Breast Cancer Research

Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer susceptibility: results from a breast cancer family registry (CFR) case-control mutation screening study

Breast Cancer Research 2011, **13**:R6 doi:10.1186/bcr2810

Florence Le Calvez-Kelm (lecalvez@iarc.fr)
Fabienne Lesueur (lesueurf@iarc.fr)
Francesca Damiola (damiolaf@fellows.iarc.fr)
Maxime Vallee (valleem@students.iarc.fr)
Catherine Voegele (voegele@iarc.fr)
Davit Babikyan (babikyand@genetics.sci.am)
Geoffroy Durand (durandg@iarc.fr)
Nathalie Forey (forey@iarc.fr)
Sandrine McKay-Chopin (chopin@iarc.fr)
Nivonirina Robinot (robinotn@iarc.fr)
Tu Nguyen-Dumont (nguyent@students.iarc.fr)
Alun Thomas (alun.thomas@utah.edu)
Graham B Byrnes (byrnesg@iarc.fr)
The Breast Cancer Family Registry (j.hopper@unimelb.edu.au)
John L Hopper (j.hopper@unimelb.edu.au)
Melissa C Southey (msouthey@unimelb.edu.au)
Irene L Andrulis (andrulis@lunenfeld.ca)
Esther M John (esther.john@cpic.org)
Sean V Tavtigian (sean.tavtigian@hci.utah.edu)

ISSN	1465-5411
Article type	Research article
Submission date	2 September 2010
Acceptance date	18 January 2011
Publication date	18 January 2011
Article URL	http://breast-cancer-research.com/content/13/1/R6

Breast Cancer Research

Breast Cancer RESEARCH

Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer susceptibility: results from a breast cancer family registry (CFR) case-control mutation screening study

Florence Le Calvez-Kelm[1]*, Fabienne Lesueur[1]*, Francesca Damiola[1], Maxime Vallée[1], Catherine Voegele[1], Davit Babikyan[2], Geoffroy Durand[1], Nathalie Forey[1], Sandrine McKay-Chopin[1], Nivonirina Robinot[1], Tù Nguyen-Dumont[1], Alun Thomas[3], Graham B Byrnes[1], Breast Cancer Family Registry[4,5,6], John L Hopper[4], Melissa C Southey[7], Irene L.Andrulis[5], Esther M John[6,8], Sean V Tavtigian[9]†

*Contributed equally

† Corresponding author. Email: sean.tavtigian@hci.utah.edu.

1 International Agency for Research on Cancer, 150 Cours Albert Thomas, Lyon CEDEX 08, 69372, France.

2 Laboratory of Cancer Genetics, Center of Medical Genetics and Primary Health Care, 4 Tigran Mets Avenue, Yerevan, 375010, Armenia.

3 Department of Internal Medicine, University of Utah School of Medicine, 391 Chipeta Way, Suite D, Salt Lake City, UT 84108, USA.

4 Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, The University of Melbourne, 723 Swanston Street, Melbourne, Victoria 3010, Australia.

5 Cancer Care Ontario, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Department of Molecular Genetics, University of Toronto, 60 Murray Street, Toronto, ON M5T 3L9, Canada.

6 Cancer Prevention Institute of California, 2201 Walnut Avenue, Suite 300, Fremont, CA 94538, USA.

7 Department of Pathology, The University of Melbourne, Medical Building 181, Melbourne, Victoria 3010, Australia.

8 Stanford University School of Medicine, 300 Pasteur Drive, Stanford, CA 94305, USA.

9 Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah School

of Medicine, 2000 Circle of Hope, Salt Lake City, UT 84112, USA.

ABSTRACT

Introduction: Both protein truncating variants and some missense substitutions in *CHEK2* confer increased risk of breast cancer. However, no large-scale study has used full open reading frame mutation screening to assess the contribution of rare missense substitutions in *CHEK2* to breast cancer risk. This absence has been due in part to a lack of validated statistical methods for summarizing risk attributable to large numbers of individually rare missense substitutions.

Methods: Previously, we adapted an *in silico* assessment of missense substitutions used for analysis of unclassified missense substitutions in *BRCA1* and *BRCA2* to the problem of assessing candidate genes using rare missense substitution data observed in case-control mutation screening studies. The method involves stratifying rare missense substitutions observed in cases and/ or controls into a series of grades *a priori* ordered from least to most likely to be evolutionarily deleterious, followed by a logistic regression test for trend to compare the frequency distributions of the graded missense substitutions in cases versus controls. Here we used this approach to analyze *CHEK2* mutation screening data from a population based series of 1303 female breast cancer cases and 1109 unaffected female controls.

Results: We found evidence of risk associated with rare, evolutionarily unlikely *CHEK2* missense substitutions. Additional findings were: (1) the risk estimate for the most severe grade of *CHEK2* missense substitutions (denoted C65) is approximately equivalent to that of *CHEK2* protein truncating variants; (2) the population attributable fraction and familial relative risk explained by the pool of rare missense substitutions were similar to those explained by the pool of protein truncating variants; and (3) *post-hoc* power calculations implied that scaling case-control mutation screening up to examine entire biochemical pathways would require roughly 2,000 cases and controls to achieve acceptable statistical power.

Conclusions: This study shows that *CHEK2* harbors many rare sequence variants that confer increased risk of breast cancer, and a substantial proportion of these are missense

substitutions. The study validates our analytic approach to rare missense substitutions and provides a method to combine data from protein truncating variants and rare missense substitutions into a one degree of freedom per gene test.

INTRODUCTION

Familial clustering of breast cancer is well recognized, having been described over 140 years ago [1]; the familial relative risk of breast cancer is on average about 2-fold, and higher among relatives of early onset cases [2, 3]. Three classes of breast cancer susceptibility sequence variants with different levels of risk and prevalence in the population are now well established [4, 5]: rare high-risk variants such as protein truncating mutations in *BRCA1*, *BRCA2*, *PTEN*, and *TP53* (MIM #s 113705, 600185, 601728, and 191170, respectively); rare intermediate-risk variants such as protein truncating mutations in *ATM* [6, 7], *BRIP1* [8], *CHEK2* [9], and *PALB2* [10, 11] (MIM #s 208900, 605882, 604373, and 610355 respectively); and, more recently, common modest penetrance variants such as the risk-SNPs detected by genome-wide association study (GWAS) in *FGFR2*, *TOX3* (*TNRC9*), *MAP3K1*, and *LSP1* [12-14] (MIM #s 176943, 611416, 600982, and 153432, respectively). High-risk variants in the known major breast cancer susceptibility genes *BRCA1*, *BRCA2*, *TP53*, and *PTEN* account for approximately 20-25% of the familial risk of breast cancer, and adding in the known intermediate-risk genes increases this by perhaps one percent for each gene [15]. Moreover, the panoply of known modest risk SNPs account for about 8% of the familial relative risk [16]. Thus known genetic effects account for about 1/3 of the familial relative risk of breast cancer, leaving 2/3 unaccounted for – a phenomenon referred to as the "problem of missing heritability". Some of this so-called missing “heritability” is of course due to the familial component of environmental risk factors; the measured surrogates for these probably explain about 5% of the familial relative risk, but if measured more specifically and more precisely they may explain considerably more familial aggregation [17].

The gene *CHEK2* encodes a serine/ threonine kinase, CHK2, that functions in the signaling pathways activated by DNA damage, particularly DNA double strand breaks [18]. Inheritance of a *CHEK2* protein truncating mutation such as the relatively well investigated Northern European founder mutation c.1100delC confers a 2-3 fold increased risk of breast cancer, an increased risk of a number of other cancer types, and perhaps a decreased risk of some smoking-related cancers [9, 19-21]. Some missense substitutions in *CHEK2* also alter cancer risk, as exemplified by the Ashkenazi *CHEK2* missense substitution p.S428F and the Slavic substitution p.I157T [22-26]. Most large-scale genetic studies of *CHEK2* conducted to date have focused on genotyping known variants, such as founder mutations. Consequently, there has been little opportunity to assess the role of the potentially more numerous rarer variants in this gene.

