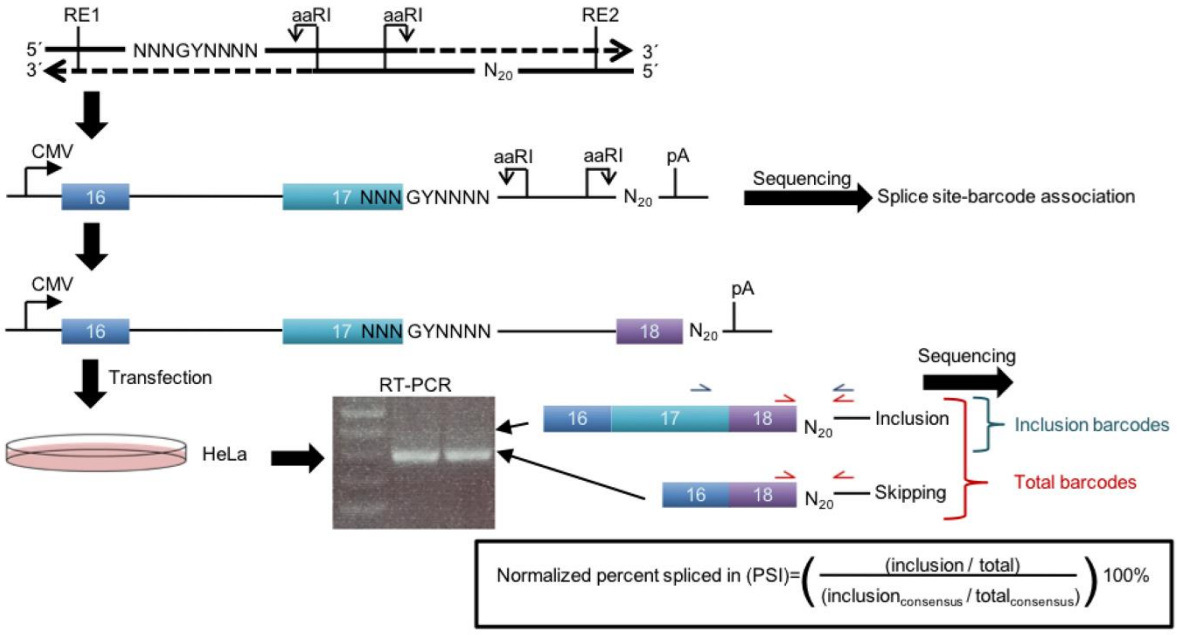


Figure S1. Detailed diagram of high-throughput method used to assess all 5′ss sequences. Related to Figure 2A.
RE1 and RE2 represent restriction enzyme sites used to insert the minigenes into the pcDNA5 expression vector, which has a cytomegalovirus (CMV) promoter and a bGH polyadenylation site (pA).

