al. (2018) Validation of CZECANCA (CZEch CAncer paNel for Clinical Application) for targeted NGS-based analysis of hereditary cancer syndromes. PLoS ONE 13(4): e0195761. https://doi.org/10.1371/journal.pone.0195761

**Editor:** Obul Reddy Bandapalli, German Cancer Research Center (DKFZ), GERMANY

**Received:** October 19, 2017

**Accepted:** March 28, 2018

**Published:** April 12, 2018

**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Funding:** This study was supported by the grants of Ministry of Health (www.mzcr.cz) 15-27695A (JS), 15-28830A (ZK), 16-29959A (ZK), and DRO MMCI, 00209805 (LF), grants of Charles University (www.cuni.cz) PROGRES Q28/LF1 (ZK), and SVV2017/260367 (SK) and Ministry of Education, Youth and Sports (www.msmt.cz) project CZ.02.1.01/0.0/0.0/16_013/0001634 (SK).

**Competing interests:** The authors have declared that no competing interests exist.

accuracy. The panel demonstrates a reliable detection of both single nucleotide and copy number variants. Inter-laboratory, intra- and inter-run replicates confirmed the robustness of our approach.

## Conclusion

The objective of CZECANCA is a nationwide consolidation of cancer-predisposition genetic testing across various clinical indications with savings in costs, human labor and turnaround time. Moreover, the unified diagnostics will enable the integration and analysis of genotypes with associated phenotypes in a national database improving the clinical interpretation of variants.

## Introduction

Hereditary cancer syndromes are heterogeneous diseases characterized by the development of various cancer types in carriers of rare germline mutations in cancer susceptibility genes. These genes dominantly code for tumor suppressor proteins negatively regulating mitotic signals and cell cycle progression, activating apoptotic pathways, or executing DNA repair processes [1].

In general, it is considered that around 5% of all cancer diagnoses arise in hereditary cancer form. However, the percentage of hereditary cancers varies by cancer type, ranging from less than 3% in lung cancer to over 30% in pheochromocytoma [2, 3]. Important features distinguishing hereditary and sporadic cancers include an increased lifetime cancer risk with early disease onset, an increased risk of cancer multiplicity, the accumulation of cancer diagnoses in affected families, and a 50% risk of disease trait transmission to the offspring [1]. Considering these attributes and their consequences in terms of decreased life expectancy, decreased quality of life and increased medical expenses, patients carrying mutations in cancer susceptibility genes and their relatives represent a medically important subgroup with specific needs for increased cancer surveillance, a tailored follow-up and therapy, and rational prevention. However, the primary need is an unequivocal identification of the causative germline variant.

Although cancer inheritance has been suggested for over 150 years, the first gene conferring an increased cancer risk (*Rb*) was discovered only 30 years ago [4]. Hundreds of predisposing or candidate genes have been characterized since then, including the clinically most important "major" cancer susceptibility genes with high penetrance representing a subset of genes whose germline variants confer a high cancer risk (with relative risk (RR) > 5.0) in a substantial proportion of hereditary cancer patients. Pathogenic germline variants in "major" genes occur most commonly in patients with breast, ovarian, and colorectal cancers with variable proportions across populations worldwide. The group of cancer susceptibility genes with moderate penetrance is more extensive and growing steadily [5]. However, the clinical utility for many moderate penetrance genes is currently limited by the insufficient evidence about the degree of cancer risks associated with their germline variants.

The rapid improvement and availability of next-generation sequencing (NGS) technologies enable efficient simultaneous analyses of many cancer susceptibility genes in oncology patients or asymptomatic individuals at risk in routine diagnostics. NGS offers multiple approaches for the investigation of cancer predisposition, including the sequencing of whole genomes, exomes or transcriptomes. At present, however, the most widely used method of detecting clinically informative genetic alterations in the clinical setting is targeted panel NGS, analyzing selected subsets of genes of interest [6]. Nevertheless, the numbers of genes included in panels differ substantially among laboratories and depend on healthcare systems. While some cancer-specific or multi-cancer panels include only the "major" predisposition genes for which substantial literature exists with regard to their diagnostic relevance, others include larger gene sets consisting of all clinically relevant genes and additional genes for which the evidence of cancer predisposition is still unclear.

NGS-based cancer testing has been rapidly adopted by routine clinical laboratories [7]. Their primary choice resides in the decision whether to use a commercially available NGS panel, or to design custom-made systems. The decision is influenced by clinical demand determining the set of targeted genes, by the spectrum of cancer diagnoses that will be analyzed, by the expected number of analyzed samples, and by costs of the analyses.

Our aim was to develop a universal diagnostic approach suitable for contributing genetic laboratories and allowing sample batching across multiple cancer indications. We focused on i) designing a custom-made multi-cancer panel with the desired sequencing quality and uniformity permitting a reliable variant identification, ii) the development of a robust analytical procedure limiting inter-run and inter-laboratory differences, and iii) the optimization of the bioinformatics pipeline enabling unified variant calling and annotation. The data collected from analyses of high-risk individuals performed in contributing laboratories will be used to create a nationwide genotype–phenotype database improving clinical variant interpretation in high-risk individuals.

# Methods

## Validation samples

**Patient DNA samples.** Validation of CZECANCA pipeline included analyses of 389 samples previously tested for the presence of germline variants available from DNA repository of the Institute of Biochemistry and Experimental Oncology. First Faculty of Medicine, Charles University. Of these, 137 samples carried pathogenic SNVs or short indels (in *BRCA1/2*, *PALB2*, *CHEK2*, *ATM*, *NBN*, *DPYD*, *PPM1D*, *RAD51C*, *RAD51D*, or *TP53*), 217 had been tested negatively using previous gene-by-gene analyses based on Sanger sequencing or a protein truncation test (PTT) [8–16], and 35 samples carried intragenic rearrangements in *BRCA1*, *CHEK2*, *PALB2*, or *TP53*, identified by the MLPA (multiplex ligation-dependent probe amplification) analysis [10, 17, 18]. All blood-isolated DNA samples were obtained from individuals that gave their written informed consent with mutation analyses of cancer susceptibility genes and who agreed to use their genetic material for research purposes. The study was approved by Ethics Committee of the First Medical Faculty, Charles University and General University Hospital in Prague. All used samples were anonymized prior analysis.

**Human genome reference standards.** Five commercially available DNA reference standards (NA12878, NA24149, NA24385, NA24631 and NA24143) were obtained from Coriell Institute for Medical Research. Well described genotypes, including high confident calls for variant and wild-type alleles, is the major advantage of these reference standards. The genotypes and variants in reference samples identified by CZECANCA analysis and obtained from reference variant-call format (VCF) files (available from the Genome in a Bottle (GIAB) website; http://jimb.stanford.edu/giab/), respectively, were compared to compute CZECANCA sensitivity, specificity, and accuracy, as described by Hardwick et al. [19].

## Panel design

The multi-cancer panel CZECANCA was designed using the online NimbleDesign software utility (NimbleGen, Roche; http://sequencing.roche.com/products/software/nimbledesign-software.html). For enrichment, we selected genes with a known predisposition for hereditary breast, ovarian, colorectal, pancreatic, gastric, endometrial, kidney, prostate and skin cancers, together with known DNA repair genes associated (or potentially associated) with cancer susceptibility (a list of 219 selected genes is provided in S1 Table), considering the results of our previous NGS analysis with a broad panel of 581 genes [20]. The primary gene target for probe coverage was represented by all exons (in case of known cancer susceptibility genes) or all coding exons (in other genes), including 10 bases from adjacent intronic regions. The design considered all transcription variants of selected genes available at UCSC website (https://genome.ucsc.edu/; accessed 2015-05-21). The promoter regions of the *BRCA1* and *BRCA2* genes were included into the primary target. The probes were designed using *continuous design* under strict conditions–minimal and maximal *close matches* (number of times in which a probe sequence matches the genome with either $\leq$ 5 insertions or deletions, or gap of $\leq$ 5 bp) were one and three, respectively, allowing us to hybridize the probes up to three targets across the genome. Because of the strict design conditions, some clinically relevant regions were left untargeted for technical reasons such as repeats and homologous regions (see S1 Table). The final panel target size reached 628,069 bases.

## Library preparation

Five hundred ng of genomic DNA isolated from peripheral blood and dissolved in TE buffer was used for preferred ultrasound shearing using Covaris E220 (Covaris Inc). As an alternative DNA fragmentation method, we tested enzymatic digestion using Fragmentase (KAPA Biosystems, Roche) with incubation for 25 min at 37˚C according to the manufacturer's instruction. The mean average size of DNA fragments targeted 200 bp. Sizing and quality was controlled using the Agilent High Sensitivity DNA kit on the Agilent 2100 Bioanalyzer System (Agilent).

Libraries were prepared using the KAPA HTP Library Preparation kit (for ultrasound-sheared DNA samples) or KAPA HyperPlus Kit (for Fragmentase-digested DNA samples) according to the manufacturer's instructions (KAPA Biosystems, Roche) with minor modifications including the use of universal in-house prepared adapters, double-indexing primers for ligation-mediated polymerase chain reaction (LM-PCR), and primers for post-capture PCR, as described further. The adapters [Adapter#1: 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATC$^*$T-3′ (“$^*$” denotes for phosphothiolate bond) and Adapter#2: 5′-pGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3′ (“p” denotes for 5’ phosphate)] were hybridized in Tris:NaCl buffer mix (50 mM Tris:HCl pH 7.5; 50 mM NaCl) in 97˚C for 2 min, followed by 72 cycles involving incubation at 97˚C for 1 min (-1˚C per cycle) and 25˚C for 5min. The barcoding of size-selected DNA fragments enabling subsequent sample pooling was performed during LM-PCR with indexing primers [Primer#1: 5′-AATGATACGGCGACCACCGAGATCTACACxxxxxxxxACACTCTTTCCCTACACGACGCTCTTCCGATC$^*$T-3′ and Primer#2: 5′-CAAGCAGAAGACGGCATACGAGATxxxxxxxxGTGACTGGAGTTCAGACGTGTGCTCTTCCGAT$^*$C-3′ (“$^*$” denotes for phosphothiolate bond; “xxxxxxxx” denotes for a sequence of particular indices same as the Illumina Truseq HT index i7 and i5)]. The number of LM-PCR cycles was reduced to six to limit the presence of PCR duplicates. Sizing and quality after the double-sided size selection and LM-PCR were controlled using the Agilent High Sensitivity DNA kit on the Agilent 2100 Bioanalyzer System.

To reach the targeted mean coverage (100X), 30 individual barcoded samples (33 ng each) were pooled for the enrichment (usually two overnight hybridizations; tested for 16–72 hours without a significant effect on enrichment efficacy) using the CZECANCA (NimbleGen SeqCap EZ Choice, Roche) to create a sequencing library. After the enrichment, the library was amplified using Primer 1: 5′-AATGATACGGCGACCACCGAGATCTACAC-3′ and Primer 2:

5′-CAAGCAGAAGACGGCATACGAGAT-3′. The number of post-capture PCR cycles was reduced to 11 to reach the optimal library concentration (2 ng/μl) and to minimalize the number of PCR duplicates.

After the enrichment control using qPCR (NimbleGen SeqCap EZ Library SR User's Guide), the final 18 pM libraries were sequenced on the MiSeq system using MiSeq Reagent Kit v3, 150 cycles (Illumina).

## Bioinformatics

**Single nucleotide variants (SNVs).** The NGS data obtained from sequencing with the CZECANCA were processed using an analysis pipeline based on standard tools. FASTQ files were generated by MiSeq. The quality of raw data was controlled using FastQC v0.11.2 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). FASTQ files were subsequently mapped using Novoalign v2.08.03 to hg19 (http://www.novocraft.com/products/novoalign/) to generate sequence alignment map (SAM) files. SAM files were transformed to binary form (BAM files) using Picard tools v1.129 (https://broadinstitute.github.io/picard/). Raw BAM files were further processed to eliminate PCR duplicates of mapped reads. The quality of mapped bases was checked and recalibrated according to default settings using Genome Analysis Toolkit (GATK) v3.3 (https://software.broadinstitute.org/gatk/). The finalized BAM file was converted using a GATK pipeline to a variant-call format (VCF) containing alternative variants only. ANNOVAR was used to annotate VCF files generated using GATK [21, 22] and to check the presence of each variant in external databases (ExAC, 1000Genome or ClinVar) [23–25]. Predictive values from selected prediction algorithms (for example SIFT [26], Mutation Analyzer [27], MutationTaster [28], LRT [29], PolyPhen-2 [30], phyloP [31], GERP [32], CADD [33] or spidex (https://www.deepgenomics.com/spidex) were added to the annotated alternative variants.

For a comparison with CZECANCA sequencing, the data from routine analyses using the TruSight cancer panel (Illumina), performed in a laboratory of the Masaryk Memorial Cancer Institute in Brno were analyzed by an identical bioinformatics pipeline [34].

The Integrative Genomics Viewer (IGV) was used for visualization and manual inspection of individual BAM files [35].

**Medium-size indels.** The detection and exact sequence determination of medium-size insertions and tandem duplications (involving approximately half of the sequence reads, depending on the sequencing chemistry used) is very challenging. The identification of these alterations was based on the method of soft-clipped bases using Pindel (http://gmt.genome.wustl.edu/packages/pindel/) [36]. The finalized BAM files served as an input for the analysis. In our case (with mean read size of 75 bp; MiSeq Reagent Kit v3, 150 cycles chemistry) insertion or duplication exceeding 35 bp was considered as a medium-size indel.

**Copy number variations (CNVs).** An analysis CNVs was performed using the CNVkit (https://pypi.python.org/pypi/CNVkit). The CNVs analysis is coverage-based and therefore required good coverage uniformity. Raw BAM files served as the input for this analysis.

**Coverage visualization.** The visualization of sequence coverage of the individual samples, enabling a fast visual inspection of coverage limit >20X (for a reliable identification of heterozygotes) across the analyzed genes, was performed by an in-house "Boudalyzer" script written in R language. The coverage is visualized from the finalized BAM files. This tool was used for the generation of manuscript figures showing coverages of the analyzed genes.

**Variant interpretation.** We used the scoring scheme outlined in ENIGMA guidelines (https://enigmaconsortium.org/) for variant interpretation to classify SNVs and indels as benign (Class 1), likely benign (Class 2), variant of unknown significance (Class 3), likely pathogenic (Class 4) and pathogenic (Class 5) [37]. Identified variants of unknown significance

(VUS) were further prioritized if their minor allele frequency was lower than 1% in ExAC, 1000Genome databases, or in a two sets of population-matched controls containing anonymized genomic data from 530 non-cancer controls analyzed by CZECANCA NGS and from 780 unselected Czech individuals analyzed by an exome sequencing (provided by the National Center for Medical Genomics; http://ncmg.cz). Potentially deleterious VUSes were selected based on concordant results obtained from above-mentioned *in silico* prediction algorithms. These priorized VUS variants were enrolled into the list of variants for subsequent segregation analyses or functional *in vitro* testing performed in selected genes.

The CZECANCA contains 22 genes that are listed in the ACMG recommendation (S1 Table) for the reporting of secondary findings [38].

## Results

### Target gene coverage

The NGS analysis with CZECANCA targeting the coding sequences of 219 genes (S1 Table) displayed high coverage uniformity. Under standard conditions for routine analyses, we targeted sequencing coverage 100X. In these settings, more than 85% of the targeted regions were covered 100X, 98% of the targeted regions were covered at least 50X and less than 0.2% of targeted regions had coverage below 20X (Fig 1A). The entire coding sequence was fully covered at least 100X in 144/219 targeted genes (65.8%), at least 50X in 190/219 genes (86.8%), and at least 20X in 207/219 targeted genes (94.5%; Fig 2). Coverage did not exceed 300X in any of the captured targets.

Coverage was uniform among samples independently analyzed in the participating laboratories using the described protocol (Fig 3), and also among samples sequenced using separately-synthesized CZECANCA lots (data not shown). The equal coverage uniformity was independent of coverage depth (Fig 1B). The coverage uniformity was partially influenced by the DNA fragmentation approach with better results obtained by ultrasound fragmentation in comparison with enzymatic DNA cleavage. The improved results (more random DNA shearing) obtained with the ultrasound fragmentation protocol were indicated by an analysis of terminal (di)nucleotides in reads from samples prepared by both DNA fragmentation methods, regardless of the laboratory site (Figs 1C and 3). The CZECANCA coverage uniformity substantially surpassed that of the Illumina TruSight Cancer Panel (Fig 3F).

Low-covered regions (uncovered or with coverage ≤20X) were constantly observed in 12/219 genes (5.5%; Fig 2, S1 Table). In nine genes, the low–covered regions were mostly limited to a single exon (typically the first exon) representing usually a small fraction of the coding sequence. In three incompletely covered genes (*CHEK2*, *MDC1*, *NF1*), single or several exons were omitted from the CZECANCA design (see Panel design in Methods). The remaining low-covered regions were GC-rich regions with mean GC content of 76.88% (S2 Table) while the average GC content of the CZECANCA targets is 47%.

Sequencing quality was partially influenced by the particular MiSeq sequencer. In standard runs, more than 99% of bases reached a Phred score >20 (i.e. 99% accuracy) and approximately 97% of bases overcame a Phred score of 30 (i.e. 99.9% accuracy). A decrease in PCRs cycles during library preparation reduced the number of PCR duplicates, which finally represented 7–9% of reads. The mean off-target (reads mapped to distance exceeding 250 bp from the nearest bait) across the performed runs was constantly less than 12% of reads.

### Reproducibility, specificity and sensitivity analysis

The reproducibility of variant calls was tested using intra-, inter-run, and inter-laboratory replicates. During the sequencing of intra-run replicates, we also evaluated the impact of coverage depth on coverage uniformity and reproducibility.

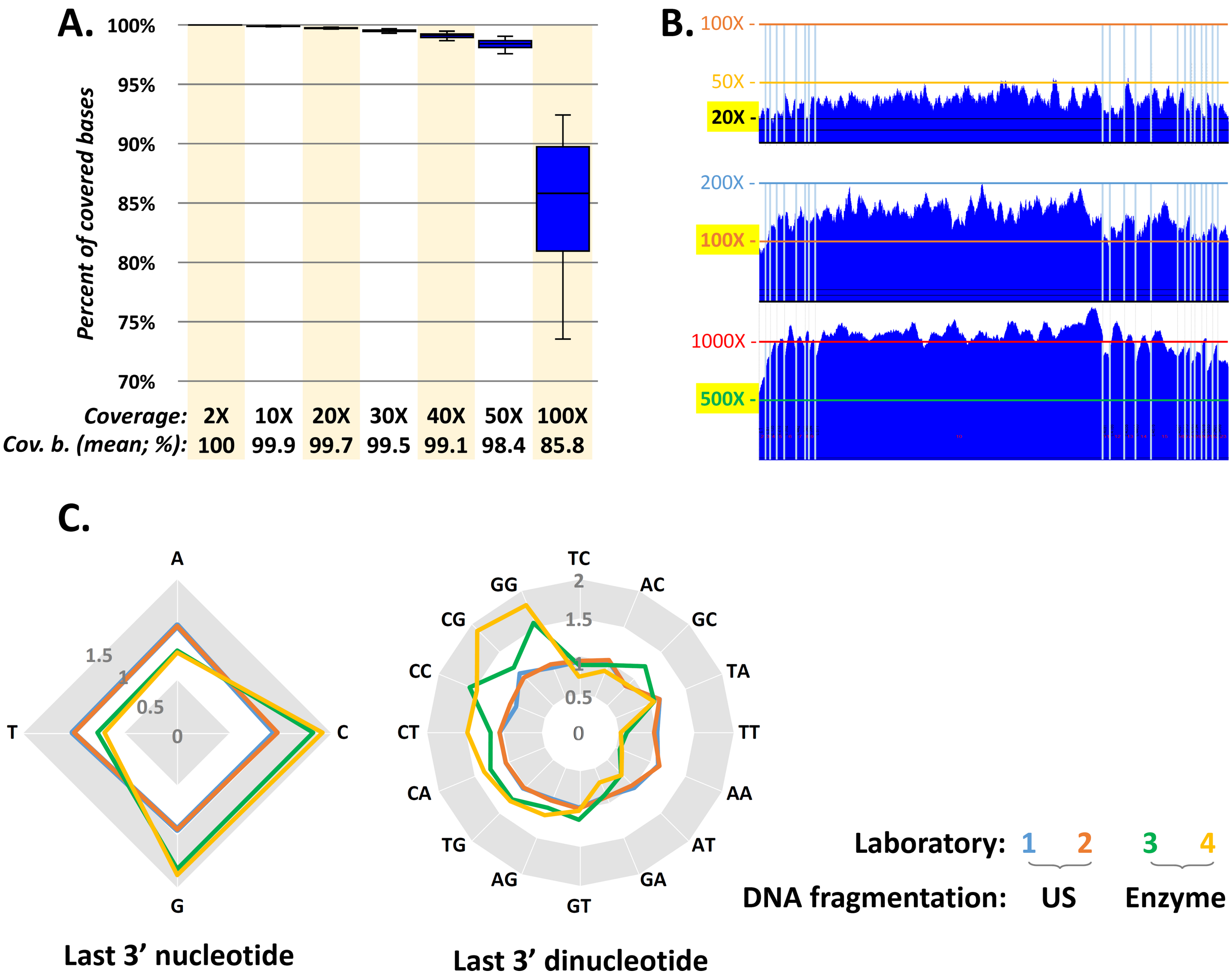


**Fig 1. Coverage parameters from CZECANCA sequencing.** (A) The chart expresses the percentages of covered target bases (cov. b.) obtained from 25 analyzed samples from a standard run targeting sequencing coverage 100X. (B) The coverage (at y-axis) of *BRCA1* coding sequence (NM_007294; x-axis; vertical lines represent exon boundaries) in three independent runs targeting sequencing coverages 20X, 100X, or 500X demonstrates coverage uniformity, not influenced by coverage depth. (C) The "randomness" of the DNA shearing approach using ultrasound (US) and enzymatic cleavage was compared by an analysis of the distribution of ending nucleotides and dinucleotides in reads completely mapped to the large exon 11 (chr17:41243452–41246877; 3426bp) in the *BRCA1* gene, representing one of the largest continuous genomic fragments targeted by CZECANCA probes. The chart displays the relativized distribution of terminal nucleotides and dinucleotides in the analyzed region from 12 samples from each laboratory normalized to the average nucleotide and dinucleotide content of the analyzed region. The distribution of last nucleotides and dinucleotides in fragments from samples processed by US oscillate closer to a normalized value (1) than in fragments of samples prepared by the enzymatic cleavage.

https://doi.org/10.1371/journal.pone.0195761.g001

Three individually bar-coded replicates were pooled for enrichment in amounts corresponding to 33 ng (considered as 100%), 24.75 ng (75%), and 16.5 ng (50%), respectively. The subsequent bioinformatics of these samples, considering variants with GATK quality >100 in the targeted regions (exon sequences with 12 bp from adjacent introns), revealed 293 (100%), 292 (99.7%) and 290 (99.0%) variants, respectively (S3 Table). Altogether, 289/293 (98.6%) variants were identified in all replicates, while four variants not detected in DNA-reduced samples were variant homozygotes located in low-covered regions or had GATK quality <100. The

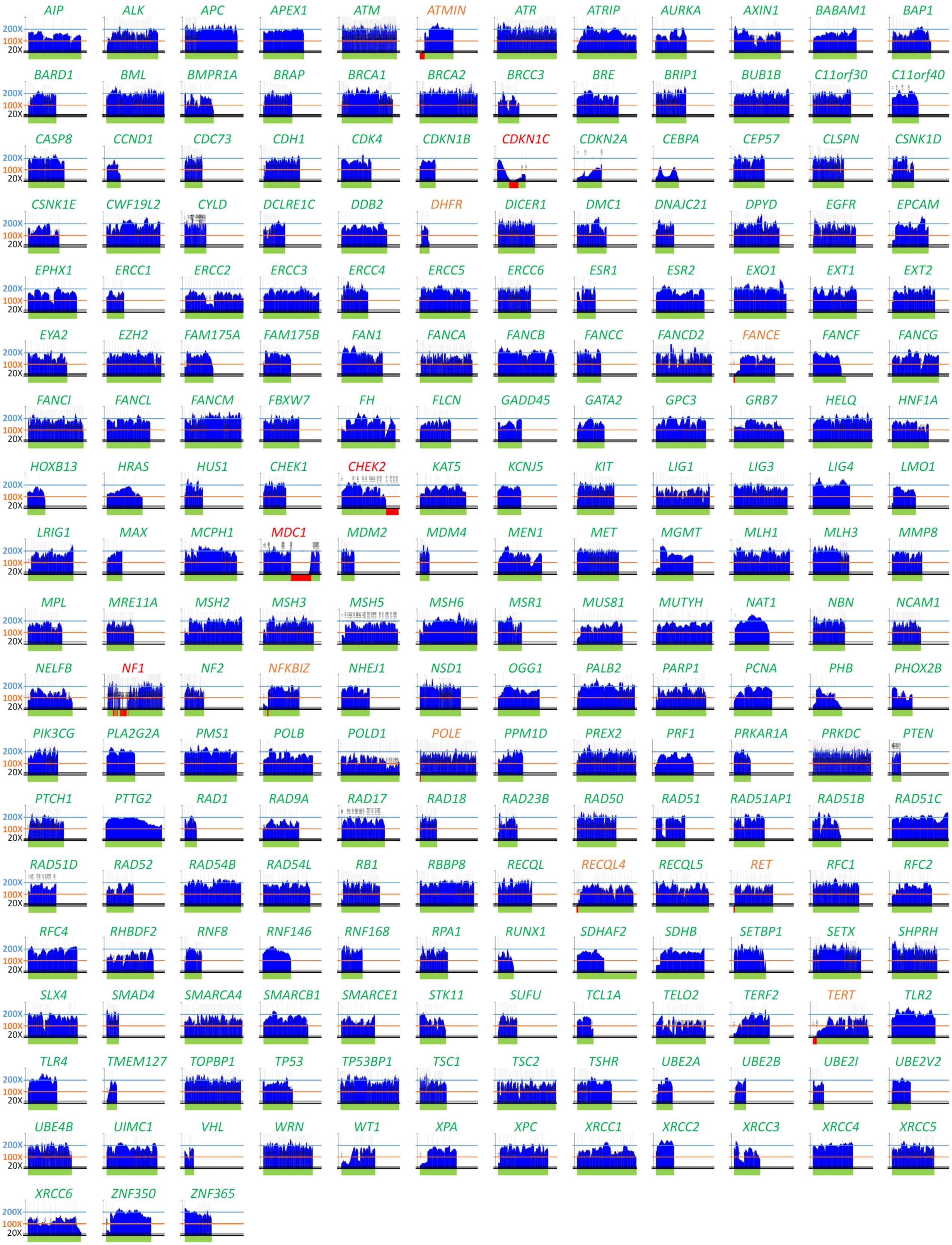

AIP
ALK
APC
APEX1
ATM
ATMIN
ATR
ATRIP
AURKA
AXIN1
BABAM1
BAP1
200X
100X
20X
BARD1
BML
BMPR1A
BRAP
BRCA1
BRCA2
BRCC3
BRE
BRIP1
BUB1B
C11orf30
C11orf40
200X
100X
20X
CASP8
CCND1
CDC73
CDH1
CDK4
CDKN1B
CDKN1C
CDKN2A
CEBPA
CEP57
CLSPN
CSNK1D
200X
100X
20X
CSNK1E
CWF19L2
CYLD
DCLRE1C
DDB2
DHFR
DICER1
DMC1
DNAJC21
DPYD
EGFR
EPCAM
200X
100X
20X
EPHX1
ERCC1
ERCC2
ERCC3
ERCC4
ERCC5
ERCC6
ESR1
ESR2
EXO1
EXT1
EXT2
200X
100X
20X
EYA2
EZH2
FAM175A
FAM175B
FAN1
FANCA
FANCB
FANCC
FANCD2
FANCE
FANCF
FANCG
200X
100X
20X
FANCI
FANCL
FANCM
FBXW7
FH
FLCN
GADD45
GATA2
GPC3
GRB7
HELQ
HNF1A
200X
100X
20X
HOXB13
HRAS
HUS1
CHEK1
CHEK2
KAT5
KCNJ5
KIT
LIG1
LIG3
LIG4
LMO1
200X
100X
20X
LRIG1
MAX
MCPH1
MDC1
MDM2
MDM4
MEN1
MET
MGMT
MLH1
MLH3
MMP8
200X
100X
20X
MPL
MRE11A
MSH2
MSH3
MSH5
MSH6
MSR1
MUS81
MUTYH
NAT1
NBN
NCAM1
200X
100X
20X
NELFB
NF1
NF2
NFKBIZ
NHEJ1
NSD1
OGG1
PALB2
PARP1
PCNA
PHB
PHOX2B
200X
100X
20X
PIK3CG
PLA2G2A
PMS1
POLB
POLD1
POLE
PPM1D
PREX2
PRF1
PRKAR1A
PRKDC
PTEN
200X
100X
20X
PTCH1
PTTG2
RAD1
RAD9A
RAD17
RAD18
RAD23B
RAD50
RAD51
RAD51AP1
RAD51B
RAD51C
200X
100X
20X
RAD51D
RAD52
RAD54B
RAD54L
RB1
RBBP8
RECQL
RECQL4
RECQL5
RET
RFC1
RFC2
200X
100X
20X
RFC4
RHBDF2
RNF8
RNF146
RNF168
RPA1
RUNX1
SDHAF2
SDHB
SETBP1
SETX
SHPRH
200X
100X
20X
SLX4
SMAD4
SMARCA4
SMARCB1
SMARCE1
STK11
SUFU
TCL1A
TELO2
TERF2
TERT
TLR2
200X
100X
20X
TLR4
TMEM127
TOPBP1
TP53
TP53BP1
TSC1
TSC2
TSHR
UBE2A
UBE2B
UBE2I
UBE2V2
200X
100X
20X
UBE4B
UIMC1
VHL
WRN
WT1
XPA
XPC
XRCC1
XRCC2
XRCC3
XRCC4
XRCC5
200X
100X
20X
XRCC6
ZNF350
ZNF365
200X
100X
20X

**Fig 2. Coverage (y-axis) of coding sequences (x-axis) of 219 CZECANCA target genes from a routine, randomly selected run targeting 100X coverage.** Note: Fully covered genes are depicted in green letters, genes with coverage <20X in a single exon are in orange letters, and genes with uncovered regions exceeding single exon or >10% of coding sequence are in red letters. Green horizontal bars (below individual graphs constructed using "Boudalyzer" script) indicate coverage ≥ 20X; red horizontal bars indicate regions covered <20X and uncovered regions.

https://doi.org/10.1371/journal.pone.0195761.g002

analysis demonstrated that alternative nucleotides could still be reliably detected in samples with reduced overall coverage, showing the robustness of the analysis in samples with unequal DNA input (Fig 4A).