During the 1990s, linkage analysis proved to be an effective genome-wide approach for finding high-risk susceptibility genes for breast and colon cancer. Over the last few years, GWAS has proved to be an effective genome-wide approach to finding common, not necessarily causal, SNPs associated with modest-risk. Case-control mutation screening – or its quantitative trait homolog of comparative mutation screening of subjects from the opposite ends of a trait spectrum – is emerging as a useful strategy for identifying and characterizing intermediate-risk susceptibility genes [6-8, 10, 27-29]. While case-control mutation screening has been, to date, too technically demanding to examine a whole biochemical pathway, let alone the entire exome, one can imagine combining exon hybridization-capture and massively parallel sequencing to accomplish such a study design. Beyond the laboratory challenge imposed by the implied scale of resequencing, a second challenge would be to conduct a statistically powerful analysis of the large number of rare sequence variants that would be revealed if such a study design were applied to a common disease such as breast or colon cancer. Previously, we used data from mutation screening of *ATM* in breast cancer cases and controls to demonstrate the ability to detect evidence of pathogenicity from both truncating and splice junction variants (T+SJV) and rare missense substitutions (rMS) [7]. Here we apply the same analytic strategy to *CHEK2* and then extrapolate the results to determine requirements for much larger-scale studies.

MATERIALS AND METHODS

Ethics Statement. *CHEK2* mutation screening studies and analyses described here were approved by the Institutional Review Board (IRB) of the International Agency for Research on Cancer (IARC), the University of Utah IRB, and the local IRBs of the Breast Cancer Family Registry (Breast CFR) centers from which we received samples. All participants gave written informed consent.

Subjects were selected from women ascertained by population-based sampling by the Breast CFR at three centers (Cancer Care Ontario, the Cancer Prevention Institute of California (formerly the Northern California Cancer Center), and the University of Melbourne) [30]. Subjects were recruited between 1995 and 2005.

Selection criteria for cases (N=1313) were diagnosis at or before age 45 years and self-reported race/ethnicity plus grandparents' country of origin information consistent with Caucasian, East Asian, Hispanic/Latino, or African American racial or ethnic heritage.

The controls (N=1123) were frequency matched to cases within each center on racial/ethnic group, with age at ascertainment not more than ten years beyond the age at diagnosis range of the cases from the same center. Due to the shortage of available controls in some ethnic groups and age groups, the frequency matching was not one-to-one in all subgroups.

Mutation screening started from whole-genome amplified (WGA) DNA for coding exons 1-9, and from genomic DNA for exons 10-14. A nested PCR strategy was used, followed by High Resolution Melting curve analysis (HRM) [31, 32], and then dye-terminator resequencing of samples that contained a melt curve aberration indicative of the presence of a sequence variant. For *CHEK2* amplicons harboring a SNP(s) with frequency ≥1% in either dbSNP or initial amplicon testing, we applied a simultaneous mutation scanning and genotyping approach using HRM analysis to improve the sensitivity and the efficiency of the mutation screening [33]. The laboratory process was as described in detail for our recent case-control mutation screening of

ATM [7], except that primary PCR of CHEK2 exons 10-14 (which are involved in a sub-telomeric repeat) relied on a long-range PCR as described by Sodha *et al*. [34].

All exonic sequence variants, plus splice junction consensus sequence variants that reduced splice junction sequence similarity to the standard consensus sequences AG^GTRRGT (donor) or Y_{16}NYAG^ (acceptor) (where ^ indicates the position of the splice junction), were re-amplified from genomic DNA for confirmation of the presence of the variant. Because of the presence of pseudogenes that partially match the sequence of the *CHEK2* long-range PCR exons (exons 10-14), sequence variants identified within these exons were subsequently tested using allele specific PCR assays for the primary PCR reaction to confirm that the sequence variants initially identified were true *CHEK2* variants. To insure amplification of *CHEK2* DNA sequence and not amplification of the potentially interfering *CHEK2* pseudogenes, positions of the specific primers were chosen so that the 3' extremity bases perfectly matched to the *CHEK2* wild-type sequence, while they mismatched the corresponding position of the pseudogenes.

All samples that failed either at the primary PCR, secondary PCR, or sequencing reaction stage were re-amplified from WGA DNAs or genomic DNAs. Samples that still did not provide satisfactory mutation screening results for at least 80% of the *CHEK2* coding sequence were excluded from further analyses (n=24). Process and data management of the mutation screening were as described by Voegele *et al*. [35]. Primer and probe sequences are available from FLCK upon request.

Alignments and scoring of missense substitutions. Previously, we used the TCoffee suite of alignment tools to prepare a CHK2 protein multiple sequence alignment in which the most diverged sequence was from sea urchin (*Strongylocentrotus purpuratus*) to analyze a small number of *CHEK2* missense substitutions and in-frame deletions [36, 37]. We updated this alignment by replacing the partial pufferfish (*Tetraodon nigroviridis*) sequence with a full-length zebrafish (*Danio rerio*) sequence and including predicted CHK2 sequences from elephant (*Loxodonta africana*), platypus (*Ornithorhynchus anatinus*), tunicate (*Ciona intestinalis*), and fruit fly (*Drosophila melanogaster*). The alignment was characterized by determining percent sequence identity between each pair of sequences in the alignment, using the Protpars routine of Phylip to make a maximum parsimony estimate of the number of substitutions that occurred along each clade of the underlying phylogeny, and by recording SIFTs "Median sequence conservation score" [38, 39]. The alignment, or updated versions thereof, is available at the Align-GVGD website [40]

Missense substitutions observed during our mutation screening of *CHEK2* were scored using Align-GVGD and SIFT with our curated alignments, and with PolyPhen-2 using its pre-compiled alignment [39-45].

Statistical analysis and power calculations. To assess risk associations using the case-control frequency distribution of T+SJVs and rMSs, we constructed a single table with one entry per subject, zero or one rare sequence variant per subject, annotations for type of sequence variants, study center, case-control status, race/ethnicity, and age. For the two subjects who carried more than one rare variant of interest (one case carried p.I448S (C15) plus p.E394D (C35), and one case carried p.E239K (C15) plus p.R346H (C25)), only the variant belonging to the more likely evolutionarily deleterious grade (i.e., higher C-number as scored by Align-GVGD) was considered.

Most analyses were performed by multivariable unconditional logistic regression using STATA version 11 (StataCorp). Differences in the case-control ratio between ethnic groups and age categories were accounted for by including categorical variables for each age category and ethnic group. Adjustment was also made for study center. We explored the possibility of interactions between ethnic group and study center, checking both improvement of model fit by the likelihood ratio statistic and comparing the estimates of the parameter of interest (log(OR) per Align-GVGD grade) in different models. Adjustment for ethnic group should also capture confounding of genetic and social factors, with interaction terms allowing that this confounding may be different for the broadly labeled ethnic groups in different centers. Because the Breast CFR matched cases and controls for age in 5 year categories, and because the maximum age of Breast CFR cases included in this study was 45, all subjects aged 41 or older (at diagnosis for cases, at ascertainment for controls) were combined into a single age category.

Logistic regression trend tests were formatted such that subjects who did not carry any T+SJV or any rMS, and carriers of the seven grades of rMSs (C0, C15, C25, C35, C45, C55 and C65) defined by Align-GVGD [42] were assigned the default row labels 0,1,2,3,4,5,6, and 7, respectively. These row labels were then used as a continuous variable in the logistic regressions. Regression coefficients and trend test P-values ("P_{trend}") were estimated from the resulting ln-ORs using the logit function of STATA. Carriers of T+SJVs were analyzed against the same non-carrier group defined above. Two strategies were used to combine evidence of association with T+SJV and rMS variants: (i) carriers of T+SJVs were combined with carriers of C65 rMSs in category 7, and (ii) T+SJV carriers were assigned row label 8. We used the Fisher's exact test (FET) to obtain the lower bound of the 95% confidence interval for associations with categories that contained one or more cases but zero controls.

