Open access • Journal Article • DOI:10.1038/NG.955

Mutations in BRIP1 confer high risk of ovarian cancer — Source link

Thorunn Rafnar, Daniel F. Gudbjartsson, Patrick Sulem, Aslaug Jonasdottir ...+43 more authors

Institutions: deCODE genetics, University of Iceland, University of Zaragoza, Radboud University Nijmegen Medical Centre ...+2 more institutions

Published on: 01 Nov 2011 - Nature Genetics (Nature Publishing Group)

Topics: Cancer, Frameshift mutation, Ovarian cancer, Allele frequency and Breast cancer

Mutations in *BRIP1/FANCJ* confer high risk of ovarian cancer

Thorunn Rafnar[1]*, Daniel F Gudbjartsson[1]*, Patrick Sulem[1]*, Aslaug Jonasdottir[1], Asgeir Sigurdsson[1], Adalbjorg Jonasdottir[1], Soren Besenbacher[1], Pär Lundin[1], Simon N Stacey[1], Julius Gudmundsson[1], Olafur Magnusson[1], Louise le Roux[1], Gudbjorg Orlygsdottir[1], Hafdis Helgadottir[1], Hrefna Johannsdottir[1], Arnaldur Gylfason[1], Laufey Tryggvadottir[2,3], Jon G Jonasson[2,3], Ana De Juan[4], Eugenia Ortega[5], Jose M. Ramon-Cajal[6], Maria D García-Prats[7], Carlos Mayordomo[8], Angeles Panadero[9], Fernando Rivera[4], Katja KH Aben[10,11], Anne M van Altena[12], Leon FAG Massuger[12], Mervi Aavikko[13], Paula M Kujala[14], Synnöve Staff[15, 16], Lauri A Aaltonen[13], Kristrun Olafsdottir[17], Johannes Bjornsson[17], Augustine Kong[1], Anna Salvarsdottir[18], Hafsteinn Saemundsson[18], Karl Olafsson[18], Kristrun R Benediktsdottir[3,17], Jeffrey Gulcher[1], Gisli Masson[1], Lambertus A Kiemeney[10,11,19], Jose I Mayordomo[20], Unnur Thorsteinsdottir[1,3], Kari Stefansson[1,3]

[1]deCODE Genetics, Sturlugata 8, 101 Reykjavik, Iceland
[2]Icelandic Cancer Registry, Skógarhlíð 8, 105 Reykjavik
[3]Faculty of Medicine, University of Iceland, 101 Reykjavik, Iceland
[4]Division of Medical Oncolology, Marques de Valdecilla University Hospital, 39008 Santander, Spain
[5]Division of Medical Oncolology, Arnau de Vilanova University Hospital, Lérida, Spain
[6]Division of Gynecology, San Jorge General Hospital, 22004 Huesca, Spain
[7]Division of Pathology, San Jorge General Hospital, 22004 Huesca, Spain
[8]University of Zaragoza, 50009 Zaragoza, Spain
[9]Division of Medical Oncolology, Ciudad de Coria Hospital, 10800 Coria, Spain
[10]Comprehensive Cancer Center The Netherlands, Location Nijmegen, 6501 BG Nijmegen, The Netherlands
[11]Department of Epidemiology, Biostatistics and HTA, Radboud University Nijmegen Medical Centre, 6500 HB Nijmegen, The Netherlands
[12]Department of Gynaecology, Radboud University Nijmegen Medical Centre, 6500 HB Nijmegen, The Netherlands
[13]Department of Medical Genetics, Genome-Scale Biology Research Program, University of Helsinki, Helsinki, Finland
[14]Department of Pathology, Centre for Laboratory Medicine, University of Tampere and Tampere University Hospital, Tampere, Finland
[15]Laboratory of Cancer Biology, Institute of Biomedical Technology, University of Tampere, Finland
[16]Department of Obstetrics and Gynecology, Tampere University Hospital, Tampere, Finland
[17]Dept. Of Pathology, Landspitali University Hospital 101, Reykjavik, Iceland
[18]Department of Obstetrics and Gynecology, Landspitali-University Hospital, 101 Reykjavik, Iceland
[19]Department of Urology, Radboud University Nijmegen Medical Centre, 6500 HB Nijmegen, The Netherlands.
[20]Division of Medical Oncology. University Hospital, 50009, Zaragoza, Spain

Supplementary Information

Contents:

Supplementary methods:

Whole-genome sequencing and SNP imputations:

SNPs were imputed based on unpublished data from the Icelandic whole genomic sequencing project (457 Icelandic individuals) selected for various neoplastic, cardiovascular and psychiatric conditions. All of the individuals were sequenced to a depth of at least 10X. Sixteen million SNPs were imputed based on this set of individuals.

1. *Sample preparation.* Paired-end libraries for sequencing were prepared according to the manufacturer's instructions (Illumina). In short, approximately 5 µg of genomic DNA, isolated from frozen blood samples, were fragmented to a mean target size of 300 bp using a Covaris E210 instrument. The resulting fragmented DNA was end repaired using T4 and Klenow polymerases and T4 polynucleotide kinase with 10 mM dNTP followed by addition of an 'A' base at the ends using Klenow exo fragment (3′ to 5′-exo minus) and dATP (1 mM). Sequencing adaptors containing 'T' overhangs were ligated to the DNA products followed by agarose (2%) gel electrophoresis. Fragments of about 400 bp were isolated from the gels (QIAGEN Gel Extraction Kit), and the adaptor-modified DNA fragments were PCR enriched for ten cycles using Phusion DNA polymerase (Finnzymes Oy) and PCR primers PE 1.0 and PE 2.0 (Illumina). Enriched libraries were further purified using agarose (2%) gel electrophoresis as described above. The quality and concentration of the libraries were assessed with the Agilent 2100 Bioanalyzer using the DNA 1000 LabChip (Agilent). Barcoded libraries were stored at –20 °C. All steps in the workflow were monitored using an in-house laboratory information management system with barcode tracking of all samples and reagents.
2. *DNA sequencing.* Template DNA fragments were hybridized to the surface of flow cells (Illumina PE flowcell, v4) and amplified to form clusters using the Illumina cBot. In brief, DNA (3–10 pM) was denatured, followed by hybridization to grafted adaptors on the flowcell. Isothermal bridge amplification using Phusion polymerase was then followed by linearization of the bridged DNA, denaturation, blocking of 3´ ends and hybridization of the sequencing primer. Sequencing-by-synthesis was performed on Illumina GAIIx instruments equipped with paired-end modules. Paired-end libraries for whole genome sequencing were sequenced using either 2×101 or 2×120 cycles of incorporation and imaging with Illumina sequencing kits, v4 or v5 (TruSeq). Each library or sample was