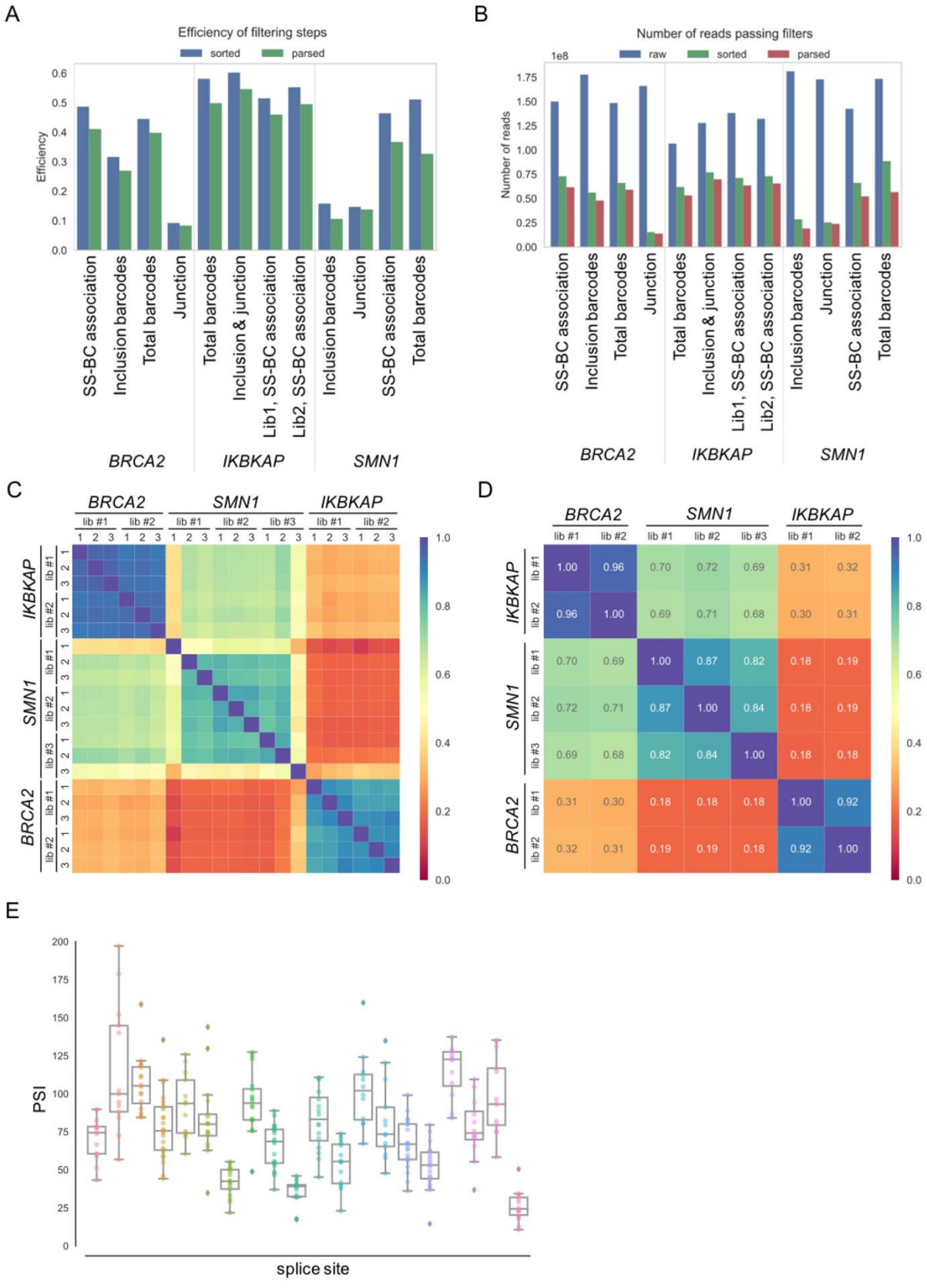

A
Efficiency of filtering steps
sorted
parsed
Efficiency
SS-BC association
Inclusion barcodes
Total barcodes
Junction
Total barcodes
Inclusion & junction
Lib1, SS-BC association
Lib2, SS-BC association
Inclusion barcodes
Junction
SS-BC association
Total barcodes
BRCA2
IKBKAP
SMN1
B
Number of reads passing filters
1e8
raw
sorted
parsed
Number of reads
C
BRCA2
SMN1
IKBKAP
lib #1
lib #2
lib #3
D
1.00 0.96 0.70 0.72 0.69 0.31 0.32
0.96 1.00 0.69 0.71 0.68 0.30 0.31
0.70 0.69 1.00 0.87 0.82 0.18 0.19
0.72 0.71 0.87 1.00 0.84 0.18 0.19
0.69 0.68 0.82 0.84 1.00 0.18 0.18
0.31 0.30 0.18 0.18 0.18 1.00 0.92
0.32 0.31 0.19 0.19 0.18 0.92 1.00
E
PSI
splice site

Figure S2. Parsing of raw high-throughput results and association matrix of the coefficient of determination comparing libraries. Related to Table and Figure 2D.

(A) The efficiency of filtering of raw sequencing data through the bioinformatics pipeline.

(B) Number of reads for analysis after filtering.

(C) Two low-quality datasets (*SMN1* library 1, replicate 1, and *SMN1* library 3, replicate 3) were removed from subsequent analyses.

(D) Coefficient of determination values show that the independently derived libraries highly correlate with each other within a context.

(E) Box plot of the PSI of individual barcodes for 20 randomly selected 5′ss. For this analysis, each barcode is required to have at least 10 counts in the total RNA sample in order to accurately calculate a PSI. The median PSI of the 5′ss is required to be ≥20. For each 5′ss, a minimum of 10 associated barcodes is necessary for the 5′ss to be included in the analysis. The central rectangle spans the first to the third quartile, with the median line segment. The vertical line presents the maximum and minimum.