Post-hoc power calculations were performed by specifying a hypothetical OR and population prevalence for each class of variant, together with the cumulative probability of breast cancer prior to age 70 years. The ORs and control carrier frequencies that we specified for the individual grades of sequence variants, relative to the non-carriers, were based on data from the population-based Breast CFR sample series. For the grades where there were a reasonable number of observations, i.e., C0, C15, C25, C65, and T+SJV, we used the adjusted ORs and observed carrier frequencies. Due to the very low numbers of observations in C35-C55, ORs for these categories were estimated from the logistic regression OR coefficient and population carrier frequencies defined so as to obtain the specified OR given the number of observations in cases. From these OR and frequency estimates, we calculated expected values and variances of the test statistics for the types of test considered: Pearson's chi-squared for the two-category tests, and the Wald statistic from a logistic regression for the trend test. We then calculated the probability of these statistics exceeding a series of desired P-value thresholds, using a normal approximation.

Attributable fractions were estimated according to Rothman and Greenland, and familial relative risks were estimated according to Goldgar [46, 47]. Both calculations used the same frequency and risk association estimates as were used for the post-hoc power calculations.

RESULTS

Number of subjects included in the analysis

Of the 2,436 Breast CFR subjects, 24 (10 cases and 14 controls) were excluded because their PCR failure rate for *CHEK2* mutation screening amplicons was greater than 20% (Table 1). The distributions of the remaining cases and controls by age, race/ethnicity, and study center are detailed in Table 2.

Analysis of protein truncating variants

Full open reading frame mutation screening of *CHEK2* revealed three distinct nonsense substitutions and four distinct small insertion deletion variants that should result in a truncated protein. One of these, c.1100delC, a well-known Northern European founder mutation that has been shown beyond any reasonable doubt to confer a moderately increased risk of breast cancer [48], was observed in 11 cases compared with three controls. The other six protein truncating variants were observed once each, always in a case (Supplementary table S1 in Additional file 1). The overall OR associated with T+SJVs was 6.18 (P= 0.005) (Table 3).

However, as 1100delC genotyping has already been reported for most of the Breast CFR subjects included in this study [48, 49], we note that the combination of the other six protein truncating variants was marginally significant by itself (P= 0.033) but as none of this set of controls were found to carry such a variant, we could not estimate the OR.

Analysis of rare missense substitutions

In the course of this mutation screening, we observed 34 distinct *CHEK2* missense substitutions (Supplementary table S1 in Additional file 1). The majority of these (24 of 34) were observed once each. The most common one, p.I448S, was observed 10 times, and none had an overall frequency of greater than 1% in this sample series. Overall, 42 of the cases carried one rMS, two of the cases carried two rMSs, and 17 controls carried one rMS. Thus, there was a significant excess of rMS carriers among the cases (OR= 2.20, P=0.010).

To make a more detailed analysis of the rMSs, we prepared and characterized a protein multiple sequence alignment containing CHK2 sequences from seven mammals, three additional vertebrates, two additional deuterostomates, and one protostomate. Ordering the non-mammalian sequences by decreasing identity to human CHK2 and sequentially assessing overall sequence diversity, the alignment exceeded a maximum parsimony estimate of an average of three substitutions per position upon inclusion of the sea urchin (*Strongylocentrotus purpuratus*) sequence (Supplementary table S2 in Additional file 1). Three substitutions per position was suggested as a criterion of sequence diversity for analysis of missense substitutions, and we have adopted it as our criterion for use with Align-GVGD in case-control mutation screening applications [7, 50, 51].

Using this alignment, we scored the 34 missense substitutions with Align-GVGD [40,41] (Supplementary table S1 in Additional file 1). Rather than generating a binary classification, Align-GVGD categorizes missense substitutions into seven grades ordered from evolutionarily most likely (C0) to least likely (C65) [42]. Align-GVGD scored 14 of the rMSs as C0; with 12 cases versus 9 controls carrying a C0 rMS (as their highest-grade *CHEK2* variant), the OR for this grade of rMS was near 1.0 (1.39, 95%CI 0.55- 3.56) (Table 3). In contrast, five different rMSs scored as C65; with nine cases versus one control carrying a C65 rMS (again, as their highest-grade *CHEK2* variant), the OR for C65 rMSs was 8.75, P= 0.044 (Table 3). Exploiting the intrinsic ordering of the Align-GVGD grades, we performed a logistic regression test for log-linear OR trend across non-carriers and carriers of the seven grades of rMSs; this yielded a ln(OR) increase of 0.33/grade (P_{trend}= 0.0055) (Table 4). Thus the statistical evidence in favor of pathogenicity from the trend test was stronger than that generated by either the binary test over all the missense substitutions or testing any individual grade of missense substitution. These results include adjustment for age category, study center and ethnic group. Neither removal of study center, nor inclusion of interactions between center and ethnic group, changed the first two digits of these estimates. The interaction terms did not significantly improve model fit (p=0.18) and were omitted. While removing study center did not significantly reduce the goodness of fit (p=0.12), this adjustment was retained on the grounds of prior plausibility.

While we emphasize that our pre-planned rMS analysis was based on rMS grading using Align-GVGD with a *CHEK2* protein multiple sequence alignment having an average of at least three substitutions per position and in which the furthest diverged sequence was from the (deuterostomate) sea urchin (*Strongylocentrotus purpuratus*), thus conforming to the conditions under which Align-GVGD was calibrated and used to grade missense substitutions in *ATM* [7, 42], we also carried out corresponding analyses based on rMS grading with SIFT and PolyPhen-2 [39, 44]. With SIFT, we set up three rMS grades: the program's standard likely neutral grade of SIFT score > 0.05, a likely deleterious grade of 0.05 ≥ SIFT score ≥ 0.01, and a more likely deleterious grade of SIFT score = 0.00. Using a *CHEK2* alignment in which the furthest diverged sequence was from the (protostomate) fruitfly (*Drosophila melanogaster*), which reached SIFT's "Median sequence conservation score" threshold of 3.25, the OR for the

SIFT score = 0.00 grade was 3.03, and the logistic regression trend test gave P_{trend}= 0.012 (Table 3). Using the slightly less informative alignment in which the most diverged sequence was from sea urchin, the logistic regression trend test gave P_{trend}= 0.014 (data not shown). PolyPhen-2 uses a combination of its own pre-compiled protein multiple sequence alignments and crystal structure information to score missense substitutions. Using PolyPhen-2, we also set up three rMS grades: the program's standard "Benign" grade, its standard "Possibly Damaging" grade, and its standard "Probably Damaging" grade. The OR for the "Probably Damaging" grade was 3.87, and the logistic regression trend test gave P_{trend}= 0.0070. The rMS gradings obtained with SIFT and PolyPhen-2 are also included in Supplementary table S1 in Additional file 1.

One question that arises from this approach to missense substitution analysis is whether the rMSs that drive the difference between cases and controls are truly evolutionarily unlikely, which is shorthand for "subject to purifying selection such that they are disproportionately unlikely ever to become fixed as major alleles". To address this question, we waited until after our primary protein multiple sequence alignment had been created and the rare human missense substitutions scored, and then we assembled an additional Mammalian *CHEK2* gene model (from Guinea pig, *Cavia porcellus*). Insertion of the *C. porcellus* CHK2 sequence into our alignment and comparison to the other placental mammal CHK2 sequences revealed 34 *C. porcellus*–specific amino acid substitutions (i.e., apparently wild-type *C. porcellus* CHK2 amino acid residues that differ from the residue(s) present at that position in the other placental mammal CHK2 sequences). We then scored these residues with Align-GVGD as if they were amino acid substitutions to the human *CHEK2* sequence. All 34 scored C0, the most evolutionarily likely grade and the grade that contributes least to the difference that we observe between breast cancer cases and controls. Simulating and scoring all possible single nucleotide substitutions to the canonical human *CHEK2* cDNA sequence, we find that 57.2% of possible missense substitutions are C0. Taking into account differing probabilities of these substitutions due to their underlying sequence contexts as estimated by dinucleotide substitution rate constants [52], 58.6% of a random draw of missense substitutions would be C0. Therefore, ignoring the effects of purifying selection, the probability that 34 of 34 C. porcellus-specific substitutions would be C0 is ~ $(0.586^{34})= 1.3 \times 10^{-8}$. Thus selection acts against the rMSs of grade >C0. As these grades have sequentially increasing leverage (towards C65) on the test for trend, evolutionarily unlikely rMSs indeed drive the observed difference between cases and controls.

