OPEN ACCESS

Short report

CRISPR–Cas9/long-read sequencing approach to identify cryptic mutations in *BRCA1* and other tumour suppressor genes

Tom Walsh ,[1] Silvia Casadei ,[1] Katherine M Munson ,[2] Mary Eng,[1] Jessica B Mandell,[1] Suleyman Gulsuner ,[1] Mary-Claire King [1]

Abstract Current clinical approaches for mutation discovery are based on short sequence reads (100–300 bp) of exons and flanking splice sites targeted by multigene panels or whole exomes. Short-read sequencing is highly accurate for detection of single nucleotide variants, small indels and simple copy number differences but is of limited use for identifying complex insertions and deletions and other structural rearrangements. We used CRISPR-Cas9 to excise complete *BRCA1* and *BRCA2* genomic regions from lymphoblast cells of patients with breast cancer, then sequenced these regions with long reads (>10 000 bp) to fully characterise all non-coding regions for structural variation. In a family severely affected with early-onset bilateral breast cancer and with negative (normal) results by gene panel and exome sequencing, we identified an intronic SINE-VNTR-Alu retrotransposon insertion that led to the creation of a pseudoexon in the *BRCA1* message and introduced a premature truncation. This combination of CRISPR–Cas9 excision and long-read sequencing reveals a class of complex, damaging and otherwise cryptic mutations that may be particularly frequent in tumour suppressor genes replete with intronic repeats.

For numbered affiliations see end of article.

Correspondence to
Professor Mary-Claire King, Departments of Medicine and Genome Sciences, University of Washington, Seattle, Washington, USA; mcking@uw.edu

Received 4 July 2020
Revised 24 August 2020
Accepted 27 August 2020
Published Online First 15 October 2020

Check for updates

To cite: Walsh T, Casadei S, Munson KM, *et al*. *J Med Genet* 2021;**58**:850–852.

INTRODUCTION

Multigene panel sequencing for inherited cancer risk was widely adopted in 2013 after the US Supreme Court invalidated the patenting of the genomic DNA sequences of *BRCA1* and *BRCA2*.[1] Panel sequencing has increased the diagnostic yield of pathogenic variants and decreased the cost of genetic testing for patients. The approach relies on DNA capture of exons and flanking intronic splice sites and highly accurate sequencing with short reads (100–300 bp).[2] However, this technology does not efficiently detect complex structural rearrangements such as inversions and mobile element insertions.[3] The *BRCA1* genomic region is particularly challenging for short-read sequencing. It is composed of 42% Alu repeats,[4] the second highest proportion in the genome, and a 30 kb tandem segmental duplication spanning its promoter and first two exons.[5] As a consequence of this unstable genomic structure, simple structural variants (deletions and duplications) of *BRCA1* exons represent more than 10% of all germline *BRCA1* mutations.[6] However, the frequency and nature of complex structural variants within introns and other non-coding regions of *BRCA1* is not yet known.

METHODS

For families severely affected with breast cancer, we applied sequencing of long DNA reads (>10 000 bp) to evaluate complete *BRCA1* and *BRCA2* genomic loci, including exons, introns, promoters and regulatory regions. Participants with DNA sequenced by this approach were the probands of 19 families with at least four relatives with young-onset breast cancer, all with negative (normal) sequence based on gene panel and whole exome sequencing. All participants provided informed consent (UW protocol 1583). For each participant, freshly grown lymphoblasts were loaded onto a high molecular weight library system (Sage Science, Beverly, Massachusetts, USA) and lysed directly in agarose gels. Pairs of CRISPR guides, designed to excise 200 kb genomic loci including *BRCA1*[7] (chr17:41,170,535–41,368,879) and *BRCA2* (chr13:32,836,996–33,026,430) were added to the gels along with Cas9 enzyme. Cut fragments were separated by field gel inversion electrophoresis (FIGE), an approach developed in the 1980s for separation of very large DNA fragments by reversing the polarity of the electrophoretic field periodically at pulse times in hundreds to thousands of milliseconds. 'Next generation' FIGE is automated to modify duration of pulse times systematically over course of the run. Separated fragments were eluted and evaluated for *BRCA1* and *BRCA2* enrichment by TaqMan qPCR. *BRCA1* and *BRCA2* fragments, which were ~200 kb in size, were sheared to ~20–30 kb by two passages through a gTUBE (Covaris, Woburn, Massachusetts, USA). Fragments were then end repaired, A-tailed and ligated to SMRTbell adapters using the Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, California, USA) following the manufacturers recommendations for low DNA input. Libraries were then sequenced on a Sequel I (Pacific Biosciences) with average read length of 9700 bp. Reads were aligned to *BRCA1* and *BRCA2* and evaluated using PALMER[8] for structural variants (deletions, duplications, insertions, inversions and translocations) >50 bp in size.

