Supplemental Material for:

Long-read DNA and cDNA sequencing identifies cancer-predisposing deep intronic variation in tumor suppressor genes

Table of Contents

Supplemental Methods

Adaptive sampling sequencing. DNA samples sequenced by adaptive sampling were barcoded with combinations of the 24 adapters from SQK-NBD114.24 (Oxford Nanopore Technologies), as outlined in **Supplemental Table S1**. During adaptive sampling, 1MB regions around 10 breast and ovarian cancer genes were targeted with a FASTA file of 10 million bases defined by the GRCh38/hg38 regions in **Supplemental Table S2**. A zipped FASTA file is available upon request.

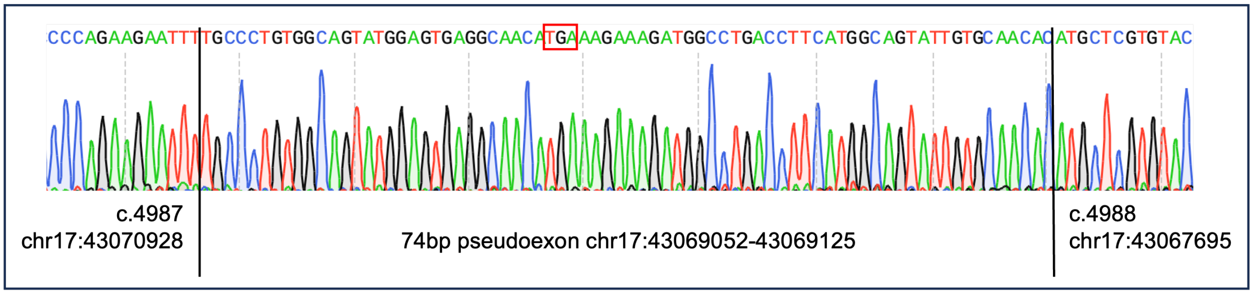
Intronic variants were defined as those >30bp from a canonical splice site. Rare variants were defined as those with ≤3 entries on gnomAD non-cancer v.3.1.2 for *BRCA1, BRCA2, PALB2, BARD1, BRIP1, RAD51C, RAD51D, TP53;* and ≤10 entries on gnomAD non-cancer v.3.1.2 for *ATM* or *CHEK2*. Rare deep intronic variants in the ten breast cancer genes in participants from severely affected families are indicated in **Supplemental Table S3**.

Gene-specific cDNA preparation. First-strand cDNA was generated with a 2uM pool of RT primers for each of the ten breast and ovarian cancer genes (**Supplementary Table S4**). Each oligo was designed to bind in the 3'UTR of its target gene, 5' of polyA signal sites and avoiding known common SNPs.

Confirmation of pseudoexon splice junctions. Long-read cDNA-derived alternate splicing events (pseudoexons) were validated by RT-PCR and Sanger sequencing. 1ug of total RNA was reverse transcribed with qScript (Quantabio) using random hexamers (Invitrogen). Following cDNA synthesis, 1ul of template was PCR-amplified with primers flanking the splice event, and PCR products were purified and Sanger sequenced as described in **Supplemental Figures S1 and S2**.

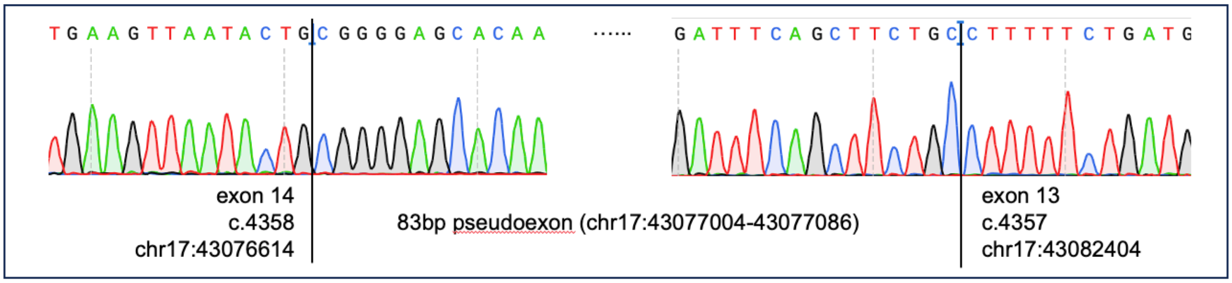
Supplementary Figure S1. Confirmation by RT-PCR and Sanger sequencing of pseudoexon splice events and junctions. All coordinates are hg38.