A subsequent analysis of inter-run replicates (performed with another DNA sample analyzed in two independent runs) revealed 356 unique variants with GATK quality >100 in at least one replicate (S4 Table). Overall, 354 (99.4%) variants were identified in both inter-run replicates with a strong coverage correlation (Fig 4B).

In addition, the inter-laboratory performance was tested by an NGS analysis of an identical DNA control sample in four laboratories participating in the panel validation (Fig 4C), which revealed 332 unique variants with GATK quality >100 in at least one laboratory, from which we identified 331 (99.7%), 327 (98.5%), 329 (99.1%), and 329 (99.1%) variants in the particular laboratory, respectively. The discordant findings were caused by variants in low-covered regions, with low base Phred quality, or GATK quality <100 (S5 Table).

Sensitivity and specificity were assessed in 354 samples previously tested for the presence of germline variants. All 137 previously identified pathogenic germline mutations in *BRCA1/2* and other susceptibility genes were detected by CZECANCA (S6 Table). Moreover, an analysis

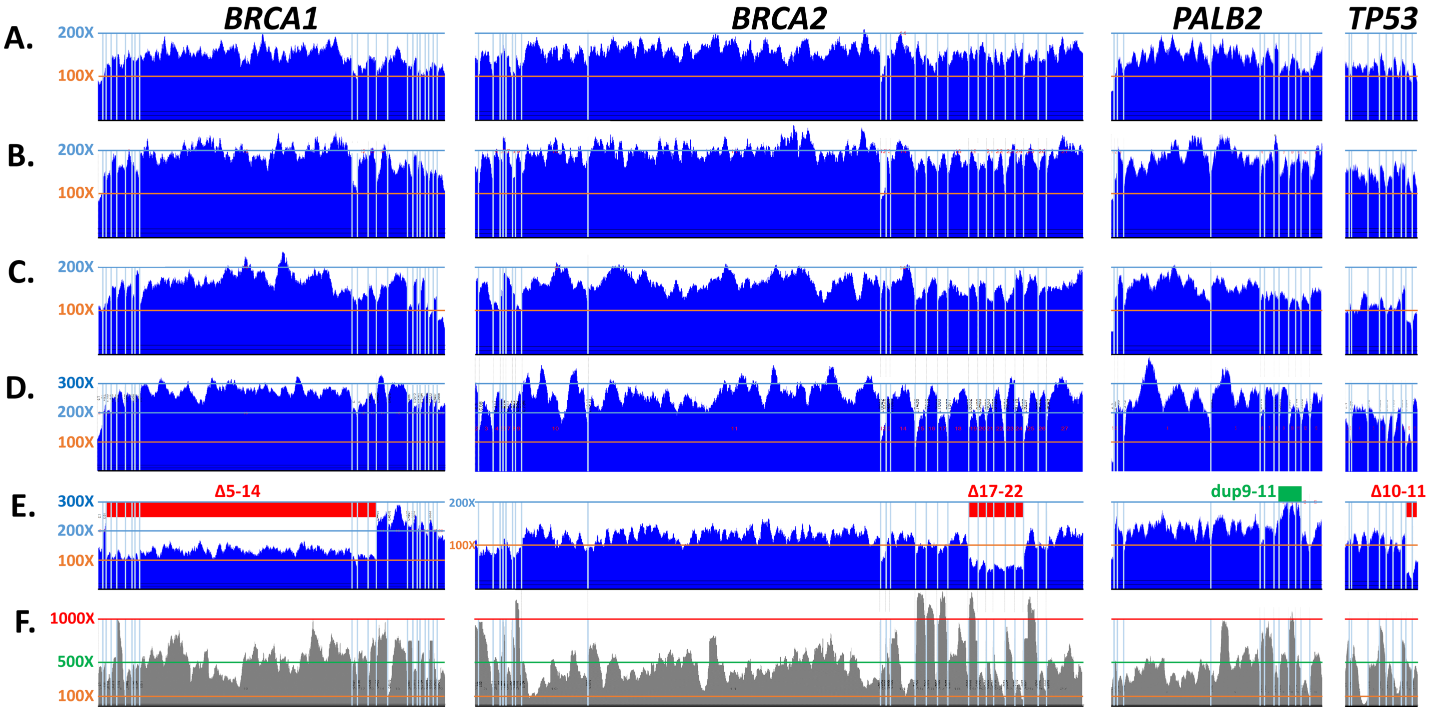


**Fig 3. Coverage of selected genes from the CZECANCA (A-E) and TruSight Cancer sequencing (F) panels.** The pictures show coverage (at y-axis) alongside the coding sequences of *BRCA1* (NM_007294), *BRCA2* (NM_000059), *PALB2* (NM_024675), and *TP53* (NM_000546), the vertical lines represent exon boundaries. Panels A–D show results obtained from a CZECANCA NGS analysis of various samples performed in four participating laboratories using the ultrasound (A, B) or enzymatic (C, D) DNA fragmentation protocol. Examples of the identified CNV aberrations in the depicted genes (deletions in *BRCA1*, *BRCA2* and *TP53* and duplication in *PALB2*) are shown in panel E. For comparison, panel F demonstrates the uneven coverage of the depicted genes by sequencing using the TruSight Cancer panel (Illumina).

https://doi.org/10.1371/journal.pone.0195761.g003

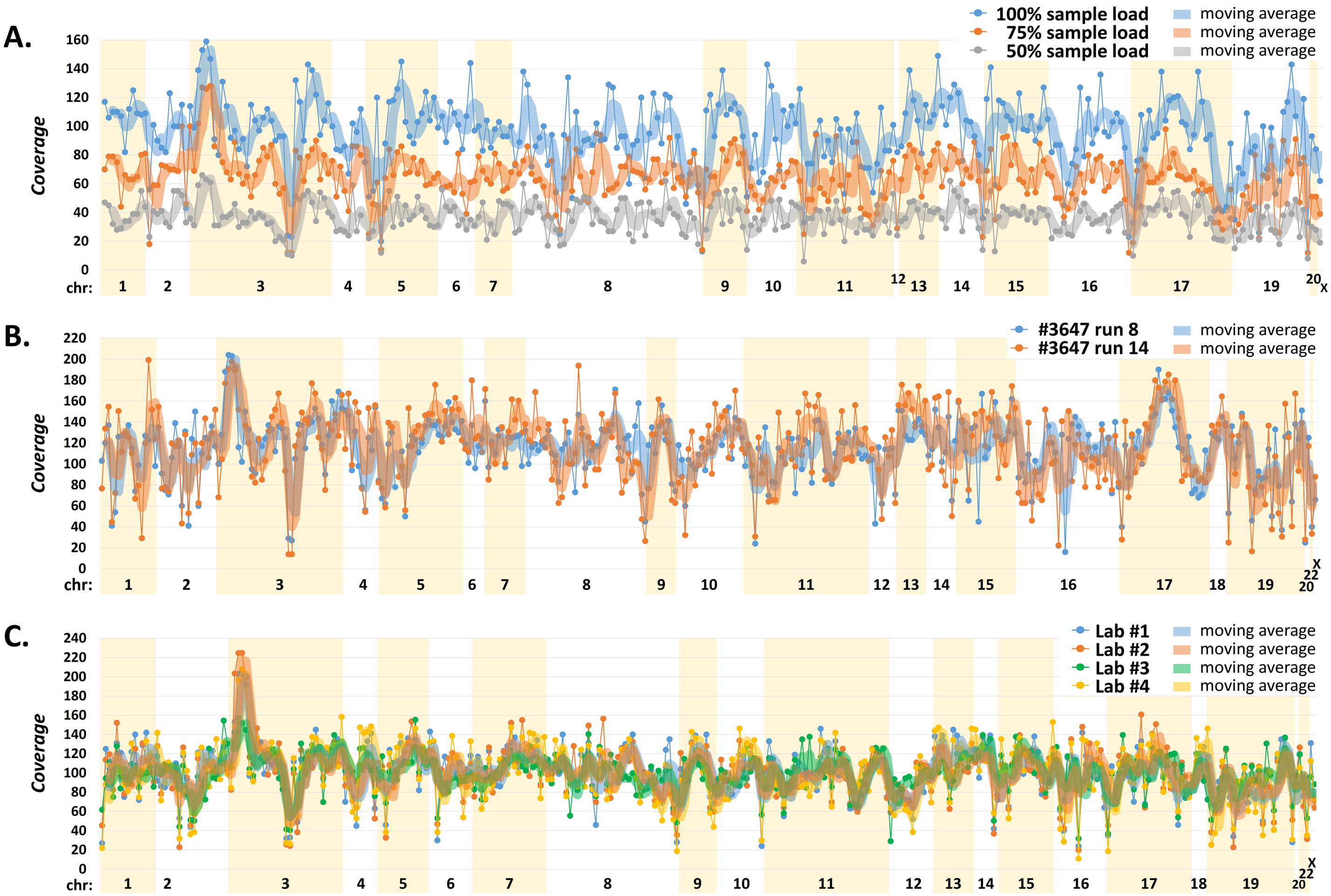


**Fig 4. Analysis of intra-run (A), inter-run (B), and inter-laboratory (C) replicates.** The panels show sequencing coverages (y-axis) of the identified variants arranged according to chromosomal localizations (x-axis). We used moving average curves (average of 3 values) to compare trends in coverages. Panel (A) describes the results of an analysis of three independently processed intra-run replicates from an identical DNA sample pooled in 33 ng (considered as 100%), 24.75 ng (75%), and 16.5 ng (50%), respectively. Panel (B) demonstrates variant coverages identified in two independent inter-run (run 8 and 14) replicates. All coverage values of sample #3647 in run 14 were corrected by a factor of 1.3880 to normalize coverages between samples (see S4 Table). Panel (C) shows coverages of variants identified in an inter-laboratory control sequenced in four laboratories (Lab) participating in panel validation (see S5 Table). The coverages of variants identified in Lab 2, 3, and 4 were normalized to the average coverage of Lab 1 for better comparisons of coverages.

https://doi.org/10.1371/journal.pone.0195761.g004

revealed nine additional *BRCA1* or *BRCA2* mutations. Of these, seven mutations were identified in samples previously tested by cDNA sequencing (they had not been detected previously, probably because of nonsense-mediated decay). The pathogenic missense mutation c.3G>A in *BRCA2* was found in a sample negatively analyzed using PTT and the pathogenic *BRCA2* mutation c.5645C>A was found in the carrier of c.5266dupC in *BRCA1* in whom the identification of a pathogenic *BRCA1* variant discontinued subsequent *BRCA2* testing.

Further, we validated the sensitivity of CNVs detection on 35 samples tested positively using the MLPA analysis (S7 Table). All CNVs including 18 samples with large *BRCA1* deletions or duplications, 12 CNVs in *CHEK2*, four in *PALB2* and one in *TP53* were detected using CNVkit software in routine settings targeting 100X coverage (Fig 5A; S8 Table). This analysis also enabled to setup CNVkit thresholds indicating the presence of a deletion or a duplication. To estimate the number of false positive and true positive CNV calls obtained from CNVkit, we further analyzed aggregated results from four consecutive runs performed in two

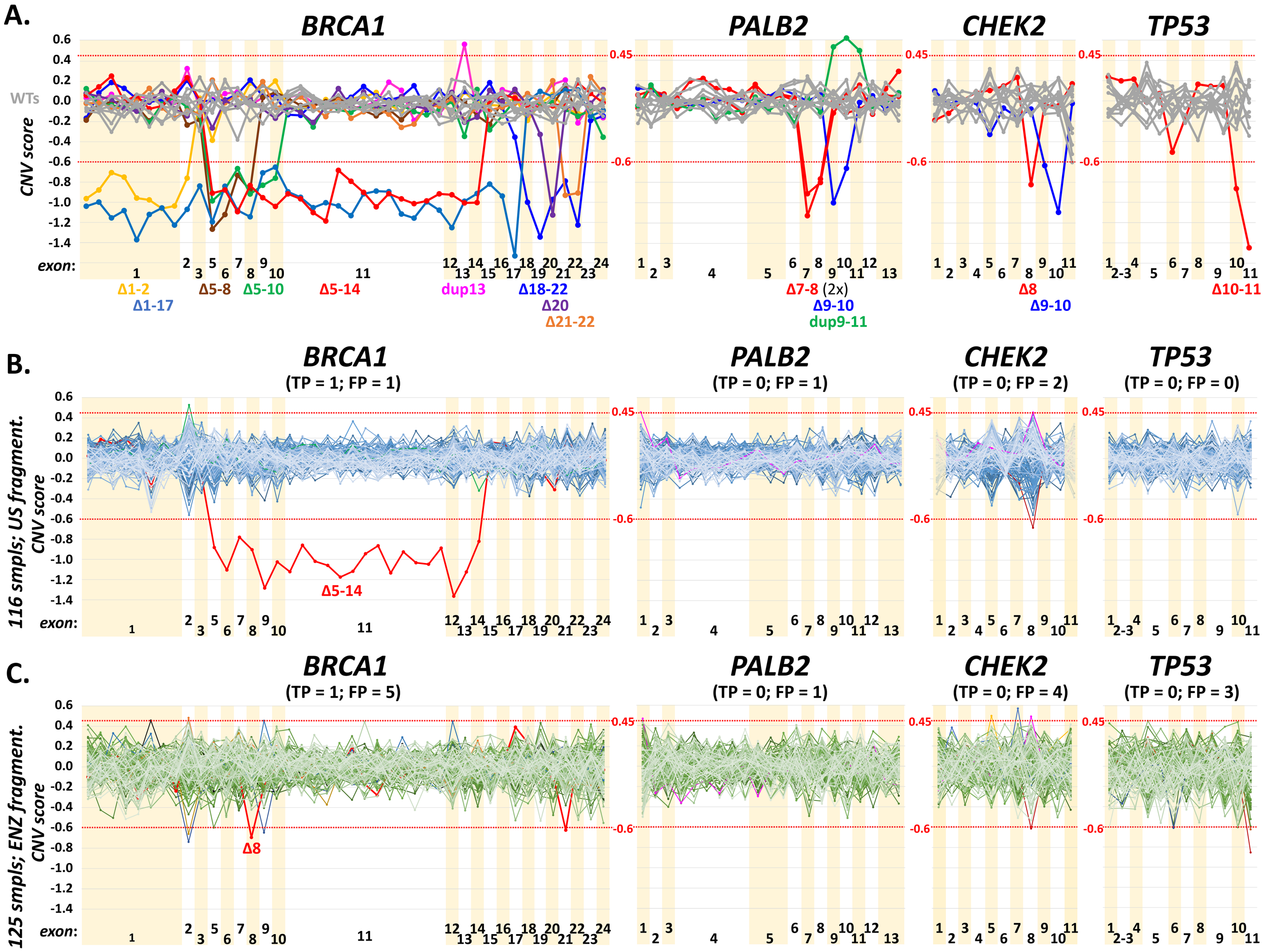


**Fig 5. The panel A show results of CNV analysis revealing large deletions or duplications in four genes in a testing set of 35 samples with previously identified CNVs.** The charts show median-normalized values of CNV scores for particular gene bins (default settings in CNVkit software; S8 Table). Values <-0.6 and >0.45 (red dotted lines) were assumed as thresholds indicating a deletion or a duplication, respectively. All shown CNVs were confirmed by MLPA previously (S7 Table). The panels B and C demonstrate frequency of true positive (TP) and false positive (FP) CNV signals from analyses performed in two participating laboratories (laboratory 1 in B and laboratory 3 in C). While 116 samples analyzed in four consecutive runs in B were prepared using the ultrasound (US) fragmentation, 125 other samples in four consecutive runs in C were prepared using the enzymatic (ENZ) fragmentation method. Samples in vivid colors highlight suspected samples that were further analyzed by MLPA analysis and samples in *BRCA1* Δ5–14 (B) and Δ8 (C) denote for true positives. The presence of putative CNVs in *PALB2*, *CHEK2*, and *TP53* were excluded by analysis that revealed heterozygotes in regions with suspected deletions or by an MLPA analysis.

https://doi.org/10.1371/journal.pone.0195761.g005

participating laboratories preparing sequencing libraries by ultrasound shearing and enzymatic digestion, respectively (Fig 5B and 5C). The CNV analysis in *BRCA1* gene revealed that two out of 116 (1.7%) ultrasound-sheared samples (from laboratory 1) and five out of other 125 (4%) enzymatically-digested samples (from laboratory 3) were scored as the samples with suspected deletion or duplication. The *BRCA1* MLPA analysis performed in all samples revealed that one suspected sample from each laboratory was true positive (exon 5–14 del in laboratory 1 and exon 8 del in laboratory 3), remaining suspected samples (one from laboratory 1 and four from laboratory 3) were false positive, and 114/116 in laboratory 1 and 120/125 in laboratory 3 were true negative *BRCA1* samples.

While the minimum coverage for a reliable detection of SNVs was estimated at 20X, the minimum coverage required for a reliable detection of CNVs is higher [39]. However, we have noticed that coverage uniformity is at least of the same importance. While the type of the DNA fragmentation protocol (ultrasound vs. enzymatic digestion) did not influence the sensitivity of SNVs detection (Fig 4C), enzymatic digestion caused difficulties in reliable CNVs detection (with an increased number of CNVkit false positives) when comparing samples with the same coverage. We suppose that the main problem of a CNVs coverage-based analysis of enzymatically fragmented samples is worse coverage uniformity caused by non-random DNA cleavage, as discussed above (Fig 1C). To evaluate the sensitivity of CNVs detection in other targeted genes and to better address the influence of DNA fragmentation protocol on the CNV analysis, we compared results of CNVkit analysis in remaining 20 ACMG genes (except *BRCA1* and *TP53* discussed above) covered by CZECANCA target (Fig 6).

The analysis revealed relative low rate of suspected CNVs (0–4 and 0–23 carriers per gene in samples prepared by ultrasound DNA fragmentation and enzymatic DNA digestion, respectively) and demonstrated that preparation of sequencing libraries using ultrasound digestion substantially decreased the need for subsequent MLPA analyses. With the exception of *BRCA2* in which MLPA analysis was performed in all suspected samples, application of MLPA analysis in remaining genes were directed by the phenotype characteristics of analyzed probands. The only CNV identified in remaining ACMG genes was exon 17 deletion in the tuberin (*TSC2*) gene in a patient with typical skin affections. The CNV analysis of the entire set of CZECANCA target genes is provided in S11 Table. The data indicate that deviations of median-normalized CNVkit values in a run of consecutive bin sets could indicate highly probable presence of a large intragenic deletion or duplication (S1 Fig). The extreme case of such situation provides the analysis of genes localized on X chromosome in male and female probands (S2 Fig) that also demonstrates the dynamic range of analysis in detection of real deletion.

For the detection of medium-size insertions and tandem duplications, we added the Pindel tool to the bioinformatics pipeline in order to identify the 64 bp tandem duplication in *BRCA1* (c.5468-11_5520dup64; NM_007294; Chr17: 41197765–41197830 on Assembly GRCh37) not detected by GATK. The sensitivity of a Pindel analysis was recently confirmed by another GATK-omitted variant, the 38 bp duplication in *CHEK2* (c.845_846+36dup38; NM_007194; Chr22: 29105958–29105995 on Assembly GRCh37), confirmed by Sanger sequencing.

Five DNA reference standards (NA12878, NA24149, NA24385, NA24631 and NA24143) with well-described genotypes were analyzed by CZECANCA pipeline to benchmark the overall workflow performance [19]. Comparison between genotypes identified in CZECANCA analysis and available as reference VCFs showed a high concordance in identification of homozygotes and heterozygotes and also high sensitivity, specificity and accuracy of CZECANCA NGS analysis (Fig 7; S9 Table). Totally, 1,722 true positive variants (332–355 per sample), 252 false positive variants (42–57 per sample), and 13 false negative variants (0–5 per sample) were scored in all analyzed DNA reference standards considering 628,069 bases of CZECANCA target region. All were localized in 84 short genomic regions that comprised in majority homopolymeric or repetitive non-coding sequences creating recurrent sequencing errors in currently used sequencing platforms, as indicated by 7/13 not identified (false negative) variants flanking to position of false positive variants. The subsequent manual IGV inspection revealed that the remaining six false negative variants (all indels) were present with allelic fraction below 15% (filtered out through the bioinformatics pipeline).

Finally, an external quality assessment of CZECANCA was performed using the pilot NGS germline mutations scheme provided by the European Molecular Genetics Quality Network (EMQN; www.emqn.org). This external quality assessment showed a 100% sensitivity of variant detection (S10 Table).

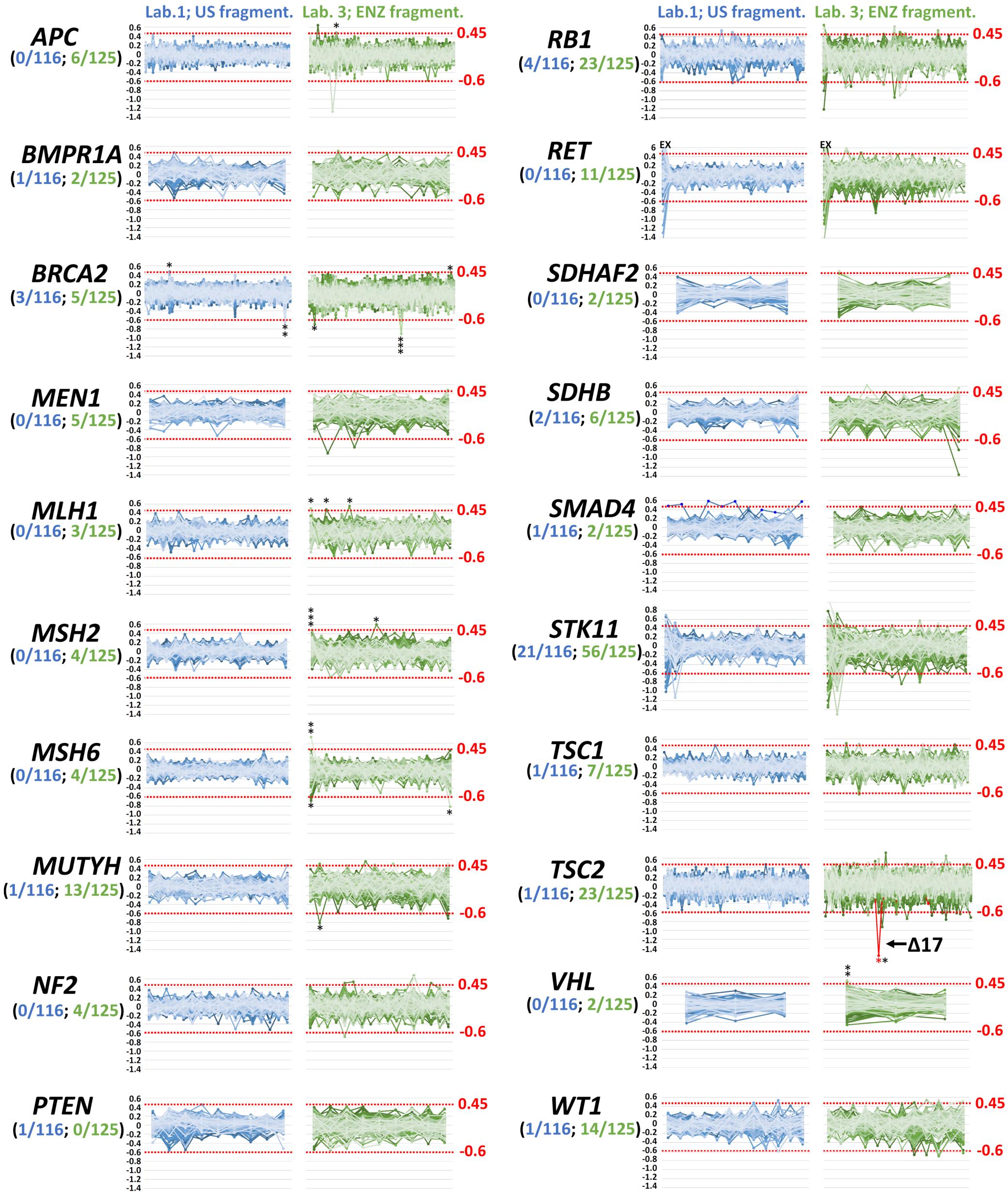
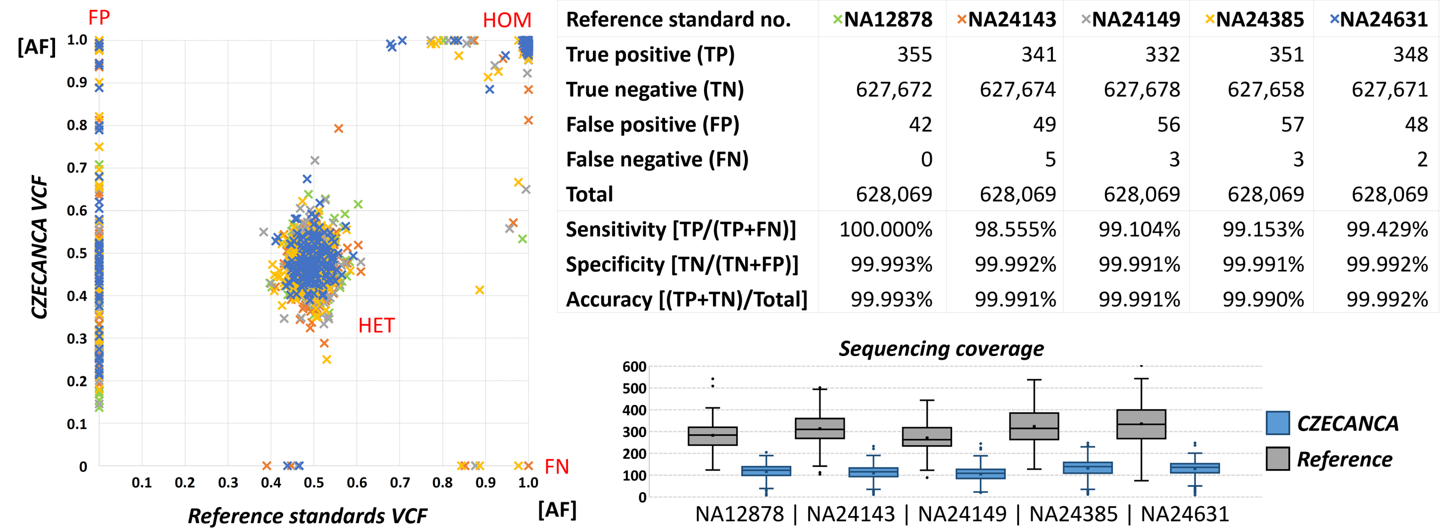


**Fig 6. CNV detection is influenced by a DNA preparation method.** Panels show analyses of remaining ACMG genes (not shown in Fig 5B and 5C) from four runs performed in laboratory 1 (116 DNA samples fragmented by ultrasound) and laboratory 3 (125 DNA samples fragmented enzymatically). The numbers in parentheses express number of samples with possible CNVs from all analyzed samples in contributing laboratories. *indicate samples analyzed by MLPA negatively (FP–black) or positively (TP–red). Bin set covering exon 1 in *RET* was excluded from the analysis due to the large coverage variability.

https://doi.org/10.1371/journal.pone.0195761.g006

## Discussion

Multi-gene panel NGS has changed the genetic landscape for hereditary cancer syndromes. At present, clinical testing prioritizes the use of smaller cancer-specific panels, usually up to 30 cancer susceptibility genes. A large number of panels is available particularly for breast/ovarian and colorectal cancers, which represent frequent diagnoses with a high contribution of genetic components influencing the disease onset, progression and treatment outcomes [40]. Analyses



| Reference standard no. | ×NA12878 | ×NA24143 | ×NA24149 | ×NA24385 | ×NA24631 |
|---|---|---|---|---|---|
| True positive (TP) | 355 | 341 | 332 | 351 | 348 |
| True negative (TN) | 627,672 | 627,674 | 627,678 | 627,658 | 627,671 |
| False positive (FP) | 42 | 49 | 56 | 57 | 48 |
| False negative (FN) | 0 | 5 | 3 | 3 | 2 |
| Total | 628,069 | 628,069 | 628,069 | 628,069 | 628,069 |
| Sensitivity [TP/(TP+FN)] | 100.000% | 98.555% | 99.104% | 99.153% | 99.429% |
| Specificity [TN/(TN+FP)] | 99.993% | 99.992% | 99.991% | 99.991% | 99.992% |
| Accuracy [(TP+TN)/Total] | 99.993% | 99.991% | 99.991% | 99.990% | 99.992% |

**Fig 7. Comparison of variant detection (shown as values of variant allelic fraction; AF) in DNA reference standards** (NA12878, NA24149, NA24385, NA24631 and NA24143) obtained from CZECANCA analysis (x-axis) and AF from VCF files for these standards downloaded from http://jimb.stanford.edu/giab/ (y-axis). The graph shows all variants with GATK quality >100 reached in CZECANCA analysis (including FP variants) and undetected (FN) variants. Heterozygote variants clustered in the center, while homozygote variants in right upper corner. Variant distribution was partially influenced by the differences in mean sequencing coverage targeting 100X and 300X in CZECANCA and DNA reference standards VCFs, respectively. The number of TP, TN, FP, FN, and total number of variant (= CZECANCA target) was used to calculate of sensitivity, specificity, and accuracy of CZECANCA analysis.

https://doi.org/10.1371/journal.pone.0195761.g007

based on smaller panels mainly simplify the clinical interpretation of the identified genotypes with a reduction of incidental findings. While their use is beneficial in clearly indicated patients with typical phenotype characteristics for a given cancer syndrome, the selection of a proper cancer-specific gene panel is not trivial in individuals with less characteristic features (e.g. patients from multi-cancer families). Moreover, our current knowledge of many cancer syndromes is based on the analyses of mostly prototypical cases, the testing criteria are changing dynamically, and the list of cancer predisposition genes with clinical utility is far less complete. Recently, Pearlman et al. analyzed 450 early-onset colorectal cancer patients and showed that a third (24/72) of mutation-positive patients did not meet the established genetic testing criteria for the gene(s) in which they had a mutation [41]. An analysis of mismatch repair (MMR) genes (traditionally linked to hereditary non-polyposis colorectal cancer) in a set of 34,981 cancer patients in a study by Espenschied et al. revealed that out of 528 patients with MMR mutations, 63 (11.9%) had breast cancer only and thus *MSH6* and *PMS2* mutation carriers may manifest with a hereditary breast and ovarian cancer phenotype [42]. In an analysis of *BRCA1* and *BRCA2* in 1,371 unselected breast cancer cohorts, Grindedal et al. showed that common guidelines identified only 45–90% of mutation carriers [43]. The ultimate solution to identify cancer risks would be an analysis of the whole exome (or even better genome) in all cancer patients; however, the implementation of such a strategy is not realistic at present [44]. We suppose that the use of larger multi-cancer panels (containing hundreds of genes) for an analysis of genetic risk in cancer patients is beneficial for several reasons. i) Such an analysis reveals a complex variation landscape of target genes in different cancers [7]. ii) It reveals carriers of concurrent pathogenic mutations and iii) it enables the testing of affected individuals from multi-cancer families with reasonable costs and turnaround time. Finally, iv) combining all genes of interest in a single panel simplifies and unifies laboratory procedures in a single workflow even if testing for different syndromes.