RESULTS

In genomic DNA of one of the 19 probands, we identified an intronic insertion event that was not

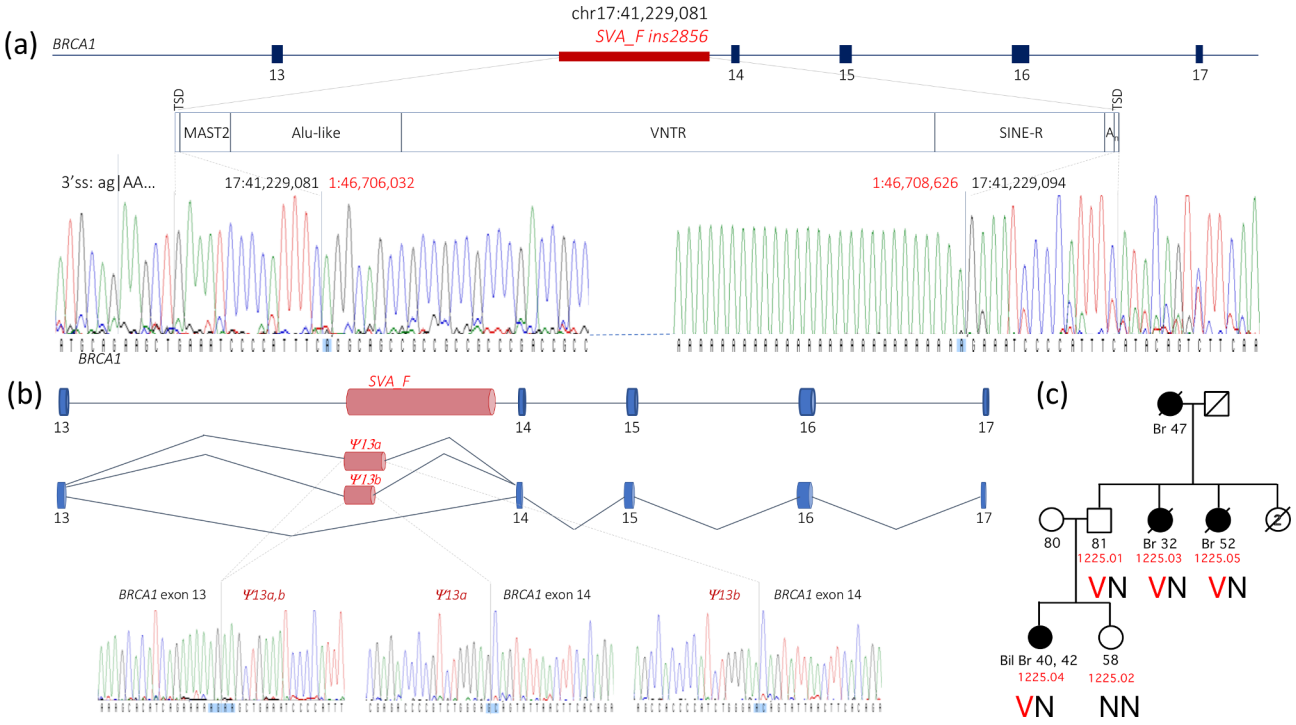

Figure 1 **(a) SVA retrotransposon insertion in BRCA1 intron 13**. In family CF1225, a 2856bp SVA retrotransposon is inserted at chr17:41,229,081. The retrosponson is flanked on 5′ and 3′ ends by a 14bp palindromic target site duplication (TSD) GAAATGGGGATTTC, produced by nuclease cleavage at the insertion site. The SVA insertion is 98.6% identical to sequence at chr1:46,706,032-46,708,626. From 5′ to 3′, the DNA elements of the SVA_F composite transposon are: (i) sequence sharing identity with MAST2 exon 1, acquired through splicing (nt 1 – 150), (ii) a domain of two antisense Alu fragments (nt 154 - 674), (iii) a GC-rich variable number tandem repeats (VNTR) (nt 675 - 2295), (iv) a SINE-R domain with sequence homology to the 3′ end of the HERV-K10 env gene and right portion of an LTR (U3, R, polyA signal), terminating with a polyA tail (An) (nt 2296 - 2842), and (v) the target-site duplication (TSD). Sequence elements of the SVA_F transposon were annotated using BLAT queries against the reference genome (GRCh37/hg19) and BLAST alignments between individual SVA regions and degenerate repeats (Alu, SINE-R, VNTR) or the reference HERV-K10 viral genome sequence. **(b) Transcriptional consequences of the SVA retrotransposon insertion.** RT-PCR across BRCA1 exons 12-15 yielded the expected size product and two larger transcripts. Sanger sequencing of the transcripts indicates that two cryptic splice donor sites within the 5′ Alu-like domain of the SVA element exploit a cryptic splice acceptor in BRCA1intron 13, resulting in exonification of segments of 509bp and 666bp in the BRCA1 message and a premature stop at codon 1558. **(c) Family 1225.** All members of the family with breast cancer had negative (normal) results from comprehensive panel testing and subsequent whole exome sequencing. Black symbols indicate patients with breast cancer (Br). Ages are age at diagnosis for cancer patients and current age for living relatives. The proband was diagnosed with bilateral breast cancer (Bil Br) at ages 40 and 42. The red 'V' indicates the BRCA1 intron 13 SVA insertion, the black 'N' indicates normal sequence at intron 13.

present in the Database of Genomic Variants,[9] or in the gnomAD V.2.1 structural variant call set[10] or in a diverse group of individuals whose whole genomes were sequenced to high depth with long reads.[11] The proband of family CF1225 harboured a 2856 bp SVA_F (SINE+VNTR+Alu) retrotransposon at chr17:41,229,081 in intron 13 of *BRCA1* (GRCh37/hg19 assembly). This participant, American of Romanian ancestry, was diagnosed with bilateral breast cancer at ages 40 and 42 years. PCR and Sanger sequencing confirmed the SVA insertion location, flanked on both ends by a palindromic 14 bp target site duplication (figure 1A). The SVA insertion shared 98.6% identity to sequence at chr1:46,706,032–46,708,626. Multiple long reads included all elements of the mutation and of wild-type flanking *BRCA1* intronic sequence, so that the mutation's position and the sequence were clear.