initially run on a single lane for validation followed by further sequencing of ≥4 lanes with targeted raw cluster densities of 500–700 k/mm2, depending on the version of the data imaging and analysis packages. Imaging and analysis of the data was performed using SCS2.6 /RTA1.6, SCS2.8/RTA1.8 or SCS2.9&RTA1.9 software packages from Illumina, respectively. Real-time analysis involved conversion of image data to base-calling in real-time.

3. *Alignment.* For each lane in the DNA sequencing output, the resulting qseq files were converted into fastq files using an in-house script. All output from sequencing was converted, and the Illumina quality filtering flag was retained in the output. The fastq files were then aligned against Build 36 of the human reference sequence using bwa version 0.5.7 [1].

4. *BAM file generation.* SAM file output from the alignment was converted into BAM format using samtools version 0.1.8 [2], and an in-house script was used to carry the Illumina quality filter flag over to the BAM file. The BAM files for each sample were then merged into a single BAM file using samtools. Finally, Picard version 1.17 (see http://picard.sourceforge.net/) was used to mark duplicates in the resulting sample BAM files.

5. *SNP calling and genotyping in whole-genome sequencing.* A two-step approach was applied. The first step was to detect SNPs by identifying sequence positions where at least one individual could be determined to be different from the reference sequence with confidence (quality threshold of 20) based on the SNP calling feature of the pileup tool samtools [2]. SNPs that always differed heterozygous or homozygous from the reference were removed. The second step was to use the pileup tool to genotype the SNPs at the positions that were flagged as polymorphic. Because sequencing depth varies and hence the certainty of genotype calls also varies, genotype likelihoods rather than deterministic calls were calculated (see below). Of the 2.5 million SNPs reported in the HapMap2 CEU samples, 96.3% were observed in the whole-genome sequencing data. Of the 6.9 million SNPs reported in the 1000 Genomes Project data, 89.4% were observed in the whole-genome sequencing data.

Long range phasing:

Long range phasing of all chip-genotyped individuals was performed with methods described previously [3,4]. In brief, phasing was achieved using an iterative algorithm which phases a single proband at a time given the available phasing information about everyone else that shares a long haplotype identically by state with the proband. Given the large fraction of the Icelandic population that has been chip-typed, accurate long range phasing is available genome-wide for all chip-typed Icelanders.

Genotype imputation:

We imputed the SNPs identified and genotyped through sequencing into all Icelanders who had been phased with long range phasing using the same model as used by IMPUTE [5]. The genotype data from sequencing can be ambiguous due to low sequencing coverage. In order to phase the sequencing genotypes, an iterative algorithm was applied for each SNP with alleles 0 and 1. We let H be the long range phased haplotypes of the sequenced individuals and applied the following algorithm:

1. For each haplotype h in H, use the Hidden Markov Model of IMPUTE to calculate for every other k in H, the likelihood, denoted $\gamma_{h,k}$, of h having the same ancestral source as k at the SNP.
2. For every h in H, initialize the parameter θ_h, which specifies how likely the one allele of the SNP is to occur on the background of h from the genotype likelihoods obtained from sequencing. The genotype likelihood L_g is the probability of the observed sequencing data at the SNP for a given individual assuming g is the true genotype at the SNP. If L_0, L_1 and L_2 are the likelihoods of the genotypes 0, 1 and 2 in the individual who carries h, then set $\theta_h = \frac{L_2+\frac{1}{2}L_1}{L_2+L_1+L_0}$.
3. For every pair of haplotypes h and k in H that are carried by the same individual, use the other haplotypes in H to predict the genotype of the SNP on the backgrounds of h and k: $\tau_h = \sum_{l\in H\setminus\{h\}}\gamma_{h,l}\theta_l$ and $\tau_k = \sum_{l\in H\setminus\{k\}}\gamma_{k,l}\theta_l$. Combining these predictions with the genotype likelihoods from sequencing gives un-normalized updated phased genotype probabilities: $P_{00} = (1-\tau_h)(1-\tau_k)L_0$, $P_{10} = \tau_h(1-\tau_k)\frac{1}{2}L_1$, $P_{01} = (1-\tau_h)\tau_k\frac{1}{2}L_1$ and

$P_{11} = \tau_h \tau_k L_2$. Now use these values to update θ_h and θ_k to $\theta_h = \frac{P_{10}+P_{11}}{P_{00}+P_{01}+P_{10}+P_{11}}$ and $\theta_k = \frac{P_{01}+P_{11}}{P_{00}+P_{01}+P_{10}+P_{11}}$.

4. Repeat step 3 when the maximum difference between iterations is greater than a convergence threshold ε. We used $\varepsilon=10^{-7}$.

Given the long range phased haplotypes and θ, the allele of the SNP on a new haplotype h not in H, is imputed as $\sum_{l \in H} \gamma_{h,l} \theta_l$.

The above algorithm can easily be extended to handle simple family structures such as parent-offspring pairs and triads by letting the P distribution run over all founder haplotypes in the family structure. The algorithm also extends easily to the X-chromosome. If source genotype data are only ambiguous in phase, such as chip genotype data, then the algorithm is still applied, but all but one of the Ls will be 0. In some instances, the reference set was intentionally enriched for carriers of the minor allele of a rare SNP in order to improve imputation accuracy. In this case, expected allele counts is biased toward the minor allele of the SNP. Call the enrichment of the minor allele E and let θ' be the expected minor allele count calculated from the naïve imputation method, and let θ be the unbiased expected allele count, then $\theta' = \frac{E\theta}{1-\theta+E\theta}$ and hence $\theta = \frac{\theta'}{E+(1-E)\theta'}$.