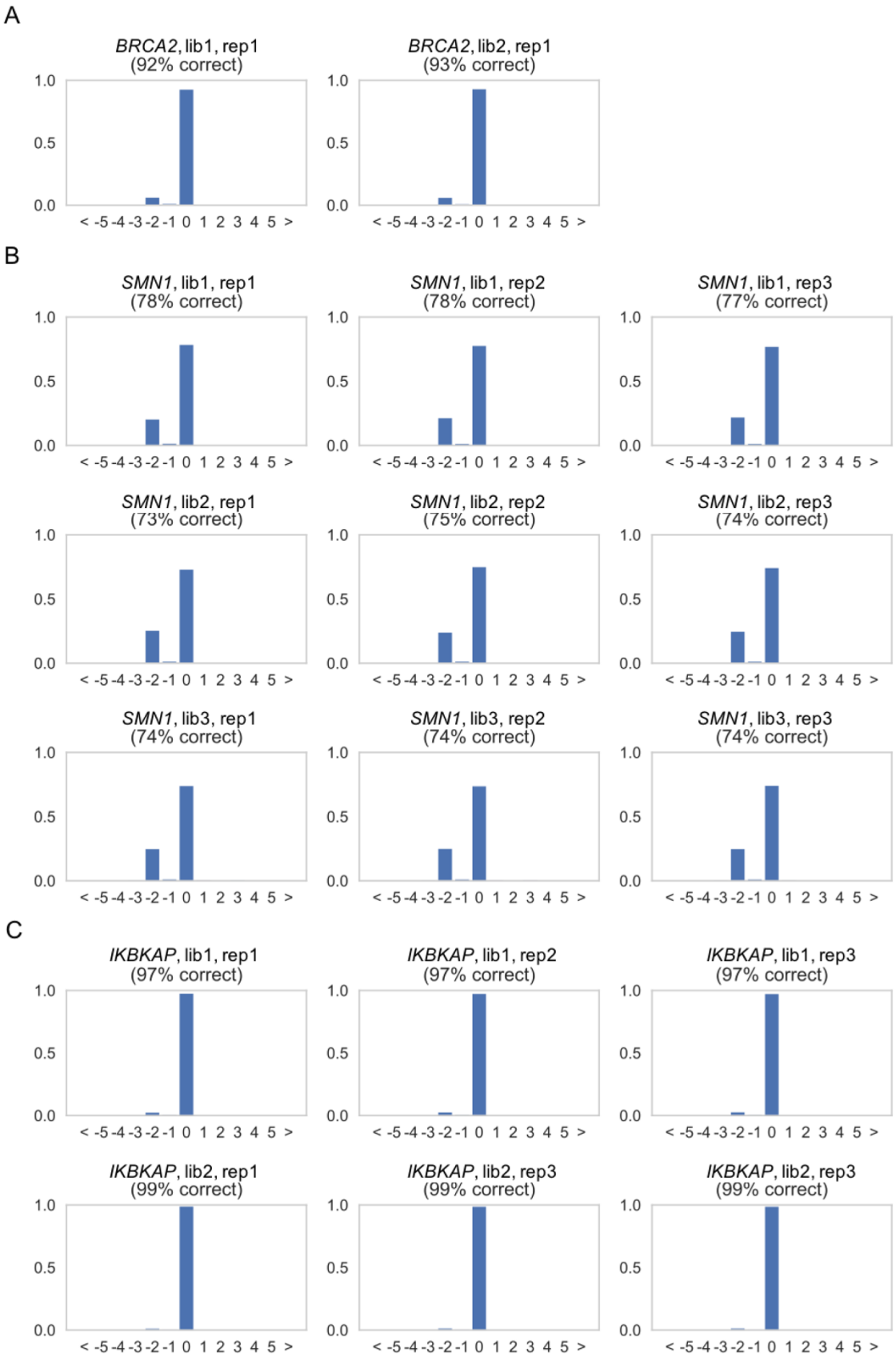

A
BRCA2, lib1, rep1
(92% correct)
BRCA2, lib2, rep1
(93% correct)
B
SMN1, lib1, rep1
(78% correct)
SMN1, lib1, rep2
(78% correct)
SMN1, lib1, rep3
(77% correct)
SMN1, lib2, rep1
(73% correct)
SMN1, lib2, rep2
(75% correct)
SMN1, lib2, rep3
(74% correct)
SMN1, lib3, rep1
(74% correct)
SMN1, lib3, rep2
(74% correct)
SMN1, lib3, rep3
(74% correct)
C
IKBKAP, lib1, rep1
(97% correct)
IKBKAP, lib1, rep2
(97% correct)
IKBKAP, lib1, rep3
(97% correct)
IKBKAP, lib2, rep1
(99% correct)
IKBKAP, lib2, rep3
(99% correct)
IKBKAP, lib2, rep3
(99% correct)
1.0
0.5
0.0
< -5 -4 -3 -2 -1 0 1 2 3 4 5 >

Figure S3. Exon-exon junction analysis of libraries. Related to Figure 2 and “high-throughput analysis of the activity of all 5′ss sequences”.

(A) Sequencing results for the exon-exon junction reveal that a secondary GU at the -2 and -1 positions is preferentially used when the GU or GC at the +1 and +2 positions escapes recognition. 5′ss sequences with the secondary -2G-1U were removed from further analysis.

(B) Same as A, but in *SMN1*.

(C) Same as A, but in *IKBKAP*.

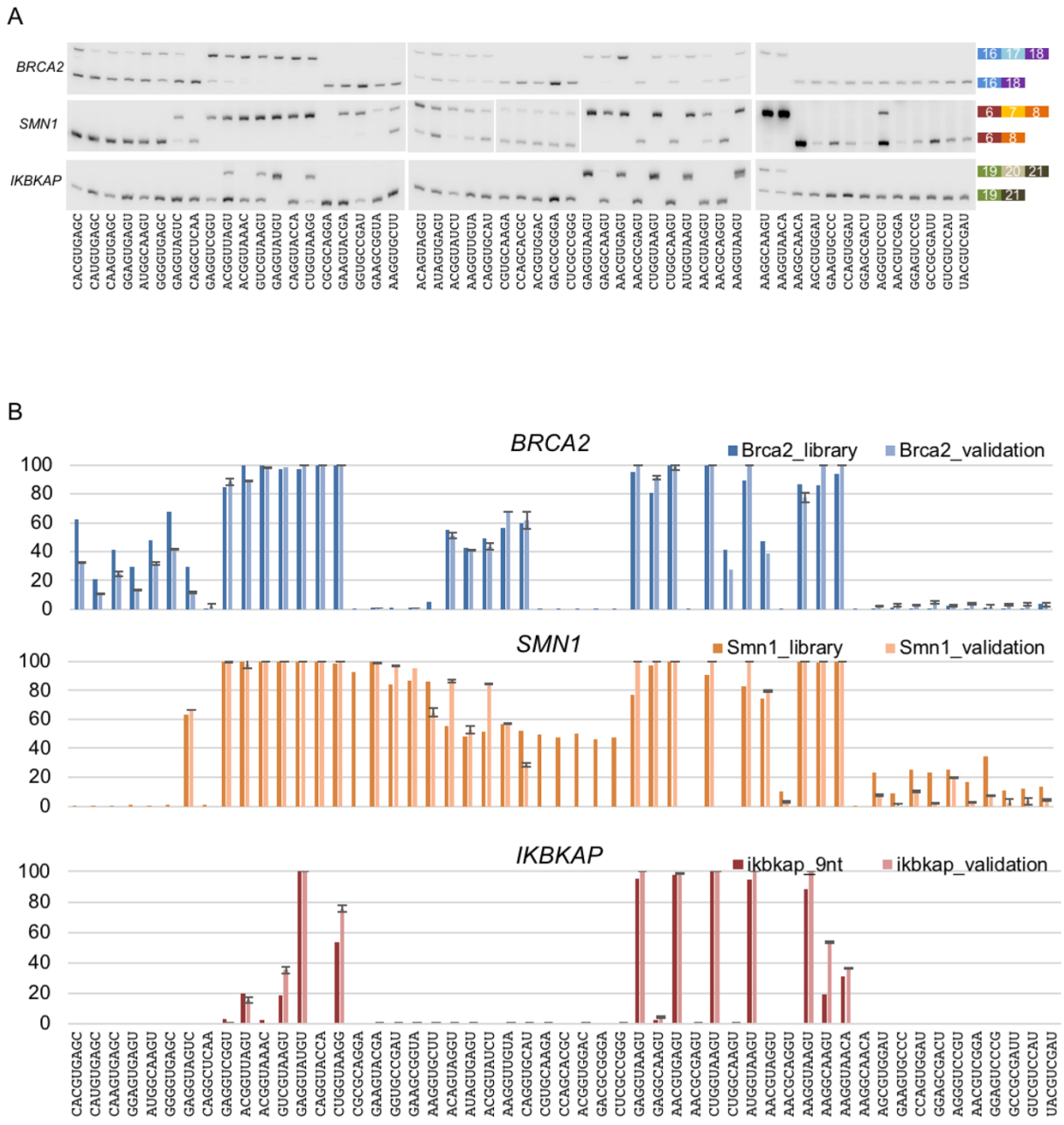


Figure S4. Manual validations of randomly-selected 5′ss in the three contexts. Related to Figure 2F.