Combined evidence

Looking forward to candidate gene studies, it could be useful to combine evidence from both T+SJVs and rMSs. The log-linear OR trend test provides a simple mechanism to do this: observations of T+SJVs can either be combined with observations of the highest grade of missense substitutions (C65s), or we can add an 8th (even higher) carrier grade for the T+SJVs. For this dataset, combining T+SJVs and C65 rMSs together in grade 7 appeared slightly more effective: ln(OR) increased by 0.29/grade ($P_{trend}= 8.8 \times 10^{-5}$) as opposed to 0.26/grade ($P_{trend}= 1.1 \times 10^{-4}$) with the alternate approach. The important point is that the data were less compatible with chance when combined than when considered as either T+SJVs or rMSs alone.

Extrapolation to pathway and whole-exome case-control mutation screening projects

Massively parallel sequencing has evolved to the point where it is being used to identify susceptibility genes for rare diseases, and one can imagine study designs where it could be used to identify or characterize intermediate-risk susceptibility genes for common diseases. Using rare variant carrier frequencies of 0.0045, 0.0018, 0.00021*, 0.00011*, 0.00090, and 0.0027 for the rMS grades C15, C25, C35*, C55*, C65, and T+SJV, respectively, and ORs of 1.82, 2.47, 3.74*, 7.24*, 8.75, and 6.18 for the same series of grades, we estimated the number of subjects required for a reasonably powered many-gene case-control mutation screening study (note that these frequency and OR values were taken or calculated directly from Table 3 and Table 4 unless marked with an asterisk; marked values were estimated from the ln(OR) regression coefficient given in Table 4 and the number of observations in cases). Setting a Bonferroni-adjusted P-value threshold of 0.0005 for a study of the ~100 genes in the DNA double strand break repair and allied cell cycle checkpoint pathways, we estimate that ~2,000 cases and a similar number of controls would be required to have 80% power in a combined analysis of T+SJVs and rMSs (Table 5). An analysis based on T+SJVs alone would require 3,400 each of cases and controls, and an analysis based on rMSs alone would require 4,700 each of cases and controls. Setting a P-value threshold of 2.5×10^{-6}, which might be considered appropriate for a whole exome study, 3,350 each of cases and controls would be required to have 80% power.

DISCUSSION

That protein-truncating variants in CHEK2 confer moderately increased risk of breast cancer is well established. The OR that we observed for T+SJVs is numerically somewhat higher than that reported in the 2004 international CHEK2 consortium study of c.1100delC [48], but not significantly so as our 95% confidence intervals do include the point estimate from that study. Moreover, as previous studies have observed higher ORs for c.1100delC in familial versus sporadic cases and in early onset versus later onset cases [9, 48], we should expect that this study's focus on early onset breast cancer cases with oversampling of familial cases would result in relatively high OR estimates.

Previous studies have shown that some CHEK2 missense substitutions are pathogenic, but the scale of their contribution to breast cancer susceptibility, relative to that of T+SJVs, was not known. Although we would hesitate to extrapolate our current data to true population attributable risks (within the age groups that we have sampled) or familial relative risks, the data do provide a basis to compare the relative contributions of these two classes of variants. Working from the control carrier frequencies and the OR point estimates (adjusted for race/ethnicity, study center, and age) observed from the population-based Breast CFR sample series, we calculate attributable fractions of 0.014 for T+SJVs as compared to 0.015 for the sum of C15-C65 rMSs. In addition, we calculate a familial relative risk to first-degree relatives of 1.036 for T+SJVs as compared to 1.033 for a product across the C15-C65 rMSs. Thus, to a first approximation, the attributable fractions and familial relative risks of truncating variants and rare missense substitutions are virtually identical. It is important to remember that these attributable fraction and familial relative risk point estimates are inflated compared to those that would be obtained from a population-based study that includes cases diagnosed in their 70s or older. In addition, as more than 25% of the T+SJVs observed in this study were nonsense and frameshift mutations other than c.1100delC, these data also speak to the importance of full open reading frame mutation screening to observe the majority of genetically relevant sequence variants in this cancer susceptibility gene.

Several of the missense substitutions observed in this study have been subjected to functional assays in one or more published works. For the 14 that Align-GVGD scored C0 and which we would consequently predict to be neutral or nearly so, assay results have been reported for four (p.P85L, p.R137Q, p.R180H, and p.T323P). Using a *Saccharomyces cerevisiae* Rad53 complementation assay, Shaag *et al.* found that p.P85L is equivalent to wild type CHEK2. While Bell *et al*. found this allele to have modestly reduced activity in an *in vitro* kinase function assay, both groups concluded that the allele is effectively neutral [22, 53]. Sodha *et al*. assayed the p.R137Q allele, and found that it encodes a protein with normal stability and normal response to DNA damage [37]. Bell *et al*. also assayed the p.R137Q allele, and found that it has normal kinase activity. In addition, Sodha *et al*. assayed the p.R180H allele; they found that it encodes a protein with slightly reduced stability but normal response to DNA damage. Thus existing functional assay results for these three variants are consistent with them being either neutral or at most weakly pathogenic. The fourth C0 substitution, p.T323P was found to have moderately reduced autophosphorylation and Cdc25C kinase activity by Wu *et al*. [54]. Classification of this substation as C0 is probably a true Align-GVGD error; the crystal structure of the protein reveals that T323 is located in an alpha helix, which would typically not be permissive of substitution to proline. The algorithmic problem is that the atomic composition and polarity of proline (the amino acid sidechain characteristics considered by the original Grantham Difference and Align-GVGD are atomic composition, polarity and volume) are intermediate between those of threonine and isoleucine, which are the two amino acids observed at position 323 in our alignment. The consequence is that proline is only slightly outside of the range of variation represented by these two wild-type residues and is consequently predicted to be neutral or nearly so. Although unpublished, misclassification of substitutions to proline that map within an alpha helix is a problem that we have observed before and is an obvious issue to bear in mind when considering missense substitution analyses made by Align-GVGD. p.I157T is perhaps the most interesting of the substitutions observed in our study that have been subjected to functional assays. Align-GVGD scores the variant as C15, indicative of modest evidence in favor of pathogenicity. Initially, Lee *et al*. found that kinase activity of the p.I157T allele was comparable to wild-type [55]. More recent studies have reported that the allele is at least partially defective in dimerization and autophosphorylation, binding and phosphorylating Cdc25, and binding BRCA1 [56-59]. In populations where p.I157T and c.1100delC are both present at appreciable frequencies and have been subject to independent risk estimates, p.I157T does appear to confer increased risk of breast cancer, but the OR or penetrance associated with the missense substitution appears to be more modest than that association with the frameshift c.1100delC [60]. At the other end of the spectrum, of the five C65 substitutions that we observed, only one, p.R117G, has been subjected to functional assays. Summing across several studies, this allele is phosphorylated by ATM in response to DNA damage, shows slightly to markedly reduced autophosphorylation, probably fails to oligomerize, and has severely compromised kinase activity towards Cdc25C [37, 54, 59]. Therefore the R117G allele encodes a functionally defective protein and is in all likelihood pathogenic. Thus for the missense substitutions observed in our mutation screening study and subjected to functional assays, there is a qualitative trend towards agreement between the Align-GVGD classification and functional assay result, consistent with the trend in OR that we observed across the Align-GVGD defined ordered series of missense substitution grades. However, noting that concordant results between *in silico* assessments and functional assays are not yet considered sufficient for formal clinical classification of missense substitutions observed in *BRCA1* and *BRCA2* [61-63], it does not appear that the state of the art of CHK2 functional assays has reached the point where concordant results from an *in silico* assessment and a functional assay would be sufficient for clinically relevant classification of a *CHEK2* missense substitution.

