Published in final edited form as:

Curr Protoc Hum Genet. 2018 April ; 97(1): . doi:10.1002/cphg.60.

In vivo and In vitro methods to identify DNA sequence variants that alter RNA Splicing

Parth N Patel^{1,*}, Joshua M Gorham², Kaoru Ito³, and Christine E. Seidman⁴

¹Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115

²Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115

³Laboratory for Cardiovascular Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan

⁴Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115 and Howard Hughes Medical Institute, Boston, MA 02115

Abstract

Identification of sequence variants that create or eliminate splice sites has proven to be a significant challenge and represents one of many roadblocks in the clinical interpretation of rare genetic variation. Current methods of identifying splice altering sequence variants exist, however, these are limited by an imperfect understanding of splice signals and cumbersome functional assays. We have recently developed a computational tool that prioritizes putative splice-altering sequence variants, and a moderate-throughput minigene assay that confirms the variants which alter splicing. This bioinformatic strategy represents a substantial increase in accuracy and efficiency of historical in vitro splicing assays. In this unit we give detailed instructions on how to organize, run, and interpret various features of this protocol. We expect that splice-altering variants revealed through this protocol can be reliably carried forward for further clinical and biological analyses.

Keywords

Splice sites; mRNA processing; bioinformatic; minigene; sequence variant

INTRODUCTION

Over the last decade, interpretation of rare genetic variation has not kept pace with accelerating data generation using high-throughput DNA sequencing. As new genomic technologies have allowed for widespread interrogation of the growing number of genes implicated in disease, novel sequence variants associated with these diseases have accumulated (Richards et al., 2015). Unsurprisingly, these extensive sequencing efforts have

*Corresponding author.

INTERNET RESOURCES

GitHub page containing further detail of all codes and usage described in this unit: https://GitHub.com/SplicingVariant/SplicingVariants_Beta/

only modestly increased the number of nonsense, frameshift, deleterious missense, insertion and deletion variants that cause loss of function of the encoded protein. As a result, a sizeable fraction of these results continue to be labeled as variants of uncertain significance (VUS).

Some of these VUS may alter splicing (Ito et al., 2017; Soemedi et al., 2017). As gene transcription occurs, precursor mRNAs undergo splicing to connect sequential exons without intervening introns (Ward & Cooper, 2010). Splice sites are defined by signals at the exon-intron boundaries of all human genes and include a 5' splice donor and 3' splice acceptor sequence that flank each intron (Figure 1). The donor site contains a nine-base pair (bp) consensus sequence with a canonical GT dinucleotide, while the acceptor site contains a 23 bp consensus sequence with a canonical AG dinucleotide. Variants that alter the GT and AG canonical residues, which are invariant in most human genes, have been known for years to prevent normal splice site usage and lead to loss of function of the encoded protein (Shapiro & Senapathy, 1987). Substitutions outside of these canonical positions can also eliminate splice sites if they occur within the remaining 7 or 21 bp of these native splice signals (i.e. splice site loss)(Krawczak et al., 2007). Additionally, new, inappropriate splice signals can be created if a mutation occurs in a nearby region and creates a novel sequence that sufficiently mimics that of the splice site consensus (i.e. splice site gain)(Scotti & Swanson, 2016). A number of VUS fall into one of these two categories.

Software programs exist that predict whether any of these VUS have an effect on splicing, however, the poor specificity of these tools prevents their reliable application to variant interpretation (Houdayer et al., 2012; Jian, Boerwinkle, & Liu, 2014; Vreeswijk et al., 2009). Functional in vitro assays through minigenes or trapping vectors can confirm the biological splicing effects of these rare mutations, but these often require costly and cumbersome laboratory methods (Cooper, 2005; Gaildrat et al., 2010). To address these issues, we have developed a bioinformatic strategy that leverages the sensitivity of computational tools with the specificity of functional assessment. In this unit, we give detailed instructions on how to organize, run, and interpret various features of this approach. We first describe use of a splice predictor to rank and prioritize putative splice-altering variants. We then outline a laboratory protocol that utilizes synthetic gene fragments, a cellular assay, and RNA sequencing to evaluate the effects of individual sequence variants on splicing (Figure 2).

BASIC PROTOCOL 1: Using MaxEnt Scoring to Select Candidate Splice-Altering Variants

The objective of this protocol is to successfully install the software, use the software to calculate splice prediction scores, and interpret the output to select candidate variants for the subsequent laboratory workflow. This software tool uses MaxEnt scoring, one of the most widely utilized algorithms for splice prediction (Yeo & Burge, 2004). In brief, MaxEnt grades 9 bp and 23 bp sequence motifs for homology to the splice donor and acceptor consensus sequence, with higher scores representing a greater likelihood of a sequence functioning as a splice site. For this computational tool, a simple list of variants is used as

input. The tool retrieves genomic coordinates for each variant and calculates MaxEnt scores for the surrounding sequences with and without the variant. The output lists the most significant score changes due to each variant and helps identify which of the original variants are most likely to lead to splice site loss or gain.

Materials

- Installed versions of R 3.3.1 or later. This will need to be run from a command line. (Bash on UNIX, Terminal on Mac, or Cygwin on Windows)
<https://www.r-project.org/>
- If genome build is hg19 (GRCh37): Bioconductor package:BSgenome.Hsapiens.UCSC.hg19
<https://bioconductor.org/packages/release/data/annotation/html/BSgenome.Hsapiens.UCSC.hg19.html>
If genome build is hg38 (GRCh38): Bioconductor package:BSgenome.Hsapiens.UCSC.hg38
<https://bioconductor.org/packages/release/data/annotation/html/BSgenome.Hsapiens.UCSC.hg38.html>
- Files required to calculate MaxEnt scores.
Create a folder named “ssfiles” and place files from ssfiles.zip (provided at https://GitHub.com/SplicingVariant/SplicingVariants_Beta) into this “ssfiles” folder location.
- Files containing genome gene and transcript information.
If genome build is hg19 (GRCh37): Please create a folder named “usehg19” in the local environment and place files from usehg19.tar.gz (provided at https://GitHub.com/SplicingVariant/SplicingVariants_Beta).
 - If genome build is hg38 (GRCh38): Please make a folder named “usehg38” in the local environment and place files from usehg38.tar.gz (provided at https://GitHub.com/SplicingVariant/SplicingVariants_Beta).
- Script to calculate MaxEnt Scores from variant list.
Please download Regress_Score.v0.97.R (available at https://GitHub.com/SplicingVariant/SplicingVariants_Beta) and save to local environment. Modifiers to the script can be found in the README_regress.score.md file on the GitHub webpage.

For the purposes of simplification, we will use hg19 (GRCh37) in all examples in this protocol.

1. Start R 3.3.1 or later and enter the following to install the bioconductor package:
try http:// if https:// URLs are not supported

```
source("https://bioconductor.org/biocLite.R")
```

```
biocLite("BSgenome.Hsapiens.UCSC.hg19")
```

2. Ensure that the “ssfiles” folder, “usehg19” or “usehg38” folder, and Regress_Score.v0.97.R are prepared in the local environment as described in the materials.
3. Prepare the variant list in the following format. No headers are required. Each row specifies the variant chromosome, position, reference base, and alternate base for each individual variant that will be assessed. As an example, we will be running Regress_Score.v0.97.R on this list of 6 variants within *LMNA* on chromosome 1. This file will be referred to as the “mutfile”.

```
chr1:156100609T>G
```

```
chr1:156105681C>G
```

```
chr1:156106824G>C
```

```
chr1:156096389A>G
```

```
chr1:156106742C>G
```

```
chr1:156107550C>T
```

4. Identify the ENSEMBL transcript against which variants will be assessed for splicing changes (i.e. ENST00000368300). This can be indicated on the command line in the following step using “-esbTranscriptID ENST00000368300”. If no specific transcript is desired, you may also simply check for splicing effects on the longest protein-coding transcript by using “-takeCanonicalTranscript” in the command line.
5. From the command line, type the following argument. The command will have to be customized as indicated by the “[]” brackets. Hit enter to run the script.

```
R -slave -vanilla -args -mutfile [path to mutfile] -output [choose name of raw output file] -
summarizeresult [choose prefix for summarized result file] -skipRegressScore -skipSRE -
refdirectory [path to folder that contains ssfiles] -usehg19 [path to folder that contains
usehg19files] -skipcDNApos -takeCanonicalTranscript < Regress_Score.v0.97.R.
```

6. If the script was run correctly, two files will now have been created. The first will be the raw output file which will have the name that was defined in the previous step on the command line. There will also be a summarized result file with a suffix of “SigLines” which is more useful for the purposes of this protocol. Interpretation of these two files is provided below but optional. If desired, Table 1 provides explanations for the most important column headers in each file.

Interpretation of the Raw Output File (optional)

The raw output file contains the score calculations for each variant in the mutfile. You will immediately see that there are multiple rows for each variant. This is because a variant at any position sits within numerous independent sequence motifs. To illustrate this point,

imagine a sliding 9 bp window in which the window contains the 8 bases upstream of the variant and the variant at the 9th position. After sliding this window one base to the right, there are now 7 bases upstream of the variant followed by the variant at the 8th position and another base at the 9th position. This continues until the variant is in the first position of the sequence. The variant therefore changes the splice site strength of 9 different putative donor sites (Figure 3). Similarly, a variant can affect the strength of 23 different putative acceptor sites.