(A) Gel showing the splicing results of the same 53 randomly-selected 5′ss in the *BRCA2*, *SMN1*, and *IKBKAP* minigenes. Gel images are representatives of triplicates.

(B) Graphs comparing the percent spliced in (PSI) of 5′ss derived from the library results and manual validations. Standard deviation is represented by error bars.

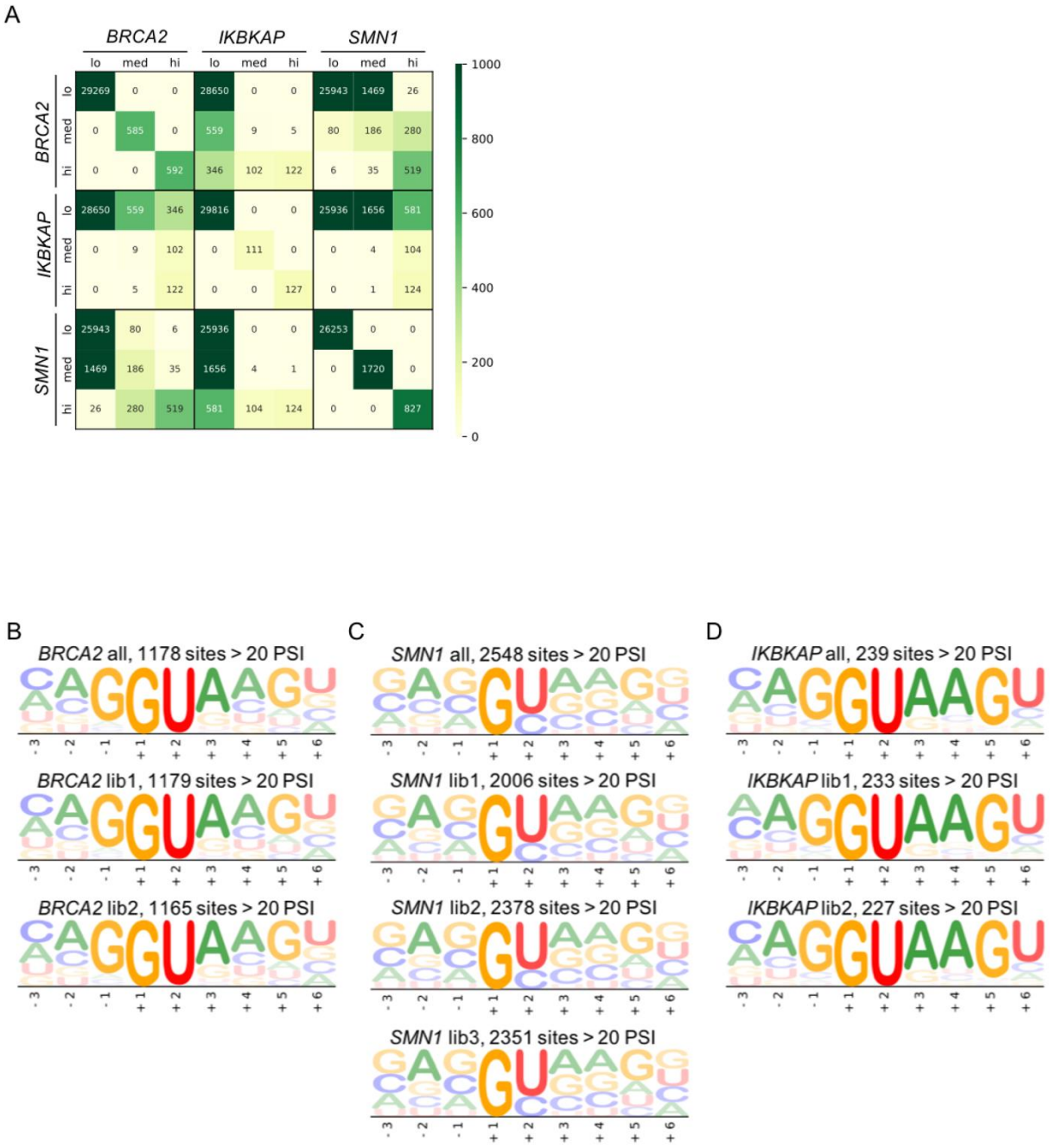


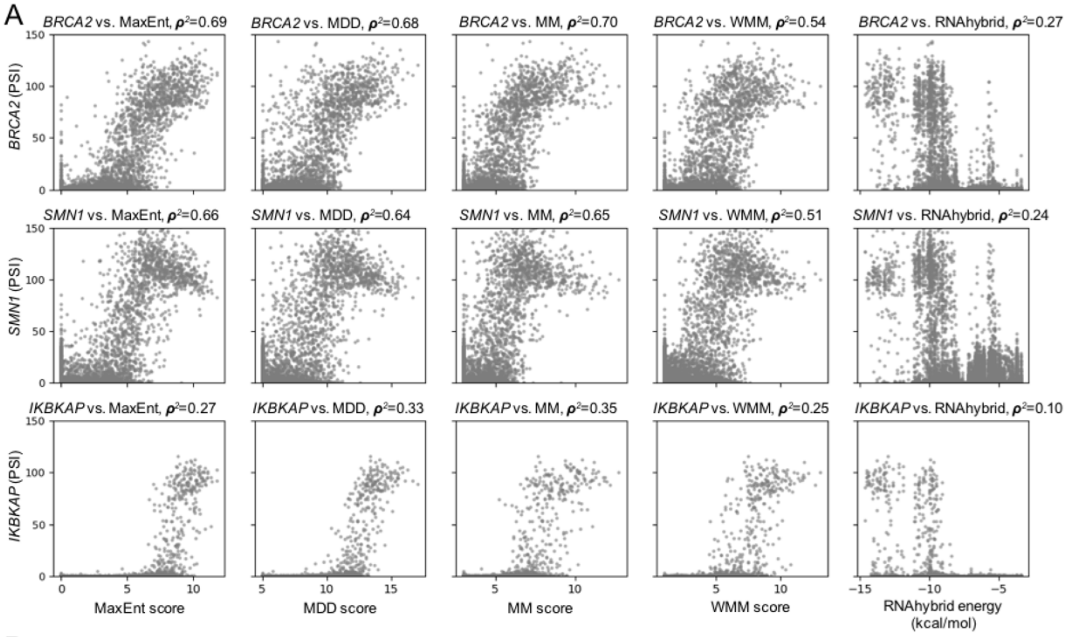
Figure S5. Sequence logo of 5′ss in each context. Related to Figure 3B and 3C.

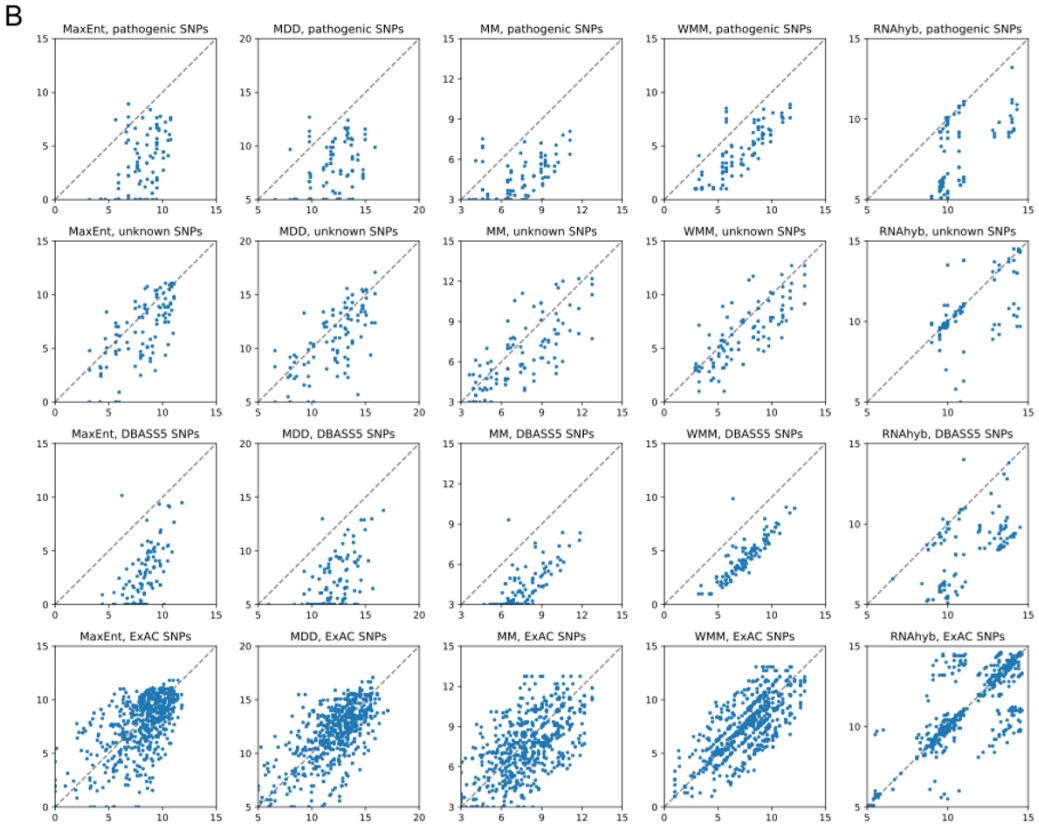
(A) Heat map illustrating the overlaps between 5′ss with lo (PSI < 20), med (20≤PSI<80) or hi (80≤PSI) activity levels among the three gene contexts.

(B) Sequence logo for all 5′ss compiled or separated by each independently-derived library in *BRCA2*.

(C) Same as A, but in *SMN1*.

(D) Same as A, but in *IKBKAP*.


A
BRCA2 vs. MaxEnt, ρ²=0.69
BRCA2 vs. MDD, ρ²=0.68
BRCA2 vs. MM, ρ²=0.70
BRCA2 vs. WMM, ρ²=0.54
BRCA2 vs. RNAhybrid, ρ²=0.27
BRCA2 (PSI)
SMN1 vs. MaxEnt, ρ²=0.66
SMN1 vs. MDD, ρ²=0.64
SMN1 vs. MM, ρ²=0.65
SMN1 vs. WMM, ρ²=0.51
SMN1 vs. RNAhybrid, ρ²=0.24
SMN1 (PSI)
IKBKAP vs. MaxEnt, ρ²=0.27
IKBKAP vs. MDD, ρ²=0.33
IKBKAP vs. MM, ρ²=0.35
IKBKAP vs. WMM, ρ²=0.25
IKBKAP vs. RNAhybrid, ρ²=0.10
IKBKAP (PSI)
MaxEnt score
MDD score
MM score
WMM score
RNAhybrid energy (kcal/mol)