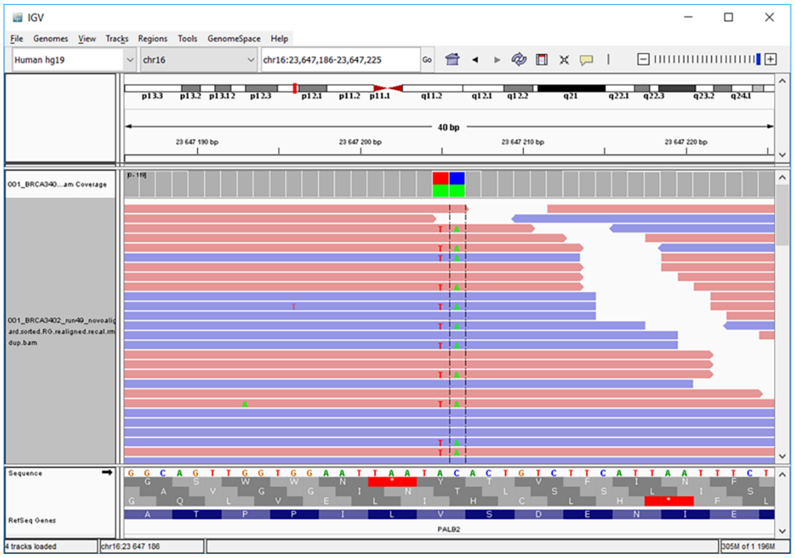
We have developed the custom-designed CZECANCA multi-cancer panel targeting the coding sequence of 219 cancer susceptibility or candidate genes, enabling the identification of a genetic predisposition in the most frequent hereditary cancer syndromes. Besides the established cancer susceptibility genes, we have decided to include also a subset of genes with low, clinically still unconfirmed utility, although their variants cannot be reported until their clinical evidence is known. These genes code for known interactors of established cancer susceptibility gene products, whose mutations may result in a similar phenotypic outcome. However, we suppose that knowledge obtained through the association of the identified genotypes with the phenotypic characteristics of the analyzed patients may substantially accelerate the process of clinical utility evaluation. Moreover, a subsidiary genetic report could be easily generated from the stored data in case of the approval of new cancer susceptibility genes included in CZECANCA. From the technical point of view, a larger genomic target has a favorable impact on panel complexity, improving its coverage uniformity [45].

The validation of the CZECANCA analytic workflow together with the bioinformatics pipeline is necessary for its implementation into routine diagnostics [46]. The presented analytical workflow was optimized for sequencing using MiSeq Illumina, representing the most frequently used NGS platform currently available in diagnostic laboratories. Genetic testing using gene panels is a cost-effective strategy [47]. The material costs for library preparation and sequencing (chemicals, kits, and disposables) using CZECANCA do not exceed €150 per patient in the standard settings (targeting sequencing coverage 100X). The CZECANCA workflow was intended mainly for medium throughput laboratories. As a universal panel, CZECANCA significantly reduces the turnaround time. The sequencing data for 30 analyzed DNA samples in one sequencing MiSeq run might be available in four days (three days for DNA fragmentation and library preparation, depending on hybridization time, and one day for MiSeq sequencing). We are aware that the low-covered or uncovered regions (affecting 12/219 CZECANCA-targeted genes) may require additional effort and time, when requested for genetic assessment.

The validation showed CZECANCA's high sensitivity, specificity, analytical robustness, and accuracy. We have demonstrated that SNVs and small/medium-size indels could be detected with high confidence. Moreover, we have shown that the uniform coverage (targeting to mean 100X coverage) of a target sequence enabled a robust identification of CNVs without the need of routine MLPA, serving as the method for independent CNVs confirmation or exclusion of false positivities. However, despite that the number of false positive calls was low and we detect no false negative sample in ACMG genes, we are aware that with caution needs to be interpreted positive CNV calls in genes for which MLPA assay (or other method) are not routinely available for confirmatory purposes. When required, presence of false positive signals can be reduced by the use of ultrasound fragmentation providing unbiased DNA shearing over enzymatic lysis and/or increased sequencing coverage.

Another advantage of NGS (over Sanger sequencing) is its ability to identify *cis* or *trans* positions of compound, closely localized heterozygous SNVs. For example, the position of double substitution in the *PALB2* gene creating a stop codon (c.661_662delinsTA; p.Val221*; NM_024675), which required further analyses (e.g. PTT) before the NGS era [10], can be identified directly from sequencing reads (Fig 8). The identification of additional pathogenic mutations during the validation procedure in negatively pre-tested samples indicated that a re-analysis is warranted for at least high-risk patients negatively tested by historical analyses based on indirect prescreening methods (e.g. PTT) or cDNA sequencing [48].

CZECANCA (CZEch CAncer paNel for Clinical Application) is intended to unify cancer predisposition testing in the Czech Republic, helping diagnostics laboratories transform the gene-by-gene strategy to NGS, even if is not a population-specific panel *per se*. NGS-based



IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Go
p13.3 p13.2 p13.12 p12.3 p12.1 p11.2 p11.1 q11.2 q12.1 q12.2 q21 q22.1 q22.3 q23.2 q24.1
40 bp
23 647 190 bp
23 647 200 bp
23 647 210 bp
23 647 220 bp
[0 - 119]
001_BRCA340...am Coverage
001_BRCA3402_run49_novoalig
ard.sorted.RG.realigned.recal.rm
dup.bam
T
A
Sequence
RefSeq Genes
G G C A G T T G G T G G A A T T A A T A C A C T G T C T T C A T T A A T T T C T
PALB2

Human hg19

chr16

chr16:23,647,186-23,647,225

4 tracks loaded

chr16:23 647 186

305M of 1 196M

**Fig 8. Identification of c.661_662delinsTA double substitution (p.Val221*) in *PALB2* (NM_024675).** The BAM file displayed in IGV shows the *cis*-position of both substitutions in approximately 50% of forward (pink bars) and reverse (blue bars) reads, respectively.

https://doi.org/10.1371/journal.pone.0195761.g008

technologies bring new challenges including technological aspects, bioinformatics processing, the management of large datasets, and clinical interpretation of results [46]. The use of a uniform analytical and bioinformatics approach improves the identification of technical and platform-specific sequencing errors, as we demonstrated in inter-run and intra-run comparisons. Moreover, validation of the panel using reference standard DNA samples with known genotypes enabled identification of genomic loci (dominantly homopolymeric regions) providing these recurrent sequencing errors, which could be subsequently easily eliminated by bioinformatics. The use of CZECANCA will help generate a global view of constitutional variants from the perspective of known cancer predisposition and candidate genes in the population. Simultaneously with the sequencing of cancer patients, we aim to sequence non-cancer controls in order to identify and establish the frequency of population-specific neutral variants. The introduction of patients' and control genotypes with associated phenotypes into a nationwide database currently being created will simplify the interpretation of variants, which remains the main challenge at present. In general, NGS-based analyses result in an increased number of incidental findings or variants of unknown significance. The patient must be informed about this possibility before the testing and must have the opt in / opt out possibility clearly formulated in the informed consent. Consensus on what incidental information should be disclosed has yet to be reached. Currently, there is general agreement on reporting mutations in known high-penetrant genes in patients with a typical personal and family cancer history [38]. However, there is no agreement on pathogenic mutations in genes with lower penetrance or on mutations related to autosomal-recessive syndromes. These questions are currently being tackled in cooperating centers on a rather individual basis, depending on the formulation of the informed consents obtained, and on the clinical experience of the indicating geneticists [49].

In conclusion, CZECANCA allows comprehensive testing for a majority of frequent hereditary cancer syndromes while mitigating potential difficulties of incidental findings in non-cancer genes as seen in exome or genome sequencing. The reliability of the procedure enables an unbiased identification of variants present in patients, which together with a correct interpretation of variants is key for the effective management of hereditary cancer patients and their relatives.

## Supporting information

**S1 Table. List of 219 CZECANCA targeted genes with basic characteristics of their protein products.** The primary gene target for the probe coverage was represented by coding sequences (cds) representing all exons (in case of known cancer susceptibility genes) or all coding exons (in other genes), including 10 bases from adjacent intronic regions. The promoter regions of the *BRCA1* and *BRCA2* genes were included into the primary target. Because of the strict design conditions, some clinically important regions were left untargeted (highlighted) for technical reasons such as repeats and homologous regions. (The characteristics of protein products were obtained from string.embl.de and/or genecards.org).
(XLSX)

**S2 Table. Regions of interest with low coverage ≤20X.** The average coverage is the mean from 10 randomly selected samples.
(XLSX)

**S3 Table. Comparison of identified variants in the targeted exonic regions and 12 bp from adjacent introns with GATK quality >100 in three intra-run replicates of sample #2268.** The DNA sample pooled for the enrichment in amounts corresponding to 33 ng (e.g. 1/30; considered as 100%), 75% and 50% of this amount, respectively. (Cov = coverage; Q = quality; discordant variants are highlighted).
(XLSX)

**S4 Table. Comparison of identified variants in the targeted exonic regions and 12 bp from adjacent introns with GATK quality >100 in two independent run replicates of sample #3647.** All values of coverages (Cov) of sample #3647 in run 14 were corrected by a factor of 1.3880 to normalize coverages between samples for presentation in Fig 4B. (Q = quality; discordant variants are highlighted).
(XLSX)

**S5 Table. Comparison of identified variants in the targeted exonic regions and 12 bp from adjacent introns with GATK quality >100 in sample #3582 analyzed independently in four participating laboratories(Lab).** All values of coverages (Cov) in Lab2, Lab3, and Lab4 were corrected to the coverage of Lab1 by a factor shown in line 336 to normalize coverages between samples for Fig 4C. (discordant variants are highlighted).
(XLSX)

**S6 Table. List of variants used for the validation of SNVs detection.**
(XLSX)

**S7 Table. List of CNVs used for the validation of a large genomic rearrangements analysis.**
(XLSX)

**S8 Table. CNV scores (from CNVkit software) of bins in *BRCA1*, *PALB2*, *CHEK2*, and *TP53*.** The numbers of samples with previously characterized CNVs are highlighted in red. The table show raw values obtained from CNVkit as well as median-normalized values. The normalized values >0.5 (highlighted in green) were indicative for the presence of a duplication, while values <-0.6 (highlighted in yellow) were indicative for a deletion. Data from this table were used for creation of Fig 5.
(XLSX)

**S9 Table. Variants identified in five Coriell Institute reference samples sequenced using CZECANCA pipeline and their comparison with VCF files obtained from GIAB website.**

The considered targeted region encompasses 628,069 bases of CZECANCA target region. False negative variants are highlighted.
(XLSX)

**S10 Table. Variant consensus analysis report from EMQN (NGS pilot 2016) for CZENCANCA sequencing of a reference sample.**
(XLSX)

**S11 Table. Results of CNV analysis performed in two validation sets consisting of four runs from Laboratory 1 (116 samples prepared using the ultrasound DNA fragmentation on Covaris) and four runs from Laboratory 3 (125 other samples prepared using the enzymatic DNA cleavage by Fragmentase).** To estimate number of false positive (FP) and false negative (FN) samples, data for CNV analysis of Coriell Institute reference samples (Coriell; 10 samples analyzed in Laboratory 1 and prepared using the ultrasound DNA fragmentation on Covaris) were added. The values in cells represent differences of CNV scores for a given cell (i.e. sample in the coordinate) from the median value of signals from particular sample group (i.e. Coriell—columns Q-Z, Laboratory 1—columns AB-EM, Laboratory 3—columns EO-JI) in a given CNVkit_bin_set_coordinate (column A). Values in cells showing individual analyzed samples from particular sample group exceeding the given CNVkit threshold value for deletion (<-0,6) and duplication (>0,45) are highlighted as red and green cells, respectively. The columns C-O provide several aggregated metrics, that include number of individual samples in which deletion (columns G-I), duplication (J-L), or deletion+duplication (M-O) was found in a given coordinate in particular sample group. Columns C-E enable identification of non-informative bin sets with suspected false positive (FP) signals (indicated by the value = 1) that include regions on X chromosome called in male samples as deletions (highlighted in blue in column B), regions with insufficient coverage or containing pseudogenes (highlighted in orange and yellow, respectively; in column B), or bin sets containing the improbable number of deletions+duplications exceeding the 4% of analyzed samples in a particular sample group.
(XLSX)

**S1 Fig. Run of consecutive bin set coordinates with values indicating a deletion (< -0.6; red) or a duplication (> 0.45; green) increases the probability of a real rearrangement.** The *BRCA1* and *BRIP1* deletions were confirmed by MLPA analyses, which are currently no available for confirmation of secondary findings in *MSR1* or *ZNF350*. (The graphs expressed normalized CNVkit values shown in S11 Table).
(TIF)

**S2 Fig. CNV analysis of genes *BRCC3*, *FANCB*, *GPC3*, and *UBE2A* localized on X chromosome enabled to demonstrate differences in normalized CNVkit values in samples carrying a real 'deletion' in samples prepared by ultrasound DNA fragmentation or enzymatic DNA lysis.** The XX and X indicates areas of samples obtained from female and male probands, respectively. (The graphs expressed normalized CNVkit values shown in S11 Table). Upper panel shows normalized CNVkit values in 116 samples analyzed in four runs in laboratory 1. Lower panel shows normalized CNVkit values in 125 other samples analyzed in four runs in laboratory 3.
(TIF)

## Acknowledgments

We would like to thank Jaroslav Vohanka and Xavier Miro (Roche) for their valuable advice in probe design, Jana Chrudimska for technical assistance with sequencing, and Jan Flemr for language editing.

## Author Contributions

**Conceptualization:** Jana Soukupova, Marketa Janatova, Zdenek Kleibl.

**Data curation:** Jana Soukupova, Petra Zemankova, Viktor Stranecky, Michal Vocka, Zdenek Kleibl.

**Formal analysis:** Jana Soukupova, Marketa Janatova, Lenka Foretova, Petra Kleiblova, Michal Vocka, Zdenek Kleibl.

**Funding acquisition:** Jana Soukupova, Lenka Foretova, Stanislav Kmoch, Zdenek Kleibl.

**Investigation:** Jana Soukupova, Petra Zemankova, Klara Lhotova, Marketa Janatova, Marianna Borecka, Lenka Stolarova, Filip Lhota, Eva Machackova, Spiros Tavandzis, Petra Kleiblova, Michal Vocka.

**Methodology:** Jana Soukupova, Petra Zemankova, Klara Lhotova, Marketa Janatova, Lenka Foretova, Viktor Stranecky, Petra Kleiblova, Hana Hartmannova, Katerina Hodanova, Stanislav Kmoch, Zdenek Kleibl.

**Project administration:** Jana Soukupova.

**Resources:** Michal Vocka.

**Software:** Petra Zemankova, Viktor Stranecky.

**Supervision:** Jana Soukupova, Marketa Janatova, Lenka Foretova, Viktor Stranecky, Hana Hartmannova, Katerina Hodanova, Stanislav Kmoch, Zdenek Kleibl.

**Validation:** Jana Soukupova, Petra Zemankova, Klara Lhotova, Marianna Borecka, Lenka Stolarova, Filip Lhota, Eva Machackova, Spiros Tavandzis.

**Visualization:** Petra Zemankova, Zdenek Kleibl.

**Writing – original draft:** Jana Soukupova, Zdenek Kleibl.

**Writing – review & editing:** Jana Soukupova, Petra Zemankova, Klara Lhotova, Marketa Janatova, Marianna Borecka, Lenka Stolarova, Filip Lhota, Lenka Foretova, Eva Machackova, Viktor Stranecky, Spiros Tavandzis, Petra Kleiblova, Michal Vocka, Hana Hartmannova, Katerina Hodanova, Stanislav Kmoch, Zdenek Kleibl.

## References

1. Kulkarni A, Carley H. Advances in the recognition and management of hereditary cancer. Br Med Bull. 2016; 120(1):123–38. https://doi.org/10.1093/bmb/ldw046 PMID: 27941041
2. Stoffel EM, Cooney KA. Advances in inherited cancers: Introduction. Semin Oncol. 2016; 43(5):527. https://doi.org/10.1053/j.seminoncol.2016.09.003 PMID: 27899182
3. Rahman N. Mainstreaming genetic testing of cancer predisposition genes. Clin Med. 2014; 14(4):436–9. https://doi.org/10.7861/clinmedicine.14-4-436 PMID: 25099850
4. Rahman N. Realizing the promise of cancer predisposition genes. Nature. 2014; 505(7483):302–8. https://doi.org/10.1038/nature12981 PMID: 24429628
5. Foulkes WD. Inherited susceptibility to common cancers. N Engl J Med. 2008; 359(20):2143–53. https://doi.org/10.1056/NEJMra0802968 PMID: 19005198
6. Feero WG. Clinical application of whole-genome sequencing: proceed with care. JAMA. 2014; 311(10):1017–9. https://doi.org/10.1001/jama.2014.1718 PMID: 24618961
7. Shah PD, Nathanson KL. Application of Panel-Based Tests for Inherited Risk of Cancer. Annu Rev Genomics Hum Genet. 2017; 18(1):201–27. https://doi.org/10.1146/annurev-genom-091416-035305 PMID: 28504904

8. Pohlreich P, Stribrna J, Kleibl Z, Zikan M, Kalbacova R, Petruzelka L, et al. Mutations of the BRCA1 gene in hereditary breast and ovarian cancer in the Czech Republic. Med Princ Pract. 2003; 12(1):23–9. https://doi.org/10.1159/000068163 PMID: 12566964

9. Pohlreich P, Zikan M, Stribrna J, Kleibl Z, Janatova M, Kotlas J, et al. High proportion of recurrent germline mutations in the BRCA1 gene in breast and ovarian cancer patients from the Prague area. Breast Cancer Res. 2005; 7(5):R728–R36. https://doi.org/10.1186/bcr1282 PMID: 16168118

10. Janatova M, Kleibl Z, Stribrna J, Panczak A, Vesela K, Zimovjanova M, et al. The PALB2 Gene Is a Strong Candidate for Clinical Testing in BRCA1- and BRCA2-Negative Hereditary Breast Cancer. Cancer Epidemiol Biomarkers Prev. 2013; 22(12):2323–32. https://doi.org/10.1158/1055-9965.EPI-13-0745-T PMID: 24136930

11. Kleibl Z, Havranek O, Hlavata I, Novotny J, Sevcik J, Pohlreich P, et al. The CHEK2 gene I157T mutation and other alterations in its proximity increase the risk of sporadic colorectal cancer in the Czech population. Eur J Cancer. 2009; 45(4):618–24. https://doi.org/10.1016/j.ejca.2008.09.022 PMID: 18996005

12. Soukupova J, Dundr P, Kleibl Z, Pohlreich P. Contribution of mutations in ATM to breast cancer development in the Czech population. Oncol Rep. 2008; 19(6):1505–10. PMID: 18497957

13. Borecka M, Zemankova P, Vocka M, Soucek P, Soukupova J, Kleiblova P, et al. Mutation analysis of the PALB2 gene in unselected pancreatic cancer patients in the Czech Republic. Cancer Genet. 2016; 209(5):199–204. https://doi.org/10.1016/j.cancergen.2016.03.003 PMID: 27106063

14. Kleibl Z, Fidlerova J, Kleiblova P, Kormunda S, Bilek M, Bouskova K, et al. Influence of dihydropyrimidine dehydrogenase gene (DPYD) coding sequence variants on the development of fluoropyrimidine-related toxicity in patients with high-grade toxicity and patients with excellent tolerance of fluoropyrimidine-based chemotherapy. Neoplasma. 2009; 56(4):303–16. https://doi.org/10.4149/neo_2009_04_303 PMID: 19473056

15. Kleiblova P, Shaltiel IA, Benada J, Sevcik J, Pechackova S, Pohlreich P, et al. Gain-of-function mutations of PPM1D/Wip1 impair the p53-dependent G1 checkpoint. J Cell Biol. 2013; 201(4):511–21. https://doi.org/10.1083/jcb.201210031 PMID: 23649806

16. Janatova M, Soukupova J, Stribrna J, Kleiblova P, Vocka M, Boudova P, et al. Mutation Analysis of the RAD51C and RAD51D Genes in High-Risk Ovarian Cancer Patients and Families from the Czech Republic. PLoS One. 2015; 10(6):e0127711. https://doi.org/10.1371/journal.pone.0127711 PMID: 26057125

17. Ticha I, Kleibl Z, Stribrna J, Kotlas J, Zimovjanova M, Mateju M, et al. Screening for genomic rearrangements in BRCA1 and BRCA2 genes in Czech high-risk breast/ovarian cancer patients: high proportion of population specific alterations in BRCA1 gene. Breast Cancer Res Treat. 2010; 124(2):337–47. https://doi.org/10.1007/s10549-010-0745-y PMID: 20135348

18. Havranek O, Kleiblova P, Hojny J, Lhota F, Soucek P, Trneny M, et al. Association of Germline CHEK2 Gene Variants with Risk and Prognosis of Non-Hodgkin Lymphoma. PLoS One. 2015; 10(10): e0140819. https://doi.org/10.1371/journal.pone.0140819 PMID: 26506619

19. Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. Nat Rev Genet. 2017; 18(8):473–84. https://doi.org/10.1038/nrg.2017.44 PMID: 28626224

20. Lhota F, Zemankova P, Kleiblova P, Soukupova J, Vocka M, Stranecky V, et al. Hereditary truncating mutations of DNA repair and other genes in BRCA1/BRCA2/PALB2-negatively tested breast cancer patients. Clin Genet. 2016. https://doi.org/10.1111/cge.12748 PMID: 26822949

21. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38(16):e164. https://doi.org/10.1093/nar/gkq603 PMID: 20601685

22. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–303. https://doi.org/10.1101/gr.107524.110 PMID: 20644199

23. Das R, Ghosh SK. Genetic variants of the DNA repair genes from Exome Aggregation Consortium (EXAC) database: significance in cancer. DNA Repair (Amst). 2017; 52:92–102. https://doi.org/10.1016/j.dnarep.2017.02.013 PMID: 28259467

24. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422):56–65. https://doi.org/10.1038/nature11632 PMID: 23128226

25. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014; 42(Database issue):D980–5. https://doi.org/10.1093/nar/gkt1113 PMID: 24234437

26. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009; 4(7):1073–81. https://doi.org/10.1038/nprot.2009.86 PMID: 19561590

27. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Human Mutat. 2008; 29(1):6–13. https://doi.org/10.1002/humu.20654 PMID: 18000842

28. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010; 7(8):575–6. https://doi.org/10.1038/nmeth0810-575 PMID: 20676075

29. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009; 19(9):1553–61. https://doi.org/10.1101/gr.092619.109 PMID: 19602639

30. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7(4):248–9. https://doi.org/10.1038/nmeth0410-248 PMID: 20354512

31. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010; 20(1):110–21. https://doi.org/10.1101/gr.097857.109 PMID: 19858363

32. Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005; 15(7):901–13. https://doi.org/10.1101/gr.3577405 PMID: 15965027

33. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46(3):310–5. https://doi.org/10.1038/ng.2892 PMID: 24487276

34. Machackova E, Hazova J, Stahlova Hrabincova E, Vasickova P, Navratilova M, Svoboda M, et al. [Retrospective NGS Study in High-risk Hereditary Cancer Patients at Masaryk Memorial Cancer Institute]. Klin Onkol. 2016; 29 Suppl 1:S35–45. PMID: 26691941.

35. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013; 14(2):178–92. https://doi.org/10.1093/bib/bbs017 PMID: 22517427

36. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009; 25 (21):2865–71. https://doi.org/10.1093/bioinformatics/btp394 PMID: 19561018

37. 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. https://doi.org/10.1038/gim.2015.30 PMID: 25741868

38. Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet Med. 2017; 19(2):249–55. https://doi.org/10.1038/gim.2016.190 PMID: 27854360

39. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 2013; 14 (11):S1. https://doi.org/10.1186/1471-2105-14-s11-s1 PMID: 24564169

40. Easton DF, Pharoah PD, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, et al. Gene-panel sequencing and the prediction of breast-cancer risk. New Engl J Med. 2015; 372(23):2243–57. https://doi.org/10.1056/NEJMsr1501341 PMID: 26014596

41. Pearlman R, Frankel WL, Swanson B, et al. Prevalence and spectrum of germline cancer susceptibility gene mutations among patients with early-onset colorectal cancer. JAMA Oncol. 2017; 3(4):464–71. https://doi.org/10.1001/jamaoncol.2016.5194 PMID: 27978560

42. Espenschied CR, LaDuca H, Li S, McFarland R, Gau C-L, Hampel H. Multigene Panel Testing Provides a New Perspective on Lynch Syndrome. J Clin Oncol. 2017; 35(22):2568–75. https://doi.org/10.1200/JCO.2016.71.9260 PMID: 28514183

43. Grindedal EM, Heramb C, Karsrud I, Ariansen SL, Maehle L, Undlien DE, et al. Current guidelines for BRCA testing of breast cancer patients are insufficient to detect all mutation carriers. BMC cancer. 2017; 17(1):438. https://doi.org/10.1186/s12885-017-3422-2 PMID: 28637432

44. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? J Med Genet. 2011; 48(9):580–9. https://doi.org/10.1136/jmedgenet-2011-100223 PMID: 21730106

**45.** Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014; 15(2):121–32. https://doi.org/10.1038/nrg3642 PMID: 24434847

**46.** Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, et al. Guidelines for diagnostic next-generation sequencing. Eur J Human Genet. 2016; 24(1):2–5. https://doi.org/10.1038/ejhg.2015.226 PMID: 26508566

**47.** Azimi M, Schmaus K, Greger V, Neitzel D, Rochelle R, Dinh T. Carrier screening by next-generation sequencing: health benefits and cost effectiveness. Mol Genet Genomic Med. 2016; 4(3):292–302. https://doi.org/10.1002/mgg3.204 PMID: 27247957

**48.** Moran O, Nikitina D, Royer R, Poll A, Metcalfe K, Narod SA, et al. Revisiting breast cancer patients who previously tested negative for BRCA mutations using a 12-gene panel. Breast Cancer Res Treat. 2017; 161(1):135–42. https://doi.org/10.1007/s10549-016-4038-y PMID: 27798748

**49.** Soukupová J, Zemánková P, Kleiblová P, Janatová M, Kleibl Z. [CZECANCA: CZEch CAncer paNel for Clinical Application—Design and Optimization of the Targeted Sequencing Panel for the Identification of Cancer Susceptibility in High-risk Individuals from the Czech Republic]. Klin Onkol. 2016; 29 Suppl 1: S46–54. Czech. PMID: 26691942.

IJC
International Journal of Cancer

# Identification of deleterious germline *CHEK2* mutations and their association with breast and ovarian cancer

Petra Kleiblova[1,2]*, Lenka Stolarova[1]*, Katerina Krizova[3], Filip Lhota[1], Jan Hojny[1], Petra Zemankova[1], Ondrej Havranek[4,5], Michal Vocka[6], Marta Cerna[1], Klara Lhotova[1], Marianna Borecka[1], Marketa Janatova[1], Jana Soukupova[1], Jan Sevcik[1], Martina Zimovjanova[6], Jaroslav Kotlas[2], Ales Panczak[2], Kamila Vesela[2], Jana Cervenkova[7], Michaela Schneiderova[8], Monika Burocziova[3], Kamila Burdova[3], Viktor Stranecky[9], Lenka Foretova[10], Eva Machackova[10], Spiros Tavandzis[11], Stanislav Kmoch[9], Libor Macurek[3] and Zdenek Kleibl[1]

[1] Institute of Biochemistry and Experimental Oncology, First Faculty of Medicine, Charles University, Prague, Czech Republic
[2] Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic
[3] Laboratory of Cancer Cell Biology, Institute of Molecular Genetics of the ASCR, Prague, Czech Republic
[4] BIOCEV, First Faculty of Medicine, Charles University, Prague, Czech Republic
[5] Department of Hematology, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic
[6] Department of Oncology, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic
[7] Department of Radiology, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic
[8] First Department of Surgery, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic
[9] Research Unit for Rare Diseases, Department of Pediatrics and Adolescent Medicine, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic
[10] Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, Czech Republic
[11] Department of Medical Genetics, AGEL Laboratories, AGEL Research and Training Institute, Novy Jicin, Czech Republic

**Germline mutations in checkpoint kinase 2 (*CHEK2*), a multiple cancer-predisposing gene, increase breast cancer (BC) risk; however, risk estimates differ substantially in published studies. We analyzed germline *CHEK2* variants in 1,928 high-risk Czech breast/ovarian cancer (BC/OC) patients and 3,360 population-matched controls (PMCs). For a functional classification of VUS, we developed a complementation assay in human nontransformed RPE1-*CHEK2*-knockout cells quantifying CHK2-specific phosphorylation of endogenous protein KAP1. We identified 10 truncations in 46 (2.39%) patients and in 11 (0.33%) PMC ($p = 1.1 \times 10^{-14}$). Two types of large intragenic rearrangements (LGR) were found in 20/46 mutation carriers. Truncations significantly increased unilateral BC risk (OR = 7.94; 95%CI 3.90–17.47; $p = 1.1 \times 10^{-14}$) and were more frequent in patients with bilateral BC (4/149; 2.68%; $p = 0.003$), double primary BC/OC (3/79; 3.80%; $p = 0.004$), male BC (3/48; 6.25%; $p = 8.6 \times 10^{-4}$), but not with OC (3/354; 0.85%; $p = 0.14$). Additionally, we found 26 missense VUS in 88 (4.56%) patients and 131 (3.90%) PMC ($p = 0.22$). Using our functional assay, 11 variants identified in 15 (0.78%) patients and 6 (0.18%) PMC were scored deleterious ($p = 0.002$). Frequencies of functionally intermediate and neutral variants did not differ between patients**

**Key words:** breast cancer, ovarian cancer, germline mutations, CHEK2, VUS, KAP1, functional assay

**Abbreviations:** BC: breast cancer; CHEK2: checkpoint kinase 2; CI: confidence interval; CNV: copy number variant(s); CPG: cancer-predisposing genes; CZECANCA: CZEch CAncer paNel for Clinical Application; DHPLC: denaturing high performance liquid chromatography; EGFP: enhanced green fluorescent protein; ExAC: exome aggregation consortium; HRMA: high-resolution melting analysis; IGV: integrative genomics viewer; KAP1: KRAB-associated protein 1; KO: knockout; LGR: large genomic rearrangement(s); MLPA: multiplex ligation-dependent probe amplification; NFE: non-Finnish European; NGS: next-generation sequencing; OC: ovarian cancer; OR: odds ratio; PMC: population-matched controls; RPE1: retinal pigmented epithelial cells; SNV: single nucleotide variant

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

**Conflict of interest:** The authors declare no potential conflicts of interest.

*P.K. and L.S. authors contributed equally to this work

**Grant sponsor:** Akademie Věd České Republiky; **Grant sponsor:** Strategie AV21, Qualitas; **Grant sponsor:** Grantová Agentura, Univerzita Karlova; **Grant numbers:** GAUK 762216, PRIMUS/17/MED/9, PROGRES Q26/LF1, PROGRES Q28/LF1, SVV2019/260367, UNCE/MED/016; **Grant sponsor:** Ministerstvo Školství, Mládeže a Tělovýchovy; **Grant numbers:** NPU II LQ1604, LM2015091; **Grant sponsor:** Ministerstvo Zdravotnictví Ceské Republiky; **Grant numbers:** NV15-27695A, NV15-28830A, NV16-29959A, NV19-03-00279

**DOI:** 10.1002/ijc.32385

**History:** Received 6 Feb 2019; Accepted 24 Apr 2019; Online 3 May 2019

**Correspondence to:** Zdenek Kleibl, Institute of Biochemistry and Experimental Oncology, First Faculty of Medicine, Charles University, U Nemocnice 5, 120 00 Prague 2, Czech Republic, E-mail: zdekleje@lf1.cuni.cz; or Libor Macurek, Institute of Molecular Genetics of the ASCR, Videnska 1083, 142 20 Prague 4, Czech Republic, E-mail: libor.macurek@img.cas.cz

and PMC. Functionally deleterious *CHEK2* missense variants significantly increased BC risk (OR = 3.90; 95%CI 1.24–13.35; $p = 0.009$) and marginally OC risk (OR = 4.77; 95%CI 0.77–22.47; $p = 0.047$); however, carriers low frequency will require evaluation in larger studies. Our study highlights importance of LGR detection for *CHEK2* analysis, careful consideration of ethnicity in both cases and controls for risk estimates, and demonstrates promising potential of newly developed human nontransformed cell line assay for functional *CHEK2* VUS classification.