For a single transcript, the output file will therefore contain 23 rows for each variant. Each row represents a different start position for the sliding window, beginning with 9 bp and 23 bp windows that begin 22 bases upstream of the variant. As can be expected, the first 14 rows are inconsequential for the 9 bp splice donor site calculation, because these windows do not contain the variant. The last 9 windows are of greater interest. The sequence score with the reference base is compared to the sequence score with the variant base in each of these individual windows. Significantly different scores within each row may suggest splice site loss (window contains a functioning splice site and variant sequence score is lower than the reference sequence score) or splice site gain (window does not contain a functioning splice site and variant sequence score is higher than the reference sequence score).

Interpretation of the Summarized Result “SigLines” File (optional)—The summarized result file selects of the most significant rows from the raw output file for each variant. It is much easier to use and interpret and is therefore the file most relevant for the following steps of this protocol. In addition to a selection of rows from the raw output file, there are 3 columns at the end describing the kind of splicing change predicted (“Kind”), the possibility that the given score change will lead to abnormal splicing (“Possibility”), and our recommended decision whether to carry the variant forward to the next steps of the protocol (“Decision”). The variants given the designation “Decision” = YES are most likely to alter splicing.

In our example of 6 LMNA variants, 4 rows were given the designation “Decision” = YES and a subset of columns from these lines is shown in Table 2. Interpretation of the variants is as follows:

chr1:156100609T>G: This variant was selected because it had a row that predicted a significant score change that led to 5′ splice donor site gain (“Kind” = 5SS_Gain). According to the SS5 scores, the reference sequence had a score of −2.91, and the alternate sequence had a score of 7.29, creating a difference of +10.20 as a result of the variant. The variant led to a change that significantly increased the potential for this 9 bp sequence to be a functional splice site.

chr1:156105681C>G: This variant was selected because it had a row that predicted a significant score change that led to 3′ splice donor site loss (“Kind” = 3SS_Loss). First, the script recognized that this original 23 bp acceptor site sequence was an annotated splice site in the ENSEMBL transcript. According to the SS3 scores, the reference sequence had a score of 9.23, and the alternate sequence had a score of 2.91, creating a difference of −6.32

as a result of this variant. The variant led to a change that significantly decreased the potential of this 23 bp sequence to be functional.

chr1:156106824G>C: This variant was selected because it had a row that predicted a significant score change that led to 5' splice donor site loss ("Kind" = 5SS_Loss). In addition to the script recognizing that this original 9 bp donor site sequence was an annotated splice site in the ENSEMBL transcript, the SS5 MaxEnt scores changed from 11.45 in the reference sequence to 5.33 in the variant sequence for a difference of -6.12. The variant led to a change that significantly decreased the potential of this 9 bp sequence to be functional.

chr1:156096389A>G: This variant was selected because it had a row that predicted a significant score change that led to 3' splice donor site gain ("Kind" = 3SS_Gain). According to the SS3 scores, the reference sequence had a score of 8.46, and the alternate sequence had a score of 9.08, creating a difference of 0.62 as a result of the variant. The variant led to a change that significantly increased the potential of this 9 bp sequence to behave as a splice site.

7. Select the variants in the summarized result file with designation in the "Decision" column = YES. The "Decision" is located in the final column of the summarized output file and provides our recommendation on whether to carry the variant forward to Basic Protocol 2.

BASIC PROTOCOL 2: Design Constructs for Minigene Assay

The objective of this protocol is to computationally design oligonucleotides that will work as functional exon-intron-exon test sequences for the variants identified at the end of Basic Protocol 1. Use of this script retrieves the surrounding sequence context for each variant from online resources and allocates these sequences onto a 500 bp template. These sequences will be ordered as synthetic gene fragments and later built into a functional minigene in the later protocols.

Materials

- Two separate variant input files. This script uses input files similar to the mutfile used in the previous protocol, this time with the variants for which construct design is needed. One of the files will contain the 5' splice site changing variants and the other will contain the 3' splice site changing variants.
- ConstructDesigner.v.0.95.R.

Please download (available at https://GitHub.com/SplicingVariant/SplicingVariants_Beta) and save to local environment.

1. Of the variants selected from the final step in Basic Protocol 1, identify the variants that affect 5' splice sites (5SS_Loss or 5SS_Gain) and those that affect 3' splice sites (3SS_Loss or 3SS_Gain). This information is contained in the "Kind" column of the SigLines summarized result file.

2. 2. Prepare two separate files for the 5' splice donor changing variants and the 3' splice acceptor changing variants as shown below. The variant name and the transcript ID should be separated by a space. The necessary transcript ID can be found in the output of the SigLines file. No headers are needed.

```
chr1:156100609T>G ENST00000368299
```

```
chr1:156106824G>C ENST00000368299
```

3. 3. From the command line, enter the following and run the script to design constructs for 5' splice site changing variants. Additional modifiers can be found on the GitHub webpage.

```
R -slave -vanilla -args -mutfile [path to 5' splice variant mutfile] -output [name of
construct designer output file.txt] -gblocksubmit [name of gBlock order file.txt] -ss5 -
refdirectory [path to folder that contains ssfiles] -barcode
AA,AT,AG,AC,TA,TT,TG,TC,GA,GT,GG,GC,CA,CT,CG,CC -seq5
TTACGCCAAGTTATTTAGGTGACA -seq3
XXATCTAGATAACTGATCATAATCAGCCATACCACATTTGT -id 6,2 -limitintron 200 -
limit2ndexon 120 -usehg19 [path to folder that contains usehg19files] <
ConstructDesigner.v.0.95.R.
```

4. From the command line, repeat the entry with parameters running the script on 3' splice site changing variants. Additional modifiers can be found on the GitHub webpage.

```
R -slave -vanilla -args -mutfile [path to 3' splice variant mutfile] -output [name of
construct designer output file.txt] -gblocksubmit [name of gBlock order file.txt] -ss3 -
refdirectory [path to folder that contains ssfiles] -barcode
AA,AT,AG,AC,TA,TT,TG,TC,GA,GT,GG,GC,CA,CT,CG,CC -seq5
TTACGCCAAGTTATTTAGGTGACA -seq3
XXATCTAGATAACTGATCATAATCAGCCATACCACATTTGT -id 6,2 -limitintron 200 -
limit1stexon 120 -usehg19 [path to folder that contains usehg19files] < ConstructDesigner.v.
0.95.R.
```

5. Review the output files resulting from these commands. If the scripts were run correctly, two files will now have been created for each script.

- I. The first file is the gBlock order file. This file lists the name of the gBlock construct and the nucleotide composition of the test sequence. Note that there is a reference ("REF") and alternate ("ALT") construct for each variant. These two constructs are identical except for the variant and a 2 bp barcode at the end of the second exon. These will be the test sequences used in the subsequent protocols of this unit.
- II. Note: Each sequence contains a head region (lowercase), first exon (uppercase), intron (lowercase), second exon (uppercase), and a tail region (lowercase). The head and tail regions are identical for all constructs and necessary for PCR and library amplification in the later steps of the protocol. The exon-intron-exon sequence features are unique to each variant.

III. The second file is the Construct Designer Output File. This file will be used in the final analysis of sequencing data, outlined in Basic Protocol 8. Please check both 5' splice site and 3' splice site files and make sure that no two construct IDs are the same (the sequences in the "Ref_ID" and "Alt_ID" columns). These sequences serve as identifiers for each reference and alternate test construct, and therefore must be unique for each construct to be tested. If there are any replicates, these specific constructs should be kept separate during pooling of DNA at the end of Basic Protocol 3. Please make note of any variants which need to be separated.

6. Order the constructs within both gBlock order files (5' splice site and 3' splice site) from the Integrated DNA Technologies gBlocks service

(<https://www.idtdna.com/pages/products/genes/gblocks-gene-fragments>). Use the names and sequences provided in the order files. These test sequences will be used in the laboratory portion of this protocol.

BASIC PROTOCOL 3: Assembly PCR to Build Full Minigenes

The goal of this protocol is to place the exon-intron-exon test sequences designed in Basic Protocol 2 into a functional minigene. This is accomplished through two rounds of polymerase chain reaction (PCR, Figure 4). The first round of PCR attaches a SV40PA sequence to the end of the test sequence to create a PCR #1 product. The PCR #1 product is used in a second round of PCR that attaches a CMV promoter in front of the construct. The end result is a minigene in which the functional exon-intron-exon test sequence is under the expression of a ubiquitous promoter and stabilized by a transcriptional termination unit.