(**A**) *BRCA1 c.4987-1352A>G* (chr17:43,069,047 T>C) in families CF3679 and CF6196.



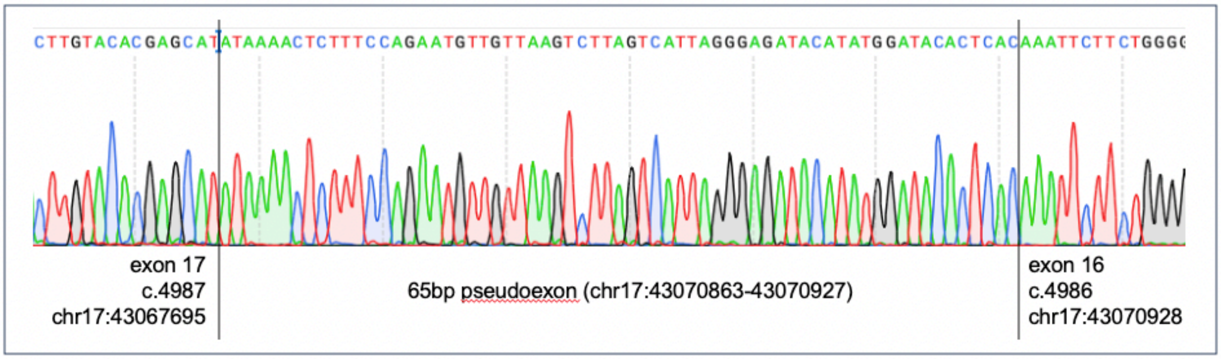
The electropherogram above illustrates the transcribed 74bp pseudoexon in *BRCA1*. Random-hexamer-primed cDNA generated from participant whole blood RNA was amplified with *BRCA1* primers: Forward, TCAGAAAAAGCAGTATTAACTTCA at c.4461–c.4484; Reverse, GGTCACCCAGAAATAGCTAAC at c.5273-c.5253 (NM_007294.4). PCR products were purified and Sanger sequenced. The pseudoexon introduces a stop at codon 1673.

(**B**) *BRCA1* c.4358-473T>G (chr17:43,077,087 A>C) in family CF4358.



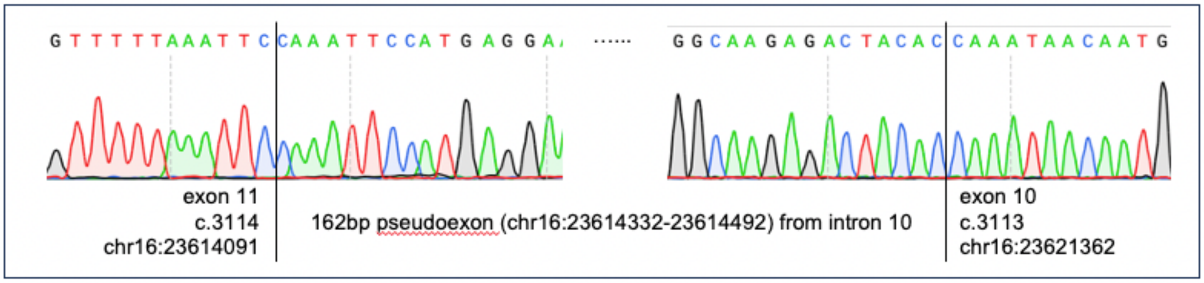
The electropherogram above illustrates the transcribed 83bp pseudoexon in *BRCA1*. Random hexamer-primed cDNA generated from participant whole blood RNA was amplified with *BRCA1* primers: Forward, GATTCAAACTTAGGTGAAGCAG at c.4197–c.4218; Reverse, TGTACACGAGCATAAATTCTTC at c.5112-c.5091 (NM_007294.4). PCR products were purified, and Sanger sequenced. The pseudoexon introduces three new codons followed by a stop after residue 1452.

(**C**) *BRCA1* c.4986+69G>A (chr17:43,070,858 C>T) in family CF4455.