**What's new?**

The tumor suppressor gene checkpoint kinase 2 (*CHEK2*) encodes a protein that serves an important role in DNA repair. However, *CHEK2* is also vulnerable to mutations that potentially impact breast cancer risk. Using a functional cell-based assay, the authors of the present study show that truncating and missense *CHEK2* variants are associated with risk of both breast and ovarian cancer. One-third of truncating mutations involved large genomic rearrangements. In addition, *CHEK2* mutations predisposed women to specific breast cancer types, and *CHEK2* mutation carriers with a family history of cancer were at increased risk of developing second primary cancers.

## Introduction

Approximately 10% of breast cancer (BC) and 20% of ovarian cancer (OC) cases arise as a hereditary disease in patients carrying a pathogenic mutation in BC/OC-predisposing genes.[1,2] The clinical utility of pathogenic mutations in major BC/OC genes (*BRCA1* and *BRCA2*) is well established but it remains less certain for a growing group of cancer-predisposing genes (CPG) whose germline mutations confer a moderate cancer risk (*ATM*, *CHEK2*, *PALB2*).[3] This problem is becoming even more critical with the introduction of multigene panel next-generation sequencing (NGS) into the routine genetic analysis of high-risk BC/OC individuals.[4]

Germline *CHEK2* mutations have been linked with susceptibility to several malignancies including BC.[5] The *CHEK2* gene codes for serine/threonine CHK2 kinase involved in DNA damage response (DDR). Activated by a DNA lesion, ATM kinase catalyzes CHK2 T68 phosphorylation promoting CHK2 homodimerization through its forkhead-associated domains and kinase domain autophosphorylation.[6,7] Activated CHK2 phosphorylates multiple proteins involved in DNA repair and DDR, including BRCA1/BRCA2 and p53.[8,9] Another CHK2 substrate is KRAB-associated protein 1 (KAP1, alias TIF1β, TRIM28) a universal corepressor required for transcriptional repression mediated by the KRAB protein superfamily. CHK2-mediated KAP1 S473 phosphorylation reduces its transcription repression resulting in wide effects on gene expression.[10] Although the role of the ATM–CHK2–p53 pathway in the DNA damage-induced cell cycle checkpoint is redundant, CHK2 participates in p53-dependent cell death.[11–14]

The association of germline *CHEK2* variants with BC was assessed early in studies genotyping European founder mutations including the truncating mutation c.1100delC and the missense variant c.470T>C (p.I157T).[5] Subsequent meta-analyses demonstrated that while c.1100delC represents a moderate-risk variant for unselected (OR = 2.7; 95% confidence interval [CI] 2.1–3.4), early onset (OR = 2.6; 95%CI 1.3–5.5) and familial BC (OR = 4.8; 95% CI 3.3–7.2),[15] p.I157T is a low-risk variant with OR <1.5 for all BC subgroups.[16] Other founder variants include the spliceogenic mutation c.444+1G>A (IVS2+1G>A) and a large genomic rearrangement (LGR) with exon 9–10 deletion (c.909-2028_1095+330del5395) identified in Slavic populations,[17] and the Ashkenazi Jewish founder missense mutation c.1283C>T (p.S428F).[18]

Only few early studies analyzed the entire *CHEK2* coding sequence and revealed that c.1100delC and p.I157T represent only a fraction of *CHEK2* variants in BC patients.[19–22] Recent panel NGS analyses in large cohorts have shown that the *CHEK2* mutation rate is one of the highest among non-*BRCA1/BRCA2* genes in BC in individuals of Ashkenazi Jewish or European ancestry.[23–26] However, the classification of most missense variants remains uncertain,[27] their assessment is problematic,[4] and nearly one-third of *CHEK2* variants are reported discordantly.[28]

In contrast to BC, the association of *CHEK2* germline variants with OC risk is disputable. While several case–control studies have not significantly associated the c.1100delC mutation with OC development,[29,30] recent panel NGS analyses in 4,439 and 6,001 OC samples from the US identified *CHEK2* as the third most frequently affected susceptibility gene.[31,32]

In our study, we identified germline *CHEK2* variants in 1,928 high-risk BC/OC patients and 3,360 population-matched controls (PMCs). Subsequently, we have developed a cell-based assay utilizing a human RPE1 cell line model with endogenous *CHEK2* knockout to functionally classify the identified variants of unknown significance (VUS). This strategy enabled us to identify deleterious germline *CHEK2* mutations, to evaluate cancer risk in their carriers and to describe the clinical and histopathological characteristics of breast tumors in mutation carriers.

## Methods

Detailed information is provided in Supporting Information Methods.

### Subjects

The patient group included 1,928 BC/OC patients (herein denoted as *all patients*) referred by clinical geneticists for a CPG-mutation analysis performed at the Laboratory of Oncogenetics, First Faculty of Medicine, Charles University, in 1997–2017. Overall, 424/1,928 patients carried a mutation in other (i.e., non-*CHEK2*) cancer-predisposing gene for BC (*BRCA1*, *BRCA2*, *PALB2*, *TP53*) or OC (*BRCA1*, *BRCA2*, *RAD51C*, *RAD51D*, *MLH1*, *MSH2*, *MSH6*) and were denoted herein as *other CPG-mutated*. Remaining 1,504/1,928 patients were negative for mutations in aforementioned genes (herein denoted as *other CPG-wt*). All participants signed an informed consent approved by the local ethical committee. Clinical and histopathological data (Supporting Information Table S1) were obtained during genetic counseling or retrieved from the patients' records.

The set of 3,360 adult PMCs comprised 720 samples of noncancer individuals, 369 samples of adult blood donors, 609 noncancer controls aged >60 years without cancer in first-degree relatives and 1,662 individuals analyzed by exome sequencing at the National Center for Medical Genomics (http://ncmg.cz). In total, PMC set included 1,593 female (with median age 66 years, range 20–98 years) and 1,767 male (with median age 60 years, range 18–94 years) controls. All patients and controls were Caucasians, of the Czech origin.

### Mutation analyses

Until 2015, mutation analyses of the entire *CHEK2* coding sequence in BC patients were performed by a high-resolution melting analysis (HRMA) of all coding exons. LGRs were analyzed by a multiplex ligation-dependent probe amplification (MLPA), as described previously.[33] All OC patients' samples, samples from BC patients enrolled since 2015, and samples from all identified *CHEK2* variant carriers were analyzed by a CZECANCA panel (CZEch CAncer paNel for Clinical Application; custom-made SeqCap EZ choice panel, Roche) targeting 219 genes with MiSeq (Illumina) NGS as described recently.[34] The coverage uniformity enabled to evaluate CNVs at 100× average coverage. *CHEK2* variants identified in patients were also sequenced at the mRNA (cDNA) level to determine a potential impact on splicing. NGS-analysis performed in 2,271/3,360 (67.6%) PMC samples (609 noncancer controls and 1,662 NCMG controls) included SNV/indels and CNV analyses. In remaining 1,089/3,360 (32.4%) PMC samples (720 noncancer individuals and 369 blood donors), entire *CHEK2* coding sequence was analyzed by HRMA, similarly as in patients and mutation-specific PCR/HRMA was used for identification of two *CHEK2* LGRs identified in our population (see Supporting Information Methods for details). The consequences of the identified missense variants were predicted by *in silico* tools: Align-GVGD, MutationTaster, CADD, SIFT, PolyPhen-2, Spidex and GERP.

### Cell lines

To generate RPE1-*CHEK2*-KO cells, hTERT-RPE1 cells were transfected with a *CHEK2*-CRISPR/Cas9-KO plasmid (Santa Cruz Biotechnology, Santa Cruz, CA; sc-400,438) and a *CHEK2*-HDR plasmid (1:1) and selected by puromycin (7.5 μg/ml) for 3 weeks. The integration of an HDR cassette into the *CHEK2* locus was confirmed by sequencing and a loss of CHK2 expression by immunoblotting (all used antibodies are described in Supporting Information Methods). To remove the HDR cassette, cells were transfected with Cre vector (Santa Cruz, sc-418,923) and RFP-negative cells were selected by flow cytometry. For stable complementation of CHK2, RPE1-*CHEK2*-KO cells were transfected with a linearized pcDNA4-EGFP-*CHEK2* plasmid, selected with zeocin for 3 weeks and single clones were expanded. Plasmid DNA was transfected using polyethylenimine HCl MAX (MW 40000, Polysciences, Warrington, PA) at a 1:5 ratio and growth media were changed after 3 hr. Silencer Select siRNA oligonucleotides (5 nM, Ambion) were transfected using RNAiMAX (Life Technologies, Carlsbad, CA) according to the manufacturer's instructions.

### Plasmids

*CHEK2* mutants were generated using QuickChange II Site-Directed Mutagenesis (Agilent Technologies, Santa Clara, CA). Wild-type or mutated *CHEK2* was amplified by PCR and cloned in frame into pcDNA4-EGFP or pGEX-6P-1 plasmids using a Gibson assembly kit (NEB). All mutants were verified by Sanger sequencing. A DNA fragment corresponding to the GVKRSRSGEGEV peptide (containing S473) from human KAP1 was ligated in frame into a pGEX-6P-1 plasmid. Alternatively, a fragment corresponding to T2A-EGFP was ligated into the XbaI site of pcDNA4, and subsequently a fragment corresponding to wild-type or mutant FLAG-*CHEK2* was cloned into *Hind*III/*Xho*I sites resulting in a plasmid for bicistronic expression of FLAG-CHK2 and EGFP.

### Immunofluorescence microscopy, cell-based assay for the detection of CHK2 activity

RPE1-*CHEK2*-KO cells transfected with an empty EGFP plasmid, wild-type or mutant EGFP-*CHEK2* were seeded on glass coverslips and fixed by 4% paraformaldehyde 48 hr after transfection. Cells were permeabilized by 0.2% Triton X-100 in PBS for 20 min and blocked with 3% BSA in PBS at room temperature. The coverslips were incubated with the KAP1-pS473 antibody for 1 hr at room temperature, three times washed with PBS and incubated with the goat-antimouse Alexa568 antibody and DAPI. After the PBS washing, the coverslips were mounted using Vectashield H-1000 and imaged using a Scan^R microscope (Olympus, Waltham, MA) equipped with an ORCA-285 camera and a 40×/1.3 NA objective. The total intensity of the KAP1-pS473 signal per nucleus was determined in cells expressing low levels of GFP. Three independent experiments were performed and >300 cells were quantified per condition in each experiment. The KAP1-pS473 signal in cells expressing only EGFP typically reached <10% of the signal in cells expressing wild-type CHK2 and was subtracted as a background. The KAP1-pS473 signal measured in cells expressing mutant CHK2 was normalized to wild-type CHK2-expressing cells. The activities of the analyzed variants were classified as normal, intermediate or deleterious based on mean pS473 reaching >50%, 25–50% and <25% of wild-type CHK2, respectively.

### *In vitro* kinase assays

*Escherichia coli* BL21 transformed with wild-type or mutant pGEX-6P-1-*CHEK2* plasmids were induced at $A_{600} = 0.6$ by 0.2 mM IPTG and grown for 5 hr at 37°C. The bacteria were lysed in ice-cold PBS supplemented with 0.1% TX-100 and 1 mM PMSF and sonicated 2 × 30 sec. Cleared lysates were incubated with Glutathione Sepharose 4 Fast Flow beads (GE Healthcare, Chicago, IL) for 5 hr at 4°C. Bound proteins were eluted with 10 mM reduced glutathione in 50 mM Tris pH 8.0 and mixed with 30% glycerol. Protein concentration was determined by a BCA assay (Pierce, Puyallup, WA). Purified CHK2 was incubated in a kinase buffer (10 mM HEPES pH 7.4, 2.5 mM β-glycerolphosphate, 2 mM EDTA, 1 mM EGTA, 4 mM $MgCl_2$, 100 μM ATP) with GST-KAP1 substrate (2 μg) for 20 min at 30°C and its phosphorylation was detected by immunoblotting using KAP1-pS473 antibody. Alternatively, wild-type or mutant EGFP-CHK2 was immunoprecipitated from transfected HEK293 cells using GFP-Trap (Chromotek, Munich, Germany), treated with λ-phosphatase (200 U/reaction, Santa Cruz). Beads were washed three times with PBS and incubated for 20 min at 30°C with GST-KAP1 in the kinase buffer supplemented with PhosSTOP inhibitor (Roche, Basel, Switzerland). Alternatively, CHK2 kinase activity was measured in crude bacterial lysates *in vitro* using Omnia kinase assay kit (Life Technologies) as described previously.[19]

### Statistical analysis

The patients were stratified according to (*i*) functional classes of germline *CHEK2* variants (deleterious, intermediate, neutral), (*ii*) the presence of a mutation in other (i.e., non-*CHEK2*) CPG and (*iii*) cancer and histopathological characteristics. Associations between the *CHEK2* mutation status and cancer diagnoses were analyzed using 3,360 PMC. The strength of the associations was estimated by the odds ratio (OR) in Fisher's exact test and *p* values <0.05 were considered significant.

## Results

### Germline CHEK2 variants are more frequent in cancer patients than in PMC

We analyzed germline *CHEK2* variants in 1,928 high-risk Czech BC/OC patients and 3,360 PMCs. We identified 36 distinct nonsynonymous variants (Table 1) in 131/1,928 (6.79%) patients and 142/3,360 (4.23%) PMC ($p = 7.4 \times 10^{-5}$).

Ten different frame-shift and splicing mutations ("All truncations" in Table 1) were found in 46 patients (2.39%) and 11 PMC (0.33%; $p = 1.3 \times 10^{-11}$). The most prevalent alterations were LGRs, present in 20 (1.04%) patients and four PMC (0.12%). LGRs included a recurrent exon 9–10 (5,395 bp) deletion and a novel exon 8 (5,601 bp) deletion. The c.1100delC mutation was found in seven (0.36%) patients and three PMC (0.09%). We identified three spliceogenic variants altering the mRNA sequence: c.444+1G>A, recurrent, population-specific c.846+4_846+7delAGTA (resulting in in-frame exon 7 skipping), and c.1260-8A>G (splice acceptor-shift with 7b exonization; Supporting Information Fig. S1). Variants reported as pathogenic in the ClinVar database, causing a frame-shift or truncating the kinase domain were considered pathogenic. Five of 46 patients with a truncating *CHEK2* mutation (four with female BC and one with double primary BC/OC) carried an additional pathogenic mutation in *BRCA1* or *BRCA2* (but not in another CPG). These patients were assigned into a group of 424 other CPG-mutation carriers.

Twenty-six distinct missense variants were found in 88 (4.56%) patients and 131 (3.90%) PMC (*p* = 0.22; Table 1). The most frequent variant was p.I157T with comparable prevalence in patients (58 carriers; 3.01%) and PMC (104 carriers; 3.10%; *p* = 0.93). Functional consequences of the detected missense variants predicted *in silico* yielded contradictory results (Supporting Information Table S2). While MutationTaster, CADD, and GERP predicted all SNVs as deleterious (except a maximum of 3/26 scored as neutral), the remaining four prediction tools, Align-GVGD, SIFT, PolyPhen2 and Spidex, were 100% and ≥75% concordant for 4/26 and 16/26 variants, respectively. Since the clinical significance of the detected SNVs was described as uncertain or conflicting in the ClinVar database (Table 1), we subjected them to subsequent functional analyses.

### Functional assays identified deleterious CHEK2 missense variants

To evaluate the enzymatic activity of the identified CHK2 protein variants, we developed a cell-based assay quantifying KAP1-S473 phosphorylation in nontransformed human RPE1 cells. First, we verified the specificity of a monoclonal antibody against phosphorylated KAP1-S473 by immunoblotting and immunofluorescence microscopy (Supporting Information Fig. S2A). Next, we used the CRISPR/Cas9 technology to inactivate *CHEK2* in RPE1 cells (RPE1-*CHEK2*-KO; Fig. 1*a*, Supporting Information Figure S2B). A complete loss of CHK2 as well as RNAi-mediated CHK2 depletion impaired KAP1-S473 phosphorylation in RPE1 cells after ionizing radiation exposure. In contrast, CHK2 loss did not affect the phosphorylation of KAP1 at S824, an established ATM kinase site (Fig. 1*a*). A similar effect was also observed after treating the cells with neocarzinostatin and etoposide (Supporting Information Fig. S2C), suggesting that CHK2 phosphorylates KAP1 at S473 after the induction of DNA damage in general. A stable expression of EGFP-CHK2 in RPE1-*CHEK2*-KO cells rescued the phosphorylation of KAP1 at S473 after exposure to ionizing radiation, further confirming that CHK2 phosphorylates KAP1 after genotoxic stress (Fig. 1*b*). Finally, we transiently expressed the wild-type or mutant CHK2 isoforms in RPE1-*CHEK2*-KO cells and quantified the level of KAP1-S473 phosphorylation by immunofluorescence microscopy (Fig. 2*a*). We supplemented this cell-based model with a semiquantitative measurement of KAP1-pS473 in a cell-free *in vitro* assay using purified CHK2 and GST-KAP1 peptide as a substrate (Fig. 2*b*).

Table 1. The prevalence of *CHEK2* germline variants

| Variant; cDNA (reference: NM_007194.3) | Variant; protein | rs number | ClinVar class | Unilateral FBC ($n$ = 1,298) | Bilateral FBC ($n$ = 149) | MBC ($n$ = 48) | BC and OC ($n$ = 79) | OC only ($n$ = 354) | All patients ($n$ = 1,928) | PMC ($n$ = 3,360) |
|---|---|---|---|---|---|---|---|---|---|---|
| TRUNCATING mutations ([a]frame-shift; [b]in-frame) | | | | | | | | | | |
| c.100–101delCA[a] | p.Q34Vfs*42 | NA | NA | – | – | – | 1 | – | 1 | – |
| c.277delT[a] | p.W93Gfs*17 | rs786203458 | 5 | 3 | 2 | – | – | – | 5 | – |
| c.283C>T[a] | p.R95* | rs587781269 | 5 | – | – | – | – | – | – | 1 |
| c.366delA[a] | p.E122Dfs*8 | rs1555927302 | 5 | 1 | – | – | – | – | 1 | – |
| c.444+1G>A[a] | p.E149Ifs*6 | rs121908698 | 5 | 4 | 1 | – | – | – | 5 | 2 |
| c.846+4_846+7delAGTA[b] | p.D265_H282del | rs764884641 | 3 | 7 | – | – | – | – | 7 | – |
| c.846+1888_908+987del5601[a] | p.P283Dfs*8 | NA | NA | 2 | – | – | – | – | 2 | – |
| c.909-2028_1095+330del5395[a] | p.M304Lfs*16 | NA | 5 | 11 | 1 | 3 | 1 | 2 | 18 | 4 |
| c.1100delC[a] | p.T367Mfs*15 | rs555607708 | 5 | 5 | – | – | 1 | 1 | 7 | 3 |
| c.1260-8A>G[a] | p.L421Ifs*4 | rs863224747 | 3 | 1 | – | – | – | – | 1 | 1 |
| All truncations (%) | | | | 33 (2.54)[1] | 4 (2.68) | 3 (6.25) | 3 (3.80) | 3 (0.85) | 46 (2.39)[1] | 11 (0.33) |
| *p*-value (Fisher exact test) | | | | $9.4 \times 10^{-11}$ | 0.003 | $8.6 \times 10^{-4}$ | 0.004 | 0.14 | $1.3 \times 10^{-11}$ | Ref. |
| Missense *CHEK2* mutations classified as DELETERIOUS | | | | | | | | | | |
| c.190G>A | p.E64K | rs141568342 | 3–4 | 3 | – | – | – | 1 | 4 | 2 |
| c.503C>T | p.T168I | rs730881684 | 3 | – | – | 1 | 1 | – | 2 | – |
| c.520C>G | p.L174V | rs876659400 | 3 | 1 | – | – | – | – | 1 | – |
| c.917G>C | p.G306A | rs587780192 | 3–4 | 1 | – | – | – | – | 1 | 2 |
| c.980A>G | p.Y327C | rs587780194 | 3 | 1 | – | – | – | – | 1 | – |
| c.1037G>A | p.R346H | rs730881688 | 3 | 1 | – | – | – | – | 1 | – |
| c.1180G>A | p.E394K | rs587780169 | 3 | 1 | – | – | – | – | 1 | – |
| c.1183G>C | p.V395 L | rs587780170 | 3 | – | – | – | – | 1 | 1 | – |
| c.1270T>C | p.Y424H | rs139366548 | 3 | – | 1 | – | – | – | 1 | – |
| c.1274C>T | p.P425L | rs1555913537 | 3 | 1 | – | – | – | – | 1 | – |
| c.1421G>A | p.R474H | rs121908706 | 3 | – | – | – | – | 1 | 1 | 2 |
| All deleterious missense variants (%) | | | | 9 (0.69) | 1 (0.67) | 1 (2.08) | 1 (1.27) | 3 (0.85) | 15 (0.78) | 6 (0.18) |
| *p*-value (Fisher exact test) | | | | 0.009 | 0.26 | 0.09 | 0.15 | 0.047 | 0.002 | Ref. |
| Missense *CHEK2* variants classified as INTERMEDIATE | | | | | | | | | | |
| c.470T>C | p.I157T | rs17879961 | 3–5 | 38 | 6 | 2 | 3 | 9 | 58 | 104 |
| c.688G>T | p.A230S | rs748636216 | 3 | – | – | – | – | – | – | 1 |
| c.715G>A | p.E239K | rs121908702 | 3 | – | – | – | – | – | – | 2 |
| c.1067C>T | p.S356L | rs121908703 | 3 | – | – | – | – | – | – | 1 |
| c.1217G>A | p.R406H | rs200649225 | 2–3 | – | – | – | – | – | – | 1 |
| All intermediate missense variants (%) | | | | 38 (2.93)[2] | 6 (4.03)[2] | 2 (4.17) | 3 (3.80) | 9 (2.54) | 58 (3.01)[2] | 109 (3.24) |
| *p*-value (Fisher exact test) | | | | 0.64 | 0.63 | 0.67 | 0.74 | 0.63 | 0.68 | Ref. |

*(Continues)*

**Table 1.** The prevalence of *CHEK2* germline variants (Continued)

| Variant; cDNA (reference: NM_007194.3) | Variant; protein | rs number | ClinVar class | Unilateral FBC (*n* = 1,298) | Bilateral FBC (*n* = 149) | MBC (*n* = 48) | BC and OC (*n* = 79) | OC only (*n* = 354) | All patients (*n* = 1,928) | PMC (*n* = 3,360) |
|---|---|---|---|---|---|---|---|---|---|---|
| Missense *CHEK2* variants classified as NEUTRAL | | | | | | | | | | |
| c.7C›T | p.R3W | rs199708878 | 3 | – | – | 1 | – | – | 1 | – |
| c.538C›T | p.R180C | rs77130927 | 1–3 | 1 | – | – | – | – | 1 | 3 |
| c.539G›A | p.R180H | rs137853009 | 3 | 1 | – | – | – | – | 1 | 1 |
| c.541C›T | p.R181C | rs137853010 | 3 | – | – | – | – | – | – | 3 |
| c.542G›A | p.R181H | rs121908701 | 3 | 1 | – | 1 | – | – | 2 | – |
| c.1091T›C | p.I364T | rs774179198 | 3 | 1 | – | – | – | – | 1 | – |
| c.1309A›G | p.K437E | rs764238637 | 3 | 1 | – | – | – | – | 1 | – |
| c.1312G›T | p.D438Y | rs200050883 | 3 | 3 | – | – | – | 1 | 4 | 2 |
| c.1427C›T | p.T476M | rs142763740 | 3–4 | 2 | – | – | – | 1 | 3 | 3 |
| c.1525C›T | p.P509S | rs587780179 | 3 | 1 | – | – | – | 1 | 2 | 4 |
| All neutral missense variants (%) | | | | 11 (0.85) | – | 2 (4.17) | – | 3 (0.85) | 16 (0.83) | 16 (0.48) |
| *p*-value (Fisher exact test) | | | | 0.14 | | 0.03 | | 0.42 | 0.14 | Ref. |
| All CHEK2 missense variants (%) | | | | 58 (4.47) | 7 (4.70) | 5 (10.42) | 4 (5.06) | 14 (3.95)[3] | 88 (4.56)[3] | 131 (3.90) |
| *p*-value (Fisher exact test) | | | | 0.38 | 0.52 | 0.04 | 0.55 | 0.96 | 0.22 | Ref. |

The prevalence of individual variants (divided into subgroups of truncating mutations and missense variants classified according to the results of an RPE1-*CHEK2*-KO cell-based analysis as deleterious, intermediate and neutral; Fig. 2*a*). It is displayed for all patients, their subgroups (unilateral female BC [FBC], bilateral FBC, male BC [MBC], double primary BC and OC and OC only) and population-matched controls (PMC; used as the reference). NA, not available.

[1] Include a FBC compound heterozygote of c.277delT and c.444+1G›A.

[2] Include two p.I157T homozygotes (with unilateral and bilateral FBC all diagnosed at ‹50 years, respectively).

[3] Four other compound heterozygotes in patients group were carriers of p.D265_H282del+p.D438Y, c.5601del+p.I157T, c.1100delC+p.I157T and p.E64K+p.I157T. The NM_007194.3 CHEK2 transcription variant A was used as the reference.