In order to determine if the intronic SVA insertion altered *BRCA1* transcription, we grew lymphoblasts of CF1225.04 in puromycin to inhibit nonsense mediated decay, then evaluated cDNA of *BRCA1* by RT-PCR. Sequencing cDNA across *BRCA1* exons 12–15 yielded three transcripts: one of the expected size and two larger. Sanger sequence of these larger products revealed that the naturally occurring splice acceptor of *BRCA1* intron 13 was paired with each of two cryptic splice donor sites in the Alu portion of the SVA insertion, yielding pseudoexons of sizes 509 bp and 666 bp in the *BRCA1* message (figure 1B). Both pseudoexons included premature stop codons, predicted to truncate the BRCA1 protein at codon 1558 of the 1863 full length protein. The SVA insertion segregated with breast cancer in family CF1225 (figure 1C). All relatives and their adult children have been recontacted for genetic and clinical follow-up. We also hope to determine if this SVA retrotransposon could represent a founder allele in the Romanian population.

DISCUSSION

The genomic regions harbouring tumour-suppressor genes are replete with repeats and segmental duplications. Indeed, these features yield the tumour suppressor phenotype, in that they lead to frequent somatic mutation and complete loss of gene function among persons carrying an inherited damaging allele at the same locus. Given these genomic structures, it is possible, even likely, that complex mutations are common at tumour suppressor genes. We suggest that complex mutations have thus far been rarely encountered, because they are difficult to detect

with existing approaches. A recent whole genome sequencing study of triple negative breast tumours, with targeted analysis of mobile elements, identified an SVA insertion in *BRCA1* intron 2 in a tumour with independent loss of the wild-type *BRCA1* allele, leading to reduced expression of the *BRCA1* message.[12] Insofar as we know, the only other tumour-suppressor gene previously known to harbour an SVA insertion is *PMS2*, in a case discovered by Southern blotting.[13] The genomic approach described here, integrating CRISPR–Cas9 excision of critical loci with long-read sequencing, yields complete sequence of targeted loci and thus can detect all classes of complex non-coding structural variants. The frequency of these classes of mutations could be determined by offering this approach on a research basis to families severely affected with breast, ovarian or prostate cancer but with negative gene panel and exome sequencing results.

Author affiliations
[1]Departments of Medicine (Medical Genetics) and Genome Sciences, University of Washington, Seattle, Washington, USA
[2]Department of Genome Sciences, Unversity of Washington, Seattle, Washington, USA

Acknowledgements We would like to thank Chris Boles for technical advice and help.