The genetics results described in this work, combined with the above functional assay summary, have implications for potential clinical genetic susceptibility tests that might include *CHEK2* and other genes with similar mutation profiles. From the 2003 American Society of Clinical Oncology Policy Statement Update on Genetic Testing for Cancer Susceptibility, the second and third "indications for genetic testing for cancer susceptibility" were that "2) the genetic test can be adequately interpreted, and 3) the test results will aid in diagnosis or influence the medical or surgical management of the patient or family members at hereditary risk of cancer" [64]. Towards the third criterion, some investigators have argued that, in the context of a high-risk family, the difference in risk between carriers and non-carriers of clearly pathogenic *CHEK2* sequence variants is sufficient to justify a difference in cancer surveillance strategies [65-67]. However, our results plus similar work with *ATM* [7, 68] point towards an issue under the second criterion. If roughly one-half of the genetically relevant risk that the test can pick up actually resides in rare missense substitutions that will be considered unclassified variants at their initial detection, it may not currently be possible to adequately interpret the test results. Therefore, while it is now technically feasible to design a massively parallel sequencing based test that can accurately and relatively inexpensively identify mutations in a panel of breast cancer susceptibility genes that includes *ATM* and *CHEK2* [69], it may be inappropriate to introduce such a test into widespread use before a clinically validated method of assessing unclassified missense substitutions in these genes has been developed.

The rare missense substitution analysis model combining Align-GVGD with the logistic regression test for trend grew out of the *in silico* analysis of missense substitutions that has now become a standard component in the integrated evaluation of unclassified variants in *BRCA1* and *BRCA2* [62, 70]. We proposed the model based on clinical *BRCA1* and *BRCA2* mutation screening data and then demonstrated its effectiveness by an analysis of ATM case-control mutation screening data [7, 42]. Thus the *CHEK2* analysis presented here stands as a methodological confirmation of our approach to inclusion of rare missense substitution data in case-control mutation screening studies. The logistic regression test for trend that we have used also provides a simple approach to combining evidence from rare missense substitutions with evidence from protein truncating sequence variants to build a more complete and statistically powerful approach to assessing case-control mutation screening data than would be afforded by either alone. From a technological perspective, we can envision combining exon capture and massively parallel sequencing to extend case-control mutation screening to entire biochemical pathways and beyond. Based on our *post-hoc* power calculations, at least 2,000 cases and 2,000 controls would be required for a whole pathway (such as DNA double strand break repair and allied cell cycle checkpoints) study, and 3,300 cases and controls to undertake a whole-exome study. On the one hand, these numbers could be an underestimate because *CHEK2* might be among the most important (in terms of familial relative risk) of the intermediate-risk class of breast cancer susceptibility genes. On the other hand, it could turn out that a test based on observations of evolutionarily unlikely sequence variants has an intrinsically lower false positive rate than anonymous marker GWAS and consequently not require a full Bonferroni multiple testing correction in order to reasonably constrain the rate of false positives.

CONCLUSIONS

This case-control mutation screening study of *CHEK2* shows that the gene harbors many different rare pathogenic sequence variants, a substantial proportion of which are missense substitutions. From a clinical perspective, the risk of breast cancer conferred by some pathogenic sequence variants in *CHEK2* may be great enough to be of use in a clinical cancer genetics setting, and we note that the technical capability to offer a multi-gene breast cancer susceptibility testing panel at relatively low per gene laboratory cost is in place. Yet our results with both *CHEK2* and *ATM* suggest that such a test would create a severe burden of unclassified missense substitutions, and that a large fraction of the genetically relevant risk would reside in those unclassified missense substitutions. Paradoxically, from the research perspective of susceptibility gene identification and characterization, this study validates our approach to analysis of rare missense substitutions observed during case-control mutation screening and provides a method to combine data from protein truncating variants and rare missense substitutions into a one degree of freedom per gene test.

ABBREVIATIONS

Breast CFR, Breast Cancer Family Registry; Center for MEGA Epidemiology, Center for Molecular, Environmental, Genetic & Analytic Epidemiology; FET, Fisher's exact test; GWAS, Genome-wide association study; IARC, International Agency for Research on Cancer; IRB, Institutional Review Board; OR, Odds ratio; PCR, Polymerase chain reaction; rMS, rare missense substitution; SNP, Single nucleotide polymorphism; T+SJV, Truncating and splice junction variant; WGA, Whole genome amplification.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHOR'S CONTRIBUTIONS

FLCK contributed to study design, led the laboratory team, and helped to draft the manuscript. FL contributed to study design, led the data analysis, and helped to draft the manuscript. FD contributed to the mutation screening and data analysis, and helped to refine the laboratory platform. MV contributed to the sequence alignment and data analysis. CV was responsible for data management throughput the project and helped to refine the laboratory platform. DB contributed to the sequence alignment and method for analysis of rare missense substitutions. GD contributed to the mutation screening and data analysis, and helped to refine the laboratory platform. NF contributed to the mutation screening and data analysis, and helped to refine the laboratory platform. SMC contributed to the mutation screening and data analysis, and helped to refine the laboratory platform. NR contributed to the mutation screening and data analysis, and helped to refine the laboratory platform. TND contributed to the sequence alignment and data analysis. AT contributed to statistical analyses and helped to draft the manuscript. GBB contributed to statistical analyses and helped to draft the manuscript. JLH was responsible for subjects ascertained through the University of Melbourne and helped to draft the manuscript.

MCS contributed to study design and contributed to management of samples ascertained through the University of Melbourne. ILA was responsible for subjects ascertained through Cancer Care Ontario. EMJ was responsible for subjects ascertained through the Northern California Cancer Center (now the Cancer Prevention Institute of California). SVT was responsible for overall study design, contributed to data analysis, and helped to draft the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

We would like to thank David Goldgar for help with familial relative risk calculations and note that a Microsoft Excel-based familial relative risk calculator is available from him at <david.goldgar@hsc.utah.edu>. We would also like to thank James McKay for help with mutation screening amplicon design, and Mia Hashibe and Deborah Neklason for helpful comments on the manuscript. This work was supported by the National Cancer Institute, National Institutes of Health (NIH) grant RO1 CA121245 and Canadian Institute for Health Research (CIHR) grant CRN-87521-IC089832. The Breast CFR was funded by the NIH under RFA-CA-06-503, and through cooperative agreements with Breast CFR members including Cancer Care Ontario (U01 CA69467), the Northern California Cancer Center (U01 CA69417), and the University of Melbourne (U01 CA69638). The content of this manuscript does not necessarily reflect the views or policies of the NCI or any of the collaborating centers in the Breast CFR, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the Breast CFR.

REFERENCES

1 Broca PP: **Traite des tumeurs**. Paris: Asselin; 1866.

2 Goldgar DE, Easton DF, Cannon-Albright L, Skolnick MH: **Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands**. *J Natl Cancer Inst* 1994, **86:**1600-1608.

3 Amundadottir LT, Thorvaldsson S, Gudbjartsson DF, Sulem P, Kristjansson K, Arnason S, Gulcher JR, Bjornsson J, Kong A, Thorsteinsdottir U, Stefansson K: **Cancer as a Complex Phenotype: Pattern of Cancer Distribution within and beyond the Nuclear Family.** *PLoS Med* 2004, **1:**e65.