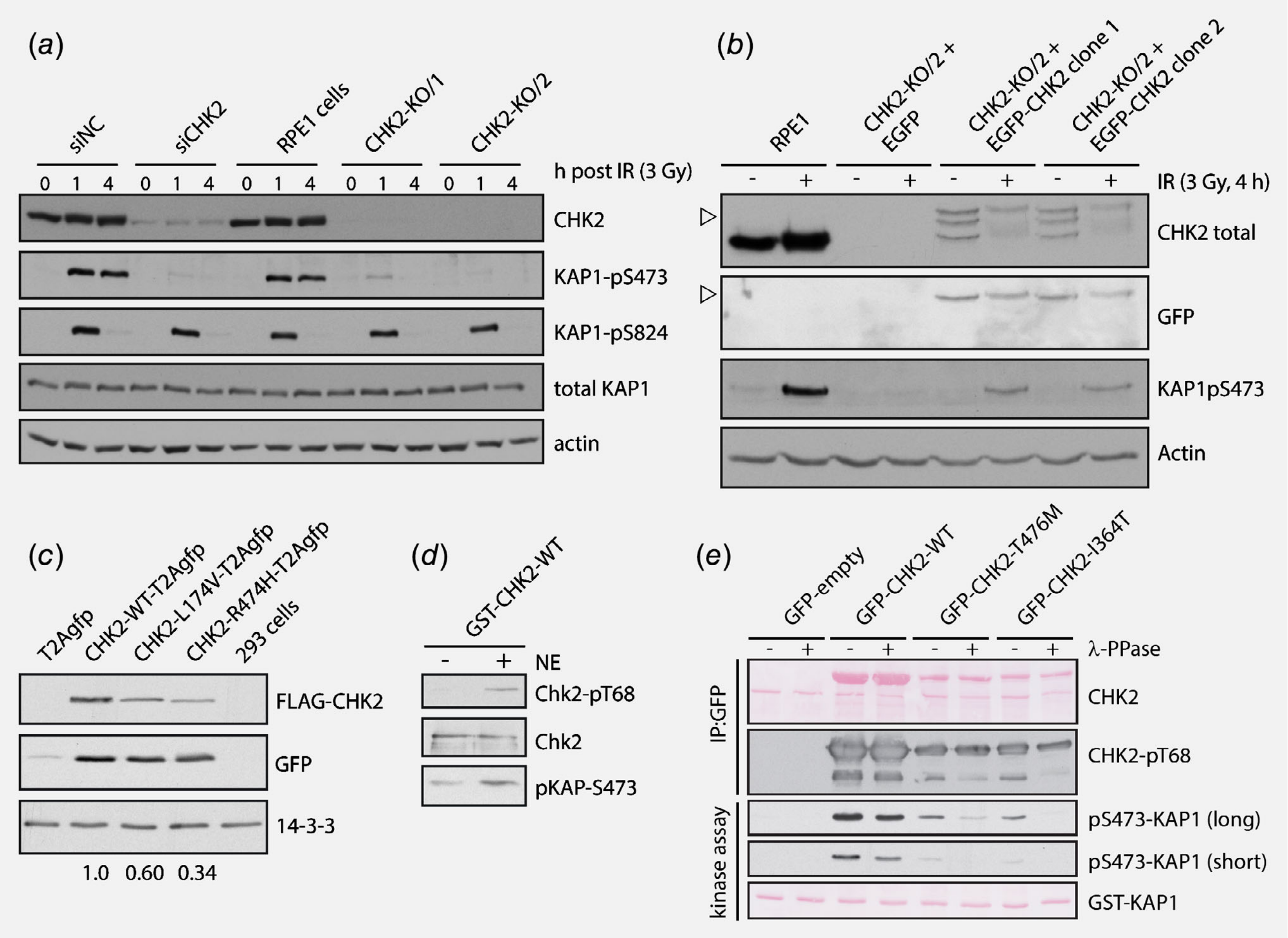


**Figure 1.** Characterization of a model system for functional analysis of *CHEK2* variants in RPE1-*CHEK2*-KO cells. Comparison of CHK2 depletion and knockout (*a*). RPE1 cells were transfected with control (siNC) or CHK2 siRNA (siCHK2) and assayed in parallel with parental RPE1 (RPE1 cells) and two clones of RPE1-*CHEK2*-KO cells (KO/1 and KO/2, respectively). Cells were harvested 0, 1 or 4 hr after exposure to IR (3 Gy) and analyzed by immunoblotting. Rescue of the *CHEK2* knockout (*b*). RPE1-*CHEK2*-KO cells were transfected with EGFP or EGFP-*CHEK2*-WT plasmids, selected with zeocin and exposed or not to IR. Parental RPE1 cells are shown for comparison. Arrowheads indicate the position of EGFP-CHK2. Impact of *CHEK2* mutations on protein stability (*c*). HEK293 cells were mock-treated or transfected with plasmids (1 μg) coding T2Agfp, CHK2-WT-T2Agfp, CHK2-L174V-T2Agfp or CHK2-R474H-T2Agfp and whole cell lysates were harvested after 20 hr. Numbers indicate the level of FLAG-CHK2 normalized to the level of GFP. Impact of CHK2 phosphorylation on its activity *in vitro* (*d*). Wild-type CHK2 purified from bacteria was incubated or not with nuclear extract from HCT116 cells (NE) in the presence of ATP at 30°C. After the addition of PBS, CHK2 was purified again using glutathione beads. Eluted CHK2 was incubated with KAP1 substrate and phosphorylation was assayed by KAP1-pS473 antibody. Impact of CHK2 phosphorylation on its activity (*e*). HEK293 cells were transfected with plasmids coding EGFP, EGFP-CHK2-WT, EGFP-CHK2-T476M, EGFP-CHK2-I364T. After 48 hr proteins were immunoprecipitated by GFP-Trap and treated or not with λ-phosphatase. Kinase activity was measured in the presence of phosphatase inhibitors and using GST-KAP1 as a substrate (shown in short and long exposition, respectively). [Color figure can be viewed at wileyonlinelibrary.com]

The results of the KAP1 cell-based analysis were in full agreement with KAP1 *in vitro* assays for 13 out of 26 tested missense variants, deleterious mutations (p.D265_H282del and c.1100delC) and wild-type CHK2 (Fig. 2, Supporting Information Table S2). Another five SNVs agreed partially between these two KAP1 assays (being intermediate in one and deleterious or neutral in the complementary assay). Eight variants were discrepant between cell based and *in vitro* KAP1 assays. As an example of discrepant results, the p.L174V variant showed only slightly decreased catalytic activity *in vitro*, but failed to phosphorylate KAP1 in cells. A comparison of the expression levels of CHK2-V174 and wild-type CHK2 both expressed from the bicistronic vector together with GFP (Fig. 1*c*) showed a suppressed expression of p.L174V to ~60% of wild-type CHK2, most probably reflecting impaired folding and/or reduced protein stability. Surprisingly, some variants with low *in vitro* activity were still able to phosphorylate KAP1 in human cells to a similar extent as wild-type CHK2. We hypothesized that the CHK2 kinase activity in human cells is influenced by its posttranslational modifications and, therefore, may differ from bacterially expressed CHK2. Indeed, preincubation of CHK2 purified from bacteria with nuclear extract led to CHK2-T68 phosphorylation. Subsequently, modified CHK2 showed higher ability to phosphorylate KAP1-S473 compared to

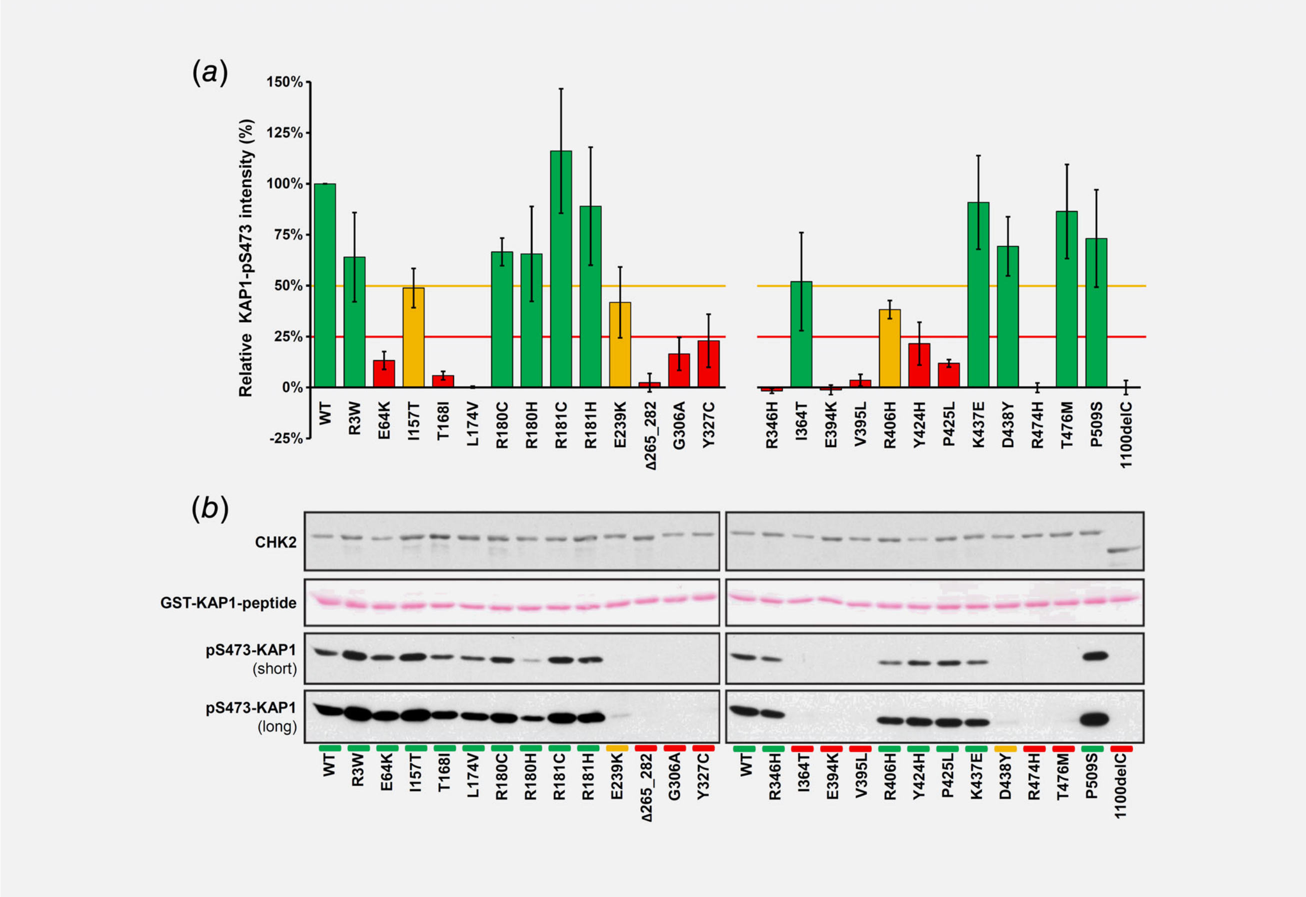


Figure 2. Functional classification of *CHEK2* germline variants was based on RPE1-*CHEK2*-KO cell-based assay. The chart describes relative levels of CHK2-dependent KAP1-S473 phosphorylation in RPE1-*CHEK2*-KO cells (*a*) for detected CHK2 variants. Variants were scored according to the WT (100%) and c.1100delC (0%) CHK2 kinase activity: ›50% as "neutral" (green), 25–50% as "intermediate" (yellow) and ‹25% as "deleterious" (red). Error bars represent standard deviations (SD). Immunoblotting of phosphorylated GST-purified KAP1-peptide at S473 by purified CHK2 isoforms *in vitro* (*b*) was used to complement the assay in RPE1 cells. The individual panels show amounts of particular CHK2 isoforms and GST-KAP1-peptide, and intensity of KAP1-pS473 staining after incubation with purified CHK2 (in short and long exposition, respectively). Colors bars represent classifications from *a*; Δ265_282 means p.D265_H282del. (See online version for color images). *Note*: Variants p.A230S and p.S356L found in PMC (exome samples; not shown in this figure) were functionally classified by the RPE1-CHEK2-KO cell-based assay as intermediate (Supporting Information Table S2). [Color figure can be viewed at wileyonlinelibrary.com]

unmodified CHK2 (Fig. 1*d*). Conversely, the phosphatase treatment of CHK2 immunoprecipitated from HEK293 cells suppressed the *in vitro* activity of p.T476M and p.I364T variants that originally scored well in the cell-based assay (Fig. 1*e*). Our results suggest that posttranslational modifications substantially modulate CHK2 kinase activity and thus the human cell-based assay may better reflect the real CHK2 kinase activity *in vivo*. We also functionally analyzed detected VUS using commercial Omnia kinase *in vitro* assay that fully or partially corresponded to a principally comparable KAP1 *in vitro* assay for 23/26 VUS (Supporting Information Table S2, Fig. S3); however, was unable to dissect VUS discordant between KAP1 assays. Therefore, results from our cell-based assay (Fig. 2*a*), that reflects *in vivo* behavior of analyzed CHK2 variants more appropriately, led us to use solely this assay for the final functional VUS classification (Table 1).

The cell-based assay revealed strongly reduced kinase capacity (<25% of wild-type CHK2) for 11/26 missense variants that were classified as deleterious (Fig. 2*a*). These variants were significantly enriched in patients over PMC (Table 1). A significantly reduced kinase activity was also observed in recurrent c. 846+4_846+7delAGTA (in-frame exon 7 deletion; p.D265_H282del) eliminating the structurally important αC helix (residues 269–280) in the kinase domain.[7] The available pedigrees of patients with deleterious missense variants and c.846+4_846+7delAGTA are provided in the Supporting Information Figure S4. Five missense variants (p.I157T and four VUS identified only in PMC) were functionally classified as intermediate, with kinase activity at 25–50% of wild-type CHK2 in the cell-based assay. Ten missense variants with normal or mildly reduced catalytic activity (retaining >50% of wild-type CHK2) were considered neutral.

### CHEK2 mutations are associated with BC and OC risk

We evaluated the association of *CHEK2* germline variants and cancer risk in diagnosis subgroups, considering all 1,928 patients and separately 1,504 patients without other CPG mutation. Regardless of the presence of other CPG mutations, truncating *CHEK2* variants significantly increased cancer risk in all analyzed subgroups except patients with OC only (Table 2). The most significant association was identified for group of 1,298 unilateral female BC patients that included 33 carriers (2.54%) of *CHEK2* truncations (OR = 7.94; 95%CI 3.90–17.47; $p = 9.4 \times 10^{-11}$). Truncations in *CHEK2* had the third highest mutation rate in this subgroup, preceded by *BRCA1* (153 carriers; 11.79%) and *BRCA2*

**Table 2.** Risk associated with germline *CHEK2* truncating and functionally classified missense variants (deleterious, intermediate and neutral) in all analyzed patients and in a subgroup of patients negatively tested for mutations in other cancer-predisposing genes against frequencies of *CHEK2* variants found in Czech population-matched controls PMC, Table 1

| Group of patients | All patients | | | Other cancer-predisposing genes wt patients | | |
|---|---|---|---|---|---|---|
| ***CHEK2* variant group** | **Carriers; *N* (%)** | **OR (95%CI)** | ***p*-value** | **Carriers; *N* (%)** | **OR (95%CI)** | ***p*-value** |
| Unilateral female BC (I) | *n* = 1,298 | | | *n* = 1,065 | | |
| Truncations | 33 (2.54) | **7.94 (3.90–17.47)** | **$9.4 \times 10^{-11}$** | 29 (2.72) | **8.52 (4.11–18.97)** | **$1.2 \times 10^{-10}$** |
| Deleterious missense | 9 (0.69) | **3.90 (1.24–13.35)** | **0.009** | 8 (0.75) | **4.23 (1.28–14.82)** | **0.008** |
| Intermediate missense | 38 (2.93) | 0.90 (0.60–1.32) | 0.64 | 34 (3.19) | 0.98 (0.64–1.47) | 0.99 |
| Neutral missense | 11 (0.84) | 1.79 (0.75–4.11) | 0.14 | 10 (0.94) | 1.98 (0.80–4.66) | 0.11 |
| Bilateral female BC (II) | *n* = 149 | | | *n* = 104 | | |
| Truncations | 4 (2.68) | **8.39 (1.92–28.74)** | **0.003** | 4 (3.85) | **12.15 (2.77–41.94)** | **$8.1 \times 10^{-4}$** |
| Deleterious missense | 1 (0.67) | 3.77 (0.08–31.42) | 0.26 | 1 (0.96) | 5.42 (0.12–45.31) | 0.19 |
| Intermediate missense | 6 (4.03) | 1.25 (0.44–2.88) | 0.63 | 5 (4.81) | 1.51 (0.47–3.74) | 0.39 |
| Neutral missense | 0 (0) | – | – | 0 (0) | – | – |
| Male BC (III) | *n* = 48 | | | *n* = 39 | | |
| Truncations | 3 (6.25) | **20.21 (3.50–80.00)** | **$8.6 \times 10^{-4}$** | 3 (7.69) | **25.23 (4.34–101.34)** | **$4.7 \times 10^{-4}$** |
| Deleterious missense | 1 (2.08) | 11.87 (0.25–100.83) | 0.10 | 1 (2.56) | 14.66 (0.31–125.29) | 0.08 |
| Intermediate missense | 2 (4.17) | 1.30 (0.15–5.07) | 0.67 | 2 (5.13) | 1.61 (0.19–6.39) | 0.37 |
| Neutral missense | 2 (4.17) | **9.07 (0.98–40.41)** | **0.03** | 2 (5.13) | **11.26 (1.21–50.79)** | **0.02** |
| BC and OC (IV) | *n* = 79 | | | *n* = 40 | | |
| Truncations | 3 (3.80) | **11.99 (2.11–46.6)** | **0.004** | 2 (5.00) | **15.97 (1.67–77.08)** | **0.01** |
| Deleterious missense | 1 (1.27) | 7.15 (0.15–59.97) | 0.15 | 0 (0) | – | – |
| Intermediate missense | 3 (3.80) | 1.18 (0.24–3.67) | 0.74 | 1 (2.50) | 0.76 (0.02–4.61) | 0.99 |
| Neutral missense | 0 (0) | – | – | 0 (0) | – | – |
| OC only (V) | *n* = 354 | | | *n* = 256 | | |
| Truncations | 3 (0.85) | 2.60 (0.46–9.91) | 0.14 | 3 (1.17) | 3.61 (0.64–13.78) | 0.07 |
| Deleterious missense | 3 (0.85) | **4.77 (0.77–22.47)** | **0.047** | 3 (1.17) | **6.62 (1.07–31.22)** | **0.02** |
| Intermediate missense | 9 (2.54) | 0.78 (0.34–1.55) | 0.63 | 8 (3.13) | 0.96 (0.40–1.99) | 0.99 |
| Neutral missense | 3 (0.84) | 1.79 (0.33–6.28) | 0.42 | 2 (0.78) | 1.65 (0.18–7.06) | 0.37 |
| Any female BC (I + II + IV) | *n* = 1,526 | | | *n* = 1,209 | | |
| Truncations | 40 (2.62) | **8.19 (4.11–17.75)** | **$4.1 \times 10^{-12}$** | 35 (2.90) | **9.07 (4.49–19.87)** | **$2.4 \times 10^{-12}$** |
| Deleterious missense | 11 (0.72) | **4.06 (1.37–13.39)** | **0.006** | 9 (0.74) | **4.19 (1.33–14.34)** | **0.006** |
| Intermediate missense | 47 (3.08) | 0.95 (0.66–1.35) | 0.79 | 40 (3.31) | 1.02 (0.69–1.49) | 0.92 |
| Neutral missense | 11 (0.72) | 1.52 (0.64–3.49) | 0.30 | 10 (0.83) | 1.74 (0.70–4.10) | 0.18 |
| Any OC (IV + V) | *n* = 433 | | | *n* = 296 | | |
| Truncations | 6 (1.39) | **4.28 (1.29–12.69)** | **0.009** | 5 (1.69) | **5.23 (1.41–16.45)** | **0.007** |
| Deleterious missense | 4 (0.92) | **5.21 (1.08–22.06)** | **0.02** | 3 (1.01) | **5.72 (0.92–26.94)** | **0.03** |
| Intermediate missense | 12 (2.77) | 0.85 (0.42–1.56) | 0.77 | 9 (3.04) | 0.94 (0.41–1.87) | 0.99 |
| Neutral missense | 3 (0.69) | 1.46 (0.27–5.12) | 0.47 | 2 (0.68) | 1.42 (0.16–6.09) | 0.65 |

The calculations were performed in individual diagnostic subgroups (Roman numerals I–V) and in aggregated groups of any female BC (subgroups I, II and IV) and any OC patients (subgroups IV and V). "Other CPG-wt" group consists of patients without germline mutations in genes predisposing for BC (*BRCA1*, *BRCA2*, *PALB2*, *TP53*) or OC (*BRCA1*, *BRCA2*, *RAD51C*, *RAD51D*, *MLH1*, *MSH2*, *MSH6*). Significant association of *CHEK2* variants with cancer risk is highlighted (in bold). Both aggregated subgroups (Any FBC and Any OC) include patients with double primary BC and OC (IV).

(56 carriers; 4.31%), and followed by *PALB2* (21 carriers; 1.62%) and *TP53* (3 carriers; 0.23%). We also observed a significantly higher prevalence of *CHEK2* truncations in small subgroups of patients with bilateral female BC (4/149; 2.68%; $p$ = 0.003), male BC (3/48; 6.25%; $p = 8.6 \times 10^{-4}$) and with double primary BC/OC (3/79; 3.80% $p$ = 0.004); however, the low number of patients and mutations limits relevance of calculated ORs. The analysis of two aggregated subgroups of "any female BC" and "any OC" patients (overlapping in patients diagnosed with double primary BC/OC; Table 2) reflected clinically relevant overall risk for BC and OC development in females with *CHEK2* truncations. We found significant associations with both cancer types, which was substantially higher and more significant for "any female BC" (OR = 8.19; 95%CI 4.11–17.75; $p = 4.1 \times 10^{-12}$) than for "any OC" (OR = 4.28; 95%CI 1.29–12.69; $p$ = 0.009) subgroups in all patients as well as in patients after excluding those with mutations in other CPG (OR = 9.07; 95%CI 4.49–19.87; $p = 2.4 \times 10^{-12}$ and OR = 5.23; 95%CI 1.41–16.45; $p$ = 0.007, respectively).

While the frequencies of functionally deleterious SNV were significantly more frequent in unilateral female BC, OC, any female BC and also any OC subgroups (Tables 1 and 2), the frequencies of functionally neutral or intermediate SNVs did not differ from PMC in any patient subgroup (except for neutral SNVs in a small subgroup of 48 male BC patients). Risks associated with functionally deleterious SNV were lower than risks associated with truncations, except that in OC patients. However, low number of functionally deleterious SNV carriers makes our findings only suggestive but not conclusive.

Twelve out of 54 *BRCA1*/*BRCA2*-negative *CHEK2* mutation carriers had a VUS in other genes, in which further modification of cancer risk cannot be ruled out (Supporting Information Table S3).

### CHEK2 mutations predispose to specific BC types and multiple cancer development

We evaluated histopathological tumor characteristics in 1,209 *other CPG-wt* female BC patients. Breast tumors in *CHEK2* mutation carriers differed from noncarriers, tended to be more frequently of luminal A and less frequently of basal BC subtype, with lower grade and with nonsignificant tendency toward lower clinical stage (Fig. 3; Supporting Information Table S4). Histology, menopausal status and indication criteria for testing did not differ among *CHEK2* mutation carriers and noncarriers. Although the most frequent p.I157T variant did not affect BC risk, its carriers had a similar tendency for BC subtype distribution. Phenotypical characteristics of functionally deleterious missense and truncating *CHEK2* mutation carriers were similar (Supporting Information Table S5).

Second primary cancers (other than BC/OC; Supporting Information Table S3) were diagnosed in *CHEK2* mutation carriers more frequently (10/54; 18.5%) than in carriers of other CPG mutations (25/424; 5.9%; $p$ = 0.003) or noncarriers (110/1,403; 7.8%; $p$ = 0.01). All 10 *CHEK2* mutation carriers with second cancer (developing 13 tumors together including two cases each of colon, thyroid, renal, head/neck cancers or hematological malignancy, and one case each of lung, urinary bladder or endometrial cancer) had a positive family cancer history.

## Discussion

The frequency of germline truncating and splice site *CHEK2* mutation carriers in our study strongly prevailed in all patients over PMC (2.39% *vs.* 0.33%; $p = 1.3 \times 10^{-11}$) but the frequencies of missense variants were comparable (4.56% *vs.* 3.90%; $p$ = 0.22). Most missense variants, especially in moderate risk genes (including *CHEK2*) are interpreted as inconclusive VUS, lacking clearly defined risk estimates and representing a major drawback for multigene testing in diagnostic settings.[26,27] Only several reports have described a functional characterization of *CHEK2* VUS by *in vitro*[19,22] or yeast models.[35,36] The *in vitro* assays measure CHK2 kinase catalytic activity over artificial substrate but do not reflect changes in CHK2 intracellular targeting, stability and posttranslational modifications. Moreover, transient CHK2 overexpression can cause its autophosphorylation even in the absence of DNA damage, bypassing necessity for CHK2-T68 phosphorylation and participation of FHA domain on CHK2 activation *in vivo*.[37] Yeast analyses are based on functional complementation of *RAD53*-defective *Saccharomyces cerevisiae* cells by human CHK2 homolog. A growth rate of the yeast cells upon DNA damage correlates with functional competence of the analyzed *CHEK2* variant in this assay. In contrast, our newly developed RPE1-*CHEK2*-KO cell-based assay allowed us to quantify catalytic activity of analyzed *CHEK2* variants in nontransformed human cells in the presence of CHK2 natural upstream activators and downstream substrates.

Altogether, results of functional analysis for 18/26 (69%) of analyzed missense VUS were in full agreement or partially overlapped between our KAP1 cell-based and *in vitro* analyses. Remaining eight variants (p.E64K, p.T168I, p.L174V, p.R346H, p.I364T, p.Y424H, p.P425L, p.T476M) scored discordantly. In subsequent analyses of p.L174V, p.I364T and p.T476M variants, we demonstrated that discordance between results of cell-based and *in vitro* assays resulted from their fundamental differences (Figs. 1*c*–1*e*). Variant p.L174V only mildly decreased KAP1 phosphorylation *in vitro*, but failed to phosphorylate KAP1 in cells. Further analysis revealed that this variant impairs intracellular protein stability explaining its functional defect in cells. This rare FHA domain variant was described once in ClinVar. We identified p.L174V in BC patient diagnosed at 35 years carrying also a pathogenic *BRCA1* mutation (Supporting Information Fig. S4). Variant p.I364T showed low KAP1 phosphorylation *in vitro* but was able to phosphorylate KAP1 in cells. Subsequent analysis demonstrated that CHK2-T364 protein was phosphorylated at T68 when immunoprecipitated from cells and that removing this modification by λ-phosphatase treatment strongly reduced its catalytic activity (Figs. 1*d* and 1*e*) comparable to that in wild-type CHK2. Moreover, Chrisanthar *et al.* described normal dimerization and autophosphorylation, and only mildly reduced kinase activity for p.I364T, concluding a nonaffected

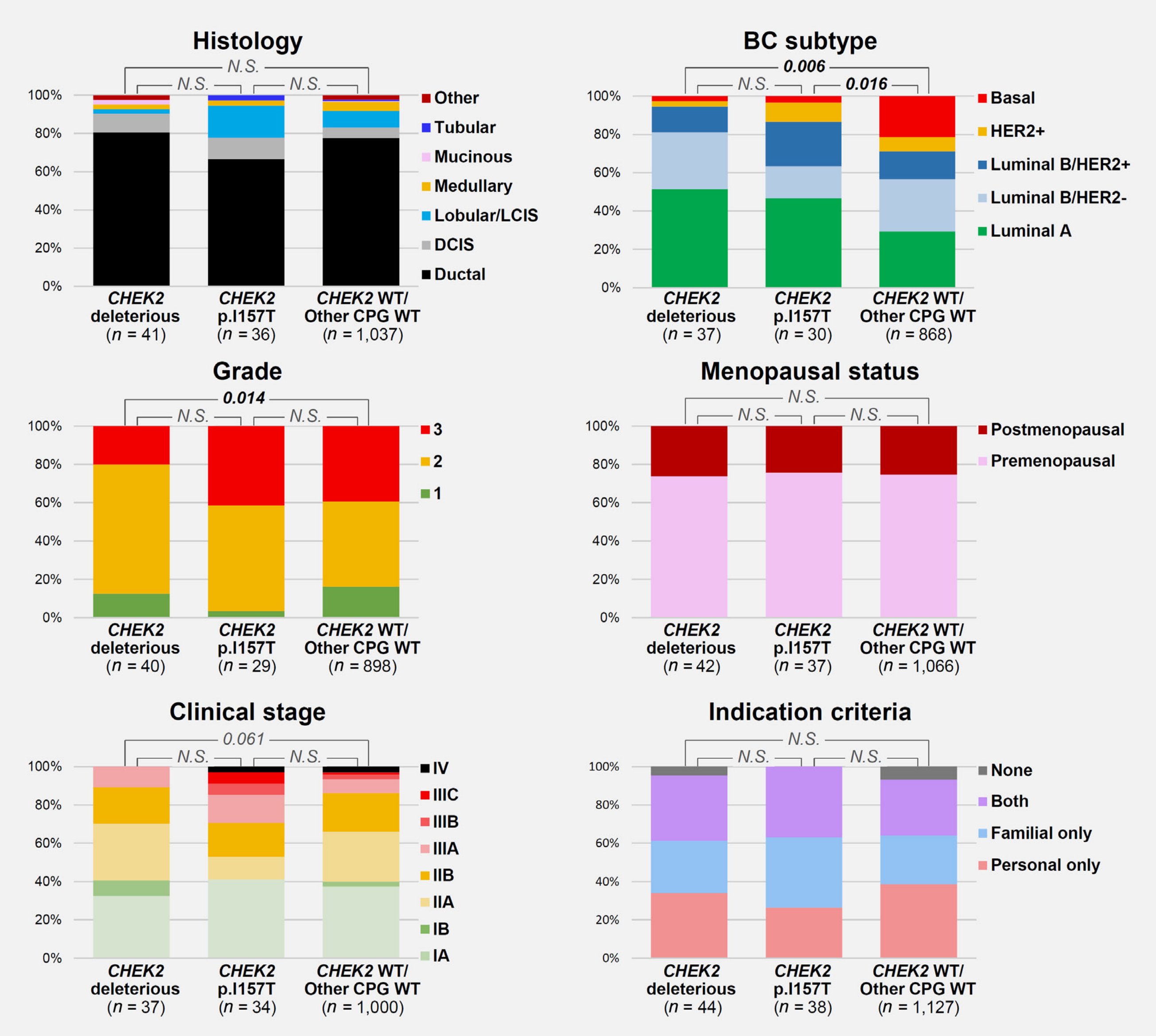


Figure 3. Clinical and histopathological characteristics of female BC patients. A subgroup of 1,209 other CPG-wt patients with any BC were stratified according to the presence of germline deleterious *CHEK2* mutation (truncating or pathogenic missense; *n* = 44), p.I157T (*n* = 38) and *CHEK2*-wt patients (*n* = 1,127), respectively. Significant differences between groups are highlighted in bold (N.S. denoted for not significant differences with $p < 0.1$). Numbers in parenthesis (*n*) characterize number of individuals with known values for particular characteristic. *Note*: "Other CPG-wt" group consists of patients without germline mutation in genes predisposing for BC (*BRCA1*, *BRCA2*, *PALB2*, *TP53*) or OC (*BRCA1*, *BRCA2*, *RAD51C*, *RAD51D*, *MLH1*, *MSH2*, *MSH6*). [Color figure can be viewed at wileyonlinelibrary.com]

kinase function;[38] Delimitsou *et al.* recently scored p.I364T by *S. cerevisiae* assay functionally intermediate (Supporting Information Table S2).[36] We identified this variant in premenopausal BC patient with no cancer diagnosed in first or second-degree relatives. The p.T476M variant behaved similarly as p.I364T, with T68 phosphorylation-dependent kinase activity (Fig. 1*e*). This variant was classified by Delimitsou intermediate, but previous analyses by Roeb *et al.*[35] and Desrichard *et al.*[19] (Supporting Information Table S2) scored p.T476M deleterious by yeast and *in vitro* assays, respectively. We found this variant in three patients and three PMC. Moreover, in concordance with our cell-based assay, the p.T476M was classified as likely benign by Myriad using history weighting algorithm.[39]

Another five discrepant variants were scored in our cell-based assay functionally deleterious. The p.E64K variant affecting SQ/TQ domain was previously analyzed by Wu *et al.*[40] who described its reduced autophosphorylation, CDC25C phosphorylation and severely impaired T68 phosphorylation and concluded that p.E64K alters SQ/TQ domain conformation impairing CHK2 activation. Two later independent analyses showed mutually opposite results in yeast assays (Supporting Information Table S2).[35,36] We found p.E64K in one OC and three BC patients, including a carrier who developed three primary tumors (Supporting Information Fig. S4); however, two carriers were also identified in PMC, including a male (aged 68) and female (aged 63). We found no additional functional data for p.T168I, a variant localized to the FHA domain, functionally defective also in our Omnia kinase assay (Supporting Information Table S2). We detected p.T168I in a patient carrying a *BRCA2* mutation diagnosed with BC and OC (Supporting Information Fig. S4). Variant p.R346H, affecting kinase domain, was functionally classified deleterious also by Delimitsou *et al.*[36] and our Omnia kinase assay (Supporting Information Table S2). Moreover, in a BCAC study, Southey found an increased BC risk (OR = 5.06; 95%CI 1.09–23.5; $p = 0.017$) for p.R346C variant at the same position[41] and we observed a segregation of p.R346H with BC in analyzed HBC family (Supporting Information Fig. S4). The p.Y424H kinase domain variant was classified functionally defective by two out of three previous yeast-based analyses and in our Omnia kinase assay (Supporting Information Table S2). We detected p. Y424H in patient with double primary premenopausal BC with multiple cancers in family members. The p.P425L variant, affecting P425 participating in CHK2 kinase domain dimerization,[7] showed also partially reduced Omnia kinase assay activity. We found this variant in BC patients diagnosed at 47 years; however, no other relatives were available for the genetic analysis.