This adjustment was applied to all imputations based on enriched imputations sets. We note that if θ' is 0 or 1, then θ will also be 0 or 1, respectively.

In silico (genealogy-based) genotyping:

In addition to imputing sequence variants from the whole genome sequencing effort into chip genotyped individuals, we also performed a second imputation step where genotypes were imputed into relatives of chip genotyped individuals, creating *in silico* genotypes. The inputs into the second imputation step are the fully phased (in particular every allele has been assigned a parent of origin) imputed and chip type genotypes of the available chip typed individuals. The algorithm used to perform the second imputation step consists of:

1. For each ungenotyped individual (the proband), find all chip genotyped individuals within two meiosis of the individual. The six possible types of two meiosis relatives of the

proband are (ignoring more complicated relationships due to pedigree loops): Parents, full and half siblings, grandparents, children and grandchildren. If all pedigree paths from the proband to a genotyped relative go through other genotyped relatives, then that relative is excluded. e.g. if a parent of the proband is genotyped, then the proband's grandparents through that parent are excluded. If the number of meiosis in the pedigree around the proband exceeds a threshold (we used 12), then relatives are removed from the pedigree until the number of meiosis falls below 12, in order to reduce computational complexity.

2. At every point in the genome, calculate the probability for each genotyped relative sharing with the proband based on the autosomal SNPs used for phasing. A multipoint algorithm based on the hidden Markov model Lander-Green multipoint linkage algorithm using fast Fourier transforms is used to calculate these sharing probabilities [6,7]. First single point sharing probabilities are calculated by dividing the genome into 0.5cM bins and using the haplotypes over these bins as alleles. Haplotypes that are the same, except at most at a single SNP, are treated as identical. When the haplotypes in the pedigree are incompatible over a bin, then a uniform probability distribution was used for that bin. The most common causes for such incompatibilities are recombinations within the pedigree, phasing errors and genotyping errors. Note that since the input genotypes are fully phased, the single point information is substantially more informative than for unphased genotyped, in particular one haplotype of the parent of a genotyped child is always known. The single point distributions are then convolved using the multipoint algorithm to obtain multipoint sharing probabilities at the center of each bin. Genetic distances were obtained from the most recent version of the deCODE genetic map [8].
3. Based on the sharing probabilities at the center of each bin, all the SNPs from the whole genome sequencing are imputed into the proband. To impute the genotype of the paternal allele of a SNP located at x, flanked by bins with centers at x_{left} and x_{right}. Starting with the left bin, going through all possible sharing patterns v, let I_v be the set of haplotypes of genotyped individuals that share identically by descent within the pedigree with the proband's paternal haplotype given the sharing pattern v and $P(v)$ be the probability of v at the left bin – this is the output from step 2 above – and let e_i be the expected allele count of the SNP for haplotype i. Then $e_v = \frac{\sum_{i \in I_v} e_i}{\sum_{i \in I_v} 1}$ is the expected allele count of the paternal haplotype of the proband given v and an overall estimate of the

allele count given the sharing distribution at the left bin is obtained from $e_{left} = \sum_v P(v)e_v$. If I_v is empty then no relative shares with the proband's paternal haplotype given v and thus there is no information about the allele count. We therefore store the probability that some genotyped relative shared the proband's paternal haplotype, $O_{left} = \sum_{v,I_v=\emptyset} P(V)$ and an expected allele count, conditional on the proband's paternal haplotype being shared by at least one genotyped relative: $c_{left} = \frac{\sum_{v,I_{v\neq\emptyset}} P(v)e_v}{\sum_{v,I_{v\neq\emptyset}} P(v)}$. In the same way calculate O_{right} and c_{right}. Linear interpolation is then used to get an estimates at the SNP from the two flanking bins:

$$O = O_{left} + \frac{x - x_{left}}{x_{right} - x_{left}}\left(O_{right} - O_{left}\right),$$

$$c = c_{left} + \frac{x - x_{left}}{x_{right} - x_{left}}\left(c_{right} - c_{left}\right).$$

If θ is an estimate of the population frequency of the SNP then $Oc + (1 - O)\theta$ is an estimate of the allele count for the proband's paternal haplotype. Similarly, an expected allele count can be obtained for the proband's maternal haplotype.

Case-control association testing:

Logistic regression was used to test for association between SNPs and disease, treating disease status as the response and expected genotype counts from imputation or allele counts from direct genotyping as covariates. Testing was performed using the likelihood ratio statistic. When testing for association based on the *in silico* genotypes, controls were matched to cases based on the informativeness of the imputed genotypes, such that for each case C controls of matching informativeness where chosen. Failing to match cases and controls will lead to a highly inflated genomic control factor, and in some cases may lead to spurious false positive findings. The informativeness of each of the imputation of each one of an individual's haplotypes was estimated by taking the average of

$$a(e,\theta) = \begin{cases} \dfrac{e-\theta}{1-\theta}, & e \geq \theta \\ \dfrac{\theta - e}{\theta}, & e < \theta \end{cases}$$

over all SNPs imputed for the individual, where e is the expected allele count for the haplotype at

the SNP and θ is the population frequency of the SNP. Note that $a(\theta,\theta) = 0$ and $a(0,\theta) = a(1,\theta) = 1$. The mean informativeness values cluster into groups corresponding to the most common pedigree configurations used in the imputation, such as imputing from parent into child or from child into parent. Based on this clustering of imputation informativeness, we divided the haplotypes of individuals into seven groups of varying informativeness which created 27 groups of individuals of similar imputation informativeness; 7 groups of individuals with both haplotypes having similar informativeness, 21 groups of individuals with the two haplotypes having different informativeness, minus the one group of individuals with neither haplotype being imputed well. Within each group we calculate the ratio of the number of controls and the number of cases, and choose the largest integer C that was less than this ratio in all the groups. For example, if in one group there are 10.3 times as many controls as cases and if in all other groups this ratio was greater, then we would set $C = 10$ and within each group randomly select ten times as many controls as there are cases.