4 Stratton MR, Rahman N: **The emerging landscape of breast cancer susceptibility.** *Nat Genet* 2008, **40:**17-22.

5 **Genetic Susceptibility.** In *World Cancer Report 2008*. Edited by Boyle P, Levin B. Lyon, France: International Agency for Research on Cancer (IARC); 2008:182-185.

6 Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, North B, Jayatilake H, Barfoot R, Spanova K, McGuffog L, Evans DG, Eccles D, Easton DF, Stratton MR, Rahman N: **ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles.** *Nat Genet* 2006, **38:**873-875.

7 Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F, Byrnes GB, Chuang SC, Forey N, Feuchtinger C, Gioia L, Hall J, Hashibe M, Herte B, McKay-Chopin S, Thomas A, Vallee MP, Voegele C, Webb PM, Whiteman DC, Sangrajrang S, Hopper JL, Southey MC, Andrulis IL, John EM, Chenevix-Trench G: **Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer.** *Am J Hum Genet* 2009, **85:**427-446.

8 Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, Chagtai T, Jayatilake H, Ahmed M, Spanova K, North B, McGuffog L, Evans DG, Eccles D, Easton DF, Stratton MR, Rahman N: **Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles.** *Nat Genet* 2006, **38:**1239-1241.

9 Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M, Elstrodt F, van Duijn C, Bartels C, Meijers C, Schutte M, McGuffog L, Thompson D, Easton D, Sodha N, Seal S, Barfoot R, Mangion J, Chang-Claude J, Eccles D, Eeles R, Evans DG, Houlston R, Murday V, Narod S, Peretz T, *et al.*: **Low-penetrance susceptibility to breast cancer due to**

CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet* 2002, **31:**55-59.

10. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T, Jayatilake H, McGuffog L, Hanks S, Evans DG, Eccles D, Easton DF, Stratton MR: **PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene.** *Nat Genet* 2006, **39:**165-167.

11. Erkko H, Dowty JG, Nikkila J, Syrjakoski K, Mannermaa A, Pylkas K, Southey MC, Holli K, Kallioniemi A, Jukkola-Vuorinen A, Kataja V, Kosma VM, Xia B, Livingston DM, Winqvist R, Hopper JL: **Penetrance analysis of the PALB2 c.1592delT founder mutation.** *Clin Cancer Res* 2008, **14:**4667-4671.

12. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, *et al.*: **Genome-wide association study identifies novel breast cancer susceptibility loci.** *Nature* 2007, **447:**1087-1093.

13. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JFJ, Hoover RN, Thomas G, Chanock SJ: **A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer.** *Nat Genet* 2007, **39:**870-874.

14. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, Masson G, Jakobsdottir M, Thorlacius S, Helgason A, Aben KK, Strobbe LJ, Albers-Akkers MT, Swinkels DW, Henderson BE, Kolonel LN, Le Marchand L, Millastre E, Andres R, Godino J, Garcia-Prats MD, Polo E, Tres A, Mouy M, Saemundsdottir J, Backman VM, Gudmundsson L, Kristjansson K, Bergthorsson JT, Kostic J, *et al.*: **Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer.** *Nat Genet* 2007, **39:**865-869.

15. Thompson D, Easton D: **The genetic epidemiology of breast cancer genes.** *J Mammary Gland Biol Neoplasia* 2004, **9:**221-236.

16. Mavaddat N, Pharoah PD, Blows F, Driver KE, Provenzano E, Thompson D, Macinnis RJ, Shah M, Search SO, Easton DF, Antoniou AC: **Familial relative risks for breast cancer by**

pathological subtype: a population-based cohort study. *Breast Cancer Res* 2010, **12:**R10.

17. Hopper JL, Carlin JB: **Familial aggregation of a disease consequent upon correlation between relatives in a risk factor measured on a continuous scale.** *Am J Epidemiol* 1992, **136:**1138-1147.

18. Antoni L, Sodha N, Collins I, Garrett MD: **CHK2 kinase: cancer susceptibility and cancer therapy - two sides of the same coin?** *Nat Rev Cancer* 2007, **7:**925-936.

19. Bell DW, Varley JM, Szydlo TE, Kang DH, Wahrer DC, Shannon KE, Lubratovich M, Verselis SJ, Isselbacher KJ, Fraumeni JF, Birch JM, Li FP, Garber JE, Haber DA: **Heterozygous germ line hCHK2 mutations in Li-Fraumeni syndrome.** *Science* 1999, **286:**2528-2531.

20. Cybulski C, Gorski B, Huzarski T, Masojc B, Mierzejewski M, Debniak T, Teodorczyk U, Byrski T, Gronwald J, Matyjasik J, Zlowocka E, Lenner M, Grabowska E, Nej K, Castaneda J, Medrek K, Szymanska A, Szymanska J, Kurzawski G, Suchy J, Oszurek O, Witek A, Narod SA, Lubinski J: **CHEK2 is a multiorgan cancer susceptibility gene.** *Am J Hum Genet* 2004, **75:**1131-1135.

21. Cybulski C, Masojc B, Oszutowska D, Jaworowska E, Grodzki T, Waloszczyk P, Serwatowski P, Pankowski J, Huzarski T, Byrski T, Gorski B, Jakubowska A, Debniak T, Wokolorczyk D, Gronwald J, Tarnowska C, Serrano-Fernandez P, Lubinski J, Narod SA: **Constitutional CHEK2 mutations are associated with a decreased risk of lung and laryngeal cancers.** *Carcinogenesis* 2008, **29:**762-765.

22. Shaag A, Walsh T, Renbaum P, Kirchhoff T, Nafa K, Shiovitz S, Mandell JB, Welcsh P, Lee MK, Ellis N, Offit K, Levy-Lahad E, King MC: **Functional and genomic approaches reveal an ancient CHEK2 allele associated with breast cancer in the Ashkenazi Jewish population.** *Hum Mol Genet* 2005, **14:**555-563.

23. Laitman Y, Kaufman B, Lahad EL, Papa MZ, Friedman E: **Germline CHEK2 mutations in Jewish Ashkenazi women at high risk for breast cancer.** *Isr Med Assoc J* 2007, **9:**791-796.

24. Cybulski C, Gorski B, Huzarski T, Byrski T, Gronwald J, Debniak T, Wokolorczyk D, Jakubowska A, Kowalska E, Oszurek O, Narod SA, Lubinski J: **CHEK2-positive breast cancers in young Polish women.** *Clin Cancer Res* 2006, **12:**4832-4835.

25. Cybulski C, Wokolorczyk D, Kladny J, Kurzawski G, Suchy J, Grabowska E, Gronwald J, Huzarski T, Byrski T, Gorski B, D Ecedil Bniak T, Narod SA, Lubinski J: **Germline CHEK2**

mutations and colorectal cancer risk: different effects of a missense and truncating mutations? *Eur J Hum Genet* 2007, **15:**237-241.

26. Brennan P, McKay J, Moore L, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Chow WH, Rothman N, Chabrier A, Gaborieau V, Odefrey F, Southey M, Hashibe M, Hall J, Boffetta P, Peto J, Peto R, Hung RJ: **Uncommon CHEK2 mis-sense variant and reduced risk of tobacco-related cancers: case-control study.** *Hum Mol Genet* 2007,

27. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH: **Multiple rare alleles contribute to low plasma levels of HDL cholesterol.** *Science* 2004, **305:**869-872.

28. Kanetsky PA, Rebbeck TR, Hummer AJ, Panossian S, Armstrong BK, Kricker A, Marrett LD, Millikan RC, Gruber SB, Culver HA, Zanetti R, Gallagher RP, Dwyer T, Busam K, From L, Mujumdar U, Wilcox H, Begg CB, Berwick M: **Population-based study of natural variation in the melanocortin-1 receptor gene and melanoma.** *Cancer Res* 2006, **66:**9330-9337.

29. Fernandez L, Milne R, Bravo J, Lopez J, Aviles J, Longo M, Benitez J, Lazaro P, Ribas G: **MC1R: three novel variants identified in a malignant melanoma association study in the Spanish population.** *Carcinogenesis* 2007, **28:**1659-1664.