Conceptual differences in functional *CHEK2* assays contribute to discrepant findings for individual VUS, especially in variants sensitive to posttranslational CHK2 modifications. Hence, we think that our assay performed in human nontransformed cells provides an opportunity for realistic functional *CHEK2* VUS analysis. Estimated BC risks associated with functionally deleterious, intermediate and neutral variants (Table 2) revealed a lack of risk association for the latter two groups, supporting our correct functional classification. Altogether, functionally deleterious missense mutations were identified in 15 out of 88 *CHEK2* missense variant carriers (Table 1) constituting 20–25% of pathogenic *CHEK2* mutation in BC patients and 40% in OC patients. However, low number of carriers of functionally deleterious variants limited validity of presented data. The extension of our assay to large-scale *CHEK2* VUS analyses with evaluation of clinical data in their carriers will be required to validate our findings, including lower risk associated with functionally deleterious missense variants in comparison to truncations.

To calculate cancer risk for carriers of deleterious *CHEK2* mutations, we considered *all high-risk patients* and, in parallel, a subgroup of *CPG-wt patients*. The *all high-risk patients* group revealed the real proportion of *CHEK2* mutation carriers and associated cancer risk in a realistic context of all individuals indicated for genetic testing according to current guidelines. The analysis of the *CPG-wt* subgroup (raising the proportion of *CHEK2* mutation carriers by excluding 424 other CPG-mutation carriers of whom 90% carried a *BRCA1/BRCA2* mutation) allows to compare our findings with studies analyzing *BRCA1/BRCA2*-wt patients (Table 3).

We are aware that risk calculations have their specific limitations. Analyzed patients' groups were enriched in high-risk patients from multiple cancer families and, in contrast, PMC group share higher proportion of older noncancer individuals. Both factors can contribute to an overestimated risks found in our study. Other *CHEK2* studies also demonstrated higher OR found in analyses involving patients with familial BC (Table 3) indicating that a precise risk estimation will require a representative number of analyzed individuals and appropriately selected PMC. Higher cancer risks found in our study was affected also by high frequency of LGRs whose identification by panel NGS has been considered problematic[34] or omitted[26] in comparable analyses. Our data urge its careful evaluation in *CHEK2* analyses. Although the OR values calculated in our study must be interpreted with caution (especially in case of missense variants), our data clearly show that germline *CHEK2* mutations carriers are significantly enriched especially in the largest group of female BC patients. Interestingly, deleterious *CHEK2* mutations increased risk of male BC. *CHEK2* was the second most frequently mutated CPG in this small subgroup, preceded by *BRCA2* and followed by *BRCA1*, and *PALB2* (data not shown), indicating that germline *CHEK2* mutations contribute to male BC, as suggested previously.[51,53,54]

Deleterious *CHEK2* mutations were associated with a moderately increased OC risk in our study. However, due to the limited numbers of analyzed OC individuals with *CHEK2* mutations (10 in all patients, 4 in the CPG-negative subgroup), these observations need further validation. A substantial proportion of deleterious missense mutations (4/10) in OC patients indicates that their functional classification will be necessary for proper OC risk assessment.

Our analysis confirmed proposed "*CHEK2* mutation-specific" tumor phenotype, characterized by premenopausal, ductal, grade 2, luminal A or luminal B/HER2-negative tumors, reported in other studies.[25,26,46,55] These tumor characteristics lost in carriers of coincidental *BRCA1/BRCA2* mutations having a stronger effect on tumor phenotype. Nurmi *et al.*[42] identified an additive effect of mutations in moderate-penetrance genes, including *CHEK2*, increasing BC risk in Finnish *BRCA1/BRCA2* mutation carriers. The effect of coincidental alterations in other moderate-penetrance CPG with *CHEK2* mutations are unknown; however, the influence of a polygenic risk score on c.1100delC penetrance has been recently documented.[56]

A strongly increased frequency of second cancers of various origin in *CHEK2* mutation carriers and tumors in their relatives corresponds to documented multiorgan cancer susceptibility in *CHEK2* mutations carriers[5,25] and indicates that family cancer history associated with *CHEK2* mutations must be reconsidered to facilitate the selection of potential *CHEK2* mutation carriers for genetic analyses.

The p.I157T variant did not increase cancer risk in our study; an observation we have previously reported for sporadic BC

**Table 3.** Analyses of germline variants in the *CHEK2* gene or analyses of selected germline *CHEK2* variants in studies (upper part) and selected meta-analyses (lower part) calculating odds ratio for breast cancer development in mutation carriers

| References | Pop. | P: patientsC: controls | Analysis | Odds ratio (95%CI); *p*—evaluated group or *CHEK2* variant |
|---|---|---|---|---|
| Nurmi *et al.*[42] | FI | P: 3156 BC or OC patients<br>C: 2089 PMC | c.319+2T›A | 5.40 (1.58–18.45); 0.007—unselected BC<br>6.04 (1.65–22.10); 0.007—familial BC |
| Girard *et al.*[43] | FR | P: 1,207 *BRCA1/2*-negative BC females having sister with BC<br>C: 1,199 noncancer PMC | *CHEK2* (panel NGS) | 3.0 (1.9–5.0); 1 × $10^{-5}$—any variant<br>5.8 (2.0–16.9); 0.001—loss of function variant<br>2.4 (1.4–4.3); 0.002—likely deleterious missense |
| Hauke *et al.*[26] | DE | P: 5,589 *BRCA1/2*-negative BC<br>C: 2,189 noncancer PMC | *CHEK2* (panel NGS) | 3.72 (1.99–6.94); ‹0.0001—truncations |
| Couch *et al.*[24] | US | P: 29,090 BC<br>C: ExAC-NFE non-TCGA | *CHEK2* (panel NGS) | 2.31 (1.88–2.85); 3.04 × $10^{-17}$—c.1100delC<br>2.26 (1.89–2.72); 1.75 × $10^{-20}$—pathogenic variants (p.I157T, p.S428F excluded)<br>1.48 (1.31–1.67); 1.75 × $10^{-10}$—any variant (p.I157T, p.S428F included)<br>1.35 (1.12–1.63); 0.0002; bilateral BC |
| Decker *et al.*[44] | UK | P: 13,087 BC<br>C: 5,488 PMC | *CHEK2* (4 genes) | 3.11 (2.15–4.69); 5.6 × $10^{-11}$—truncations<br>1.36 (0.99–1.87); 0.066—all rare missense<br>1.51 (1.02–2.24); 0.047—rare missense in any domain<br>3.27 (1.66–5.83); 0.0014—bilateral BC |
| Slavin *et al.*[45] | US | P: 2,266 *BRCA1/2*-neg. Fam. BC<br>C: ExAC | *CHEK2* | 1.62 (1.03–2.51); 0.004—truncations |
| Schmidt *et al.*[46] | BCAC | 44,777 BC<br>42,977 PMC | c.1100delC | 2.26 (1.90–2.69); 2.3 × $10^{-20}$—invasive BC<br>2.55 (2.10–3.10); 4.9 × $10^{-21}$—ER-positive BC<br>1.32 (0.93–1.88); 0.12—ER-negative BC |
| Southey *et al.*[41] | BCAC | P: 42,671<br>C: 42,164 PMC | c.349A›G (p.R117G)<br>c.538C›T (p.R180C)<br>c.715G›A (p.E239K)<br>c.1036C›T (p.R346C)<br>c.1312G›T (p.D438Y) | 2.26 (1.29–3.95); 0.003—for variant p.R117G<br>1.33 (1.05–1.67); 0.016—for variant p.R180C<br>1.70 (0.73–3.93); 0.210—for variant p.E239K<br>5.06 (1.09–23.5); 0.017—for variant p.R346C<br>1.03 (0.62–1.71); 0.910—for variant p.D438Y |
| Cybulski *et al.*[47] | PL | P: 7,494 *BRCA1*-negative BC<br>C: 4,346 PMC | c.1100delC, c.444+1G›A, del5395 | 3.6 (2.6–5.1)—BC<br>3.3 (2.3–4.7)—patients with no BC family history<br>5.0 (3.3–7.6)—patients with BC in first or second degree relatives<br>7.3 (3.2–16.8)—patients with BC in first and second degree relatives |
| Desrichard *et al.*[19] | FR | P: 507 *BRCA1/2*-negative BC<br>C: 513 noncancer PMC | *CHEK2* | 4.15 (1.38–12.50); 0.007—all *CHEK2* variants<br>5.18 (1.49–18.00); 0.004—*CHEK2* mutations (p.K244R ex) |
| Le Calvez-Kelm *et al.*[20] | US, AU | P: 1303 BC ≤45 years<br>C: 1,109 noncancer females | *CHEK2* | 6.18 (1.76–21.8)—truncations<br>2.20 (1.20–4.01)—rare missense |
| Weischer *et al.*[48] | DK | P: 1,101 BC<br>C: 4,665 PMC | c.1100delC | 3.2 (1.0–9.9)—BC (prospective study)<br>2.6 (1.3–5.4)—BC (case–control study) |

(*Continues*)

**Table 3.** Analyses of germline variants in the *CHEK2* gene or analyses of selected germline *CHEK2* variants in studies (upper part) and selected meta-analyses (lower part) calculating odds ratio for breast cancer development in mutation carriers (Continued)

| References | Pop. | P: patientsC: controls | Analysis | Odds ratio (95%CI); *p*—evaluated group or *CHEK2* variant |
|---|---|---|---|---|
| Cybulski *et al.*[5] | PL | P: 1,017 BC<br>C: 4,000 PMC | c.1100delC; c.444+1G›A; p.I157T | 2.2; $p$ = 0.02—for c.1100delC and c.444+1G›A<br>1.4; $p$ = 0.02—for p.I157T |
| Dufault *et al.*[21] | DE | P: 516 *BRCA1/2*-negative BC<br>C: 1,315 random PMC | *CHEK2* | 3.44 (1.19–9.95); 0.016—c.1100delC<br>3.9 (1.3–10.9)—c.1100delC and c.1214del4 |
| CHEK2 Breast Cancer Case-Control Consortium[49] | UK, NL, FI, DE, AU | P: 10,860 BC<br>C: 9,065 PMC | c.1100delC | 2.34 (1.72–3.20); 1 × $10^{-7}$<br>2.23 (1.60–3.11)—BC with no BC in first degree relative<br>3.12 (1.90–5.15)—BC with 1 BC in first degree relative<br>4.17 (1.26–13.75)—BC with ≥2 BC in first degree relative |
| Vahteristo, 2002[50] | FI | 1,035 unselected BC<br>1885 PMC | c.1100delC | 1.48 (0.83–2.65); 0.182 unselected BC<br>2.27 (1.11–4.63); 0.021 familial BC<br>6.17 (1.87–20.32); 0.007 bilateral BC |
| Liang *et al.*[51] | Meta | P: 118,735 BC<br>C: 195,807 | c.1100delC | 2.88 (2.65–3.16)—female BC<br>2.87 (1.85–4.47)—early onset BC<br>3.21 (2.41–4.29)—familial BC<br>3.13 (1.94–5.07)—male BC |
| Liu *et al.*[16] | Meta | P: 19,621 BC<br>C: 27,001 | p.I157T | 1.48 (1.31–1.66); ‹0.0001—unselected BC<br>1.48 (1.16–1.89); ‹0.0001—familiar BC<br>1.47 (1.29–1.66); ‹0.0001—early onset BC<br>4.17 (2.89–6.03); ‹0.0001—lobular BC |
| Zhang *et al.*[52] | Meta | P: 9,970/ C:7,526<br>P: 13,331/C: 10,817<br>P: 10,543/ C:10,817<br>P: 4,1,791/C: 50,910 | c.444+1G›A<br>p.I157T<br>del5395<br>c.1100delC | 3.07 (2.03–4.63); 9.82 × $10^{-8}$—for variant c.444 +1G›A<br>1.52 (1.31–1.77); 4.76 × $10^{-8}$—for variant p.I157T<br>2.53 (1.61–3.97); 6.33 × $10^{-5}$—for variant del5395<br>3.10 (2.59–3.71); ‹$10^{-20}$—for variant c.1100delC |
| Weischer *et al.*[15] | Meta | P: 26,488<br>C: 27,402 | c.1100delC | 2.7 (2.1–3.4)—unselected BC<br>2.6 (1.3–5.5)—early onset BC<br>4.8 (3.3–7.2)—familial BC |

Abbreviations: AU, Australia; BC, breast cancer; BCAC, Breast Cancer Association Consortium; CN, China; DE, Germany; EU, European union; FI, Finland; DK, Denmark; FR, France; meta, meta-analysis; NL, Nederland; PL, Poland; US, USA.

patients.[57] With OR = 1.5 reported in numerous studies (Table 3), is below the threshold considered for moderate-penetrance genes (OR > 2) and together with a high frequency in PMC it negates a clinically considerable effect on BC risk. We noticed a higher proportion of lobular BC in p.I157T carriers (Fig. 3), known from previous studies.[16,58,59] Our functional analysis classified p.I157T as an "intermediate" variant with catalytic activity reaching 48.8% of wild-type CHK2. Hence, an increased cancer risk cannot be ruled out in homozygote p.I157T carriers.

In conclusion, our study demonstrated a substantial clinical relevance of a *CHEK2* analysis in high-risk BC/OC patients, supported by the results of a cell-based functional assay markedly reducing the number of VUS. In addition, the high frequency of non-BC/OC tumors in *CHEK2* mutation carriers and their relatives warrants further investigation by collaborative international efforts.

## Acknowledgements

We thank Prof. Yves-Jean Bignon who kindly provided the *CHEK2* plasmid, Prof. Martin Haluzik and Dr. Petra Kavalkova for their help with FluoroStar analysis, Jan Flemr for language editing. We thank all patients and their families for their participation in our study. The study was supported by grants of the Ministry of Health of the Czech Republic NV15-28830A, NV15-27695A, NV16-29959A, NV19-03-00279 and the grants of Charles University PROGRES Q28/LF1 and Q26/LF1, GAUK 762216, SVV2019/260367, PRIMUS/17/MED/9 and UNCE/MED/016. We thank The National Center for Medical Genomics (LM2015091) for providing allelic frequencies in ethnically matched populations (project CZ.02.1.01/0.0/0.0/16_013/0001634) and NPU II LQ1604. LM was partially supported by the Academy of Sciences of the Czech Republic (Strategie AV21, Qualitas).

## References

1. Sun J, Meng H, Yao L, et al. Germline mutations in cancer susceptibility genes in a large series of unselected breast cancer patients. *Clin Cancer Res* 2017;23:6113–9.
2. Norquist BM, Harrell MI, Brady MF, et al. Inherited mutations in women with ovarian carcinoma. *JAMA Oncol* 2016;2:482–90.
3. Kleibl Z, Kristensen VN. Women at high risk of breast cancer: molecular characteristics, clinical presentation and management. *Breast* 2016;28: 136–44.
4. Easton DF, Pharoah PD, Antoniou AC, et al. Gene-panel sequencing and the prediction of breast-cancer risk. *N Engl J Med* 2015;372: 2243–57.
5. Cybulski C, Gorski B, Huzarski T, et al. CHEK2 is a multiorgan cancer susceptibility gene. *Am J Hum Genet* 2004;75:1131–5.
6. Matsuoka S, Rotman G, Ogawa A, et al. Ataxia telangiectasia-mutated phosphorylates Chk2 in vivo and in vitro. *Proc Natl Acad Sci U S A* 2000;97:10389–94.
7. Cai Z, Chehab NH, Pavletich NP. Structure and activation mechanism of the CHK2 DNA damage checkpoint kinase. *Mol Cell* 2009;35:818–29.
8. Zannini L, Delia D, Buscemi G. CHK2 kinase in the DNA damage response and beyond. *J Mol Cell Biol* 2014;6:442–57.
9. White DE, Negorev D, Peng H, et al. Rauscher FJ, 3rd. KAP1, a novel substrate for PIKK family members, colocalizes with numerous damage response factors at DNA lesions. *Cancer Res* 2006; 66:11594–9.
10. Hu C, Zhang S, Gao X, et al. Roles of Kruppel-associated box (KRAB)-associated co-repressor KAP1 Ser-473 phosphorylation in DNA damage response. *J Biol Chem* 2012;287:18937–52.
11. Jallepalli PV, Lengauer C, Vogelstein B, et al. The Chk2 tumor suppressor is not required for p53 responses in human cancer cells. *J Biol Chem* 2003;278:20475–9.
12. Takai H, Naka K, Okada Y, et al. Chk2-deficient mice exhibit radioresistance and defective p53-mediated transcription. *EMBO J* 2002;21:5195–205.
13. Hirao A, Kong YY, Matsuoka S, et al. DNA damage-induced activation of p53 by the checkpoint kinase Chk2. *Science* 2000;287:1824–7.
14. Cao L, Kim S, Xiao CY, et al. ATM-CHK2-p53 activation prevents tumorigenesis at an expense of organ homeostasis upon Brca1 deficiency. *EMBO J* 2006;25:2167–77.
15. Weischer M, Bojesen SE, Ellervik C, et al. CHEK2*1100delC genotyping for clinical assessment of breast cancer risk: meta-analyses of 26,000 patient cases and 27,000 controls. *J Clin Oncol* 2008;26:542–8.
16. Liu C, Wang Y, Wang QS, et al. The CHEK2 I157T variant and breast cancer susceptibility: a systematic review and meta-analysis. *Asian Pac J Cancer Prev* 2012;13:1355–60.
17. Walsh T, Casadei S, Coats KH, et al. Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. *JAMA* 2006;295:1379–88.
18. Walsh T, Mandell JB, Norquist BM, 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.
19. Desrichard A, Bidet Y, Uhrhammer N, et al. CHEK2 contribution to hereditary breast cancer in non-BRCA families. *Breast Cancer Res* 2011;13:R119.
20. Le Calvez-Kelm F, Lesueur F, Damiola F, et al. Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer susceptibility: results from a breast cancer family registry case-control mutation-screening study. *Breast Cancer Res* 2011;13:R6.
21. Dufault MR, Betz B, Wappenschmidt B, et al. Limited relevance of the CHEK2 gene in hereditary breast cancer. *Int J Cancer* 2004;110:320–5.
22. Bell DW, Kim SH, Godwin AK, et al. Genetic and functional analysis of CHEK2 (CHK2) variants in multiethnic cohorts. *Int J Cancer* 2007;121: 2661–7.
23. Leedom TP, LaDuca H, McFarland R, et al. Breast cancer risk is similar for CHEK2 founder and non-founder mutation carriers. *Cancer Genet* 2016;209:403–7.
24. Couch FJ, Shimelis H, Hu C, et al. Associations between cancer predisposition testing panel genes and breast cancer. *JAMA Oncol* 2017;3:1190–6.
25. Fan Z, Ouyang T, Li J, et al. Identification and analysis of CHEK2 germline mutations in Chinese BRCA1/2-negative breast cancer patients. *Breast Cancer Res Treat* 2018;169:59–67.
26. Hauke J, Horvath J, Gross E, et al. Gene panel testing of 5589 BRCA1/2-negative index patients with breast cancer in a routine diagnostic setting: results of the German consortium for hereditary breast and ovarian cancer. *Cancer Med* 2018;7: 1349–58.
27. Young EL, Feng BJ, Stark AW, et al. Multigene testing of moderate-risk genes: be mindful of the missense. *J Med Genet* 2016;53:366–76.
28. Espenschied C, Kleiblova P, Richardson M, et al. Classifying variants in the CHEK2 gene: the importance of collaboration. *Eur J Cancer* 2017; 72:S25.
29. Baysal BE, DeLoia JA, Willett-Brozick JE, et al. Analysis of CHEK2 gene for ovarian cancer susceptibility. *Gynecol Oncol* 2004;95:62–9.
30. Suspitsin EN, Sherina NY, Ponomariova DN, et al. High frequency of BRCA1, but not CHEK2 or NBS1 (NBN), founder mutations in Russian ovarian cancer patients. *Hered Cancer Clin Pract* 2009;7:5.
31. Carter NJ, Marshall ML, Susswein LR, et al. Germline pathogenic variants identified in women with ovarian tumors. *Gynecol Oncol* 2018;151:481–8.
32. Kurian AW, Ward KC, Howlader N, et al. Genetic testing and results in a population-based cohort of breast cancer patients and ovarian cancer patients. *J Clin Oncol* 2019. https://doi.org/10.1200/JCO.18.01854.
33. Havranek O, Kleiblova P, Hojny J, et al. Association of Germline CHEK2 gene variants with risk and prognosis of non-Hodgkin lymphoma. *PLoS One* 2015;10:e0140819. [Epub ahead of print]
34. Soukupova J, Zemankova P, Lhotova K, et al. Validation of CZECANCA (CZEch CAncer paNel for clinical application) for targeted NGS-based analysis of hereditary cancer syndromes. *PLoS One* 2018;13:e0195761. [Epub ahead of print]
35. Roeb W, Higgins J, King MC. Response to DNA damage of CHEK2 missense mutations in familial breast cancer. *Hum Mol Genet* 2012;21:2738–44.
36. Delimitsou A, Fostira F, Kalfakakou D, et al. Functional characterization of CHEK2 variants in a *Saccharomyces cerevisiae* system. *Hum Mutat* 2019;40:631–48. https://doi.org/10.1002/humu.23728.
37. Ahn JY, Li X, Davis HL, et al. Phosphorylation of threonine 68 promotes oligomerization and autophosphorylation of the Chk2 protein kinase via the forkhead-associated domain. *J Biol Chem* 2002;277:19389–95.
38. Chrisanthar R, Knappskog S, Lokkevik E, et al. CHEK2 mutations affecting kinase activity together with mutations in TP53 indicate a functional pathway associated with resistance to epirubicin in primary breast cancer. *PLoS One* 2008; 3:e3062.

39. Mundt E, Nix P, Bowles KR, et al. *Complexities in hereditary cancer variant classification: three case examples.* ACMG Annual Clinical Genetics Meeting March 21–25, 2017, Poster#154, Phoenix Convention Center Phoenix, Arizona.
40. Wu X, Dong X, Liu W, et al. Characterization of CHEK2 mutations in prostate cancer. *Hum Mutat* 2006;27:742–7.
41. Southey MC, Goldgar DE, Winqvist R, et al. PALB2, CHEK2 and ATM rare variants and cancer risk: data from COGS. *J Med Genet* 2016;53: 800–11.
42. Nurmi A, Muranen TA, Pelttari LM, et al. Recurrent moderate-risk mutations in Finnish breast and ovarian cancer patients. *Int J Cancer* 2019. https://doi.org/10.1002/ijc.32309.
43. Girard E, Eon-Marchais S, Olaso R, et al. Familial breast cancer and DNA repair genes: insights into known and novel susceptibility genes from the GENESIS study, and implications for multigene panel testing. *Int J Cancer* 2019;144:1962–74.
44. Decker B, Allen J, Luccarini C, et al. Rare, protein-truncating variants in ATM, CHEK2 and PALB2, but not XRCC2, are associated with increased breast cancer risks. *J Med Genet* 2017; 54:732–41.
45. Slavin TP, Maxwell KN, Lilyquist J, et al. The contribution of pathogenic variants in breast cancer susceptibility genes to familial breast cancer risk. *NPJ Breast Cancer* 2017;3:22.
46. Schmidt MK, Hogervorst F, van Hien R, et al. Age- and tumor subtype-specific breast cancer risk estimates for CHEK2*1100delC carriers. *J Clin Oncol* 2016;34:2750–60.
47. Cybulski C, Wokolorczyk D, Jakubowska A, et al. Risk of breast cancer in women with a CHEK2 mutation with and without a family history of breast cancer. *J Clin Oncol* 2011;29:3747–52.
48. Weischer M, Bojesen SE, Tybjaerg-Hansen A, et al. Increased risk of breast cancer associated with CHEK2*1100delC. *J Clin Oncol* 2007;25: 57–63.
49. CHEK2 Breast Cancer Case-Control Consortium. CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. *Am J Hum Genet* 2004;74:1175–82.
50. Vahteristo P, Bartkova J, Eerola H, et al. A CHEK2 genetic variant contributing to a substantial fraction of familial breast cancer. *Am J Hum Genet* 2002;71:432–8.
51. Liang M, Zhang Y, Sun C, et al. Association between CHEK2*1100delC and breast cancer: a systematic review and meta-analysis. *Mol Diagn Ther* 2018;22:397–407.
52. Zhang B, Beeghly-Fadiel A, Long J, et al. Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *Lancet Oncol* 2011;12: 477–88.
53. Hallamies S, Pelttari LM, Poikonen-Saksela P, et al. CHEK2 c.1100delC mutation is associated with an increased risk for male breast cancer in Finnish patient population. *BMC Cancer* 2017;17:620.
54. Pritzlaff M, Summerour P, McFarland R, et al. Male breast cancer in a multi-gene panel testing cohort: insights and unexpected results. *Breast Cancer Res Treat* 2017;161:575–86.
55. Schmidt MK, Tollenaar RA, de Kemp SR, et al. Breast cancer survival and tumor characteristics in premenopausal women carrying the CHEK2*1100delC germline mutation. *J Clin Oncol* 2007;25:64–9.
56. Muranen TA, Greco D, Blomqvist C, et al. Genetic modifiers of CHEK2*1100delC-associated breast cancer risk. *Genet Med* 2017;19:599–603.
57. Kleibl Z, Havranek O, Novotny J, et al. Analysis of CHEK2 FHA domain in Czech patients with sporadic breast cancer revealed distinct rare genetic alterations. *Breast Cancer Res Treat* 2008; 112:159–64.
58. Muranen TA, Blomqvist C, Dork T, et al. Patient survival and tumor characteristics associated with CHEK2:p.I157T—findings from the breast cancer association consortium. *Breast Cancer Res* 2016; 18:98.
59. Huzarski T, Cybulski C, Domagala W, et al. Pathology of breast cancer in women with constitutional CHEK2 mutations. *Breast Cancer Res Treat* 2005;90:187–9.

Gene 637 (2017) 41–49

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Methodological paper

# Multiplex PCR and NGS-based identification of mRNA splicing variants: Analysis of BRCA1 splicing pattern as a model

Jan Hojny[a], Petra Zemankova[a], Filip Lhota[a], Jan Sevcik[a], Viktor Stranecky[b], Hana Hartmannova[b], Katerina Hodanova[b], Ondrej Mestak[c], David Pavlista[d], Marketa Janatova[a], Jana Soukupova[a], Michal Vocka[e], Zdenek Kleibl[a], Petra Kleiblova[a,f,*]

[a] *Institute of Biochemistry and Experimental Oncology, First Faculty of Medicine, Charles University, Prague 12853, Czech Republic*
[b] *Institute of Inherited Metabolic Disorders, First Faculty of Medicine, Charles University and General University Hospital in Prague, Prague 120 00, Czech Republic*
[c] *Department of Plastic Surgery, First Faculty of Medicine, Charles University and Na Bulovce Hospital, Prague 180 81, Czech Republic*
[d] *Department of Obstetrics and Gynecology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Prague 120 00, Czech Republic*
[e] *Department of Oncology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Prague 120 00, Czech Republic*
[f] *Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University and General University Hospital in Prague, Prague 120 00, Czech Republic*

## ARTICLE INFO

*Keywords:*
Alternative splicing
mRNA splicing variant
BRCA1
Multiplex PCR
NGS

## ABSTRACT

Alternative pre-mRNA splicing increases transcriptome plasticity by forming naturally-occurring alternative splicing variants (ASVs). Alterations of splicing processes, caused by DNA mutations, result in aberrant splicing and the formation of aberrant mRNA isoforms. Analyses of hereditary cancer predisposition genes reveal many DNA variants with unknown clinical significance (VUS) that potentially affect pre-mRNA splicing. Therefore, a comprehensive description of ASVs is an essential prerequisite for the interpretation of germline VUS in high-risk individuals.

To identify ASVs in a gene of interest, we have proposed an approach based on multiplex PCR (mPCR) amplification of all theoretically possible exon-exon junctions and subsequent characterization of size-selected and pooled mPCR products by next-generation sequencing (NGS). The efficiency of this method is illustrated by a comprehensive analysis of *BRCA1* ASVs in human leukocytes, normal mammary, and adipose tissues and stable cell lines.

We revealed 94 BRCA1 ASVs, including 29 variants present in all tested samples. While differences in the qualitative expression of BRCA1 ASVs among the analyzed human tissues were minor, larger differences were detected between tissue and cell line samples.

Compared with other ASV analysis methods, this approach represents a highly sensitive and rapid alternative for the identification of ASVs in any gene of interest.

## 1. Introduction

Hereditary mutations in cancer-susceptibility genes are responsible for tumor development in about 5% of all cancer patients. The carriers of hereditary mutations face a high life-time risk of cancer, which often develops at an early age (Rahman, 2014). Tailored care improving life expectancy in these high-risk individuals requires an unequivocal identification of causative mutations in hundreds of known cancer-susceptibility genes. The recent introduction of next-generation sequencing (NGS) into clinical diagnostics enables a simultaneous analysis of multiple genes; however, its clinical utility is hampered by the presence of many variants with unknown significance (VUS) (Cheon et al., 2014). These genetic changes emerge as rare germline missense, silent or intronic variants with an uncertain biological and functional impact on the resulting protein isoform. The number of identified VUS rises proportionally to the length of the analyzed genomic sequence, and many of them may alter mRNA splicing processes (Tavtigian & Chenevix-Trench, 2014).

Pre-mRNA splicing controls the composition of matured mRNA by regulated intron exclusion and exon linking. A primary wild-type (wt) transcript (pre-mRNA) can be variably processed by alternative splicing into alternative mRNA variants translated into protein isoforms with

*Abbreviations:* ASV(s), alternative splicing variant(s); mPCR, multiplex PCR; NGS, next generation sequencing; SAS, splice acceptor shift; SDS, splice donor shift; VUS, variant of unknown clinical significance

[*] Corresponding author at: Institute of Biochemistry and Experimental Oncology, First Faculty of Medicine, Charles University, U Nemocnice 5, Prague 2 128 53, Czech Republic.
*E-mail address:* petra.kleiblova@lf1.cuni.cz (P. Kleiblova).

http://dx.doi.org/10.1016/j.gene.2017.09.025
Received 28 June 2017; Received in revised form 26 July 2017; Accepted 13 September 2017
Available online 14 September 2017

different biological activities (Bentley, 2014). Alternative splicing must be distinguished from aberrant splicing resulting from a dysregulation of natural splice site recognition caused by DNA mutations. The DNA sequence variants affecting pre-mRNA splicing are more prevalent than estimated up to now, and they account for at least 15% of disease-causing mutations, and for up to 50% of all mutations described in some genes (Caminsky et al., 2015; Soukarieh et al., 2016). Various splicing assays help to disclose the impact of VUS on splicing processes (Whiley et al., 2014), and variants that cause aberrant splicing are considered pathogenic. The evaluation of aberrant splicing requires a precise knowledge of alternative splicing variants (ASVs) for the analyzed primary transcript (Colombo et al., 2014). Despite large-scale RNA sequencing projects (e.g. ENCODE, GTEx), there is no precise catalogue of ASVs or validated RNAseq data for most clinically-relevant genes (Sloan et al., 2016; Baralle & Buratti, 2017).