Sibling recurrence risk ratio:

The sibling recurrence risk ratio is defined as $\lambda_{sibling} = \frac{P(A \mid B)}{P(A)} = \frac{P(AB)}{P(A)P(B)}$, where A is the event that a person gets a disease and B is the event that a particular sibling of the person gets the disease. Assuming a multiplicative model, the $\lambda_{sibling}$ accounted for by a variant with frequency f and relative risk of r is equal to $\frac{\frac{1}{4}\left[fr^2 + 1 - f + (fr + 1 - f)^2\right]^2}{(fr + 1 - f)^4}$.

Microsatellite genotyping:

The PCR amplifications were set up and pooled using Zymark SciClone ALH 500 robots. The sequences of the primers used for genotyping are listed in Supplementary table 8. The reaction volume was 5 µl, and, for each PCR, 20 ng of genomic DNA was amplified in the presence of 2 pmol of each primer, 0.14 U AmpliTaq Gold, 0.33 mmol/liter dNTPs, and 3.3 mmol/liter MgCl2. The PCR conditions were 95°C for 10 minutes, then stepdown 4 cycles of 15 s at 94°C, 30 s at 63°C -2,5°C per cycle, and 30 s at 72°C, 11 cycles of 15 s at 94°C, 30 s at 55°C, and 1 min at 72°C, at last 22 cycles of 15 s at 89°C, 30 s at 55°C, and 1 min at 72°C. The PCR products were supplemented with an internal size standard, and the fragments were separated and detected on an

Applied Biosystems model 3730 sequencer, using Genescan (v. 3.0) peak-calling software. Alleles were called using an internal allele-calling program [9].

References:

1. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
2. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
3. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* **40**, 1068-74 (2008).
4. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868-74 (2009).
5. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).
6. Lander, E.S. & Green, P. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A* **84**, 2363-7 (1987).
7. Kruglyak, L. & Lander, E.S. Faster multipoint linkage analysis using Fourier transforms. *J Comput Biol* **5**, 1-7 (1998).
8. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-103 (2010).
9. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat Genet* **31**, 241-7 (2002).

Supplementary table 1: Association between SNPs at chr17q23.2 and ovarian cancer

SNP	P adj	OR	chr	Locaction	Risk allele
rs34289250	5.7×10^{-13}	7.95	chr17	57,235,428	C
rs12938171	2.1×10^{-12}	6.59	chr17	57,335,137	A
chr17:55422245	5.6×10^{-9}	10.60	chr17	55,422,245	A
chr17:55217320	5.7×10^{-9}	10.61	chr17	55,217,320	C
rs12451939	7.1×10^{-9}	0.24	chr17	57,402,424	A
chr17:56567990	7.2×10^{-9}	10.23	chr17	56,567,990	G
chr17:56478611	7.7×10^{-9}	10.19	chr17	56,478,611	A
chr17:56505864	7.7×10^{-9}	10.19	chr17	56,505,864	C
rs12937080	8.3×10^{-9}	0.24	chr17	57,284,519	A

SNP genotypes of 68 chip-genotyped and 572 *in silico*-genotyped ovarian cancer cases were compared to genotypes of more than 40,000 controls. Shown are the SNP annotations, *P* values, Odds ratios (OR), chromosome location and the risk allele. The *P*-values have been adjusted using a genomic control correction factor.

Supplementary Table 2. Association of the BRIP mutation, c.2040_2041insTT, with 23 cancers or having any cancer in the Icelandic population.

	Directly genotyped					Imputed				Combined		
Phenotype	N	Control frequency	Case frequency	OR	P	N	Case frequency[a]	OR	P	OR (95% CI)	P	P_{het}
Bladder Cancer	754	0.41%	0.13%	0.32	0.079	720	0.40%	0.98	0.97	0.66 (0.31, 1.40)	0.28	0.17
Brain Cancer Glioma	121	0.41%	0.00%	0.00	0.18	213	0.42%	1.03	0.98	0.63 (0.07, 5.62)	0.68	0.20
Brain Cancer Meningioma	139	0.41%	0.72%	1.76	0.48	62	0.42%	1.02	0.99	1.57 (0.39, 6.37)	0.52	0.76
Breast Cancer	2,740	0.39%	0.62%	1.60	0.097	1,751	0.37%	0.94	0.86	1.28 (0.84, 1.96)	0.25	0.23
Cervix Uteri Cancer	232	0.40%	0.43%	1.08	0.92	433	0.09%	0.23	0.14	0.61 (0.18, 2.01)	0.42	0.22
Chronic Lymphocytic Leukemia	113	0.41%	0.44%	1.08	0.94	184	0.61%	1.48	0.71	1.28 (0.29, 5.66)	0.75	0.84
Colon Cancer	291	0.40%	1.20%	3.03	0.031	1,895	0.34%	0.85	0.63	1.25 (0.72, 2.16)	0.43	0.038
Endometrial Cancer	428	0.41%	0.23%	0.56	0.41	342	0.99%	2.40	0.13	1.33 (0.55, 3.19)	0.53	0.11
Esophagus Cancer	103	0.41%	0.97%	2.38	0.32	351	0.46%	1.12	0.88	1.57 (0.50, 4.91)	0.43	0.52
Gastric Cancer	354	0.41%	0.99%	2.43	0.083	1,840	0.51%	1.24	0.54	1.54 (0.87, 2.70)	0.14	0.28
Kidney Cancer	519	0.41%	0.77%	1.88	0.16	746	0.18%	0.44	0.23	1.20 (0.57, 2.49)	0.63	0.074
Liver Cancer	70	0.41%	0.00%	0.00	0.31	255	0.56%	1.37	0.72	1.15 (0.21, 6.22)	0.87	0.29
Lung Cancer	273	0.40%	0.55%	1.37	0.65	2,869	0.56%	1.40	0.19	1.39 (0.87, 2.23)	0.16	0.98
Lymphoma Hodgkin	79	0.41%	1.90%	4.71	0.041	128	0.49%	1.19	0.89	3.30 (0.92, 11.85)	0.067	0.36
Lymphoma Non Hodgkin	259	0.41%	1.16%	2.85	0.046	444	0.22%	0.54	0.45	1.74 (0.74, 4.13)	0.21	0.086
Multiple Myeloma	115	0.41%	0.43%	1.06	0.96	252	0.98%	2.40	0.21	1.89 (0.60, 5.98)	0.28	0.53
Ovarian Cancer	318	0.41%	2.36%	5.86	3.7e-06	338	4.56%	11.60	7.1e-10	8.13 (4.74, 13.95)	2.8e-14	0.22
Pancreatic Cancer	161	0.41%	0.93%	2.29	0.26	736	1.17%	2.87	0.014	2.71 (1.31, 5.58)	0.0069	0.79
Prostate Cancer	2,244	0.40%	0.45%	1.12	0.73	1,978	0.31%	0.78	0.46	0.94 (0.59, 1.49)	0.80	0.44
Rectal Cancer	77	0.40%	1.30%	3.29	0.18	736	0.83%	2.09	0.062	2.25 (1.11, 4.58)	0.025	0.64
Testicular Cancer	196	0.41%	0.51%	1.25	0.77	65	0.51%	1.24	0.90	1.25 (0.31, 4.93)	0.75	1.0
Thyroid Cancer	550	0.41%	0.36%	0.88	0.82	407	0.09%	0.22	0.16	0.66 (0.25, 1.74)	0.40	0.26
Upper Airway Cancer	35	0.41%	1.43%	3.52	0.33	300	0.01%	0.02	0.044	0.76 (0.09, 6.40)	0.80	0.026
All Cancers*	4,457	0.32%	0.56%	1.78	0.048	26,541	0.46%	1.47	5.6e-05	1.50 (1.25, 1.79)	8.9e-06	0.53