30. John EM, Hopper JL, Beck JC, Knight JA, Neuhausen SL, Senie RT, Ziogas A, Andrulis IL, Anton-Culver H, Boyd N, Buys SS, Daly MB, O'Malley FP, Santella RM, Southey MC, Venne VL, Venter DJ, West DW, Whittemore AS, Seminara D: **The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer.** *Breast Cancer Res* 2004, **6:**R375-89.

31. Reed GH, Wittwer CT: **Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis.** *Clin Chem* 2004, **50:**1748-1754.

32. Takano EA, Mitchell G, Fox SB, Dobrovic A: **Rapid detection of carriers with BRCA1 and BRCA2 mutations using high resolution melting analysis.** *BMC Cancer* 2008, **8:**59.

33. Nguyen-Dumont T, Calvez-Kelm FL, Forey N, McKay-Chopin S, Garritano S, Gioia-Patricola L, De Silva D, Weigel R, Sangrajrang S, Lesueur F, Tavtigian SV: **Description and validation of high-throughput simultaneous genotyping and mutation scanning by high-resolution melting curve analysis.** *Hum Mutat* 2009, **30:**884-890.

34. Sodha N, Houlston RS, Williams R, Yuille MA, Mangion J, Eeles RA: **A robust method for detecting CHK2/RAD53 mutations in genomic DNA.** *Hum Mutat* 2002, **19:**173-177.

35. Voegele C, Tavtigian SV, de Silva D, Cuber S, Thomas A, Le Calvez-Kelm F: **A Laboratory Information Management System (LIMS) for a high throughput genetic platform aimed at candidate gene mutation screening.** *Bioinformatics* 2007, **23:**2504-2506.

36. Wallace IM, O'Sullivan O, Higgins DG, Notredame C: **M-Coffee: combining multiple sequence alignment methods with T-Coffee.** *Nucleic Acids Res* 2006, **34:**1692-1699.

37. Sodha N, Mantoni TS, Tavtigian SV, Eeles R, Garrett MD: **Rare germ line CHEK2 variants identified in breast cancer families encode proteins that show impaired activation.** *Cancer Res* 2006, **66:**8966-8970.

38. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2)**. *Cladistics* 1989, **5:**164-166.

39. Ng PC, Henikoff S: **Accounting for human polymorphisms predicted to affect protein function.** *Genome Res* 2002, **12:**436-446.

40. Align-GVGD website. [http://agvgd.iarc.fr/].

41. Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A: **Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral.** *J Med Genet* 2006, **43:**295-305.

42. Tavtigian SV, Byrnes GB, Goldgar DE, Thomas A: **Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications.** *Hum Mutat* 2008, **29:**1342-1354.

43. SIFT website [http://sift.jcvi.org/].

44. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7:**248-249.

45. PolyPhen-2 website [http://genetics.bwh.harvard.edu/pph2/].

46. Greenland S: **Applications of stratified analysis methods.** In *Modern Epidemiology, second edition*. Edited by Rothman KJ, Greenland S. Philadelphia, USA: Lippincot Raven; 1998:281-300.

47. Goldgar DE: **Population aspects of cancer genetics.** *Biochimie* 2002, **84:**19-25.

48. 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:**1175-1182.

49. Bernstein JL, Teraoka SN, John EM, Andrulis IL, Knight JA, Lapinski R, Olson ER, Wolitzer AL, Seminara D, Whittemore AS, Concannon P: **The CHEK2*1100delC allelic variant and risk of breast cancer: screening results from the Breast Cancer Family Registry.** *Cancer Epidemiol Biomarkers Prev* 2006, **15:**348-352.

50. Greenblatt MS, Beaudet JG, Gump JR, Godin KS, Trombley L, Koh J, Bond JP: **Detailed computational study of p53 and p16: using evolutionary sequence analysis and disease-associated mutations to predict the functional consequences of allelic variants.** *Oncogene* 2003, **22:**1150-1163.

51. Cooper GM, Brudno M, Green ED, Batzoglou S, Sidow A: **Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes.** *Genome Res* 2003, **13:**813-820.

52. Lunter G, Hein J: **A nucleotide substitution model with nearest-neighbour interactions.** *Bioinformatics* 2004, **20 Suppl 1:**I216-I223.

53. Bell DW, Kim SH, Godwin AK, Schiripo TA, Harris PL, Haserlat SM, Wahrer DC, Haiman CA, Daly MB, Niendorf KB, Smith MR, Sgroi DC, Garber JE, Olopade OI, Le Marchand L, Henderson BE, Altshuler D, Haber DA, Freedman ML: **Genetic and functional analysis of CHEK2 (CHK2) variants in multiethnic cohorts.** *Int J Cancer* 2007, **121:**2661-2667.

54. Wu X, Dong X, Liu W, Chen J: **Characterization of CHEK2 mutations in prostate cancer.** *Hum Mutat* 2006, **27:**742-747.

55. Lee SB, Kim SH, Bell DW, Wahrer DC, Schiripo TA, Jorczak MM, Sgroi DC, Garber JE, Li FP, Nichols KE, Varley JM, Godwin AK, Shannon KM, Harlow E, Haber DA: **Destabilization of CHK2 by a missense mutation associated with Li-Fraumeni Syndrome.** *Cancer Res* 2001, **61:**8062-8067.

56. Cai Z, Chehab NH, Pavletich NP: **Structure and activation mechanism of the CHK2 DNA damage checkpoint kinase.** *Mol Cell* 2009, **35:**818-829.

57. Falck J, Mailand N, Syljuasen RG, Bartek J, Lukas J: **The ATM-Chk2-Cdc25A checkpoint pathway guards against radioresistant DNA synthesis.** *Nature* 2001, **410:**842-847.

58. Li J, Williams BL, Haire LF, Goldberg M, Wilker E, Durocher D, Yaffe MB, Jackson SP, Smerdon SJ: **Structural and functional versatility of the FHA domain in DNA-damage signaling by the tumor suppressor kinase Chk2.** *Mol Cell* 2002, **9:**1045-1054.

59. Chrisanthar R, Knappskog S, Lokkevik E, Anker G, Ostenstad B, Lundgren S, Berge EO,

Risberg T, Mjaaland I, Maehle L, Engebretsen LF, Lillehaug JR, Lonning PE: **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.

60. Nevanlinna H, Bartek J: **The CHEK2 gene and inherited breast cancer susceptibility.** *Oncogene* 2006, **25:**5912-5919.
61. Couch FJ, Rasmussen LJ, Hofstra R, Monteiro AN, Greenblatt MS, de Wind N: **Assessment of functional effects of unclassified genetic variants.** *Hum Mutat* 2008, **29:**1314-1326.
62. Goldgar DE, Easton DF, Byrnes GB, Spurdle AB, Iversen ES, Greenblatt MS: **Genetic evidence and integration of various data sources for classifying uncertain variants into a single model.** *Hum Mutat* 2008, **29:**1265-1272.
63. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, Tavtigian SV: **Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results.** *Hum Mutat* 2008, **29:**1282-1291.
64. American Society of Clinical Oncology: **American Society of Clinical Oncology Policy Statement Update: Genetic Testing for Cancer Susceptibility**. *Journal of Clinical Oncology* 2003, **21:**2397-2406.
65. Johnson N, Fletcher O, Naceur-Lombardelli C, dos Santos Silva I, Ashworth A, Peto J: **Interaction between CHEK2*1100delC and other low-penetrance breast-cancer susceptibility genes: a familial study.** *Lancet* 2005, **366:**1554-1557.
66. Byrnes GB, Southey MC, Hopper JL: **Are the so-called low penetrance breast cancer genes, ATM, BRIP1, PALB2 and CHEK2, high risk for women with strong family histories?** *Breast Cancer Res* 2008, **10:**208.
67. Narod SA: **Testing for CHEK2 in the cancer genetics clinic: ready for prime time?** *Clin Genet* 2010, **78:**1-7.
68. Bernstein JL, Haile RW, Stovall M, Boice JDJ, Shore RE, Langholz B, Thomas DC, Bernstein L, Lynch CF, Olsen JH, Malone KE, Mellemkjaer L, Borresen-Dale AL, Rosenstein BS, Teraoka SN, Diep AT, Smith SA, Capanu M, Reiner AS, Liang X, Gatti RA, Concannon P: **Radiation exposure, the ATM Gene, and contralateral breast cancer in the women's environmental cancer and radiation epidemiology study.** *J Natl Cancer Inst* 2010, **102:**475-483.