Materials

Reagents and Solutions (individual sequences for gBlocks and primers can be found in Table 3)

PCR #1

- Individual gBlocks Gene Fragments. These were designed and ordered in Basic Protocol 2. There should be a pair of gBlocks (REF and ALT) for each variant.
- SV40 Polyadenylation Signal (SV40PA-N500) gBlock, 225 bp.
- Tris EDTA Suspension Buffer (10 mM Tris, 0.1 mM EDTA, pH 8.0)
- Nuclease-Free H₂O (ThermoFisher Scientific Cat. No. AM9932)
- Primers for PCR #1 reaction: N500-RTPCR-F2, SV40PA-N500R
- Phusion High-Fidelity PCR Master Mix with GC Buffer (NEB Biolabs Cat. No. M0532S)
- Agencourt AMPure XP Beads (Beckman Coulter A63882)
- 80% Ethanol in Nuclease-Free H₂O

- 10mM Tris Acetate, pH 8.0 - made from Trizma base (Sigma T1503) and acetic acid (Sigma 537020)
- D1000 ScreenTape (Agilent 5067-5582) and Reagents (Agilent 5067-5583)

PCR #2

- CMV Promoter (CMVp-N500) gBlock, 600 bp
- Primers for PCR #2 reaction: CMVp-N500F, SV40PA-N500R
- D5000 ScreenTape (Agilent 5067- 5588) and Reagents (Agilent 5067- 5589)

Equipment

- 0.2 mL PCR strip tubes with caps
- Benchtop vortex and microcentrifuge
- 50 °C Incubator
- PCR Thermocycler
- TapeStation Instrument with Software Packages (Agilent Part Number G2991AA)

Note: All gBlocks, DNA intermediates, and primers in this basic protocol should be stored at −20 °C between uses.

1. Elute test sequence gBlocks by Tris-EDTA (TE). Add TE to each gBlock tube for a final concentration of 10 ng/uL. Vortex briefly. Incubate this solution at 50 °C for 20 min. Briefly vortex again. gBlocks are now ready for use.

PCR #1 to attach SV40polyA to each test sequence

2. Each individual gBlock eluted in step 1 should be attached to the SV40PA-N500 construct in an individual PCR reaction. It is easiest to perform these reactions in parallel in separate 0.2 mL PCR tubes. In each individual tube, add the PCR reagents at the specified concentrations and quantities as outlined in Table 4.
3. Pipette up and down to mix. Then cap the tubes and place in a PCR thermocycler. Run a program with the following parameters.

PCR Settings

98 °C 30 sec

98 °C 10sec 56 °C 20sec 72 °C 30sec (× 21-24 cycles)

72 °C 10min

4 °C Infinite

4. Purify PCR #1 product using 1.8× (36 uL) of Agencourt AMPure XP Beads (See Support Protocol 1). Elute into 20 uL of 10 mM Tris-Acetate. This is your purified PCR #1 product.

5. Measure concentrations of samples using the Agilent TapeStation D1000 ScreenTape assay. The attachment of a gBlock test sequence (500bp) to SV40PA (225bp) should create a product ~700bp in size (Figure 5). rd concentrations for each sample and dilute each PCR #1 product to 1 ng/uL using nuclease-free H₂O. This will be used in the next steps.

PCR #2 to attach CMV-Promoter to the PCR #1 product

6. As before, each PCR #1 product should be attached to the CMVp-N500 segment in an individual PCR reaction. It is easiest to perform these reactions in parallel in separate 0.2 mL PCR tubes. In each individual tube, add the PCR reagents at the specified concentrations and quantities as outlined in Table 5.

7. Pipette up and down to mix. Then cap the tubes and place in a PCR thermocycler. Run a program with the following parameters.

PCR Settings

98 °C 30 sec

98 °C 10sec 56 °C 20sec 72 °C 90sec (× 24-26 cycles)

72 °C 10min

4 °C Infinite

8. Purify PCR #2 product using 0.8× (20 uL) Agencourt AMPure XP Beads (see Support Protocol 1). Elute into 20 uL of 10 mM Tris-Acetate. This is your purified PCR #2 product.

9. Measure concentrations of samples using the Agilent TapeStation D5000 ScreenTape assay. The attachment of a CMV promoter (600bp) to the PCR #1 product (700bp) should create a product ~1250bp in size (Figure 6). Make sure to measure the concentration of this band specifically. This is the full minigene construct ready for transfection.

Note: If amplification was inefficient, please see the Troubleshooting section at the end of the unit.

10. Individual constructs can now be combined together into a pool of DNA. During this step, please keep all pairs of test sequences together when creating separate pools. For example, the REF and ALT test sequences for a given variant must go into the same pool. Additionally, ensure that test sequences with replicate Ref_ID or Alt_ID in the Construct Designer Order File (Basic Protocol 2) are placed in separate pools.

11. Add the volume equivalent to 100ng of full-length DNA construct for 20 samples into the same tube. This should create a single pool of approximately 2000 ng of DNA (100ng × 20 samples). If more than 20 samples are desired, create additional pools as needed. Be sure to note which test sequences were placed into each pool, as this knowledge will be important for downstream analysis. The total volume for each

pool of DNA should ideally be less than 125 μ L. This pooled DNA is ready for cellular transfection and should be stored at -20°C .

BASIC PROTOCOL 4: Preparation and Transfection of HEK293 Cells

At this point of the protocol, we have identified variants most likely to alter splicing (Basic Protocol 1), designed test sequences for these highest priority variants (Basic Protocol 2), and inserted these test sequences into functional minigenes using PCR (Basic Protocol 3). The minigenes have been pooled together in equivalent amounts and are ready to be placed in a cellular environment. In this protocol we will describe the basic cell culture procedures required to transfect these linear DNA fragments into cells.

Materials

Reagents and Solutions

- 293 [HEK-293] cellular line (ATCC CRL1573)
- HEK293 cell growth medium: 450 mL DMEM: (+) 4.5 g/L D-Glucose, (+) L-Glutamine, (+) 110mg/L Sodium Pyruvate (Fisher Scientific Cat No. MT10013CV), with 50 mL FBS and 1 mL Gentamycin.
- 70% ethanol spray
- $1\times$ Phosphate-Buffered Saline (Gibco Cat No. 10010023)
- TrypLE Express Enzyme ($1\times$), phenol red (ThermoFisher Cat No. 12605036)
- Lipofectamine 2000 (ThermoFisher Cat No. 11668027)
- Opti-MEM reduced serum medium (ThermoFisher Cat No. 31985062)

Equipment

- 37°C water bath
- Falcon 15mL and 50 mL Conical Centrifuge Tubes
- Corning tissue-culture treated culture dishes, 10 cm (Sigma-Aldrich CLS430167)
- Corning Costar TC-Treated Multiple Well Plates (Sigma-Aldrich CLS3516)
- Centrifuge
- Aseptic cell culture hood
- Microcentrifuge Tubes, 1.5 mL
- 37°C cell culture incubator with 5% CO_2 air atmosphere

Preparation and establishment of HEK293 cell line

1. Pre-warm medium to 37°C .
2. Thaw the HEK293 vial by gentle agitation in a 37°C water bath. To reduce the possibility of contamination, keep the O-ring and cap out of the water. Thawing should be rapid (approximately 2 minutes).

3. As soon as contents are thawed, remove the vial from the water and decontaminate by spraying with 70% ethanol. From this point, all steps should be carried out under strict aseptic conditions.
4. Transfer the vial contents to a centrifuge tube containing 9 mL of pre-warmed growth medium. Spin at approximately $125 \times g$ for 5 minutes.
5. Aspirate supernatant. Using the specific batch information, resuspend cellular pellet with the recommended amount of growth medium. Dispense 10 mL onto a 10 cm cell culture dish.
6. Incubate the culture at 37 °C in a 5% CO₂ air atmosphere. Change media every 2-3 days until 70% confluent.

Transfer of HEK 293 cells to 6 well plate once 70% confluent

7. Pre-warm medium and 1× PBS.
8. Aspirate medium from 10 cm cell culture dish.
9. Wash cells by 10 mL pre-warmed 1× PBS twice. Add the PBS gently and do not shake the plate.
10. Add 2.5 mL of (room temperature) Tryp LE Express to the dish, make sure it covers all the cells, and aspirate it immediately. This enzyme lifts your cells off the plate.
11. Incubate the cells in the dish at 37 °C for 5 minutes.
12. Add 10 mL pre-warmed medium and mix cells, then transfer this solution to a 50 mL Falcon tube.
13. Add 15 more mL of pre-warmed medium to the tube for a total of 25 mL.
14. Mix this solution up and down using a pipette.
15. Apply 2.5 mL of the cell-medium mix to each individual well of a 6 well plate (2.5 mL $\times$ 6 wells = 15 mL, 10 mL will be left over).
16. Dispense this remaining 10 mL on a fresh 10 cm cell culture dish for the succeeding subculture.
17. Incubate all cells at 37 °C. Transfection will occur once cells are 70-90% confluent each well of this 6 well plate. This should occur within 24 hours.

Transfection of HEK293 Using Lipofectamine 2000

Note: If cells are less than 70% confluent, they are more susceptible to the toxicity of Lipofectamine 2000 Reagent. For best cellular health, wait until cells reach 70%.

18. Each pool of DNA created at the end of Basic Protocol 3 will be a separate transfection reaction. If your DNA pool is less than 125 uL, add Opti-MEM to the pool to increase the total volume to 125 uL.

19. For each pool, create a separate Lipofectamine-Opti-MEM mix for each pool containing 10 uL Lipofectamine and 115 uL of Opti-MEM.
20. Add the 125 uL DNA-medium mix and the 125 uL Lipofectamine-medium mix together for a total volume of 250 uL for each transfection reaction. Pipette up and down to mix.
21. Allow this mix to incubate at room temperature for 10 minutes.
22. Add this 250 uL DNA-Lipofectamine mixture to a single well of HEK 293 cells on a 6 well plate. Repeat for each pool of DNA which requires transfection. Be sure to note again which test sequences were included in each transfection.
23. Incubate the plate at 37 °C for 24 hours.