B
MaxEnt, pathogenic SNPs
MDD, pathogenic SNPs
MM, pathogenic SNPs
WMM, pathogenic SNPs
RNAhyb, pathogenic SNPs
MaxEnt, unknown SNPs
MDD, unknown SNPs
MM, unknown SNPs
WMM, unknown SNPs
RNAhyb, unknown SNPs
MaxEnt, DBASS5 SNPs
MDD, DBASS5 SNPs
MM, DBASS5 SNPs
WMM, DBASS5 SNPs
RNAhyb, DBASS5 SNPs
MaxEnt, ExAC SNPs
MDD, ExAC SNPs
MM, ExAC SNPs
WMM, ExAC SNPs
RNAhyb, ExAC SNPs

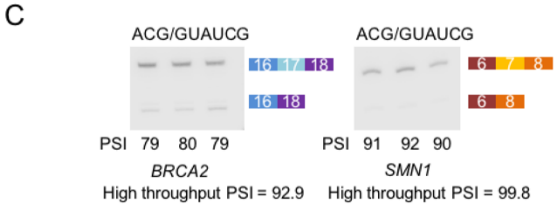

C
ACG/GUAUCG
16 17 18
16 18
PSI 79 80 79
BRCA2
High throughput PSI = 92.9
ACG/GUAUCG
6 7 8
6 8
PSI 91 92 90
SMN1
High throughput PSI = 99.8

Figure S6. Comparison between library results and several conventional prediction algorithms. Related to Figure 3D, 3E, and 6.

(A) Scatter plots comparing the predicted values for each 5′ss using maximum entropy (MaxENT; Yeo et al., 2004), maximum dependence decomposition (MDD; Burge et al., 1998), first-order Markov model (MM; Krogh et al., 1994), weight matrix model (WMM), and RNAhybrid (RNAhyb; Kruger et al., 2006) versus the experimentally derived library results.

(B) Scatter plots, corresponding to the computational scoring matrices used above, comparing the predicted PSI of the WT 5′ss sequences to mutant 5′ss sequences known to be pathogenic in *BRCA1* and *BRCA2* (pathogenic SNPs), with unclassified or uncertain significance in *BRCA1* and *BRCA2* (unknown SNPs), 5′ss mutations across a broad range of genes and diseases (DBASS5), and for 5′ss SNPs with >10% frequency found in the human population (ExAC SNPs).

(C) RT-PCR validation of the usage of the 5′ss with the sequence ACG/GUAUCG, which showed high inclusion ratio in *BRCA2* and *SMN1* library results but does not occur naturally as a 5′ss in the human transcriptome. Gel image is representative of triplicates. Percent spliced in (PSI) is indicated below each lane.