For the individual cancers, directly typed cases are compared to 3,913 directly typed controls, in each case excluding the known cases of the cancer being tested from the controls (Ns, control and case frequencies, ORs and P values given). The remaining cancer cases that have been imputed (either chip typed or in silico genotyped) are compared to over 40,000 controls (Ns, estimated case frequencies, ORs and P values given) and then combined using the Mantel-Haenszel model (OR with 95% CI and P value given). A P-value for testing for the difference between the ORs in the two groups is given in the Phet column. *Due to the relatively small number of directly genotyped controls we removed the chip typed cases from the list of directly genotyped cases when testing for association with any cancer type, thus including all chip typed cases in the second group. The P-values and 95%CIs have been adjusted using a genomic control correction factor for testing each phenotype based on the chip typed and in silico genotype data. [a]The allele frequency in the imputed cases is estimated from the directly genotyped control frequency and the OR for the imputed individuals.

Supplementary Table 3. Association of the BRIP mutation, c.2040_2041insTT, with 22 cancers or having any cancer in the Icelandic population after removal of all ovarian cancer cases.

	Directly genotyped					Imputed				Combined		
Phenotype	N	Control frequency	Case frequency	OR	P	N	Case frequency[a]	OR	P	OR (95% CI)	P	P_{het}
Bladder Cancer	752	0.41%	0.13%	0.32	0.079	720	0.41%	1.00	1.0	0.66 (0.31, 1.42)	0.29	0.16
Brain Cancer Glioma	121	0.41%	0.00%	0.00	0.18	213	0.43%	1.05	0.96	0.74 (0.12, 4.65)	0.75	0.19
Brain Cancer Meningioma	136	0.41%	0.74%	1.80	0.47	62	0.43%	1.04	0.98	1.61 (0.40, 6.53)	0.51	0.76
Breast Cancer	2,712	0.39%	0.61%	1.57	0.12	1,738	0.38%	0.97	0.93	1.28 (0.84, 1.96)	0.26	0.28
Cervix Uteri Cancer	230	0.40%	0.43%	1.09	0.91	428	0.10%	0.24	0.15	0.62 (0.19, 2.06)	0.44	0.23
Chronic Lymphocytic Leukemia	112	0.41%	0.45%	1.09	0.94	183	0.63%	1.54	0.68	1.30 (0.29, 5.79)	0.73	0.82
Colon Cancer	286	0.40%	1.22%	3.07	0.029	1,891	0.35%	0.87	0.68	1.27 (0.73, 2.22)	0.39	0.040
Endometrial Cancer	421	0.41%	0.24%	0.57	0.43	334	0.92%	2.23	0.19	1.24 (0.50, 3.05)	0.64	0.14
Esophagus Cancer	103	0.41%	0.97%	2.38	0.32	351	0.47%	1.15	0.86	1.59 (0.51, 4.98)	0.42	0.53
Gastric Cancer	352	0.41%	0.99%	2.43	0.082	1,836	0.53%	1.28	0.48	1.58 (0.89, 2.78)	0.12	0.30
Kidney Cancer	518	0.41%	0.77%	1.87	0.16	744	0.19%	0.45	0.24	1.21 (0.58, 2.52)	0.61	0.081
Liver Cancer	69	0.41%	0.00%	0.00	0.32	255	0.57%	1.39	0.70	1.18 (0.23, 6.14)	0.84	0.29
Lung Cancer	267	0.40%	0.56%	1.40	0.63	2,864	0.54%	1.36	0.24	1.36 (0.85, 2.19)	0.20	0.97
Lymphoma Hodgkin	79	0.41%	1.90%	4.69	0.041	128	0.50%	1.21	0.88	3.32 (0.92, 11.94)	0.066	0.37
Lymphoma Non Hodgkin	258	0.41%	1.16%	2.85	0.046	442	0.23%	0.56	0.48	1.77 (0.74, 4.19)	0.20	0.093
Multiple Myeloma	113	0.41%	0.44%	1.08	0.95	252	1.00%	2.46	0.20	1.93 (0.61, 6.12)	0.26	0.52
Pancreatic Cancer	158	0.41%	0.63%	1.54	0.61	734	1.20%	2.95	0.012	2.58 (1.22, 5.46)	0.013	0.49
Prostate Cancer	2,244	0.40%	0.45%	1.12	0.73	1,978	0.32%	0.80	0.51	0.95 (0.60, 1.51)	0.83	0.48
Rectal Cancer	75	0.40%	1.33%	3.38	0.17	734	0.86%	2.15	0.053	2.32 (1.14, 4.73)	0.020	0.65
Testicular Cancer	196	0.41%	0.51%	1.24	0.78	65	0.52%	1.26	0.90	1.25 (0.32, 4.93)	0.75	0.99
Thyroid Cancer	544	0.41%	0.37%	0.89	0.83	404	0.09%	0.23	0.17	0.67 (0.25, 1.77)	0.42	0.27
Upper Airway Cancer	33	0.41%	0.00%	0.00	0.49	300	0.01%	0.02	0.046	0.02 (0.00, 0.76)	0.035	0.92
All Cancers*	4,205	0.32%	0.48%	1.51	0.18	26,137	0.43%	1.35	0.0023	1.36 (1.14, 1.64)	0.00092	0.73