BASIC PROTOCOL 5: RNA Extraction

In Basic Protocol 4, functional minigene test sequences were transfected into HEK293 cells. These cells provide the biological machinery that allow for both transcription and splicing of mRNA molecules derived from these DNA constructs. In order to evaluate the amounts of normal and aberrant splicing in transcripts derived from the reference- and variant-containing constructs, RNA from these cells will need to be extracted and sequenced. In this protocol, we will outline the steps required to successfully isolate RNA from cell culture.

Materials

Reagents and Solutions

- 1× Phosphate-Buffered Saline (Gibco Cat No. 10010023)
- TRIzol Reagent (ThermoFisher Cat No. 15596026)
- Chloroform (ThermoFisher Scientific Cat No. C298-500)
- Nuclease-free H₂O (ThermoFisher Scientific Cat No. AM9932)
- 75% ethanol
- 100% isopropyl alcohol
- RNA ScreenTape (Agilent 5067- 5576) with RNA ScreenTape Sample Buffer (Agilent 5067- 5577)

Equipment

- RNase-Free Microcentrifuge Tubes, 1.5 mL (ThermoFisher Cat. No. AM12400)
- Temperature-regulated centrifuge
- 2200 TapeStation Instrument (Agilent G2964AA)
- Agilent Software packages (2200 TapeStation Controller Software, and TapeStation Analysis Software)
- PCR Thermocycler

RNA Extraction from HEK293 cells using TRIzol reagent

1. Put 1× PBS on ice as it will need to be ice-cold during the procedure.
2. Aspirate and discard HEK cell medium.
3. Gently apply 2.5 ml of ice-cold, 1× PBS to each well. Do not agitate and aspirate quickly as cells will begin to be lifted from the wells.
4. Apply 2.6 ml TRIzol reagent for each well. Pipette up and down vigorously into all corners of the well until cells have been homogenized. Avoid any cross contamination between wells.
5. Collect TRIzol-cell mix and distribute the mix from each well evenly across two RNase-free microcentrifuge tubes (1.3 ml per tube). Do this for each well of your 6 well plate. There should now be twice as many tubes as there were transfection wells.

RNA Purification

Note: The remaining steps should be done on an RNase-free benchtop.

6. Incubate homogenized samples at room temperature for 5 minutes.
7. Place nuclease-free H₂O and 75% ethanol on ice.
8. Add 260 uL chloroform to each tube containing 1.3 ml of TRIzol-cell mixture (this creates a 1:5 chloroform: TRIzol ratio). Shake vigorously by hand for 15 seconds.
9. Incubate at room temperature for 2-3 minutes.
10. Centrifuge samples at 12,000g for 15 minutes at 4 °C
11. Transfer the aqueous phase from each tube into a fresh 1.5 mL tube. 600 uL of aqueous phase should be able to be pulled transferred. Pay close attention to only isolating the clear aqueous phase, as any impurities will decrease the quality of RNA. If the aqueous layer is not transparent, add another 200 uL chloroform and spin again.
12. Add an equal proportion (600 uL) of room temperature isopropyl alcohol to the 600 uL of transferred aqueous solution. Vortex thoroughly.
13. Incubate this mix at room temperature for 10 minutes.
14. Centrifuge this mix at 12,000g for 10 minutes at 4 °C. This should create a white pellet of RNA.
15. Remove all but 10-15 uL of the supernatant, making sure not to disturb the RNA pellet.
16. Add 500 ml of cold 75% ethanol to each tube. When doing so, forcefully eject the ethanol through the pipette down the side of the tube such that the RNA pellet is lifted from the bottom. Do not disrupt the pellet with the pipette. Close the cap and vortex thoroughly.
17. Centrifuge at 10,000g for 5 minutes at 4 °C. The white pellet should again be visible.

18. The remaining steps should all be performed on ice to prevent RNA denaturation unless otherwise specified. From each individual tube, remove all ethanol while making sure to keep the RNA pellet intact. Once the ethanol is off, suspend the pellet immediately in 30 uL of ice-cold nuclease free H₂O. The cold water should be ejected from the pipette with enough force that it lifts the pellet from the bottom. If this occurs, the pellet should naturally dissolve into the solution.

19. Vortex each tube. Since there are two tubes that originated from the same transfection, they may now be combined into a single sample. Continue to keep on ice.

Measure the concentration and quality of each RNA sample by TapeStation RNA ScreenTape

Note: If TapeStation RNA integrity numbers (RIN) are poor in this assay this could be due to large quantities of RNA. In our experience, a 1:20 dilution of RNA prior to the ScreenTape assay allows for the best measurements.

20. Check product concentrations of each sample using the TapeStation RNA ScreenTape assay. Follow manufacturer's instructions for sample preparation and sample denaturation before loading the TapeStation instrument. Make note of the RIN score for quality of your RNA. 9.0 to 10.0 is ideal, however, scores above 7.0 are sufficient.

An example of the output is shown in Figure 7.

21. The next protocol requires 12 ug of RNA in 50 uL of nuclease-free water for each sample. Use the measured concentrations to create this quantity and volume accordingly for the next step.

22. RNA should be stored at -80 °C.

BASIC PROTOCOL 6: cDNA Synthesis

This protocol will describe the process required to transform RNA into cDNA. First, total RNA will undergo further purification to select for messenger RNA (mRNA) only. Next, a "first-strand" of cDNA will be reverse transcribed from the mRNA using the 500-RTPCR-R1 primer (Table 4), which is complementary to the tail region of all exon-intron-exon test sequences designed in Basic Protocol 2. This allows for minigene-specific reverse transcription. Finally, a "second-strand" of cDNA will be synthesized, thereby creating stable, double stranded DNA molecules.

Materials

Reagents and Solutions

- Nuclease-Free H₂O
- Dynabeads mRNA DIRECT Kit (ThermoFisher Scientific Cat. No. 61012)
- SuperScript III cDNA synthesis kit (Invitrogen Cat. No. 18090-051)

- 5× Second Strand Buffer (Invitrogen Cat No. 10812-014)
- DNA Polymerase I (Invitrogen Cat No. 18010025)

Materials

- Microcentrifuge Tubes, 1.5 mL
- DynaMag-2 Magnet (ThermoFisher Scientific Cat No. 12321D)
- Heat Blocks at 62 °C and 72 °C

1. Prepare RNA input sample. On ice, measure the quantity of RNA sample that contains 12 ug of RNA. Then dilute this quantity up to 50 uL using H₂O. Vortex and spin down and place on ice.

Preparation of Dynabeads Oligo (dT)₂₅ beads

2. Assign one new tube for each sample. In each tube, add 100 uL of Dynabeads Oligo (dT)₂₅ beads. Place on a magnet stand. Aspirate and discard solution once clear.
3. Add 50 uL of ice-cold Lysis/Binding Buffer and dissolve beads. Vortex gently. Place back on magnet stand, wait until clear, and discard solution once clear.
4. Add 50 uL of ice-cold Lysis/Binding Buffer and dissolve beads. Keep these at room temperature. These are now ready for use in the next step.

1st Cycle Through RNA Purification

5. For each sample, put a fresh tube on ice. You will be taking the final solution off of a magnet stand and placing this volume in these chilled tubes later.
6. Add 50 uL of ice-cold Lysis/Binding Buffer to the 50 uL of each RNA sample on ice. There should not be any beads here in your 100 uL total. Vortex gently.
7. Place each sample on a 62 °C heatblock for 2 minutes.
8. Add the samples to the prepared beads which have been sitting at room temperature (now 150 uL total).
9. Incubate samples for 10 minutes at room temperature. Vortex in the beginning, middle, and end of the incubation period. Spin down lightly to ensure contents incubate at the bottom of the tube.
10. After 10 minutes, place the samples on a magnet stand. Wait until clear, and aspirate off the solution. Keep beads.
11. Add 200 uL of room temperature Washing Buffer B to each sample.
12. Vortex to dissolve beads. Then spin down lightly.
13. Place samples on a magnet stand and wait until solution clears. Aspirate and discard the clear solution. Keep beads.

14. Add 200 uL of Washing Buffer B a second time. Vortex again to dissolve beads. Spin down lightly.
15. Place samples on a magnet stand. This time the solution should look less clear and more smeared on the back wall of the tube. Aspirate this solution carefully and discard, making sure the beads remain. Note: if the solution is not cloudy, and instead still clear, repeat with one more Buffer B wash.
16. Add 50 uL of ice-cold nuclease free H₂O to the beads. Vortex and spin down briefly. Vortex and spin down again. Keep on ice.
17. **Note: the timing of this step is crucial, so do not attempt to perform this step on more than 4 samples at once.** Place the first 4 samples on a 72 °C heat block for 2 minutes.
18. Place the 4 tubes on the magnet stand. Transfer the cleared solution to an ice-cold 1.5 ml tube on ice. At this point, there should be 12 ug of purified RNA in 50 uL of H₂O. Do not discard the tubes containing the beads.
19. Repeat this prior 72 °C heat block step for all remaining samples.
20. The beads left in the tubes will be re-used for the second cycle of RNA purification. Add 50 uL of Lysis/Binding Buffer to the beads. Vortex, spin down gently, and place back on magnet stand.
21. Wait for solution to clear, and then remove clear solution.
22. Resuspend washed beads in 50 uL of Lysis/Binding Buffer. Keep these at room temperature. These are ready for use in the next step.