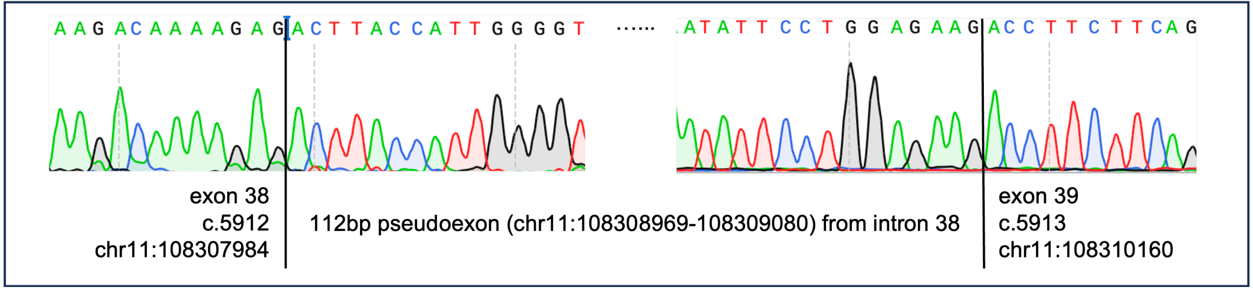
The electropherogram above illustrates the transcribed 65bp pseudoexon in *BRCA1*. Random hexamer-primed cDNA generated from participant whole blood RNA was amplified with BRCA1 primers: Forward, AAGGTCATCCCCTTCTAAAT at c.4595–c.4614; Reverse, TGTACACGAGCATAAATTCTTC at c.5198-c.5177 (NM_007294.4). PCR products were purified, and Sanger sequenced. The pseudoexon introduces 13 new codons followed by a stop after residue 1662.

(**D**) *PALB2* c.3114-239A>T (chr16:23614330 T>A) in family CF3302.



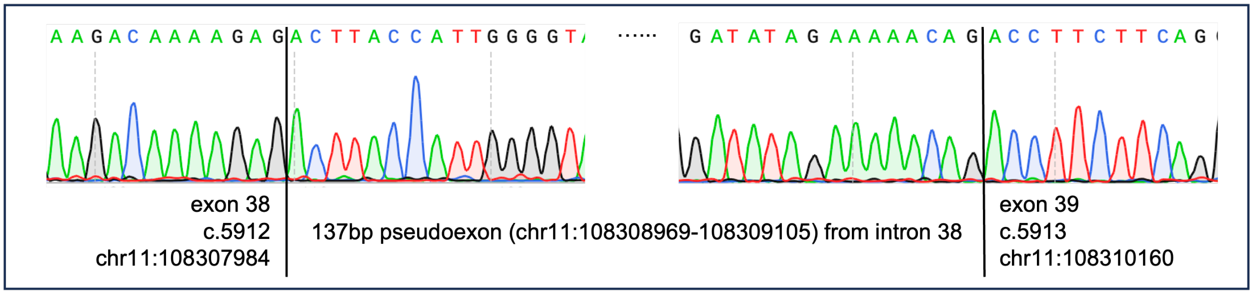
The electropherogram above illustrates the transcribed 162bp pseudoexon in *PALB2*. Random hexamer-primed cDNA generated from participant whole blood RNA was amplified with PALB2 primers: Forward, AGAGAGATCAGGGCATTGTTTT at c.2977–c.2998; Reverse, CTTCCAGGAACCTGCCAG at c.3514-c.3497 (NM_024675.4). PCR products were purified, and Sanger sequenced. The pseudoexon introduces 20 new codons followed by a stop after residue 1038

(**E**) *ATM c.5763-1080A>G* (chr11:108309080 A>G) in family CF5431.



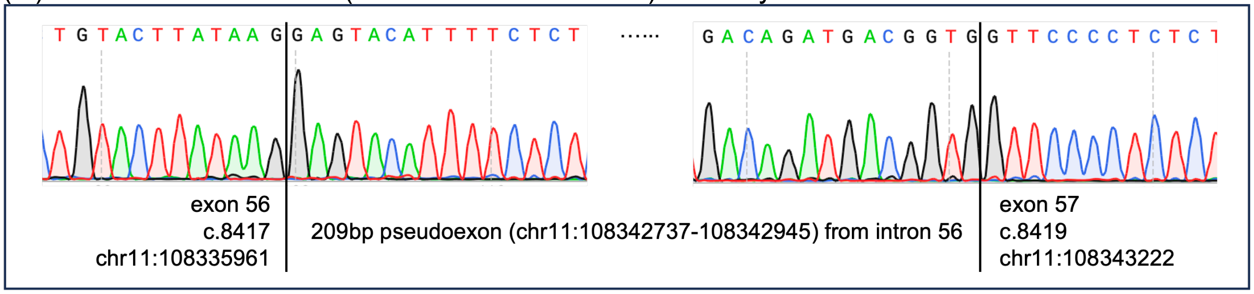
The electropherogram above illustrates the flanking splice junctions of the transcribed 112bp pseudoexon between *ATM* exons 38 and 39. Random-hexamer-primed cDNA generated from participant whole blood RNA was amplified with *ATM* primers: Forward, TAGAAGATTGTGTCAAAGTTCG at c.5318-c.5339, spanning exons 34/35; Reverse, GCAAGACTTCTTTTCTCTTGAT at c.6077-c.6056, spanning exons 39/40 (NM_000051.4). PCR products were purified and Sanger sequenced. The pseudoexon introduces a stop at codon 1929.