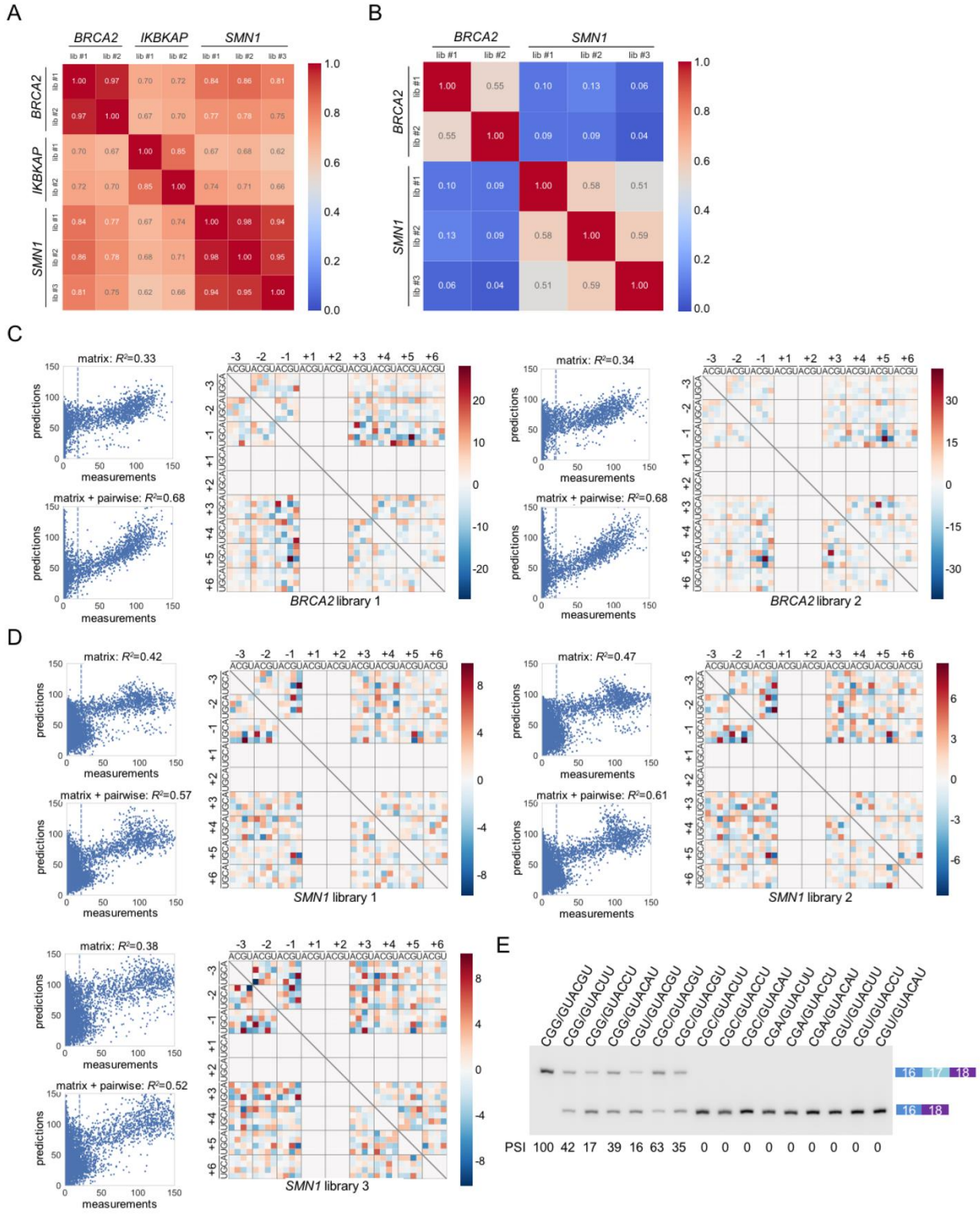

A
BRCA2
IKBKAP
SMN1
lib #1
lib #2
lib #3
1.00 0.97 0.70 0.72 0.84 0.86 0.81
0.97 1.00 0.67 0.70 0.77 0.78 0.75
0.70 0.67 1.00 0.85 0.67 0.68 0.62
0.72 0.70 0.85 1.00 0.74 0.71 0.66
0.84 0.77 0.67 0.74 1.00 0.98 0.94
0.86 0.78 0.68 0.71 0.98 1.00 0.95
0.81 0.75 0.62 0.66 0.94 0.95 1.00
1.0 0.8 0.6 0.4 0.2 0.0
B
BRCA2
SMN1
1.00 0.55 0.10 0.13 0.06
0.55 1.00 0.09 0.09 0.04
0.10 0.09 1.00 0.58 0.51
0.13 0.09 0.58 1.00 0.59
0.06 0.04 0.51 0.59 1.00
C
matrix: R²=0.33
matrix + pairwise: R²=0.68
predictions
measurements
BRCA2 library 1
matrix: R²=0.34
matrix + pairwise: R²=0.68
BRCA2 library 2
D
matrix: R²=0.42
matrix + pairwise: R²=0.57
SMN1 library 1
matrix: R²=0.47
matrix + pairwise: R²=0.61
SMN1 library 2
matrix: R²=0.38
matrix + pairwise: R²=0.52
SMN1 library 3
E
CGG/GUACGU
CGG/GUACUU
CGG/GUACCU
CGG/GUACAU
CGU/GUACGU
CGC/GUACGU
CGC/GUACUU
CGC/GUACCU
CGC/GUACAU
CGA/GUACUU
CGA/GUACCU
CGA/GUACAU
CGU/GUACUU
CGU/GUACCU
CGU/GUACAU
16 17 18
16 18
PSI 100 42 17 39 16 63 35 0 0 0 0 0 0 0 0 0

Figure S7. Accounting for pairwise associations between nucleotides improves the predictability of the results. Related to Figure 4.

(A) Association matrix showing correlations of parameters in the linear matrix model.

(B) Same as A, but in pairwise associations. The number of active 5′ss was insufficient to analyze pairwise associations for *IKBKAP*.

(C) Scatter plot showing the predicted PSI versus the measured PSI when considering a linear matrix model only (top) and a matrix model together with pairwise interaction (bottom). The heat map shows pairwise associations between nucleotides at different positions in *BRCA2*, separated by libraries. Red indicates a positive interaction, and blue indicates a negative interaction.

(D) Same as C, but for *SMN1*.

(E) Validation of a comprehensive series of mutations at -1 and +5 positions of the 5'ss CGG/GUACGU, in the *BRCA2* context. Gel image is representative of triplicates. Percent spliced in (PSI) is indicated below each lane.

Table S1 - 5′ss sequences that have activity in library results, but do not occur naturally in the human transcriptome. Related to Figure 3E.

BRCA2	***SMN1***
ACGGUAUCG	ACGGUAUCG
CGCGUACGU	CGCGUACGU
AAAGCGCUG	AACGUACGG
CCAGUACCG	CACGUACGC
CCCGCCGUG	CCGGUAUAC
GCGGUAAAC	CGCGUACGA
GUGGCAUCG	GAAGCGGUA
	GACGCGGUA
	GACGUACGA
	GGCGCGUAU

Table S2 – Predicted strength of the upsteam and downstream 3′ss. Related to Figure 5.

	Gene (input sequence, 5´→3´)	MAXENT	MM	WMM
Upstream 3′ss	*BRCA2* (TTCTACTTTTATTTGTTCAGGGC)	8.33	9.46	10.78
	SMN1 (TTCCTTTATTTTCCTTACAGGGT)	10.92	13.08	15.51
	IKBKAP (ACTGCTTTAATTTATTTAAGATG)	6.36	6.51	4.57
Downstream 3′ss	BRCA2 (ATTTTTGTTTTCACTTTTAGATA)	11.50	12.16	12.62
	SMN1 (TTCTAATTTCTCATTTGCAGGAA)	10.77	10.86	10.52
	IKBKAP (CTTTCTCTGTCTTCTCACAGACT)	11.96	12.06	13.35

Table S3 – Primer sequences. Related to STAR methods.