Table legend is the same as for Supplementary Table 2 except ovarian cancer cases have been removed from both cases and controls.

Supplementary table 4. Association of the *BRIP* mutation, c.1702_1703del, with cancer in a Spanish population.

Phenotype	N individuals		N alleles	Allele freq.	OR	P
	Carriers	Non-carriers	Mut/Wt			
Controls	1	1,779	1/3,559	0.03%	1.00	
Ovarian cancer	2	142	2/286	0.70%	25.00	0.016
Breast cancer	6	921	6/1,848	0.32%	11.17	0.0079
Lung cancer	1	514	1/1,029	0.10%	3.34	0.40
Colorectal cancer	0	497	0/994	0.00%	0.00	
Endometrial cancer	0	130	0/260	0.00%	0.00	
Bladder cancer	0	238	0/476	0.00%	0.00	
Prostate cancer	0	699	0/1,398	0.00%	0.00	
Basal cell carcinoma	0	222	0/444	0.00%	0.00	
Cutaneous melanoma	0	278	0/556	0.00%	0.00	
Thyroid cancer	0	90	0/180	0.00%	0.00	
Kidney cancer	0	89	0/178	0.00%	0.00	

Supplementary table 5. Allele counts of coding variants in *BRIP1* in ovarian cancer cases and controls from Iceland, The Netherlands and Spain

Variant	Iceland # alleles				The Netherlands # alleles				Spain # alleles			
	Cases		Controls		Cases		Controls		Cases		Controls	
	Variant	Wild type	Variant	Wild type	Variant	Wild type	Variant	Wild type	Variant	Wild type	Variant	Wild type
P47A	0	370	0	482	1	533	2	1,188	0	284	*	*
R173C	2	434	23	2,713	1	549	7	1,439	0	288	*	*
V193I	4	524	13	2,721	4	546	7	1,437	0	288	*	*
L195P	0	436	2	2,554	1	549	1	1,445	0	286	*	*
K209R	0	436	0	2,554	1	549	0	1,446	0	284	*	*
C214S	0	436	0	2,538	1	549	0	1,366	0	274	0	848
E262G	0	438	0	2,542	0	552	0	1,394	0	272	0	858
M299I	0	438	0	2,544	0	552	0	1,376	0	274	0	856
K297R	1	437	0	2,550	1	551	3	1,373	0	274	0	856
R419Q	0	378	0	496	1	547	0	1,280	0	288	*	*
Q741H	1	435	3	2,181	0	540	*	*	0	286	5	813
D745D	0	436	1	2,177	0	540	*	*	1	285	1	821
R798X	1	435	7	3,051	0	544	*	*	0	284	*	*
Q944E	0	378	0	2,854	0	538	*	*	0	282	*	*
P1034L	0	436	0	898	0	540	*	*	0	282	*	*

Variants were genotyped by Sanger sequencing of the relevant exons. Shown are the number of mutant and wild type alleles among cases and controls in each population. *Indicates that genotyping was not done in control groups since no variant alleles were found in the cases.

Supplementary table 6. Whole-genome sequencing. Number of mutant and wild-type sequence reads in the area of c.2040_2041insTT

	Germ-line DNA			Tumor DNA		
	Coverage	# WT alleles	# Mut alleles	Coverage	# WT alleles	# Mut alleles
Case 6	44	28	29	23	5	23
Case 11	15	16	14	27	7	37

Results from whole genome sequencing of germ-line DNA and tumor DNA from two heterozygous carriers of c.2040_2041insTT. Shown is the sequence coverage for each of the samples, the number of sequence reads for each allele, wild-type (WT)and mutant (Mut). Pathology review of the tumor samples estimated the tumor cell percentage to be 50-60% for case 6 and 60-70% for case 11