2nd Cycle Through RNA Purification

23. For each sample, put a fresh tube on ice. You will be taking the final solution off a magnet stand and placing this quantity in these cold tubes later.
24. Add 50 uL of ice-cold Lysis Binding Buffer to the 50 uL of your RNA sample on ice. There should not be any beads here in your 100 uL total.
25. Vortex gently.
26. Place samples on 62 °C heatblock for 2 minutes.
27. Add each sample to the second set of prepared beads which have been sitting at room temperature (now 150 uL total).
28. Incubate for 10 minutes at room temperature. Vortex in the beginning, middle, and end of the incubation period. Spin down lightly to make sure contents are incubating at the bottom of the tube.
29. After 10 minutes, place this solution on a magnet stand. Wait until clear, and aspirate off the solution. Keep beads.
30. Add 200 uL of Buffer B to each sample.
31. Vortex to dissolve beads. Spin down lightly.

32. Place samples on a magnet stand. Wait until solution clears. Aspirate and discard the solution. Keep beads.
33. Add 200 uL of Buffer B a second time. Vortex again to dissolve beads. Spin down lightly.
34. Place samples on a magnet stand. This time the solution should look less clear and more smeared on the back wall of tube. If this is not the case and the sample still appears clear, repeat with another Buffer B wash before aspirating and discarding the solution.
35. Add **12 uL** of ice-cold nuclease free H₂O. Vortex and spin down briefly. Pipette up and down to mix. Keep samples on ice.
36. Put samples on 72 °C heat block for 2 minutes. As before, limit this step to a maximum of 4 samples at any given time.
37. Place the 4 tubes on the magnet stand. Transfer the solution to an ice-cold 1.5 ml tube on ice. Now you should have 12 ug of purified RNA in 12 uL of H₂O.
38. Repeat this prior 72 °C heat block step for all remaining samples.

First Strand Synthesis by SuperScript III Reverse Transcriptase Kit

39. To each sample of 12 uL of purified mRNA, add the following:
- 2 uL – 500-RTPCR-R1 primer (10uM), see Table 3.
 - 1 uL – 10mM dNTPs, from SuperScript III kit
40. Transfer these samples to 0.2 mL PCR tubes and incubate at 65 degrees C for 4 minutes using a thermocycler. Immediately place on ice following completion.
41. To each sample add the following reagents found in the SuperScript III kit:
- 2 uL – 10x RT Buffer
 - 2 uL – DTT (0.1 M)
 - 4 uL – MgCl₂ (25 mM)
 - 1 uL – Rnase Out (enzyme, do not vortex)
 - 1 uL – SuperScript III RT (enzyme, do not vortex)
42. Cap tubes, mix by shaking by hand. Spin down and incubate samples in a thermocycler using the following program:
- 25 °C × 10 minutes/50 °C × 50 minutes/85 °C × 5 minutes/4 °C hold

Second Strand Synthesis by DNA Polymerase I

43. To each sample add the following reagents:
- 20 uL – 5× Second Strand Buffer
 - 3 uL – 10 mM dNTPs (from SuperScript III kit)

1 uL – RNase H (from SuperScript III kit)

4 uL – DNA polymerase I

52 uL – nuclease-free H₂O

44. Cap tubes, shake by hand, and spin down. Incubate samples in a thermocycler using the following program:

16 °C for 2.5 hours/4 °C hold

45. Purify with 3.0× AMPure beads (315 uL) according to manufacturer's protocol. Elute in 30 uL 10mM Tris Acetate pH 8.0. These samples now contain purified cDNA. Store at –20 °C.

BASIC PROTOCOL 7: Library Preparation for Illumina MiSeq

This protocol will describe the process required to prepare the current cDNA libraries for an Illumina MiSeq experiment. First, cDNA from all transfections will be selectively amplified to enrich for molecules containing the original test sequences. Libraries will then undergo end repair, A-tailing, and Illumina multiplexing adapter ligation. Lastly, there will be a final amplification step that will index each individual transfection and prepare multiplexed libraries for sequencing. Guidebooks that describe the Illumina multiplexing technology can be found online.

Materials

Reagents and Solutions

Enrichment of Test Sequences

- Primers for PCR reaction: N500-RTPCR-F2, N500-RTPCR-R1
- Nuclease-free H₂O
- Phusion High-Fidelity PCR Master Mix with GC Buffer
- DMSO (contained within NEB Biolabs Cat. No. M0532S)
- SYBR Green I Nucleic Acid Gel Stain 10,000x in DMSO (Sigma Aldrich S9430), diluted 1:1000

End Repair, A-tailing, and Adapter Ligation

- END-IT kit (Epicenter ER 0720)
- Klenow Fragment [3'–5' exo–] (NEB Biolabs M0212S)
- dATP (ThermoFisher Scientific Cat No. 10216018), diluted to 5 mM
- Quick Ligation Kit (New England Biolabs, Cat. No. M2200L) containing 2× Ligation Buffer and Quick Ligase
- Multiplex Adapter (see Table 3)
- Multiplexing Primer 1.0 (see Table 3)

Equipment

Enrichment of Test Sequences

- MicroAmp Fast Optical 96-well Reaction Plate with Barcode (Applied Biosystems Ref#4346906) and Optical Adhesive Covers (Applied Biosystems Part No. 4360954)
- Applied Biosystems 7500 Fast Real Time PCR System (ThermoFisher Scientific Cat No. 4351106)

End Repair, A-tailing, and Adapter Ligation

- 1.5 mL microcentrifuge tubes
- 0.2 mL PCR strip tubes with caps
- PCR Thermocycler

Enrichment of Test Sequences

This portion of the protocol will contain two parts. First, a qPCR will identify the number of cycles at which a given sample reaches saturation. Then a regular PCR using this number of cycles will be performed to amplify the libraries. These two steps are essential to prevent over-amplification of libraries, which at this stage could bias final results. For example, selective amplification of shorter cDNA fragments (those derived from spliced mRNA transcripts) over longer cDNA fragments (those derived from unspliced mRNA transcripts) can produce sequencing results that overestimate the amount of normal splicing that has occurred.

The primers used throughout this enrichment step are complementary to the head and tail regions of the original gBlocks oligonucleotides designed in Basic Protocol 2. Therefore, this amplification will select for cDNA molecules derived from transcripts containing the original exon-intron-exon test sequences.

1. Using a MicroAmp Fast Optical 96-well Reaction Plate with Barcode, plate separate PCR reactions for each purified cDNA sample as outlined in Table 6. Plate PCR reactions in duplicate. Pipette up and down to mix. For best practices, also plate a negative control containing nuclease-free water instead of cDNA.
2. Seal the plate using an Optical Adhesive Cover and insert this plate into the machine.
3. Start the 7500 Fast Real-Time PCR system and open the software program on the computer. Full instructions on use of the instrument and software can be found in the manual for the system. Under “New Experiment”, name the file and then choose the following settings:

Instrument: 7500 Fast (96 wells)

Experiment: Quantitation/Standard Curve

Detection Reagent: SYBR Green (No melt curve required)

Ramp Speed: Standard (about ~2 hours to complete run).

Designate the location of samples and identify these wells as targets in the program. Select the option where Passive Reference = None.

4. Adjust the run settings to the following:

PCR Settings

98 °C 30sec

98 °C 10sec 56 °C 20sec 72 °C 1 min [collect data here] (× 35 cycles)

72 °C 10min

4 °C Infinite

5. Start the PCR experiment.

6. Once complete, open and view the multicomponent plots for each sample. To determine the number of PCR cycles that allows for optimal amplification, identify the cycle number that corresponds to final segment of linear phase of the multicomponent plot (Figure 8). Above this cycle number, amplification is less efficient and may introduce biases into our final experimental result. In our experience, this number tends to be anywhere from 14 to 20 cycles. The products of this qPCR experiment can be discarded.

7. Use this cycle number determined for each sample to appropriately amplify the purified cDNA constructs. The PCR reactions should be plated as indicated in Table 7. This is an identical reaction as the qPCR just completed, except the SYBR Green is now replaced with an equivalent volume of water. To maximize yield of product, run 8 identical PCR reactions in parallel (i.e., 8 tubes of an identical 25 uL PCR reaction). Repeat for each cDNA sample.

8. Pipette up and down to mix, seal tubes, and place the samples in a thermocycler with the following settings:

PCR Settings

98 °C 30sec

98 °C 10sec 56 °C 20sec 72 °C 1 min (repeat to reach cycle number from qPCR)

72 °C 10min

4 °C Infinite

9. Once PCR is complete, combine identical reactions together. Purify with 3× AMPure XP Beads (600 uL) as per manufacturer's instructions. Elute in 100 uL of 10mM Tris Acetate.

10. Check the quality of the experiment at this point by running each sample on a TapeStation D1000 ScreenTape assay. A series of bands between 200bp and 500 bp for each sample suggests a good quality experiment (Figure 9). These fragments

correspond to amplified cDNA derived from spliced (< 500 bp bands) and unspliced (500 bp bands) transcripts for the original test sequences.