Name	Sequence (5′→3′)
FRT F	CTGGCTAACTAGAGAACCCACTGC
BRCA2 18R	GCTGTGTCATCCCTTTCCATTATC
BRCA2 7R	GAGCACAGTAGAACTAAGGGTGG
SMN1 R	TAGTGGTGTCATTTAGTGCTGC
IKBKAP R	GATTGATTCTCAGCTTTCTCATGC
BRCA2 Bsu36I ss top	CATCATCCTAAGGAATTTGCTAATAGATGCCTAAGCCCAGAAAGGG TGCTTCTTCAACTAAAATANNNGNNNNNTTAAAGCAGCAGGTGGAT GCACATGATGACATAAT
BRCA2 NotI bc bot	TACTACCGCCGGCGNNNNNNNNNNNNNNNNNNNNTCTAGAATGCA GGTGATTATGTCATCATGTGCATC
BRCA2 PE1 top	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG CTCTTCCGATCTNNNNNNNNNNT
BRCA2 PE1 bot	CATGTNNNNNNNNNNTCTAGCCTTCTCGCAGCACATCCCTTTCTCA CATCTAGAGCCACCAGCGGCATAGTAA
BRCA2 PE2 top	CGTNNNNNNNNNNAGATCGGAAGAGCGGTTCAGCAGGAATGCCGA GACCGATCTCGTATGCCGTCTTCTGCTT
BRCA2 PE2 bot	TTCGTCTTCTGCCGTATGCTCTAGCCAGAGCCGTAAGGACGACTTG GCGAGAAGGCTAGANNNNNNNNNNT
BRCA2 insert F	CATCATCACCTGCAGAGTTAAAGCATTACATTACG
BRCA2 insert R	TACTACCACCTGCACACTCTAGAATTACTACTTTAAC
BRCA2 17F	AGATGCCTAAGCCCAGAAAG
BRCA2 18F	GGCTCTCCTGATGCCTGTAC
BRCA2 18R	GCTGTGTCATCCCTTTCCATTATC
BRCA2 BC R	GGCAACTAGAAGGCACAGTCG
BRCA2 BC-LID F	CCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGGCTCTCCTGAT GCCTGTAC
BRCA2 BC-LID R	CATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGGCAACTA GAAGGCACAGTCG
BRCA2 JNCT F	CCCTACACGACGCTCTTCCGATCTNNNNNNNNNNAGATGCCTAAGC CCAGAAAG
BRCA2 JNCT R	CATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGCTGTGTC ATCCCTTTCCATTATC
SMN1 BseRI ss top	CATCATGAGGAGCTTAAATTAANNNGYNNNNCTGCCAGCATGCAGG TGGATGCACATGATGACATAA
SMN1 NotI bc bot	ATGATGGCGGCCGCNNNNNNNNNNNNNNNNNNNNTCTAGAATGCA GGTGATTATGTCATCATGTGCATC
SMN1 PE1 top	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG CTCTTCCGATCTNNNNNNNNNNGTGCT
SMN1 PE1 bot	CNNNNNNNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA GATCTCGGTGGTCGCCGTATCATT

SMN1 PE2 top	GGCCGCNNNNNNNNNNAGATCGGAAGAGCGGTTCAGCAGGAATG CCGAGACCGATCTCGTATGCCGTCTTCTGCTT
SMN1 PE2 bot	AAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAA CCGCTCTTCCGATCTNNNNNNNNNNGC
SMN1 insert F	CATCATCACCTGCTAGGGCCAGCATTATGAAACTGAATC
SMN1 insert R	ATGATGCACCTGCCCTATCTAGAATAACGCTTCACATTCCAGATC
SMN1 7F	GAAGGAAGGTGCTCACATTC
SMN1 8F	GACACCACTAAAGAAACGATCAG
SMN1 8R	CGCTTCACATTCCAGATCTG
SMN1 BC R	GGCAACTAGAAGGCACAGTCG
SMN1 BC-LID F	CCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGACACCACTAAA GAAACGATCAG
SMN1 BC-LID R	CATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGGCAACTA GAAGGCACAGTCG
SMN1 JNCT F	CCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGAAGGAAGGTG CTCACATTC
SMN1 JNCT R	CATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCGCTTCAC ATTCCAGATCTG
IKBKAP BseRI ss top	CATCATGAGGAGAGTGGTTGGANNNGYNNNNGCCATTGTGCAGGT GGATGCACATGATGACATAAT
IKBKAP XhoI bc bot	ATGATGCTCGAGNNNNNNNNNNNNNNNNNNNTCTAGAATGCAGG TGATTATGTCATCATGTGCATC
IKBKAP PE1 top	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG CTCTTCCGATCTNNNNNNNNNNTTAAT
IKBKAP PE1 bot	TAANNNNNNNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTG TAGATCTCGGTGGTCGCCGTATCATT
IKBKAP PE2 top	TCGAGNNNNNNNNNNAGATCGGAAGAGCGGTTCAGCAGGAATG CCGAGACCGATCTCGTATGCCGTCTTCTGCTT
IKBKAP PE2 bot	AAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAA CCGCTCTTCCGATCTNNNNNNNNNNC
IKBKAP insert F	CATCGTCACCTGCAAGCGCCATTGTACTGTTTGCGACTAGTTAGC
IKBKAP insert R	ATGATGCACCTGCCATGTCTAGAACTACTTAGGGTTATGATCAT
IKBKAP 20F	GTTGTTCATCATCGAGCCCTGG
IKBKAP 21F	GCATGAGAAAGCTGAGAATC
IKBKAP 21R	GATTCTCAGCTTTCTCATGC
IKBKAP BC R	GGCAACTAGAAGGCACAGTCG
IKBKAP BC-LID F	CCCTACACGACGCTCTTCCGATCTNNNNNNNNNGCATGAGAAAGCT GAGAATC
IKBKAP BC-LID R	CATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGGCAACTA GAAGGCACAGTCG
IKBKAP JNCT F	CCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGTTGTTCATCAT CGAGCCCTGG
IKBKAP JNCT R	CATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGATTCTCA GCTTTCTCATGC
PE1_v4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG CTCTTC

PE2_v4	AAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCT
ACTB-F	AGAGCTACGAGCTGCCTGAC
ACTB-R	AGCACTGTGTTGGCGTACAG