Supplementary table 7. Primers used for long-range PCR and pool sequencing*

Coordinates (Build36/Hg18)	Size (bp)	Forward	Reverse
chr17:57,110,656-57,116,609	5954	GACCTGGTCCAAGAGGGTTTAGT	GTTTACAGCCTGTTCCCTCCTTT
chr17:57,116,402-57,122,212	5811	CCCCCAACGAATACTAGTGGATT	CAAACATGTCATCCCACTTCCTT
chr17:57,121,590-57,127,481	5892	TGACCAATATACACATTCTGGGAGT	AATGGGTAGAATTGAGGGTCCAT
chr17:57,126,905-57,132,913	6009	TCCCTAACAACCAATGATGTTGA	AGTCCCAGAAGGAGAGGAGAGAA
chr17:57,132,253-57,138,534	6282	GTTCCAAGTGAATTGCTGAACCT	AATAGTTTGGGCGTTTCAGTCCT
chr17:57,138,054-57,144,088	6035	TCAGAGAACATACTGGACTTGAGTTG	CCCTAGAACATGTCACCCACTGT
chr17:57,143,634-57,147,986	4353	AAAGGGGAGCAGGATCTCTGTAG	CTGGAACCATAGGAACTGGTATGT
chr17:57,147,964-57,150,524	2561	CATACCAGTTCCTATGGTTCCAGTT	TGGGGTACTTGAAACAATCATCA
chr17:57,150,403-57,154,644	4242	CTTGGAACCAGTTCATTCTTTGG	TTCCCAGCTGGACAGAATTTTA
chr17:57,152,646-57,158,595	5950	AACTGCGGTCAACTGTGGTC	TCCACAGCCTTATGGCTTATTTC
chr17:57,158,170-57,164,486	6317	CTCTAGAACATGCCACCATTCGT	GAATGGCCTCATCAGAGGTTAGA
chr17:57,163,475-57,169,019	5545	AGCTATATCCAGGTGGGTCTCTG	CACCTGCAGTTACCTTTACCAGTG
chr17:57,168,595-57,175,493	6899	TGAGGATTGAGCATTCCCTTCTA	AATTGGTTTGCTACCCTTTTGC
chr17:57,174,646-57,180,645	6000	GGAAGCATGCAACACTTCCTT	CCAAGTCCTTTGTGGTTGCATA
chr17:57,180,167-57,186,041	5875	CACTATGCACAGTCCTCTATGAAACA	GGCTCATGCATATAATCCCTTTG
chr17:57,185,218-57,191,364	6147	GGGAGTGGGACTAGTGGCTTTAT	CTCACTTTCAGCCTCTGTGCAT
chr17:57,190,590-57,196,637	6048	CAAGTATTGCCATCTGAATCTGCT	TTGATGCAGTGGTAGTCTCCAAA
chr17:57,196,289-57,201,867	5579	TGCGCTGGCCTTACTCTAATAAA	TGTGACCTCAGTTCTCAGATGGA
chr17:57,200,610-57,207,075	6466	AGTCTCCATCCATTCGTGTGTTT	CAAAAAGGGAGATAGGAGATGTCTG
chr17:57,205,753-57,211,377	5625	CAGAGTGAACAGCCTACAGAACATT	TACATCCACGTTCTTCCTCTGTG
chr17:57,210,738-57,217,228	6491	AAATGGACTTTTGTGCAGTAGGC	GGCACCAAATAATACCCTGATTG
chr17:57,215,764-57,222,859	7096	AGGATGAGAGGGATTGGTTGTTT	AGTGAACCATGATTGCACCACTA
chr17:57,221,958-57,226,182	4225	TGGGCATAGAAGAAGGCATAAGTA	ACCTAAGACCTGGCTCTCAGTTG
chr17:57,225,562-57,231,734	6173	CACAATTTACCCATGCCATCAC	TCTGTGTAGCCTGTGGTACCTTG
chr17:57,231,067-57,237,394	6328	GTGAACAAGACAAAAGGTTGGACT	TGTCTTGTCCTTTGATTTGCTGA
chr17:57,236,297-57,242,327	6031	TGGAGGATCTAGAAAAAGCCTCTC	GGCTATTTAATGTGGCAGCACTC
chr17:57,241,864-57,248,431	6568	GTCATTGATGGGGAGGGAAA	TACAGCGATGGGATGAAGAAAGT
chr17:57,248,034-57,254,961	6928	GCTGTTGTCCAAGCCATCTAAAC	CTGAAAACTTGTTGGTCCCCATA
chr17:57,254,346-57,260,597	6252	TGACGAATAGACTGATAATGGGATTC	TTCCAAACTGAAACACCATACCC
chr17:57,259,642-57,265,631	5990	AGGAGGATCACTTGAGACCAGAA	CTTGTTGGCCAGGCATGTT
chr17:57,265,318-57,271,242	5925	CATGTCGGATGGGTTGCTATAC	TTTTGGGTGTCATAGGGGAGTAA
chr17:57,270,274-57,276,414	6141	TGCAGGGATGAGAATTAAACCAT	TTTGGACACATACTGCCTGATCTT
chr17:57,275,785-57,282,226	6442	CCATGTGACAAGTTTGCTCTCTCT	AATTAATTCAAGCCATGGGAGTG
chr17:57,281,458-57,287,247	5790	AGAAGAAAACTGGAACCAGGATGT	CAACCAACCACCTGTCTGATTTT
chr17:57,286,597-57,292,297	5701	CCTAACACTTTGTGAGGCACAGAT	GCATAATTCTGGCACAGATTCC
chr17:57,291,847-57,297,854	6008	ACAGCATGGCTGAACCAGTC	TTTTAAACTGAGGCCTTGCGATA
chr17:57,297,436-57,302,081	4646	CAGACCTCGAATACAGTCCCTTG	GGAGAAAGCAAATGGTGTCAATC

*All primer sequences are listed in the direction 5' -> 3'

Supplementary table 8. Primers used for typing indels in Iceland and Spain*

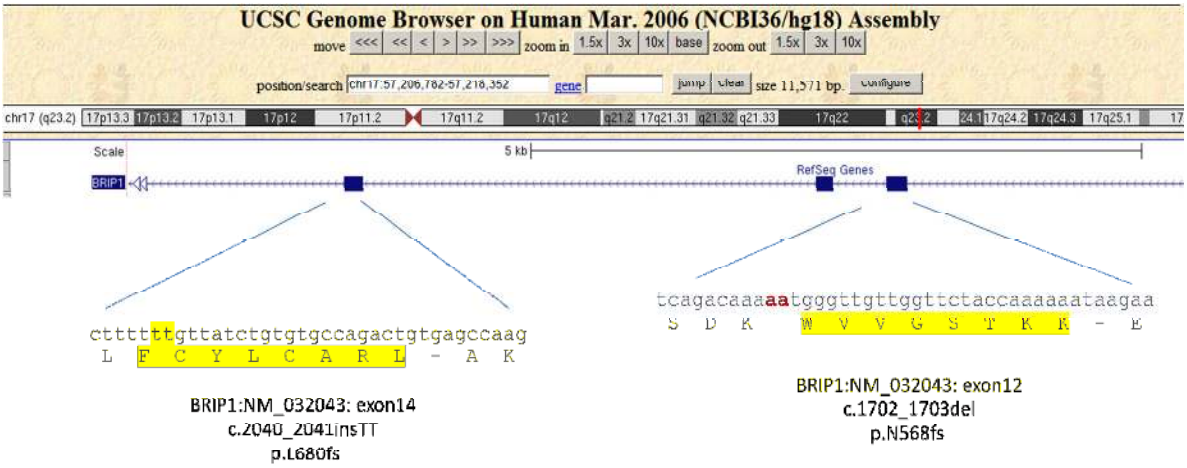
Icelandic insert, c.2040_2041insTT, exon 14		
Primers for microsatelliete genotyping	Forward	GGCAAGAAACACAAAATTCCTTG
	Reverse	CCATGCCTTTTTCAGGCATAAC
Primers for PCR and Sanger sequencing of genomic DNA from blood	Forward	GTTGCCTCTACCCTAGGAAGC
	Reverse	GAGAATTTAGTGATCTGCCTAAAGC
Primers for PCR and Sanger sequencing of genomic DNA from paraffin	Forward	TGTTGCCTCTACCCTAGGAAGC
	Reverse	TAGGTTTGGGTTGGTACCATTG
Primers for PCR and Sanger sequencing of cDNA from tumors	Forward	TGGCAAAGTTCAGACCATTGTT
	Reverse	TGACTGTCTTCACCAACTCCAG
Spanish deletion, c.1702_1703del, exon 12		
Primers for microsatelliete genotyping	Forward	TGCTGGTACTGAGCAAGAAGA
	Reverse	TGCGATTCAACAGACTTACTCC
Primers for Sanger sequencing	Forward	AGCTAAGGTGGGAGGATCG
	Reverse	CTGGTACCTTCCCATTCAAAA