End Repair, A-Tailing, Adapter Ligation of Libraries

A series of steps is required to prepare libraries of DNA fragments for multiplexed paired-end sequencing on the Illumina sequencing platform.

11. The reagents for end repair of the current DNA library are contained within the END-IT kit. To each 100 uL sample of DNA, add reagents as specified in Table 8. Pipette up and down to mix.
12. Incubate samples at room temperature for 45 minutes
13. Purify with 3× AMPureXP beads (450 uL) and elute in 40 uL of 10 mM Tris Acetate.
14. The A-tailing reaction is performed with Klenow Fragment (3'→5' exo-). To each 40 uL sample of end-repaired cDNA, add the reagents as specified in Table 9. Pipette up and down to mix.
15. Incubate the samples at 37 °C in a thermocycler for 60 minutes.
16. Prepare 1× Quick Ligase Buffer by diluting 2× Quick Ligase Buffer with an equivalent volume of nuclease-free H₂O.
17. Purify A-tailing products with 3× AMPureXP beads (150 uL) and elute samples into 20 uL of 1× Quick Ligase Buffer.
18. Illumina Multiplexing Adapter ligation is performed using the Quick Ligation Kit. To each 20 uL sample of A-tailed cDNA, add the reagents as specified in Table 10. Pipette up and down to mix.
19. Incubate samples at room temperature for 20 minutes.
20. Purify with 1.5 AMPureXP beads (60 uL) and elute into 30 uL of 10 mM Tris Acetate.

Final Amplification

A final amplification step indexes each individual sample by applying a barcode within each sequencing adapter. This allows for samples to be distinguished and sorted during data analysis. A qPCR is again used to determine the number of cycles at which the reaction approaches saturation. This is followed by a PCR that uses the number of cycles identified in the qPCR step.

21. Using a MicroAmp Fast Optical 96-well Reaction Plate with Barcode, plate separate PCR reactions for each adapter-ligated cDNA sample as outlined in Table 11. Plate PCR reactions in duplicate. Pipette up and down to mix. For best practices, also plate a negative control containing nuclease-free water instead of cDNA.
22. Seal the plate using an Optical Adhesive Cover and insert this plate into the machine.

23. Start the 7500 Fast Real-Time PCR system and open the software program on the computer. Full instructions on use of the instrument and software can be found in the manual for the system. Under “New Experiment”, name the file and then choose the following settings:

Instrument: 7500 Fast (96 wells)

Experiment: Quantitation/Standard Curve

Detection Reagent: SYBR Green (No melt curve required)

Ramp Speed: Standard (about ~2 hours to complete run).

Designate the location of samples and identify these wells as targets in the program. Select the option where Passive Reference = None.

24. Adjust the run settings to the following:

PCR Settings

98 °C 30sec

98 °C 10sec 61 °C 30sec 72 °C 1 min [collect data here] (x 35 cycles)

72 °C 5min

4 °C Infinite

25. Once complete, open and view the multicomponent plots for each sample. As before, identify the cycle number that corresponds to final segment of linear phase of the multicomponent plot (Figure 8). In our experience, this number tends to be anywhere from 4 to 10 cycles. The products of this qPCR may now be discarded.

26. Use this cycle number determined for each sample to amplify the adapter-ligated cDNA. The PCR reactions should be plated as indicated in Table 12. This is an identical reaction as the qPCR just completed, except the SYBR Green is now replaced with an equivalent volume of water. To maximize yield of product, run 8 identical PCR reactions in parallel (i.e. 8 tubes of an identical 25 uL PCR reaction). Repeat for each individual library.

27. Pipette up and down to mix, seal tubes, and place the samples in a thermocycler with the following settings:

PCR Settings

98 °C 30sec

98 °C 10sec 61 °C 30sec 72 °C 1 min (repeat to reach cycle number from qPCR)

72 °C 5min

4 °C Infinite

28. Once PCR is complete, combine identical reactions together. Purify with 1.2× (240 uL) AMPure XP Beads and elute in 20 uL of 10mM Tris Acetate.

29. Check the quality of the individual libraries by running each sample on a TapeStation D1000 ScreenTape assay. A series of bands between 200bp and 700bp for each sample suggests a good quality experiment (Figure 10).

Pooling Samples and Submitting Libraries for Illumina Sequencing

30. On the TapeStation software, calculate the region molarity for the region between 200bp and 700bp for each sample.

31. Pool the individual samples together into a combined sequencing library such that all libraries have an equimolar concentration in the final combined pool (i.e. if Library 1 has half the molarity of Library 2, then add twice as much volume of Library 1 as Library 2 in the final combined sequencing library). This ensures equal representations of all libraries in the final pool.

32. Submit the library for Illumina MiSeq sequencing. As part of the order form, be sure to name the libraries and indicate the indexing primer used for that library in the final amplification. Additionally, ensure that the sequencing parameters listed in Table 13 are selected.

BASIC PROTOCOL 8: Analysis of Fastq Files and Interpretation of Result

The sequencing data from the Illumina MiSeq sequencing experiment are contained within the output fastq files. This section will describe the procedures required to successfully analyze these files using a computational script. First, analysis input files (“input frame”) which contain the relevant information for each test sequence will be created. This is followed by analysis of fastq files using a perl script. Lastly, instructions are provided on how to properly interpret the result.

Materials

- Construct Designer Output File created from Basic Protocol 2.
- Computing cluster with perl, R, and FLASH (<https://ccb.jhu.edu/software/FLASH/>) modules loaded.
- Statistics::R module (<http://search.cpan.org/~gmpassos/Statistics-R-0.02/lib/Statistics/R.pm>) loaded for perl.
- Script to make an “input frame” file.
Please download Make.inputframe.3.pl (available at https://GitHub.com/SplicingVariant/SplicingVariants_Beta) and save to computing environment.
- Script to analyze a sequencing result.
Please download SpliceConstructSearchGrepV1.6.pl (available at https://GitHub.com/SplicingVariant/SplicingVariants_Beta) and save to computing environment.
- A pair (R1 and R2) of fastq files from MiSeq run.

1. Prior to all subsequent steps, please ensure that perl, R, FLASH, and the Statistics::R module for perl are loaded into the computing environment.
2. Identify the Construct Designer Output Files created from Basic Protocol 2. There should be a 5' splice site output file and a 3' splice site output file.
3. From the command line, enter the following and run the script to make an "input frame" file for each construct designer output file.

```
perl Make.inputframe.3.pl [name of 5' construct design output file] > [name of input frame output file]
```

```
perl Make.inputframe.3.pl [name of 3' construct design output file] > [name of input frame output file]
```

4. Review the input frame output files created. Each file will contain the construct name, the unique sequence ID for that construct, and several sequences that represent the presence of the allele, an unspliced read, normally spliced read, and aberrantly spliced read. The subsequent script will use this file and the sequences contained within it to search for the number of reads that are unspliced, normally spliced, and aberrantly spliced.

5. Combine these two files into a single input frame file. Please keep a header line in the combined file.

Note: If libraries were indexed prior to sequencing, this combined input frame file will need to be divided into several smaller input frame files, one for each indexed library (which will have its own pair of fastq files). Each input frame would then contain the header row followed by all the constructs contained within that indexed library.

6. Once the input frames are created for each pair of fastq files, enter the following and run the SpliceConstructSearchGrepV1.6.pl script to analyze fastq files and obtain the results of the experiment. Additional modifiers can be found online on the GitHub webpage.

```
perl SpliceConstructSearchGrepV1.6.pl -inputframe [name of input frame file] -output [name of output file] -R1 [path to fastq R1 file.fastq] -R2 [path to fastq R2 file.fastq]
```

7. If the Statistics::R module was successfully loaded in the environment, there will be a file in the format of "[basename of output file].stats.[file name extension of output file]" in which p-values for splice alteration are calculated. This is the most important output file and the one that will reveal the results of the experiment. Review this output file created by this command. A short explanation of the most relevant columns in this output file are listed in Table 14.