Recently, two articles have described a comprehensive analysis of naturally occurring splicing variants in *BRCA1* (Colombo et al., 2014; Romero et al., 2015), one of the most studied cancer-susceptibility genes responsible for hereditary breast and ovarian cancer (Kleibl & Kristensen, 2016). In both studies, the RNA-based analysis involved a combination of various techniques including RT-PCR, exon scanning, cloning, sequencing, and relative (semi)quantification. This experimental variability negatively affects the reproducibility of splicing variant analyses from various mRNA sources and makes the methods difficult to use in the analyses of other gene products. The large number of these analytic techniques makes the analysis laborious, may negatively affects its reproducibility, and adaptation to characterize another gene transcripts. Therefore, we aimed to develop a versatile approach suitable for the characterization of ASVs in any gene of interest based on NGS of multiplex PCR-generated amplicons covering all theoretically possible exon-exon junctions.

## 2. Materials and methods

### 2.1. An overview of experimental design

We aim to characterize the ASVs of any gene from an RNA sample, dominantly on the qualitative level. The analysis comprises four steps: i) multiplex PCR (mPCR) amplification of all theoretically possible mRNA splicing variants from a cDNA template, ii) pooling of mPCRs, purification and size selection of pooled mPCR products targeting short amplicons, iii) standard NGS library preparation from size-selected mPCR fragments followed by routine Illumina sequencing, and iv) a bioinformatics analysis. We have demonstrated the efficiency of the mPCR/NGS approach by the characterization of BRCA1 ASVs because i) the *BRCA1* gene is the most frequently altered breast cancer susceptibility gene in many countries including the Czech Republic and many BRCA1 VUS contribute to aberrant splicing, ii) 63 BRCA1 mRNA variants were recently described using conventional RT-PCR and capillary-electrophoresis by Colombo et al. (2014) and Romero et al. (2015), indicating that iii) the BRCA1 mRNA splicing isoform pattern is highly variable.

### 2.2. Alternative BRCA1 splice site nomenclature

All alternative splicing events were classified into biotypes based on previously published nomenclature (Colombo et al., 2014; Romero et al., 2015). The insertions (▼) and deletions (Δ) denote splicing events affecting a single exon (cassette) or > 1 consecutive exons (multicassette). The deletions affecting the 5′ and 3′ ends of an exon were described as an exon number with an added "p" or "q", respectively. The extension of an exon sequence into an adjacent intronic region is described as an exon number with an "a". The splice donor/acceptor shift (SDS/SAS) variants were identified and counted as NGS reads with deletions of nucleotides at the exon-exon junctions or insertions of intronic parts flanking to the 5′ or 3′ ends of an exon. The mixed biotypes denoted combinations of the above-mentioned events. The *BRCA1* exons were numbered according to the Breast Cancer Information Core Database (https://research.nhgri.nih.gov/bic/) nomenclature.

### 2.3. Patients and samples

The characterization of BRCA1 ASVs was performed in 96 RNA samples obtained from 32 individuals (Supplementary Table S1), including 16 non-cancer controls, eight breast-cancer (BC) patients without *BRCA1* mutation, and eight *BRCA1*-mutation carriers. Simultaneously-obtained tissue samples were collected during BC surgery or preventive mastectomy (in BC patients and *BRCA1* mutation carriers) or during cosmetic breast surgery (in controls). All enrolled individuals were Caucasians of a Czech origin who gave a written informed consent approved by ethical committees to participate in the study. RNA samples were isolated from the leukocytes and macroscopically dissected fresh mammary and adipose perimammary tissues of each individual. We further analyzed RNA samples from stable human cell lines (from MCF7 cells, and from pooled EM-G3, HeLa, and MDA-MB-231 cells). The cell lines were maintained as described previously (Brozova et al., 2007; Sevcik et al., 2012; Sevcik et al., 2013; Vondruskova et al., 2008).

#### 2.3.1. Total RNA isolation, quality control and cDNA synthesis

All RNA samples were processed according to MIQE guidelines (Bustin et al., 2009). Peripheral blood samples (2.5 ml) were collected into PAXgene Blood RNA tubes, incubated overnight at room temperature, and stored at − 20 °C. All stored samples were thawed and stored for 2 h at room temperature before RNA isolation performed with PAXgene Blood RNA Kit (PreAnalytiX). Solid tissue samples (~ 100 mg/sample) were submerged into 1 ml of RNAlater (Qiagen) immediately after surgical excision, processed according to the manufacturer, and after overnight incubation (2–8 °C) stored at − 80 °C until RNA isolation. Forty micrograms of thawed, RNAlater-preserved samples were homogenized using MagNA Lyser Green Beads tubes on MagNA Lyser Instrument (Roche) in the presence of 1 ml Qiazol (Qiagen). Total RNAs from the homogenated tissues and cultured cells were isolated with RNeasy Tissue Mini Kit (Qiagen).

All RNA samples were treated by DNase I, quantified on NanoDrop 1000 (Thermo Fisher Scientific) and characterized by the RNA integrity number (RIN) using Bioanalyzer 2100 with RNA 6000 Nano Kit (Agilent Technologies; tissue samples $\text{RIN}_{\text{mean}} = 7.4$; range 6.3–8.9).

Overall, 1.5 μg of RNA was used for cDNA synthesis (in a reaction volume of 20 μl). The cDNA synthesis was performed using SuperScript III Reverse Transcriptase (Thermo Fisher Scientific) and random hexamers (Roche) as described previously (Kleiblova et al., 2010). A routine PCR control of cDNA quality/integrity was performed prior to further analyses (not shown).

### 2.4. Multiplex PCR (mPCR) amplification and size selection

#### 2.4.1. Primer designing

The primers were designed to specifically cover all possible exon-exon junctions. The resulting PCR amplicons thus enable the identification of all canonical as well as alternative splicing mRNA isoforms. Forward primers targeted the 3′ region while reverse primers aimed at the 5′ region of an exon (Supplementary Fig. S1). For the analysis of a single gene transcript consisting of *N* exons, the number of *N*-2 forward and *N*-2 reverse primers is required for the amplification of all theoretically possible exon-exon junctions in at least *N*-2 mPCR reactions that are finally pooled into one mPCR pool.

For the analysis of BRCA1 ASVs, we designed 45 primers targeting 22 coding exons of the canonical BRCA1 transcript (NM_007294), and alternative exons 11q and 13A (Fig. 1A, Supplementary Table S2). All individual PCRs were optimized separately (Supplementary Fig. S2)

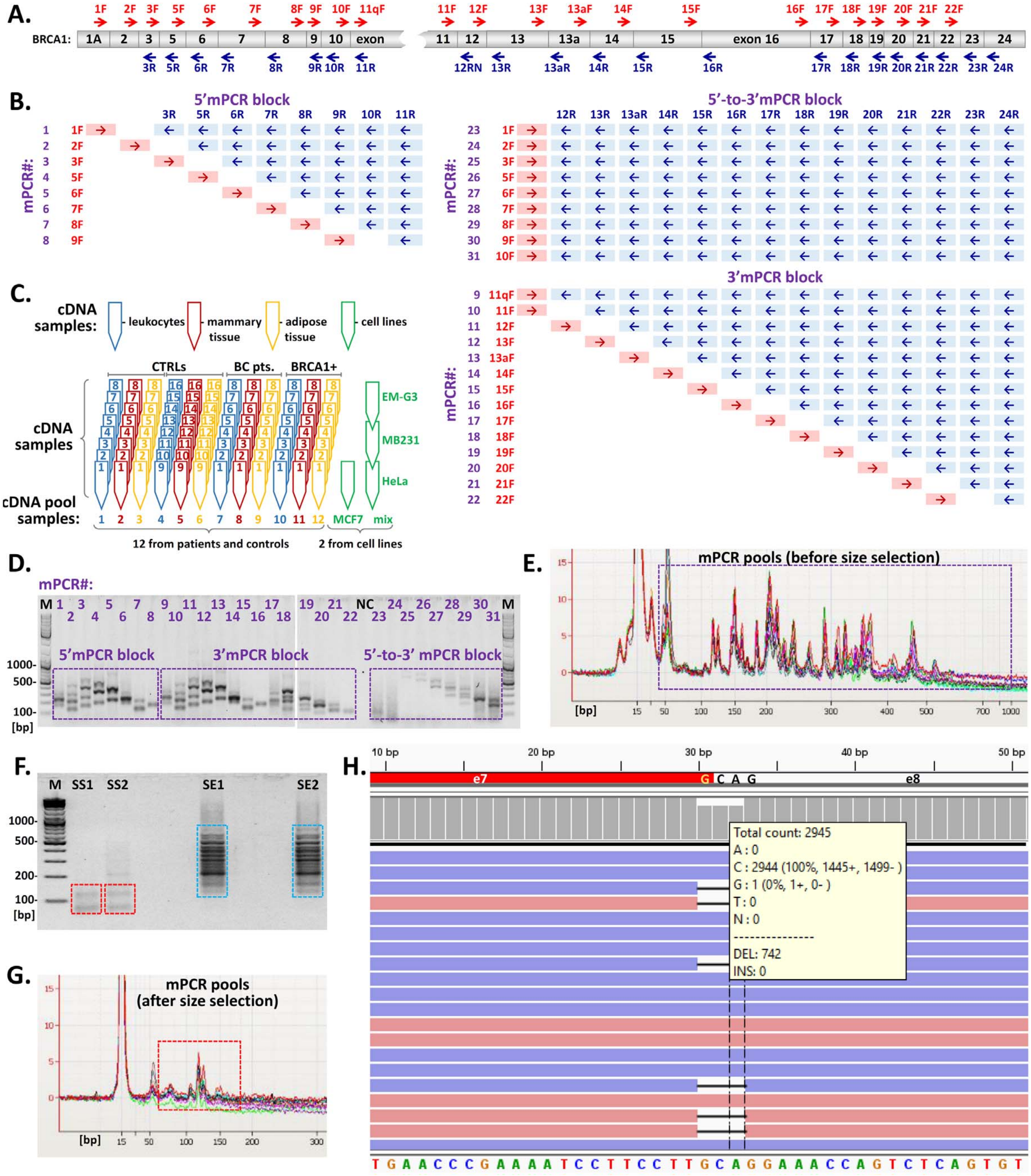


**Fig. 1.** Method overview. The chart (A) shows primer pairs (forward – red; reverse – blue) used for mPCR amplifications of all theoretically possible BRCA1 ASVs. Due to the presence of the large exon 11, 31 mPCR reactions (violet letters; B) were performed in three 'blocks'. All 31 mPCR reactions were performed with each of 14 cDNA pools (C). Twelve cDNA pools of human cDNA samples and two cDNA pools from cell line samples served as templates for 31 mPCR reactions. Agarose gel electrophoresis of 31 mPCRs (ranging in size between ~50 and ~700 bp; purple dashed line) from a single cDNA pool is shown in (D). All 31 mPCRs from each cDNA pool sample were further pooled together (to create an mPCR pool) and analyzed on Agilent Bioanalyzer (E; the overlaying electrophoretograms of 12 mPCR pools show good reproducibility). The double-sided size selection was used to enrich the short mPCR amplicons that were subsequently used for NGS library preparation. The agarose gel electrophoresis (F) displays the enrichment of size-selected fragments (SS; red dashed line boxes) while size-excluded fragments (SE; blue dashed line boxes) were discarded. The size-selected (SS) samples (ranging at 50–150 bp in length; red dashed line box in G) were verified by Agilent Bioanalyzer. The MiSeq reads were mapped to the bam files (Supplementary Table S3) containing all theoretically possible BRCA1 splicing cassette and multicassete events. Visual inspection of reads in IGV viewer enabled direct assessment and quantification of SDS/SAS as shown (H) for the ΔCAG at the 5′ end of exon 8 (erroneously mapped as ΔGCA; coverage depth is shown as grey vertical bars; forward (pink) and reverse (blue) reads are shown as vertical bars). In normal mammary tissue of BC patients (shown in H); ΔCAG accounted for 742 reads (20.1%), while 2945 sequencing reads were recorded for wt sequence (Supplementary Table S6). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and subsequently in mPCR (Supplementary Fig. S3).

### 2.4.2. mPCR amplification and preparation of mPCR pools

All theoretically possible splicing events of a gene of interest can be amplified in a few mPCRs. The number of mPCRs depends primarily on the number of exons of the analyzed gene. Every individual mPCR contains a single forward primer and a set of reverse primers targeting all consecutive exons except the exon directly flanking to the exon targeted by the forward primer (Supplementary Fig. S1). The undesired synthesis of the amplicons of several consecutive exons was reduced by a short elongation time (Supplementary Fig. S3B).

The analysis of BRCA1 ASVs in each cDNA template required 31 mPCRs (Fig. 1B). To overcome the usually limited amount of RNA (and resulting cDNA) available from human tissue samples, the cDNA template for mPCRs was prepared by the pooling of individual cDNAs (from the same tissue and sample subgroup; Fig. 1C). This may be necessary especially with multiexonic and low-expressed genes (including *BRCA1*) which require a larger number of mPCRs consuming an increased amount of cDNA. Although the cDNA pooling resulted in a loss of information about the expression of ASVs in the individual cDNA samples, it increased the chances of detecting low-expressed ASVs.

For the BRCA1 analysis, eight individual cDNA samples (16 μl each) from the same tissue type and patient group were pooled together to obtain 12 patient cDNA pools (each 128 μl; Fig. 1C). We also analyzed cDNAs from stable human cell lines (cDNA from MCF7 cells and a pool of cDNAs from EM-G3, HeLa and MD-MB-231). Each of the 12 patient cDNA pools, MCF7 cDNA, and cDNA cell line pools served as a template for 31 mPCRs. Each 40 μl mPCR contained 4 μl of pooled cDNA template (equivalent to 300 ng of RNA), a single forward primer (final concentration 225 nM), a variable set of reverse primers (final concentration 75 nM of each), and FastStart Taq DNA Polymerase (Roche) according to the manufacturer's instructions. The mPCR amplifications involved 4-minute incubation at 95 °C followed by 35 cycles (95 °C for 10 s, 62 °C for 20 s, and 72 °C for 15 s) and final extension at 72 °C for 7 min. The individual mPCR products were analyzed electrophoretically (Fig. 1D). After that, 35 μl of each of the 31 mPCRs from a single cDNA pool were mixed together to provide an mPCR pool. The resulting mPCR pools were characterized by capillary electrophoresis using 2100 Bioanalyzer and DNA 1000 Kit (Agilent Technologies; Fig. 1E).

### 2.4.3. Size selection and purification of mPCR pools

To reduce the presence of longer mPCR-amplified fragments containing short consecutive exons, the mPCRs pools were subjected to size selection using double-sided solid phase reversible immobilization with magnetic beads. The size-selected and purified amplicons served as templates to prepare a standard NGS library.

As the length of the targeted mPCR amplicons of BRCA1 ASVs was expected to range mostly at 80–90 bp, we performed size selection using magnetic beads to remove undesired amplicons (< 50 bp and > 150 bp; Fig. 1F). First, we used 1.8 × concentration of Agencourt AMPure XP reagent (Beckman Coulter) to bind and remove amplicons > 150 bp (dominantly containing PCR products amplified from the canonical BRCA1 mRNA). Subsequently, we mixed the supernatant from the first reaction with a reagent to the final 2.5 × concentration to withdraw DNA fragments > 50 bp in length. The size-selected and purified samples were characterized on Agilent Bioanalyzer with DNA 1000 Kit (Fig. 1G).

### 2.4.4. NGS library preparation, MiSeq sequencing

First, Dynazyme II DNA Polymerase (Thermo Fisher Scientific) was used to create 3′-dA overhangs in DNA fragments in size-selected and purified mPCR pools. Subsequently, Illumina sequencing adaptors were ligated using Rapid DNA Ligation Kit (Thermo Fisher Scientific). Finally, the seven-cycle PCR reaction (NEBNext High Fidelity PCR Master Mix, NEB) with a universal primer introduced a 6-bp index sequence unique for each mPCR pool. The processed samples were purified using Agencourt AMPure XP Reagent after each step of sequencing library preparation. The quality of the prepared libraries was characterized on 2100 Bioanalyzer and quantified fluorimetrically (Qubit; Thermo Fisher Scientific). To achieve sufficient sequencing coverage and sample diversity, we admixed our mPCR-prepared libraries with panel sequencing libraries of high complexity to standard MiSeq runs (one mPCR pool library represented 1/30 of sequencing capacity). Runs were sequenced with MiSeq Reagent Kit v3 (150 cycle).

### 2.4.5. Bioinformatics

Bioinformatics requires double-step mapping. First, sequencing data are mapped to a user-designed fasta file containing the exon-exon junctions of all theoretically possible splicing variants of the gene of interest. Thus we can identify all cassette and multicassette splicing deletions and variants resulting from short SDS/SAS. Second, sequencing data are mapped to the genomic sequence of the analyzed gene in order to identify the exonized intronic sequences.

In our BRCA1 analysis, the primary raw data sets (in the fastq.gz format) were processed by a routine bioinformatics pipeline using software tools specified below using the default settings (if not otherwise specified in the list of commands listed in Supplementary Table S7). The remaining sequences were trimmed (to remove adapters and low-quality bases) before mapping by Trimmomatic (ver. 0.32; http://www.usadellab.org). First, we mapped raw data sets to the prepared fasta file using Novoalign (ver. 2.08.03; Novocraft). Reads with insufficient sequencing quality were removed. This “BRCA1 splicing” fasta file contained 311 sequences (“exon-exon_computed” in Supplementary Table S3) which considered all possible combinations of known BRCA1 exons (NM_007294), including alternative exon 13A, and known SDS/SAS combinations > 10 bp (11q, and 5q). Output in the SAM format was transformed to BAM by Picard tools (ver. 1.129; https://broadinstitute.github.io/picard/) and displayed in IGV (Integrative Genomics Viewer, Broad Institute; Fig. 1H). Coverage statistics were created by SAMtools (ver. 0.1.19; http://samtools.sourceforge.net/). Since this approach ignored the presence of exonized intron sequences, we further mapped all raw data to the *BRCA1* gene sequence (81,189 bp from NG_005905 spanning sequence 92,500–173,688). The mapping results were analyzed in IGV manually to remove reads with soft-clipped bases. Mapped reads that exceeded from exons into flanking introns or reads mapped to deep intronic sequences, respectively, were recorded including sequences of unmapped nucleotides. The unmapped parts of reads were BLASTed with the *BRCA1* gene sequence (https://blast.ncbi.nlm.nih.gov/). If the unmapped sequence contained the 5′ or 3′ parts of a *BRCA1* exon, we were able to identify the entire intronic insertion. Sequences comprising the identified intronic insertions with flanking exonic sequences were added to a fasta file (“from_IGV_BLASTed” in Supplementary Table S3) and used for new mapping from the original dataset. Mapped reads were manually inspected in IGV in order to eliminate incorrectly mapped reads with soft-clipped bases.

To compare the numbers of ASV reads among the examined mPCR pools, we expressed the number of sequencing reads of an ASV as value normalized to $10^6$ reads in the given mPCR pool (Supplementary Tables S4 and S5).

## 3. Results

In order to prepare a versatile method for a direct assessment of splice junction events, we designed the mPCR/NGS-based approach enabling the identification of ASVs in a gene of interest. Using the described method, we identified 94 BRCA1 ASVs (Table 1) comprising all previously described biotypes (Colombo et al., 2014). Our analysis of simultaneously-obtained tissue RNA samples revealed that the highest number of ASVs were expressed in mammary tissue (72 variants), followed by leukocytes and perimammary adipose tissues (67 and 54 variants, respectively).

Forty-eight ASVs were identified in all examined human tissue types, with 29 of them expressed in each cDNA tissue sample (referred here as “ubiquitous”; Fig. 2; Table 1).

Only slight qualitative differences were identified among tissue samples from non-cancer controls, BC patients and *BRCA1*-mutation carriers. In contrast, the spectrum of ASVs differed between tissue and cell-line samples (Fig. 2). Altogether, 76 ASVs (11 of them exclusively) were expressed in cell lines.

The detected variants included 25 in-frame ASVs that may

**Table 1**
Description of all BRCA1 ASVs identified in this study, including variant name, systematic description of the variant at the cDNA level, functional annotation, and biotype class. The values for the expression of particular variants in analyzed cell lines and human tissue samples show normalized sequencing reads (reads per $10^6$ reads), the colours indicate normalized sequencing coverage: 0 reads = white; 1–9 reads = green; 10–99 reads = yellow; 100–999 reads = light red; 1000–9999 reads = red; > 9999 reads = dark red. The "ubiquitous" variants (expressed in all analyzed human tissue samples) are highlighted by bold letters and blue lines. The presence of variants in the analyzed human tissue samples and cell lines was compared to that reported as predominant (P), present (1), or absent (0) in studies by Colombo et al. (2014), Romero et al. (2015) and Orban and Olah (2003). The extended version of this table is shown in Supplementary Table S5.

| Variant description | HGVS description | Functional annotation | Biot ype | Analyzed cell lines | | Analyzed human tissues | | | | | | | | | References | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Leukocytes | | | Mammary | | | Adipose | | | Leuko. | | Mam. | | Tu | |
| | | | | MCF7 | mix | Non-cancer controls | BC patients | BRCA1 mut. carriers | Non-cancer controls | BC patients | BRCA1 mut. carriers | Non-cancer controls | BC patients | BRCA1 mut. carriers | Colombo (2014) | Romero (2015) | Colombo (2014) | Romero (2015) | Romero (2015) | Orban Olah (2003) |
| **1Aq** | **c.-25_-20del6** | **UTR** | **SDSΔ** | **1185** | **7854** | **3632** | **5045** | **4210** | **2981** | **11439** | **4010** | **2508** | **6035** | **4955** | **P** | **P** | **P** | **P** | **P** | **1** |
| 1Aq, 2a | c.-25_-20del6, c.-19-59_-19-1 ins59 | UTR | SDSΔ + SAS ▼ | 0 | 0 | 0 | 0 | 0 | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1Aq, Δ2** | **c.-25_80del105** | **n.c.** | **SDSΔ + CΔ** | **176** | **5032** | **40** | **101** | **150** | **46** | **191** | **28** | **59** | **82** | **205** | **1** | **1** | **1** | **0** | **0** | **0** |
| 1Aq, Δ2_3 | c.-25_134del159 | n.c. | SDSΔ + mCΔ | 0 | 1575 | 0 | 202 | 13 | 14 | 46 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| 1Aq, Δ2_5 | c.-25_212del237 | n.c. | SDSΔ + mCΔ | 0 | 438 | 0 | 0 | 0 | 2 | 0 | 83 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| 1Aq, Δ2_5, 6p | c.-25_282del307 | n.c. | SDSΔ + mCΔ + SASΔ | 0 | 350 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1Aq, Δ2_7 | c.-25_441del466 | n.c. | SDSΔ + mCΔ | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1Aq, Δ2_7, 8p | c.-25_444del469 | n.c. | SDSΔ + mCΔ + SASΔ | 0 | 0 | 0 | 34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1Aq, Δ2_10 | c.-25_670del695 | n.c. | SDSΔ + mCΔ | 0 | 438 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| 1Aq, Δ2_17 | c.-25_5074del5099 | n.c. | SDSΔ + mCΔ | 88 | 1225 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1Aq, Δ2_19 | c.-25_5193del5218 | n.c. | SDSΔ + mCΔ | 0 | 0 | 0 | 0 | 0 | 0 | 289 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1aA** | **c.-20+1_-20+89ins89** | **UTR** | **SDS ▼** | **6824** | **12098** | **321** | **219** | **1211** | **1501** | **6557** | **5429** | **940** | **427** | **349** | **1** | **1** | **1** | **0** | **P** | **0** |
| **2p** | **c.-19_-7del13** | **UTR** | **SASΔ** | **88** | **963** | **140** | **118** | **232** | **77** | **214** | **28** | **76** | **238** | **164** | **1** | **P** | **1** | **0** | **P** | **0** |
| **Δ2** | **c.-19_80del99** | **n.c.** | **CΔ** | **176** | **3982** | **482** | **1063** | **1987** | **323** | **3305** | **1337** | **481** | **1308** | **462** | **1** | **P** | **1** | **P** | **P** | **0** |
| Δ2_3 | c.-19_134del153 | n.c. | mCΔ | 0 | 1575 | 0 | 0 | 0 | 36 | 364 | 0 | 5 | 0 | 82 | 1 | P | 1 | 0 | P | 0 |
| Δ2_3, ▼4 | c.-19_134del153 + c.135-4047_135-3932ins116 | n.c. | mCΔ + Cq | 110 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 51 | 1 | 1 | 0 | 0 | 0 | 0 |
| Δ2_5 | c.-19_212del231 | n.c. | mCΔ | 88 | 1203 | 20 | 0 | 0 | 73 | 35 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| Δ2_7, 8p | c.-19_444del463 | n.c. | mCΔ + SASΔ | 0 | 2253 | 10 | 169 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ2_10 | c.-19_670del689 | n.c. | mCΔ | 0 | 88 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| Δ2_17 | c.-19_5074del5093 | n.c. | mCΔ | 0 | 2625 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ2_19 | c.-19_5193del5212 | n.c. | mCΔ | 0 | 0 | 0 | 101 | 0 | 0 | 976 | 0 | 0 | 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **▼ 145 bp int 2** | **c.81-3486_81-3342ins145** | **FS** | **C▼** | **636** | **459** | **40** | **34** | **72** | **24** | **87** | **276** | **86** | **100** | **82** | **0** | **0** | **0** | **0** | **0** | **0** |
| **Δ3** | **c.81_134del54** | **FS** | **CΔ** | **987** | **21089** | **3120** | **10293** | **8436** | **4141** | **9336** | **9438** | **2135** | **5217** | **6750** | **1** | **P** | **1** | **0** | **P** | **1** |
| Δ3, ▼4 | c.81_134del54 + c.135-4047_135-3932ins116 | FS | CΔ + C ▼ | 22 | 44 | 0 | 0 | 0 | 44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ3_5 | c.81_212del132 | IF | mCΔ | 0 | 875 | 60 | 405 | 215 | 109 | 474 | 165 | 108 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ▼ 116bp int 3 | c.134+3124_134+3239ins116 | FS | C▼ | 461 | 1663 | 20 | 0 | 26 | 61 | 23 | 83 | 38 | 212 | 62 | 0 | 0 | 0 | 0 | 0 | 0 |
| **▼ 4** | **c.135-4047_135-3932ins116** | **FS** | **C▼** | **7460** | **2188** | **281** | **489** | **238** | **367** | **306** | **276** | **470** | **279** | **574** | **1** | **1** | **1** | **0** | **P** | **0** |
| **Δ5** | **c.135_212del78** | **IF** | **C▼** | **395** | **7351** | **8016** | **13043** | **8569** | **5524** | **7476** | **10169** | **4778** | **4013** | **8823** | **P** | **1** | **P** | **P** | **P** | **1** |
| Δ5_6 | c.135_301del167 | FS | mCΔ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 634 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ5_6, 7p | c.135_307del173 | FS | mCΔ + SASΔ | 0 | 0 | 0 | 0 | 333 | 0 | 127 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ5_7 | c.135_441del307 | FS | mCΔ | 0 | 481 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ5_7, 8p | c.135_444del310 | FS | mCΔ + SASΔ | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ5_9 | c.135_593del459 | IF | mCΔ | 0 | 44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **5q** | **c.191_212del22** | **FS** | **SDSΔ** | **549** | **1269** | **3602** | **2818** | **2137** | **8023** | **4501** | **3624** | **5886** | **9339** | **4268** | **P** | **P** | **P** | **P** | **P** | **1** |
| 5q, Δ6 | c.191_301del111 | IF | SDSΔ + CΔ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Δ6 | c.213_301del89 | FS | CΔ | 44 | 0 | 40 | 0 | 0 | 65 | 35 | 0 | 0 | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **6q** | **c.293_301del9** | **IF** | **SDSΔ** | **0** | **44** | **431** | **371** | **695** | **202** | **116** | **28** | **124** | **7** | **62** | **0** | **0** | **0** | **0** | **0** | **0** |
| **8p** | **c.442_444del3** | **IF** | **SASΔ** | **197** | **7022** | **15079** | **10883** | **8566** | **7283** | **4287** | **4657** | **6356** | **5024** | **4011** | **P** | **P** | **P** | **P** | **P** | **1** |
| Δ8 | c.442_547del106 | FS | CΔ | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 55 | 92 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ8_9 | c.442_593del152 | FS | mCΔ | 0 | 0 | 0 | 0 | 20 | 141 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | P | 0 |
| Δ8_10 | c.442_670del229 | FS | mCΔ | 0 | 88 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | P | 0 |
| Δ8_16 | c.442_4986del4545 | IF | mCΔ | 88 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ▼ 94 bp int 8 | c.548-297_548-204ins94 | FS | C▼ | 110 | 481 | 251 | 202 | 457 | 71 | 17 | 69 | 238 | 0 | 154 | 0 | 0 | 0 | 0 | 0 | 0 |
| **▼ 97 bp int 8** | **c.548-300_548-204ins97** | **FS** | **C▼** | **461** | **438** | **371** | **456** | **757** | **244** | **312** | **248** | **286** | **204** | **92** | **0** | **0** | **0** | **0** | **0** | **0** |
| **Δ9** | **c.548_593del46** | **FS** | **CΔ** | **1295** | **9888** | **3642** | **12334** | **17110** | **2295** | **7303** | **9301** | **2092** | **7042** | **8689** | **P** | **1** | **P** | **0** | **P** | **1** |
| **Δ9_10*** | **c.548_670del123** | **IF** | **mCΔ** | **4542** | **19558** | **562** | **1063** | **2261** | **998** | **2155** | **3362** | **940** | **1561** | **5109** | **P** | **P** | **P** | **P** | **P** | **1** |
| Δ9_11 | c.548_4096del3549 | IF | mCΔ | 1009 | 16079 | 100 | 270 | 454 | 2 | 0 | 413 | 22 | 0 | 923 | P | 1 | P | 0 | 0 | 1 |
| Δ9_12 | c.548_4185del3638 | FS | mCΔ | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 165 | 16 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **10a** | **c.594-21_594-1ins21** | **IF** | **SAS▼** | 4827 | 2056 | 1575 | 5113 | 594 | 430 | 4957 | 1157 | 665 | 465 | 1036 | **0** | **0** | **0** | **0** | **0** | **0** |
| Δ10 | c.594_670del77 | FS | CΔ | 154 | 1684 | 30 | 34 | 0 | 12 | 0 | 0 | 27 | 7 | 0 | 1 | 1 | 1 | 0 | P | 0 |
| Δ10_11 | c.594_4096del3503 | FS | mCΔ | 0 | 4922 | 70 | 101 | 131 | 22 | 0 | 0 | 0 | 0 | 513 | 1 | 1 | 1 | 0 | 0 | 0 |
| Δ10_12 | c.594_4185del3592 | FS | mCΔ | 0 | 416 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 89 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| **Δ11** | **c.671_4096del3426** | **IF** | **CΔ** | 439 | 5163 | 60 | 337 | 104 | 69 | 185 | 303 | 265 | 71 | 246 | **1** | **1** | **1** | **0** | **0** | **1** |
| Δ11_12 | c.671_4185del3515 | FS | mCΔ | 0 | 0 | 0 | 0 | 0 | 10 | 23 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| Δ11_12, 13p | c.671_4188del3518 | FS | mCΔ + SASΔ | 0 | 175 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| **11q** | **c.788_4096del3309** | **IF** | **SDSΔ** | 6012 | 16451 | 823 | 928 | 1965 | 2971 | 1491 | 5883 | 1048 | 2557 | 7879 | **P** | **1** | **P** | **0** | **0** | **1** |
| 11 Δ3094 | c.788_3881del3094 | FS | Intronization | 241 | 219 | 0 | 0 | 39 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 Δ3110 | c.788_3897del3110 | FS | Intronization | 2041 | 700 | 40 | 67 | 7 | 32 | 0 | 220 | 32 | 223 | 41 | 1 | 1 | 0 | 0 | 0 | 0 |
| 11 Δ3240 | c.788_4027del3240 | IF | Intronization | 614 | 0 | 80 | 34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| **13p** | **c.4186_4188del3** | **IF** | **SASΔ** | 263 | 1378 | 3652 | 3881 | 9294 | 2684 | 560 | 3114 | 3297 | 847 | 6217 | **P** | **1** | **P** | **0** | **P** | **0** |
| **Δ13** | **c.4186_4357del172** | **FS** | **CΔ** | 44 | 1006 | 70 | 405 | 111 | 36 | 69 | 138 | 86 | 141 | 482 | **1** | **1** | **0** | **0** | **P** | **0** |
| Δ13, 14p | c.4186_4360del175 | FS | CΔ + SASΔ | 44 | 219 | 20 | 0 | 0 | 6 | 12 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| Δ13_14 | c.4186_4484del299 | FS | mCΔ | 0 | 2231 | 0 | 675 | 333 | 0 | 0 | 0 | 76 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ13_15 | c.4186_4675del490 | FS | mCΔ | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **▼ 13A** | **c.4358-2785_4358-2720ins66** | **IF** | **C▼** | 6495 | 24721 | 3963 | 1974 | 7267 | 873 | 1196 | 703 | 1254 | 5953 | 4309 | **1** | **1** | **1** | **0** | **P** | **0** |
| **▼ 13A, 14p** | **c.4358-2785_4358-2720ins66 + c.4358_4360del3** | **IF** | **C▼ + SASΔ** | 6604 | 40494 | 130 | 169 | 166 | 75 | 12 | 14 | 49 | 123 | 21 | **1** | **1** | **1** | **0** | **0** | **0** |
| ▼ 13A, Δ14 | c.4358-2785_4358-2720ins66 + c.4358_4484del127 | FS | C▼ + CΔ | 0 | 0 | 0 | 0 | 819 | 115 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **14p** | **c.4358_4360del3** | **IF** | **SASΔ** | 4125 | 3238 | 542 | 1282 | 917 | 333 | 335 | 358 | 303 | 892 | 533 | **P** | **P** | **P** | **P** | **P** | **0** |
| **Δ14** | **c.4358_4484del127** | **FS** | **CΔ** | 417 | 1488 | 1304 | 877 | 1429 | 607 | 1231 | 1529 | 773 | 1442 | 882 | **1** | **0** | **0** | **0** | **0** | **0** |
| Δ14_15 | c.4358_4675del318 | IF | mCΔ | 241 | 1663 | 10 | 67 | 52 | 44 | 0 | 55 | 38 | 0 | 144 | 1 | 1 | 0 | 0 | 0 | 0 |
| Δ14_17 | c.4358_5074del717 | IF | mCΔ | 132 | 919 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| **Δ15** | **c.4485_4675del191** | **FS** | **CΔ** | 285 | 766 | 642 | 911 | 1723 | 220 | 254 | 468 | 168 | 286 | 1231 | **1** | **1** | **0** | **0** | **P** | **0** |
| Δ15_16 | c.4485_4986del502 | FS | mCΔ | 0 | 1619 | 30 | 0 | 274 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ15_17 | c.4485_5074del590 | FS | mCΔ | 176 | 372 | 130 | 945 | 1214 | 8 | 162 | 55 | 103 | 0 | 718 | 1 | 1 | 1 | 0 | P | 1 |
| Δ15_19 | c.4485_5193del709 | FS | mCΔ | 88 | 744 | 0 | 0 | 228 | 4 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| Δ15_23 | c.4485_5467del983 | FS | mCΔ | 0 | 175 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16a | c.4986+1_4986+65ins65 | FS | SDS▼ | 570 | 44 | 5568 | 0 | 6687 | 944 | 2461 | 3541 | 740 | 4292 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Δ17** | **c.4987_5074del88** | **FS** | **CΔ** | 88 | 1466 | 2358 | 337 | 1951 | 1265 | 503 | 7854 | 1762 | 1988 | 6217 | **1** | **1** | **0** | **0** | **P** | **0** |
| Δ17_18 | c.4987_5152del166 | FS | mCΔ | 0 | 613 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ17_19 | c.4987_5193del207 | IF | mCΔ | 88 | 1006 | 0 | 0 | 75 | 0 | 532 | 262 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ17_20 | c.4987_5277del291 | IF | mCΔ | 0 | 0 | 0 | 0 | 0 | 107 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ18 | c.5075_5152del78 | IF | CΔ | 0 | 219 | 60 | 0 | 0 | 44 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| Δ18_19 | c.5075_5193del119 | FS | mCΔ | 0 | 0 | 10 | 67 | 0 | 4 | 81 | 28 | 11 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ18_20 | c.5075_5277del203 | FS | mCΔ | 0 | 0 | 30 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ▼ 143 bp int 19 | c.5194-1231_5194-1089ins143 | FS | C▼ | 132 | 0 | 20 | 0 | 0 | 381 | 324 | 138 | 573 | 0 | 431 | 0 | 0 | 0 | 0 | 0 | 0 |
| ▼ 146 bp int 19 | c.5194-1234_5194-1089ins146 | FS | C▼ | 373 | 197 | 80 | 101 | 653 | 40 | 0 | 289 | 32 | 19 | 421 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ20 | c.5194_5277del84 | IF | CΔ | 0 | 481 | 191 | 607 | 0 | 60 | 0 | 2618 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| **Δ21** | **c.5278_5332del55** | **FS** | **CΔ** | 1711 | 8226 | 492 | 776 | 1765 | 389 | 376 | 1226 | 195 | 22 | 1057 | **1** | **1** | **1** | **0** | **P** | **0** |
| Δ21_22 | c.5278_5406del129 | IF | mCΔ | 88 | 438 | 0 | 0 | 0 | 0 | 549 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | P | 0 |
| **▼ 129 bp int 21** | **c.5332+873_5332+1001ins129** | **FS** | **C▼** | 1711 | 263 | 592 | 641 | 1250 | 230 | 618 | 400 | 432 | 342 | 657 | **0** | **0** | **0** | **0** | **0** | **0** |
| ▼ 119 bp int 21 | c.5333-706_5333-588ins119 | FS | C▼ | 66 | 328 | 90 | 0 | 65 | 28 | 52 | 179 | 11 | 45 | 318 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Δ22** | **c.5333_5406del74** | **FS** | **CΔ** | 4739 | 16451 | 1585 | 4590 | 3064 | 4405 | 11919 | 7372 | 3464 | 3872 | 5160 | **1** | **P** | **1** | **0** | **P** | **0** |
| 23a | c.5407-9_5407-1ins9 | IF | SAS▼ | 44 | 66 | 10 | 0 | 39 | 35 | 168 | 55 | 68 | 74 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Δ23 | c.5407_5467del61 | FS | CΔ | 263 | 613 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| ***Sum of identified variants in each sample (mPCR pool)*** | | | | ***54*** | ***71*** | ***56*** | ***46*** | ***51*** | ***64*** | ***49*** | ***48*** | ***50*** | ***40*** | ***41*** | | | | | | |
| ***Number of identified variants in analyzed cell lines / tissue type*** | | | | ***76*** | | ***67*** | | | ***72*** | | | ***54*** | | | | | | | | |