(**F**) *ATM c.5763-1056A>G* (chr11:108309104 G>A) in family CF6072.

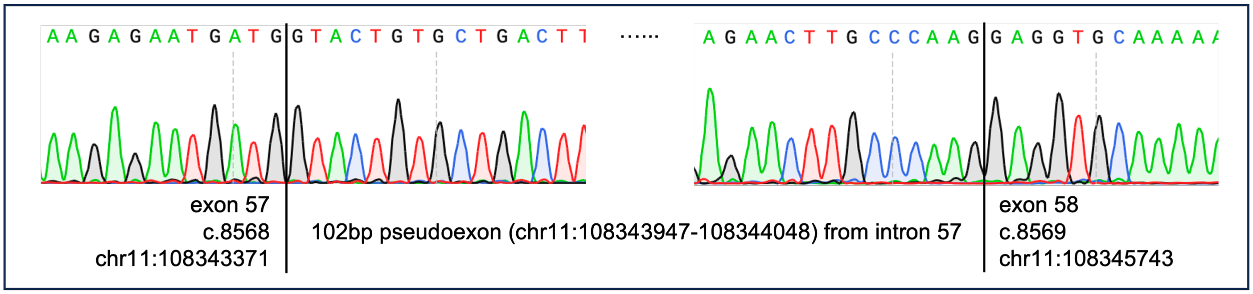


The electropherogram above illustrates the flanking splice junctions of the transcribed 137bp pseudoexon between *ATM* exons 38 and 39. Random-hexamer-primed cDNA generated from participant whole blood RNA was amplified with *ATM* primers: Forward, TAGAAGATTGTGTCAAAGTTCG at c.5318-c.5339, spanning exons 34/35; Reverse, GCAAGACTTCTTTTCTCTTGAT at c.6077-c.6056, spanning exons 39/40 (NM_000051.4). PCR products were purified, and Sanger sequenced. The pseudoexon introduces a premature stop at codon 1929 (the same as in family CF5431).

(**G**) *ATM c.8418+704G>T* (chr11:108344075 G>T) in family CF6132.



Transcript 1. The electropherogram above illustrates the flanking splice junctions of the transcribed 209bp pseudoexon between ATM exons 56 and 57. Random-hexamer-primed cDNA generated from participant whole blood RNA was amplified with *ATM* primers: Forward, AAATTAAGGTGGACCACACA at c.8153-c.8172, spanning exons 54/55; Reverse, TGTTCAAAAGCAACACCTAGA at c.8837-c.8817, spanning exons 59/60 (NM_000051.4). PCR products were purified and Sanger sequenced. The pseudoexon introduces a premature stop at codon 2762.