Supplementary table 9. Primers used for Sanger sequencing rare variants*

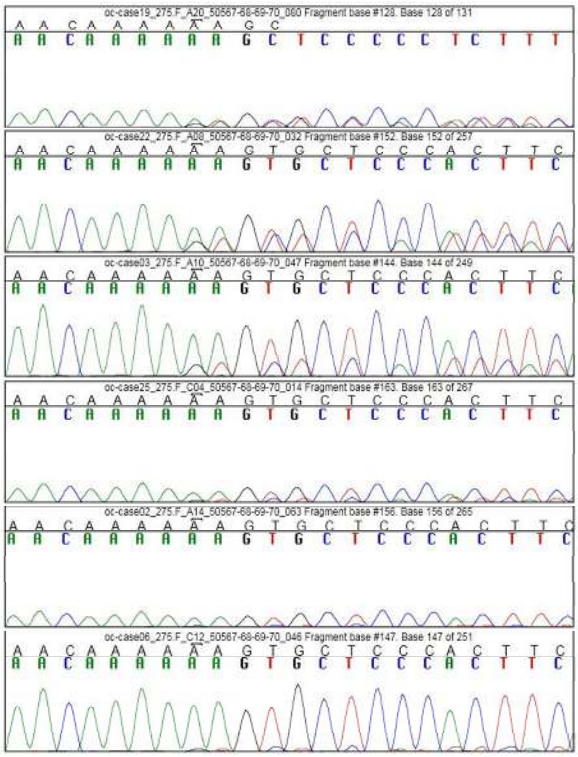
Variant	Exon		Primer sequence
P47A	3	Forward	AAACAAATATTTAAGTTAGCGACAGC
		Reverse	CATAATTCTGGCACAGATTCC
R173C, V193I, L195P, K209R	6	Forward	AGAGAGAATAGGCTTTGCCATC
		Reverse	TTGGCTAAGATTCATGAGTCACTACT
M299I, C214S, E262G, K297R	7	Forward	GCAGTTAATTTGATTTTCCGAAG
		Reverse	ATTCATTACTTACAGGTTCTGATTCC
R419Q	9	Forward	TGTGAACAAGACAAAAGGTTGG
		Reverse	GATATTTGGTTGGCCATAGTGC
Q741H, D745D	15	Forward	CCAATTTATTTTCTTTTCACTCAGG
		Reverse	CCAGCACCATTCAGATCTCTC
R798STOP	17	Forward	CATTGAATACATACCAGTTCCTATGG
		Reverse	GAAAATGTTCTATATTATGATGGTGGTTAC
Q944E	19	Forward	TTGCTCATGATCACAAAGCTAGT
		Reverse	GAACCAATATACAGGACAATGAGTCTA
P1034L	20	Forward	TCTAGTAAGGGTGGCATCAATC
		Reverse	CAAGATGAAAATCCACTAGTATTCG

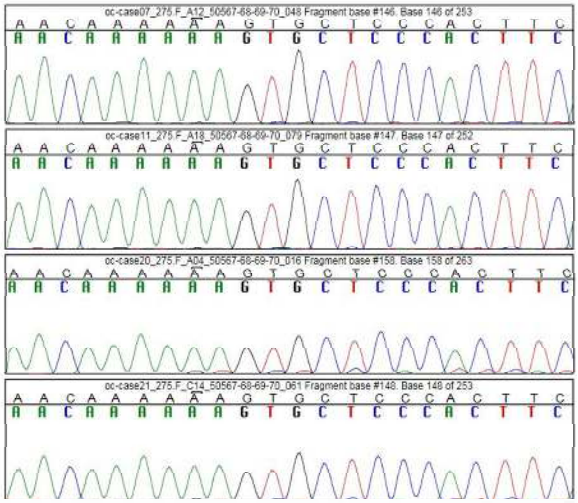
*All primer sequences are listed in the direction 5' -> 3'

Supplementary figure 1. Schematic representation of the location and consequences of the two indels found in exon 12 (Spain) and exon 14 (Iceland) of *BRIP1*. The mRNA is transcribed of the minus strand.



Supplementary figure 2. LOH in tumor samples from carriers of c.2040_2041insTT. Sequence of the region around c.2040_2041insTT in 10 tumor samples from heterozygous carriers of the insert. The tumors are ordered with the tumor showing no LOH first, followed by one tumor with partial LOH and 8 tumors with significant or complete loss of the wild-type allele.





Supplementary figure 3. Sequence traces of the region over c.2040_2041insTT in a heterozygous carrier of the insert showing loss of the wild-type allele and mRNA expression in the tumor. Top panel: Sequence of germline DNA from blood, Center panel: Sequence of tumor DNA, Bottom panel: Sequence of cDNA from tumor.