8. Repeat this command to analyze additional input frames and additional indexed libraries.

Interpretation of the Final Output File

Upon reviewing the output file, you should first notice that the first 8 columns of the output file are the same as the 8 columns that comprise the input frame. The SpliceConstructSearchGrepV1.6.pl script adds additional columns in this file, many of which are extraneous and beyond the scope of this unit. In Table 15, the same constructs which were designed in Basic Protocol 2 are shown along with a subset of the most relevant columns from the output file. Interpretation of the analysis are as follows.

chr1:156100609T>G—During Basic Protocol 1, this variant was selected due to a significant MaxEnt score change that indicated the possible creation of a 5' splice site. The first two rows of Table 15 show the results from the reference test sequence (REF_chr1_156100609T_G) and the alternate test sequence (ALT_chr1_156100609T_G) for this variant. For the reference test sequence, a total of 4689 total reads were identified in the fastq files. Of these, 4011 were normally spliced, 680 were not spliced, and 7 were aberrantly spliced. The percentages of these three groups are shown in the adjacent columns. For the alternate test sequence, a very different pattern is observed. There were 10262 reads identified from this construct in the fastq files, of which the vast majority (8970) were aberrantly spliced due to the creation of a new 5' splice site. A very small number were normally spliced (546). The p-value contained in the final column of the table is calculated after normalizing the distribution of reads in the reference group and alternate group to 100 (i.e. calculating percentages) and then comparing the distribution using a two-sided Fisher's exact test. The p-value is always printed on the line with the alternate sequence. In this case, the p-value for the comparison is 5.35E-45.

chr1:156105681C>G—In Basic Protocol 1, this variant was selected because of a MaxEnt score change that predicted the loss of a 3' acceptor site. In Table 15, the experiment demonstrated that 79.5% of reads derived from the reference test sequence (REF_chr1_156105681C_G) were normally spliced, whereas only 1.9% of the reads from the alternate construct (ALT_chr1_156105681C_G) were normally spliced. In the alternate construct, the vast majority of reads (95.4%) were not spliced. The p-value by Fisher's exact test is 1.44E-34 and can be found in the row containing the alternate construct data.

chr1:156106824G>C—In Basic Protocol 1, this variant was selected as a candidate variant leading to loss of the 5' donor site. In Table 15, the distribution of reads reveals that the reference test sequence (REF_chr1_156106824G_C) splices normally in 96.9% of cases. The alternate test sequence (ALT_chr1_156106824G_C) splices normally in 17.4% of cases. The p-value for comparison between the distribution of splice patterns is 5.53E-34 and can be found in the row containing the alternate construct data.

chr1:156096389A>G—This variant was selected due to a score change that predicted the gain of a novel 3' acceptor site. In Table 15, we see that the majority of reads derived from both the reference and alternate test sequence are normally spliced. There is almost no aberrant splicing observed in either case, and the distribution of splice patterns is the same. The p-value comparing the two distributions is 1.

9. In general, three results are possible for any variant assessed in the minigene experiment. First, a variant can be found to alter splicing. In our experience, a p-value of <0.001 represents a reliable threshold for significance. The first three variants in the above example were found to alter splicing. Second, a variant can be found to have no effect on splicing. This scenario was demonstrated by the fourth variant in the above discussion. Lastly, an experiment can be inconclusive. This occurs when the reference test sequence exhibits insufficient levels of normal splicing (i.e. $<10\%$ of all reads). This makes it especially difficult to prove splice site loss as a result of the variant. In these cases, though there may be a decrease in normal splicing, the p-value is unable to reach the threshold for significance. We prefer to consider these cases as inconclusive rather than determining that the variant has no effect.

SUPPORT PROTOCOL 1: AMPure XP Bead Cleanup

The Agencourt AMPure XP PCR purification system utilizes paramagnetic bead technology for high-throughput purification of PCR amplicons. Excess primers, nucleotides, salts, and enzymes can be removed through this support protocol. This purification step is used numerous times throughout this unit.

Materials

Input samples

Agencourt AMPure XP Beads (Beckman Coulter Product #A63882)

Depending on number of samples:

Agencourt SPRIPlate 96R - Ring Super Magnet Plate (Beckman Coulter Product #A32782), or

DynaMag-2 Magnet (ThermoFisher Scientific Cat No. 12321D)

80% Ethanol

10 mM Tris-Acetate

1. Allow AMPure XP Beads to warm to room temperature prior to use.
2. Add the specified quantity of AMPure XP beads to each sample. As an example, $1.8\times$ AMPure beads indicates 1.8 uL of beads for each 1.0 uL of sample.
3. Vortex samples or pipette up and down to mix. Gently spin down contents to bottom of tube if needed. Note: if spinning, be careful to avoid precipitation of beads to bottom of tube.
4. Incubate at room temperature for 8-10 minutes.
5. Place samples on magnet. Wait 2 minutes or until solution clears.
6. Leaving samples on magnet, discard clear solution. Do not disturb beads.
7. Apply 80% ethanol to the samples while still on the magnet.
8. Incubate for 30 seconds at room temperature.

9. Discard ethanol and reapply 80% to the samples for a second wash.
10. Incubate for 30 seconds at room temperature.
11. Discard ethanol.
12. Remove samples from magnet and dissolve beads into elution volume of 10 mM Tris Acetate. Vortex samples or pipette up and down to mix. Gently spin down contents to bottom of tube if needed.
13. Incubate at room temperature for 8-10 minutes.
14. Place samples on magnet. Wait 2 minutes or until solution clears.
15. Transfer the clear solution to a new tube. This is the purified product. The remaining beads may be discarded.

COMMENTARY

Background Information

Understanding the phenotypic consequences of rare variation is one of the key challenges in genetic medicine. Recent estimates suggest about one-third of all disease-causing mutations may alter RNA splicing, yet many of these remain classified as variants of unknown significance (Lim, Ferraris, Filloux, Raphael, & Fairbrother, 2011). In this unit, we describe a computational tool that prioritizes putative splice-altering sequence variants, and a moderate-throughput minigene assay that confirms the variants which alter splicing. Previous attempts at uncovering these splicing mutations relied on cloning of patient DNA and insertion of this template into a minigene or exon trapping vector. The development of splice prediction software and advent of solid-phase oligonucleotide synthesis technology has allowed for the creation of a protocol that completes the same job on a larger scale with greater efficiency.

The method described here has several important limitations. First, mutations affecting splicing are generally of two types: those that alter splice site strength and those that modify *cis*-acting elements (i.e. splicing silencers, enhancers). This unit is limited to sequence variants which affect the former. Second, our minigenes containing 500 bp of an exon-intron-exon test sequence only allow for comparison of normal splicing rates between the reference and alternate allele. While therefore able to detect deviations from normal splicing, it is unable to predict exon skipping versus intron inclusion in the full-length transcript. Lastly, the test sequences in this assay demonstrate insufficient normal splicing in approximately 5% of cases. We expect this is due to further features of splice site recognition that are not contained within 500 bp of surrounding genomic context.

In spite of these limitations, we estimate that the application of this protocol will increase our detection of protein-truncating variants by approximately 10-15% in most human genes. These newly identified variants will provide a significant advance to current efforts at rare variant interpretation and will be especially relevant to human genes which cause disease by haploinsufficiency.

Critical Parameters

There are two distinct segments in this protocol: a) Computational identification of variants that have a significant likelihood to alter RNA splicing and b) In vivo assessment of the RNA splicing consequences of DNA sequence variation. An essential feature for successful in vivo assessment of RNA splicing is that the 'reference construct' should produce an RNA that undergoes splicing when introduced into cultured cells. We recommend an ~500 bp exon-intron-exon fragment for in vivo splice assays. Previous studies have employed ~250 bp test sequences. In our experience 50% of these shorter test sequences fail to splice, while the ~500 bp test sequences splice >90%. If the ~500 bp test sequence fails to splice, constructing larger test sequences will generally solve this problem.

Troubleshooting

A commonly encountered problem is inefficient amplification in PCR #2 (Basic Protocol 3). From our experience, this is because IDT oligonucleotide synthesis technology is unable to provide large amounts of the CMVp-N500 gene fragment. We present two options to circumvent this issue. The first is to obtain this gene fragment from another supplier, which may allow for greater synthesis yield. The second is to amplify the CMVp-N500 gene fragment using PCR as described in Table 16. PCR settings are as follows.

98C 30 sec

98C 10sec 63C 30sec 72C 15sec (× 20 cycles)

72C 10min

4C Infinite

The PCR product should be cleaned using 1.8× AMPure XP beads and eluted into a desired volume of 10mM Tris-Acetate. This new CMVp-N500 can be used for the PCR #2 reaction.

Statistical Analyses & Understanding Results

Please refer to Basic Protocol 8 for statistical comparisons used in this unit and the interpretation of all final results.

Time Considerations

The time considerations for each basic protocol in this unit is largely dependent on the number of variants needing to be assessed. For a group of 25 variants requiring assessment, our estimate is as follows:

Basic Protocol 1: MaxEnt Scoring to Select Candidate Variants | 1-2 hours

Basic Protocol 2: Design Constructs for Minigene Assay | 1-2 hours

Basic Protocol 3: Assembly PCR to Build Full Minigenes | 1 day

Basic Protocol 4: Preparation and Transfection of HEK293 cells | 3-4 days (cell growth)

Basic Protocol 5: RNA Extraction | 4 hours

Basic Protocol 6: cDNA Synthesis | 6-7 hours

Basic Protocol 7: Library Preparation for Illumina MiSeq | 2 days

Basic Protocol 8: Analysis of Fastq Files and Interpretation of Result | 1 day

Acknowledgments

This research was supported by grants from the National Heart, Lung, and Blood Institute of the National Institutes of Health (1R01HL080494 and 1R01HL084553, C.E.S.), the Fondation Leducq (C.E.S.), and the Howard Hughes Medical Institute (C.E.S). K.I. was supported by the Banyu Fellowship Program and the Uehara Research Fellowship Program. P.N.P. was supported by a fellowship from the Sarnoff Cardiovascular Research Foundation.