potentially result in a translation of BRCA1 protein isoforms altering its biological functions, as demonstrated for Δ14_15 and Δ17_19 previously (Sevcik et al., 2012; Sevcik et al., 2013). Nine out of 11 "ubiquitous" in-frame variants (Δ5; 8p; Δ9_10; Δ11; 11q; 13p; ▼13A; ▼13A, 14p; 14p) represented known ASVs while 6q and the highly expressed 10a variant were surprisingly not scored previously. The remaining four of the 25 in-frame variants (Δ3_5; Δ9_11; Δ14_15; 23a) were detected in most of the analyzed samples. Interestingly, Δ17_19 was detected in the cell lines and mammary tissue samples of *BRCA1*-mutation carriers and BC patients, but in no control sample.

Our approach enabled a direct quantification of 29 SDS/SAS variants (including 17 mixed biotypes). Twelve SDS/SAS-only biotype variants were identified. Most of them (10 out of 12) were identified as "ubiquitous", and only the 16a and 23a splicing variants were not present in some analyzed samples. Out of 17 more complex splicing mRNA BRCA1 isoforms (mixed biotypes), only 1Aq, Δ2 and ▼13A, 14p were "ubiquitous", while the other variants occurred rather rarely (Supplementary Table S6). Besides the 11q splicing variant, which lacks 3,309 bp from exon 11 and its identification was done with a specific forward primer, the other 28 relatively short indels were co-amplified stoichiometrically alongside the corresponding canonical splicing variant (Fig. 1H) and therefore we were able to quantify their relative

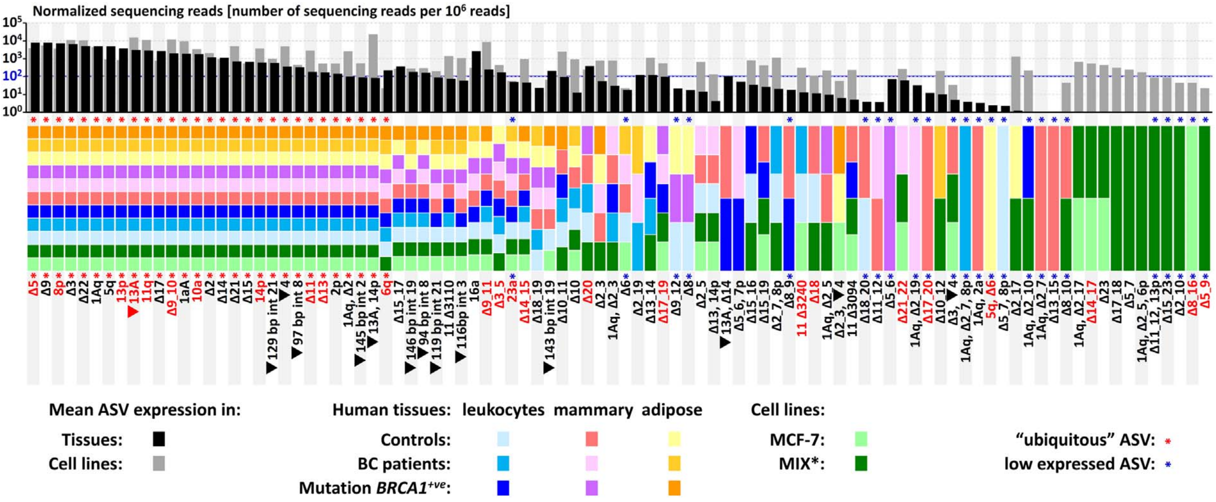


**Fig. 2.** Qualitative description of the presence of 94 identified BRCA1 ASVs (red letters indicate in-frame variants) in analyzed cDNA sample pools (colour bars). The grey-scale graph (upper part) shows the mean expression (in normalized reads per $10^6$ reads) in human tissues (black) and cell line samples (grey). Red asterisks indicate ubiquitous variants (Table 1); blue asterisks indicate variants with low expression ($< 10^2$ normalized reads averaged in both tissue and cell samples).

*Mix of cDNAs from EM-G3, HeLa, and MDA-MB-231 cell lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

expression (compared with other existing variants in the particular exon-exon junction including wt transcript) directly from sequencing reads.

To test biological reproducibility, we compared the presence of BRCA1 ASVs in two independently analyzed sets of control samples, each consisting of leukocytes, mammary tissue and adipose tissue cDNA sample pools from eight individuals (Fig. 3A). Altogether, 74/94 BRCA1 ASVs were identified in at least one tissue control sample, while 20/94 were not present in any of them. Thirty-five variants (Fig. 3B) were consistently present (or absent – Δ17_20 in leukocytes and adipose tissue, 11Δ3240 in mammary and adipose tissue, and Δ20 in adipose tissue) in the analyzed tissue type biological replicate. These 35 variants included the majority of "ubiquitous" ASVs as they were detected with high mean coverage per variant and sample (371 and 1496 in absolute and normalized sequencing reads, respectively). The remaining 39 variants (shown in Fig. 3C), discordantly expressed in paired biological replicates, represented low expressed events with low mean coverage per variant and sample (11 and 41 in absolute and normalized sequencing reads, respectively). The differences in sequencing coverage between 35 consistently present and 39 discordantly expressed variants are shown in Fig. 3D.

## 4. Discussion

An accurate description of 'naturally occurring' ASVs is a prerequisite to understanding their biological significance. RNA-sequencing (RNA-seq) of human RNA samples revealed that 90% of multi-exon genes undergo alternative splicing (Wang et al., 2015). While RNA-seq represents a superior tool for qualitative and quantitative transcriptome analyses, including ASV identification (Byron et al., 2016), it is unsuitable for small-scale projects targeting a few or a single gene. Moreover, RNA-seq analyses of low expressed transcripts require the sequencing of up to 100 million mapped reads and sophisticated bioinformatics instruments (Wang et al., 2015; Conesa et al., 2016).

A pioneering systematic description of BRCA1 ASVs was made by Orban and Olah (2003) who reviewed 23 BRCA1 ASVs known in 2003. Recently, Colombo et al. (2014) identified 63 BRCA1 ASVs by an analysis of 38 blood-derived samples and one healthy breast tissue sample. Subsequently, Romero et al. (2015) revealed 54 BRCA1 ASVs in an analysis of 70 breast tumor samples, four breast samples from healthy individuals and 72 blood-derived samples. Two later studies described the characterization of BRCA1 ASVs by capillary electrophoresis, which required further cloning or sequencing of fragments containing splice junction events in order to identify the presented peaks. However, only the in silico imputation has been used to explain the peak pattern observed in capillary electrophoresis for a subset of events (Colombo et al., 2014).

Overall, 42 out of 94 BRCA1 splicing events described by our approach had not been identified in previous studies (Supplementary Table S5) which we used to compare the obtained results (Colombo et al., 2014; Romero et al., 2015; Orban & Olah, 2003).

The most common biotypes identified in our study (59/94; 63% variants) were cassette and multicassette ASVs (Supplementary Table S5). We found 27 cassette ASVs that included all 17 variants described in Colombo's study and 10 novel variants (eight intron exonizations, Δ6 and Δ8). Of 32 multicassette biotype ASVs ascertained in our study, 16 were described previously. We did not detect four multicassette ASVs (Δ14_18; Δ14_19; Δ21_23; Δ22_23) reported by Colombo's study as minor variants.

The second most frequent biotype variants were SDS/SAS (12/94; 13% variants). Besides nine described by Colombo et al. (2014), we found another four variants containing the exonizations of adjacent intronic regions ("ubiquitous" in-frame 6q and 10a, rare in-frame 23a, and a frameshift 16a) in all analyzed patient tissue types.

We found three large intronizations affecting exon 11, including two described by Colombo et al. (2014) previously, and the sparsely expressed frameshift variant 11Δ3094. We did not target terminal modifications involving the alternative exon 1B and IRIS in our analysis.

Furthermore, we recorded 20 mixed biotype variants including two "ubiquitous" (1Aq, Δ2 and ▼13A, 14p). Nine mixed biotype variants were described previously and 11 rare were novel. We did not find six variants detected in Colombo's study, including three variants with untested alternative exon 1B, and another three (1Aq, 2p; 1Aq, Δ2_3, ▼4; and Δ10_13p) previously described as minor.

The most complex SAS/SDS events affected the non-coding 5′ untranslated region (at the exon 1A-2 junction). The 1Aa variant containing an insertion of 89 nucleotides prevailed in cell line samples. The dominant variant in all analyzed tissue samples was wt exon 1A accompanied by the shortened variant exon 1Aq (in approximately one-third of all mapped reads). The expression of three other SAS variants 8p; 14p; 13p (lacking the CAG nucleotides at the 5′-end of an exon) was ~10% of all sequencing reads in most of the analyzed samples. These variants rank among the NAGNAG tandem acceptors, a common kind of ASVs resulting in single amino acid exclusion (Sinha et al., 2009). The

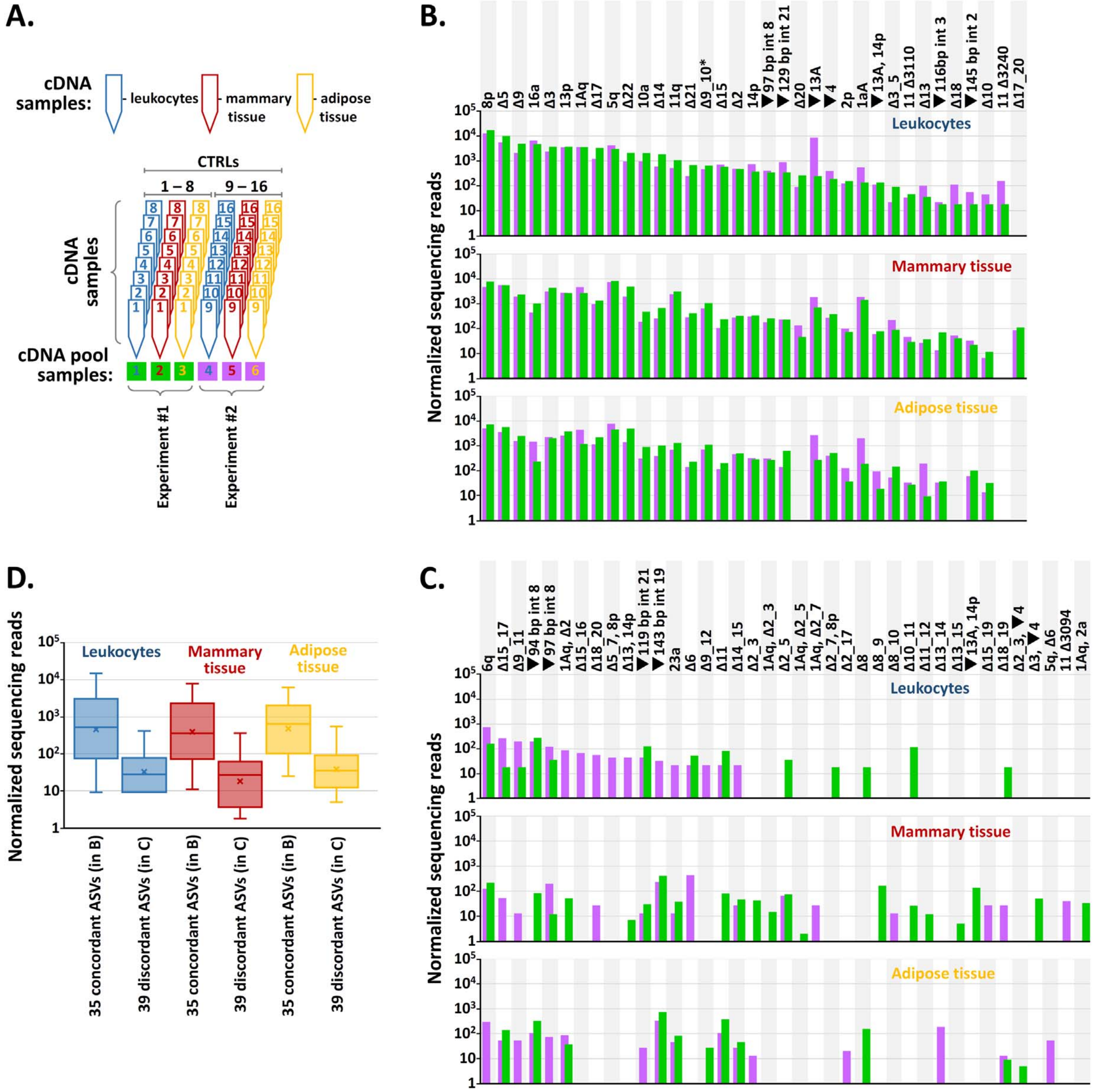


**Fig. 3.** The reproducibility test of the method in biological replicates involved an analysis of independent cDNAs from three types of tissue obtained from 16 control individuals. Panel A (adjusted from Fig. 1C) shows the arrangement of experiments #1 and #2, each consisting of NGS analyses from leukocytes, mammary tissue, and adipose tissue proceeded and sequenced independently by a pipeline described in the Method section. Graphs B and C compare the numbers of normalized sequencing reads (in log scale) for BRCA1 ASVs (listed in Supplementary Table 4) expressed in leukocytes, mammary and adipose tissues in two independent sets (Experiment #1 and #2) of control samples obtained from 16 individuals. The samples from control individuals 1–8 were analyzed in Experiment #1 (green bars), while the samples from control individuals 9–16 were analyzed in Experiment #2 (violet bars). Panel B shows the expression of 35 fully reproducible ASVs that gave concordant results in all analyzed biological duplicates. Panel C shows the expression of 39 non-fully reproducible ASVs that gave discordant results in at least one biological duplicate. The expressions of 35 fully reproducible ASVs were substantially higher than the expression of 39 non-fully reproducible ASVs as shown in panel D. The box plot charts show the values of sequencing normalized coverage (in log10 scale) for 35 fully reproducible ASVs (shown in B) and 39 non-fully reproducible ASVs (shown in C) in the analyzed tissues. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

other variants were minor or rare, with the exception of five ASVs (10a; 16a; 1Aq, Δ2; Δ13, 14p; and ▼13A, 14p) expressed with a higher proportion in cell line samples.

The presented approach showed satisfactory reproducibility documented at the level of mPCR amplification (Fig. 1E and G) and also in the analysis of biological replicates from two sets of tissue samples from control individuals (Fig. 3). We suppose that discrepancies in the detection of individual ASVs in biological replicates resulted rather from differentially expressed BRCA1 ASVs in the individual RNA samples (mixed in the mPCR pools) than from analytical variability, because reproducibility was strongly positively correlated with the level of variant expression (i.e. coverage; Fig. 3D). Our analysis, performed in native tissues and cells (with unmodified nonsense-mediated decay pathway), revealed 48 frame-shift variants. However, their ratio to in-frame variants was strongly reduced in the subset of "ubiquitous" variants (48/25 in the entire set of BRCA1 ASVs and 13/11 in "ubiquitous" variants). It has to be noted that BRCA1 mRNA is expressed at low levels, a few tenths of copies per cell in MCF7 cells (Lee et al., 2014). Many of newly identified ASVs were represented by a low number of reads (Fig. 2), indicating that they were probably expressed in a few copies per cell or present only in a subset of cDNA samples in the analyzed cDNA pool. We suppose that at least some of them may represent stochastic noise determining the number of alternative isoforms and their abundance (Melamud & Moult, 2009).

Deletions in isoform 1 (NM_007294.3) are the most frequently described BRCA1 ASVs. The longest well described ASV intron exonization is ▼4 (116 bp from intron 3). The predicted length of PCR amplicon covering ▼4 was 263 bp in our analysis, while the shortest predicted amplicons (covering variants Δ8_17; 11q, Δ12_17; Δ13_17; Δ17; Δ19_21) were 60 bp long. Therefore, the setup of mPCR and the size-selection protocol were targeted to enrich amplicons with mean fragment length of 100–150 bp in order to disclose majority of putative ASVs. We were able to identify splicing events covered by PCR fragments ranging between 60 and 287 bp. One of the shortest identified amplicons was "ubiquitous" variant Δ17, while the variant 11Δ3094 was characterized from PCR product of 287 bp (the above mentioned variant ▼4, identified from 263 bp amplicon, occurred as "ubiquitous"). These findings indicate the range of amplicon lengths (60–263 bp) which can be analyzed under defined conditions.

In conclusion, the mPCR/NGS approach enables direct identification of all biotype classes of splicing events, including mixed biotypes containing exonizations of flanking intronic sequences. The analysis of BRCA1 mRNA revealed the broadest spectrum of its splicing variants, including their distribution in the analyzed human tissue and cell line samples. Similar to most other methods (including recent RNA-seq analyses), the analysis is not able to identify possible combinations of splicing events affecting both 5′ and 3′ portions of the large BRCA1 transcripts. We are also aware that our approach could miss large deep intronic exonizations (substantially exceeding the targeted PCR amplification and/or range of size selection). We would like to emphasize that the described method can be easily adopted for an analysis of any gene of interest in order to identify its ASVs, not only in human samples. Additionally, we suppose that our approach may represent an interesting option for the functional classification of VUS introducing aberrant splicing with modified protocol using individual (instead of pooled) cDNA template for mPCR step (limited to the region of interest).

## Funding

This work was supported by the Grant Agency of the Czech Republic [grant number P301-12-1850], Charles University projects [PROGRES Q28/LF1, and SVV2017/260367], and a Ministry of Education, Youth and Sports project [The National Center for Medical Genomics LM2015091].

## Disclosure statement

The authors have no conflicts of interest to disclose.

## Acknowledgements

We thank The National Center for Medical Genomics for their instrumental and technical support with NGS sequencing and data analyses.

## Authors' contributions

Study design (ZK, PK), experimental procedures (JH, FL, HH, KH, MJ, JSo), bioinformatics (PZ, VS), sample collection and characterization (OM, DP, MV, JSe), data analysis (JH, PK), manuscript preparation (PK, ZK, JH, JSe), manuscript approval (all authors).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2017.09.025.

## References

Baralle, D., Buratti, E., 2017. RNA splicing in human disease and in the clinic. Clin. Sci. (Lond.) 131, 355–368.

Bentley, D.L., 2014. Coupling mRNA processing with transcription in time and space. Nat. Rev. Genet. 15, 163–175.

Brozova, M., Kleibl, Z., Netikova, I., Sevcik, J., Scholzova, E., Brezinova, J., Chaloupkova, A., Vesely, P., Dundr, P., Zadinova, M., et al., 2007. Establishment, growth and in vivo differentiation of a new clonal human cell line, EM-G3, derived from breast cancer progenitors. Breast Cancer Res. Treat. 103, 247–257.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., et al., 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin. Chem. 55, 611–622.

Byron, S.A., Van Keuren-Jensen, K.R., Engelthaler, D.M., Carpten, J.D., Craig, D.W., 2016. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat. Rev. Genet. 17, 257–271.

Caminsky, N., Mucaki, E., Rogan, P., 2015. Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. F1000Res 18, 282.

Cheon, J.Y., Mozersky, J., Cook-Deegan, R., 2014. Variants of uncertain significance in BRCA: a harbinger of ethical and policy issues to come? Genome Med. 6, 121.

Colombo, M., Blok, M.J., Whiley, P., Santamarina, M., Gutierrez-Enriquez, S., Romero, A., Garre, P., Becker, A., Smith, L.D., De Vecchi, G., 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.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., et al., 2016. A survey of best practices for RNA-seq data analysis. Genome Biol. 17, 13.

Kleibl, Z., Kristensen, V.N., 2016. Women at high risk of breast cancer: molecular characteristics, clinical presentation and management. Breast 28, 136–144.

Kleiblova, P., Dostalova, I., Bartlova, M., Lacinova, Z., Ticha, I., Krejci, V., Springer, D., Kleibl, Z., Haluzik, M., 2010. Expression of adipokines and estrogen receptors in adipose tissue and placenta of patients with gestational diabetes mellitus. Mol. Cell. Endocrinol. 314, 150–156.

Lee, K., Cui, Y., Lee, L.P., Irudayaraj, J., 2014. Quantitative imaging of single mRNA splice variants in living cells. Nat. Nanotechnol. 9, 474–480.

Melamud, E., Moult, J., 2009. Stochastic noise in splicing machinery. Nucleic Acids Res. 37, 4873–4886.

Orban, T.I., Olah, E., 2003. Emerging roles of BRCA1 alternative splicing. Mol. Pathol. 56, 191–197.

Rahman, N., 2014. Realizing the promise of cancer predisposition genes. Nature 505, 302–308.

Romero, A., Garcia-Garcia, F., Lopez-Perolio, I., Ruiz de Garibay, G., Garcia-Saenz, J.A., Garre, P., Ayllon, P., Benito, E., Dopazo, J., Diaz-Rubio, E., et al., 2015. BRCA1 alternative splicing landscape in breast tissue samples. BMC Cancer 15, 219.

Sevcik, J., Falk, M., Kleiblova, P., Lhota, F., Stefancikova, L., Janatova, M., Weiterova, L., Lukasova, E., Kozubek, S., Pohlreich, P., et al., 2012. The BRCA1 alternative splicing variant Delta14-15 with an in-frame deletion of part of the regulatory serine-containing domain (SCD) impairs the DNA repair capacity in MCF-7 cells. Cell. Signal. 24, 1023–1030.

Sevcik, J., Falk, M., Macurek, L., Kleiblova, P., Lhota, F., Hojny, J., Stefancikova, L., Janatova, M., Bartek, J., Stribrna, J., et al., 2013. Expression of human BRCA1Delta17-19 alternative splicing variant with a truncated BRCT domain in MCF-7 cells results in impaired assembly of DNA repair complexes and aberrant DNA damage response. Cell. Signal. 25, 1186–1193.

Sinha, R., Nikolajewa, S., Szafranski, K., Hiller, M., Jahn, N., Huse, K., Platzer, M., Backofen, R., 2009. Accurate prediction of NAGNAG alternative splicing. Nucleic Acids Res. 37, 3569–3579.

Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., et al., 2016. ENCODE data at the ENCODE portal. Nucleic Acids Res. 44, D726–732.

Soukarieh, O., Gaildrat, P., Hamieh, M., Drouet, A., Baert-Desurmont, S., Frebourg, T., Tosi, M., Martins, A., 2016. Exonic splicing mutations are more prevalent than currently estimated and can be predicted by using in silico tools. PLoS Genet. 12, e1005756.

Tavtigian, S.V., Chenevix-Trench, G., 2014. Growing recognition of the role for rare missense substitutions in breast cancer susceptibility. Biomark. Med 8, 589–603.

Vondruskova, E., Malik, R., Sevcik, J., Kleiblova, P., Kleibl, Z., 2008. Long-term BRCA1 down-regulation by small hairpin RNAs targeting the 3′ untranslated region. Neoplasma 55, 130–137.

Wang, J., Ye, Z., Huang Tim, H.M., Shi, H., Jin, V., 2015. A survey of computational methods in transcriptome-wide alternative splicing analysis. Biomol. Concepts 6, 59–66.

Whiley, P.J., de la Hoya, M., Thomassen, M., Becker, A., Brandao, R., Pedersen, I.S., Montagna, M., Menendez, M., Quiles, F., Gutierrez-Enriquez, 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.