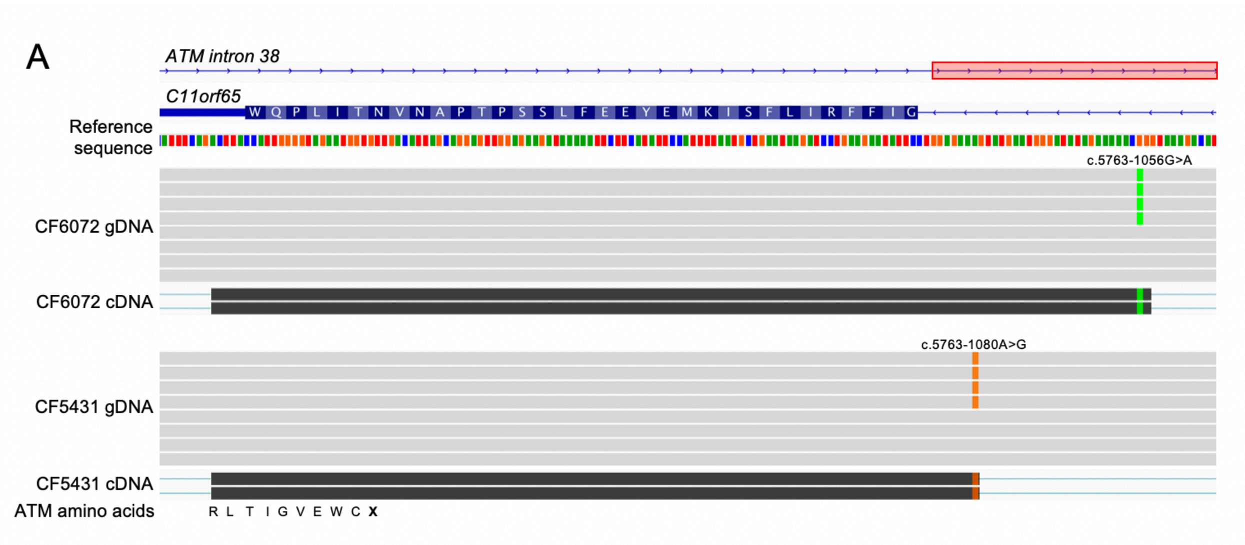
Transcript 2. The electropherogram above illustrates the flanking splice junctions of the transcribed 102bp pseudoexon between *ATM* exons 57 and 58. Random-hexamer-primed cDNA generated from participant whole blood RNA was amplified with *ATM* primers: Forward,

AAATTAAGGTGGACCACACA at c.8153-c.8172, spanning exons 54/55; Reverse, TGTTCAAAAGCAACACCTAGA at c.8837-c.8817, spanning exons 59/60 (NM_000051.4). PCR products were purified and Sanger sequenced. The pseudoexon introduces a premature stop at codon 2809.

Transcript 3. *ATM* mutant Transcript 3 includes the pseudoexon of Transcript 1 and inclusion of all of intron 57. Transcript 3 was rare but apparent from long-read cDNA sequencing, but RT-PCR and Sanger sequencing of the same cDNA samples did not reveal any copies of it. This transcript includes the premature stop at codon 2762.

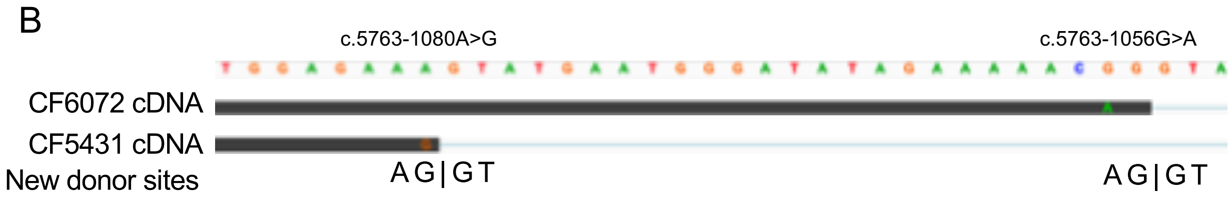
Supplemental Figure S2. *ATM* pseudoexons in families CF6072 and CF5431 in a cryptic splicing hotspot

(A) Pseudoexons in *ATM* exon 38 created by *ATM c.5763-1056G>A* in family CF6072 and *ATM c.5763-1080A>G* in family CF4531



As shown above, two variants at positions -1056 (green highlight) and -1080 (red highlight) in *ATM* intron 38 yield pseudoexons with different cryptic donors, but with the same cryptic acceptor. The red bar at the top of the figure indicates a 35bp cryptic donor hotspot predicted by SpliceAI. The hotspot is coincident with the AG-rich polypyrimidine tract in intron 6 of *C11orf65*. Pseudoexonification is possible because in addition to this feature, the 3'UTR of *C11orf65* includes a sequence that on the *ATM* strand is a near perfect acceptor motif (*TCATTCATTTCAG,* Chr11:108,308,956-108,308,968).

(**B**) Enlargement of the 3' portions of the pseudoexons shown in (**A**), indicating the basepairs of the revealed cryptic donor splice sites.


B
c.5763-1080A>G
c.5763-1056G>A
T G G A G A A A G T A T G A A T G G G A T A T A G A A A A A C G G G T A
CF6072 cDNA
CF5431 cDNA
New donor sites
AG|GT
AG|GT