69. Walsh T, Lee MK, Casadei S, Thornton AM, Stray SM, Pennil C, Nord AS, Mandell JB, Swisher EM, King MC: **Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing.** *Proc Natl Acad Sci U S A* 2010, **107:**12629-12633.

70. Spurdle AB, Lakhani SR, Healey S, Parry S, Da Silva LM, Brinkworth R, Hopper JL, Brown MA, Babikyan D, Chenevix-Trench G, Tavtigian SV, Goldgar DE: **Clinical classification of BRCA1 and BRCA2 DNA sequence variants: the value of cytokeratin profiles and evolutionary analysis--a report from the kConFab Investigators.** *J Clin Oncol* 2008, **26:**1657-1663.

Table 1: Subjects excluded because of poor mutation screening performance, by study center

Center	Cases (%)*	Controls (%)*
Breast CFR-Australia	5 (0.8%)	11 (2.1%)
Breast CFR-Canada	1 (0.3%)	2 (0.4%)
Breast CFR-Northern California	4 (1.0%)	1 (0.7%)
TOTAL	10 (0.8%)	14 (1.2%)

* Percents given are the percent of the total number of case or control DNAs provided by the indicated center.

CFR, Cancer Family Registry.

All 10 excluded cases are < 42 years old.
All 14 excluded controls are < 45 years old.

Table 2: Distribution of cases and controls* by age, race/ ethnicity, and study center

Distribution by age	Cases (%)†		Controls (%)†	
≤30	106	(8.1%)	66	(6.0%)
31-35	322	(24.7%)	171	(15.4%)
36-40	434	(33.3%)	231	(20.8%)
41-45	441	(33.8%)	199	(17.9%)
46-50	0	(0.0%)	230	(20.7%)
51-55	0	(0.0%)	212	(19.1%)
TOTAL	1,303	(100.0%)	1,109	(100.0%)
Distribution by race/ ethnicity				
Caucasian	843	(64.7%)	956	(86.2%)
East Asian	204	(15.7%)	70	(6.3%)
Latina	158	(12.1%)	47	(4.2%)
Recent African Ancestry	98	(7.5%)	36	(3.2%)
TOTAL	1,303	(100.0%)	1,109	(100.0%)
Distribution by study center				
Breast CFR-Australia	588	(45.1%)	513	(46.3%)
Breast CFR-Canada	302	(23.2%)	461	(41.6%)
Breast CFR-Northern California	413	(31.7%)	135	(12.2%)
TOTAL	1,303	(100.0%)	1,109	(100.0%)

*Not including case and controls excluded because of poor mutation screening performance.
† Percents given are the percent of the total number of case or control DNAs, in the category indicated, that met the mutation screening quality control criterion.

CFR, Cancer Family Registry.

Table 3: Analyses of rare variants, with missense substitutions stratified by Align-GVGD grade

Class	Cases	Controls	Crude OR[6]	Adjusted* OR[6]
			[95% CI[7]]	[95% CI[7]]
Noncarriers	1,242	1,089	ref	ref
T+SJV[1]	17	3	**4.97 [1.45-17.0]**	**6.18 [1.76-21.8]**
Any rMS[2]	44	17	**2.27 [1.29-4.00]**	**2.20 [1.20-4.01]**
rMSs[2] stratified by Align-GVGD grade¥				
C0	12	9	1.17 [0.49-2.79]	1.39 [0.55-3.56]
C15	14	5	2.46 [0.88-6.84]	1.82 [0.62-5.34]
C25	7	2	3.07 [0.64-14.8]	2.47 [0.45-13.49]
C35	1	0	-	
C45	0	0	-	
C55	1	0	-	
C65	9	1	**7.89 [1.00-62.4]**	**8.75 [1.06-72.2]**
rMSs[2] stratified by SIFT grade ¶				
S[3]>0.05	21	8	2.30 [1.02-5.22]	1.99 [0.83-4.77]
0.05≥S[3]>0.00	12	5	2.10 [0.74-5.99]	1.91 [0.63-5.86]
S[3]=0.00	11	4	2.41 [0.77-7.59]	3.03 [0.91-10.0]
rMSs[2] stratified PolyPhen-2 grade				
Benign	16	7	2.00 [0.82-4.89]	1.69 [0.64-4.41]
Possibly D[4]	10	6	1.46 [0.53-4.03]	1.65 [0.55-4.89]
Probably D[5]	18	4	**3.95 [1.33-11.7]**	**3.87 [1.25-12.0]**

* Adjusted for race/ethnicity (Caucasian, East Asian, African American, and Latina), study center, and age as a categorical variables.
¥Using the *CHEK2* sequence alignment through *S. purpuratus* (sea urchin).
¶Using the *CHEK2* sequence alignment through *D. melanogaster* (fruitfly).

1. T+SJVs, Protein truncating variants plus splice junction variants.
2. rMSs, Rare missense substitutions.
3. SIFT score.
4. PolyPhen-2 “Possibly Damaging”.
5. PolyPhen-2 “Probably Damaging”.
6. OR, Odds ratio.
7. CI, Confidence interval.

Table 4: Results from logistic regression tests for log-linear OR trend

Grouping of rMSs[1] and/or T+SJVs[2]	Ln(OR[3]) regression coefficient, [95% CI[4] of the regression coefficient], and P-value	
	Crude	Adjusted †
rMSs[1] only (i.e., exclude T+SJVs[2]) note that C65 is grade 7	**0.35 [0.12-0.58]** **p=0.0029**	**0.33 [0.09-0.55]** **p=0.0055**
C65 rMSs[1] and T+SJVs[2] pooled in grade 7.	**0.28 [0.14-0.43]** **p=0.00013**	**0.29 [0.14-0.43]** **p=0.000088**
C65 rMSs[1] in grade 7 and T+SJVs[2] in grade 8	**0.26 [0.12-0.39]** **p=0.00017**	**0.26 [0.13-0.40]** **p=0.00011**

†Adjusted for race/ethnicity (Caucasian, East Asian, African American, and Latina), study center, and age as a categorical variables.

1. rMSs, Rare missense substitutions.
2. T+SJVs, Protein truncating variants plus splice junction variants.
3. OR, Odds ratio.
4. CI, Confidence interval.

Table 5: Number of cases and frequency matched controls required for various scales of future intermediate-risk gene case-control mutation screening studies

Study scale	Single genes	Whole pathways*	Whole exome†
Alpha	0.05	0.0005	2.5×10^{-6}
Beta	0.80	0.80	0.80
rMSs[1] alone	1,975	4,700	7,725
T+SJVs[2] alone	1,425	3,400	5,600
rMSs[1] plus T+SJVs[2]	850	2,025	3,350

* Calculated for 100 genes, approximately the gene count of DNA double strand break repair and associated cell cycle checkpoints.

† Calculated for 20,000 genes.

1. rMSs, Rare missense substitutions.
2. T+SJVs, Protein truncating variants plus splice junction variants.

ADDITIONAL FILES

Additional file 1: Supplementary tables S1 and S2.
Supplementary table S1. Missense, nonsense, frameshift, and splice junction variants.
Supplementary table S2. CHEK2 protein multiple sequence alignment characterization.

Additional files provided with this submission:

Additional file 1: CHEK2_supplementary_v9.pdf, 150K
http://breast-cancer-research.com/imedia/2061462935486711/supp1.pdf