Contributors All the authors contributed to generating and/or analysing the data.

Funding This project was supported by National Cancer Institute grant 5R35CA197458 and by the Breast Cancer Research Foundation.

Competing interests TW discloses consulting fees from Color Genomics outside the submitted work. M-CK is an American Cancer Society Research Professor.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Tom Walsh http://orcid.org/0000-0002-8875-0310
Silvia Casadei http://orcid.org/0000-0003-2619-5662
Katherine M Munson http://orcid.org/0000-0001-8413-6498
Suleyman Gulsuner http://orcid.org/0000-0002-3897-1238
Mary-Claire King http://orcid.org/0000-0001-9426-1743

REFERENCES

1 Easton DF, Pharoah PDP, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, Devilee P, Meindl A, Couch FJ, Southey M, Goldgar DE, Evans DGR, Chenevix-Trench G, Rahman N, Robson M, Domchek SM, Foulkes WD. Gene-panel sequencing and the prediction of breast-cancer risk. *N Engl J Med* 2015;372:2243–57.
2 Toland AE, Forman A, Couch FJ, Culver JO, Eccles DM, Foulkes WD, Hogervorst FBL, Houdayer C, Levy-Lahad E, Monteiro AN, Neuhausen SL, Plon SE, Sharan SK, Spurdle AB, Szabo C, Brody LC, BIC Steering Committee. Clinical testing of *BRCA1* and *BRCA2*: a worldwide snapshot of technological practices. *NPJ Genom Med* 2018;3.
3 Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, Peluso P, Boitano M, Chin C-S, Korlach J, Wilson RK, Eichler EE. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* 2017;27:677–85.
4 Smith TM, Lee MK, Szabo CI, Jerome N, McEuen M, Taylor M, Hood L, King MC. Complete genomic sequence and analysis of 117 kb of human DNA containing the gene BRCA1. *Genome Res* 1996;6:1029–49.
5 Jin H, Selfe J, Whitehouse C, Morris JR, Solomon E, Roberts RG. Structural evolution of the BRCA1 genomic region in primates. *Genomics* 2004;84:1071–82.
6 Walsh T, Casadei S, Coats KH, Swisher E, Stray SM, Higgins J, Roach KC, Mandell J, Lee MK, Ciernikova S, Foretova L, Soucek P, King M-C. Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. *JAMA* 2006;295:1379–88.
7 Shin G, Greer SU, Xia LC, Lee H, Zhou J, Boles TC, Ji HP. Targeted short read sequencing and assembly of re-arrangements and candidate gene loci provide megabase diplotypes. *Nucleic Acids Res* 2019;47:e115.
8 Zhou W, Emery SB, Flasch DA, Wang Y, Kwan KY, Kidd JM, Moran JV, Mills RE. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res* 2020;48:1146–63.
9 MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* 2014;42:D986–92.
10 Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, Watts NA, Solomonson M, O'Donnell-Luria A, Baumann A, Munshi R, Walker M, Whelan CW, Huang Y, Brookings T, Sharpe T, Stone MR, Valkanas E, Fu J, Tiao G, Laricchia KM, Ruano-Rubio V, Stevens C, Gupta N, Cusick C, Margolin L, Taylor KD, Lin HJ, Rich SS, Post WS, Chen Y-DI, Rotter JI, Nusbaum C, Philippakis A, Lander E, Gabriel S, Neale BM, Kathiresan S, Daly MJ, Banks E, MacArthur DG, Talkowski ME, Genome Aggregation Database Production Team, Genome Aggregation Database Consortium. A structural variation reference for medical and population genetics. *Nature* 2020;581:444–51.
11 Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, Warren WC, Magrini V, McGrath SD, Li YI, Wilson RK, Eichler EE. Characterizing the major structural variant alleles of the human genome. *Cell* 2019;176:663–75.
12 Staaf J, Glodzik D, Bosch A, Vallon-Christersson J, Reuterswärd C, Häkkinen J, Degasperi A, Amarante TD, Saal LH, Hegardt C, Stobart H, Ehinger A, Larsson C, Rydén L, Loman N, Malmberg M, Kvist A, Ehrencrona H, Davies HR, Borg Åke, Nik-Zainal S. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat Med* 2019;25:1526–33.
13 van der Klift HM, Tops CM, Hes FJ, Devilee P, Wijnen JT. Insertion of an SVA element, a nonautonomous retrotransposon, in PMS2 intron 7 as a novel cause of Lynch syndrome. *Hum Mutat* 2012;33:1051–5.