LITERATURE CITED

- Cooper TA. Use of minigene systems to dissect alternative splicing elements. *Methods*. 2005; 37(4): 331–340. DOI: 10.1016/j.ymeth.2005.07.015 [PubMed: 16314262]
- Gaildrat P, Killian A, Martins A, Tournier I, Frebourg T, Tosi M. Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. *Methods Mol Biol*. 2010; 653:249–257. DOI: 10.1007/978-1-60761-759-4_15 [PubMed: 20721748]
- Houdayer C, Caux-Moncoutier V, Krieger S, Barrois M, Bonnet F, Bourdon V, Stoppa-Lyonnet D. Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. *Hum Mutat*. 2012; 33(8):1228–1238. DOI: 10.1002/humu.22101 [PubMed: 22505045]
- Ito K, Patel PN, Gorham JM, McDonough B, DePalma SR, Adler EE, Seidman JG. Identification of pathogenic gene mutations in LMNA and MYBPC3 that alter RNA splicing. *Proc Natl Acad Sci U S A*. 2017; 114(29):7689–7694. DOI: 10.1073/pnas.1707741114 [PubMed: 28679633]
- Jian X, Boerwinkle E, Liu X. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet Med*. 2014; 16(7):497–503. DOI: 10.1038/gim.2013.176 [PubMed: 24263461]
- Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, Cooper DN. Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum Mutat*. 2007; 28(2):150–158. DOI: 10.1002/humu.20400 [PubMed: 17001642]
- Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci U S A*. 2011; 108(27):11093–11098. DOI: 10.1073/pnas.1101135108 [PubMed: 21685335]
- Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Rehm HL. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015; 17(5):405–424. DOI: 10.1038/gim.2015.30 [PubMed: 25741868]
- Scotti MM, Swanson MS. RNA mis-splicing in disease. *Nat Rev Genet*. 2016; 17(1):19–32. DOI: 10.1038/nrg.2015.3 [PubMed: 26593421]
- Shapiro MB, Senapathy P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res*. 1987; 15(17):7155–7174. [PubMed: 3658675]
- Soemedi R, Cygan KJ, Rhine CL, Wang J, Bulacan C, Yang J, Fairbrother WG. Pathogenic variants that alter protein code often disrupt splicing. *Nat Genet*. 2017; 49(6):848–855. DOI: 10.1038/ng.3837 [PubMed: 28416821]
- Vreeswijk MP, Kraan JN, van der Klift HM, Vink GR, Cornelisse CJ, Wijnen JT, Devilee P. Intronic variants in BRCA1 and BRCA2 that affect RNA splicing can be reliably selected by splice-site prediction programs. *Hum Mutat*. 2009; 30(1):107–114. DOI: 10.1002/humu.20811 [PubMed: 18693280]
- Ward AJ, Cooper TA. The pathobiology of splicing. *J Pathol*. 2010; 220(2):152–163. DOI: 10.1002/path.2649 [PubMed: 19918805]
- Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*. 2004; 11(2-3):377–394. DOI: 10.1089/1066527041410418 [PubMed: 15285897]

Figure 1. Splice sites consist of 9 bp splice donor sequences and 23 bp splice acceptor sequences
 A sample gene region containing exons and introns is shown. Splice sites reside at the exon-intron junctions of all human genes. At the 5' end of each intron is a 9 bp splice donor sequence containing an invariant GT dinucleotide. At the 3' end of each intron is a 23 bp splice acceptor site containing an invariant AG dinucleotide. These two signals are essential for proper exon recognition during splicing. The size of each base within the sequence represents the level of nucleotide conservation at each position within the splice site across the genome.

Figure 2. Unit flowchart illustrating stepwise strategy to prioritize and assess sequence variants for splice disruption

There are 8 basic protocols described in this unit, intended to be performed sequentially. Basic Protocols 3 through 7 require laboratory work. The remainder of the steps can be performed computationally.

Figure 3. A single variant can change the splice donor site strength of numerous independent nine base pair sequences

Shown in this image is a variant (red) located within an arbitrary exon. The variant is a component of 9 independent 9 bp sequences, and the presence of this variant alters the splice site strength for each. Of the nine sequences that have been altered, only sequence 4 is found to now have strong homology to the splice donor. This change could possibly lead to a novel splice site at this position.

Figure 4. Two PCR reactions assemble a functional minigene test construct

In PCR #1, a 500 bp test sequence and 225 bp SV40PA sequence are added to the same reaction to form a test sequence-SV40PA combined construct. In PCR #2, the PCR #1 product and a 600 bp CMV Promoter are added to the same reaction to form the full minigene construct. Overlapping, homologous regions of these gene fragments allow for assembly to occur.

Gel Image

Figure 5. TapeStation D1000 example for PCR #1

15 lanes containing independent PCR #1 products are shown. Each band is approximately 650-700 bp in size, and represents an individual test sequence (500 bp) attached to a SV40PA signal (225 bp). Note that products seem to be migrating to the same positions in pairs, as would be expected with reference and alternate containing test sequences that have identical lengths.

Gel Image **Figure 6. TapeStation D5000 example for PCR #2**

15 lanes containing independent PCR #2 reactions are shown. The relevant band representing the full length minigene is approximately 1225 bp in length, and is the primary product from the reaction.

Gel Image

Figure 7. TapeStation RNA example for RNA extraction
5 samples originating from 5 independent transfection reactions are shown. After 1:20 dilution, samples demonstrate high RNA Integrity Numbers (RIN).

Figure 8. Multicomponent plot for enrichment of test sequence qPCR

A multicomponent plot depicts the exponential fluorescence observed during the qPCR reaction. A single sample and negative control, both plated in duplicate, are highlighted. For the sample, the linear phase of the multicomponent plot is shown to level off after cycle 16.

Gel Images

Figure 9. TapeStation D1000 example for cDNA post enrichment amplification

A gel image of 5 libraries following enrichment amplification is shown. The presence of two main features are indicative of a good-quality library. First, the vast majority of bands in each library are 200 to 500 bp in size. Second, each library appears distinct, as would be expected if test sequences within each library were unique.

Gel Image

Figure 10. TapeStation D1000 example for final amplification library pools

A gel image of 12 libraries following final indexing amplification is shown. Again, two features are indicative of a good-quality library. The first is that the vast majority of bands are 200 to 700 bp in size. Second, the gel image for each library should appear similar to the gel image after enrichment amplification (see Figure 9). The slightly larger size of the same library can be explained by the ligation of sequencing adapters between these steps.

Table 1

Explanation of Column Headers in Raw Output and Summarized Result Files

Column Header	Interpretation
Var_name	Variant name as entered in mutfile.txt
Exon	Exon or Intron number along the selected transcript
Max.Exon	Total number of exons in selected transcript
GENE	Gene corresponding to transcript
ID	ENSEMBL Transcript ID
chr	Defines start position of 9 bp or 23 bp sliding window
phypos	Defines start position of 9 bp or 23 bp sliding window
Ref_MaxEntScoreSS5	5' splice donor MaxEnt score for 9 bp sequence starting at this defined position (chr and phypos). Uses the reference base at the position of the mutation.
Ref_SS5seq	9 bp sequence starting at defined position for which score was calculated. Uppercase/lowercase represent the exon/intron boundary for this splice site. Uses the reference base at the position of the mutation
Alt_MaxEntScoreSS5	5' splice donor MaxEnt score for 9 bp sequence starting at this defined position. Uses the alternate base at the position of the mutation.
Alt_SS5seq	9 bp sequence starting at defined position for which score was calculated. Uppercase/lowercase represent the exon/intron boundary for this splice site. Uses the alternate base at the position of the mutation
Diff_MaxEntScoreSS5	Difference in MaxEnt score (Alt - Ref) for the 9 bp sequences without (Ref_SS5seq) and with (Alt_SS5seq) the variant. Negative scores indicate variant damages the strength of sequence as functioning splice site. Positive scores indicate variant increases strength of sequence as a functioning splice site.
Ref_MaxEntScoreSS3	3' splice acceptor MaxEnt score for 23 bp sequence starting at this defined position (chr and phypos). Uses the reference base at the position of the mutation.
Ref_SS3seq	23 bp sequence starting at defined position for which score was calculated. Lowercase/uppercase represent the intron/exon boundary for this splice site. Uses the reference base at the position of the mutation
Alt_MaxEntScoreSS3	3' splice acceptor MaxEnt score for 23 bp sequence starting at this defined position. Uses the alternate base at the position of the mutation.
Alt_SS3seq	23 bp sequence starting at defined position for which score was calculated. Lowercase/uppercase represent the intron/exon boundary for this splice site. Uses the alternate base at the position of the mutation
Diff_MaxEntScoreSS3	Difference in MaxEnt score (Alt - Ref) for the 23 bp sequences without (Ref_SS3seq) and with (Alt_SS3seq) the variant. Negative scores indicate variant damages the strength of sequence as functioning splice site. Positive scores indicate variant increases strength of sequence as a functioning splice site.
Kind	In SigLines file, indicates the type of splice change predicted. 5SS_Loss = 5' donor loss, 5SS_Gain = 5' donor gain, 3SS_Loss = 3' acceptor loss, 3SS_Gain = 3' acceptor gain.
Possibility	In SigLines file, ranks the probability that this variant will alter splicing in the subsequent assay. Grades are ranked using terminology No/Low/Mid/High. In our experience, No and Low designations are unlikely to alter splicing.
Decision	In SigLines file, recommends whether the variant should be carried forward to the next steps of the protocol.

Table 2

Interpreting the Scoring of High Priority Variants (Decision = YES) in the SigLines File

Var_name	Ref_ MaxEnt Score SS5	Alt_ MaxEnt Score SS5	Diff_ MaxEnt Score SS5	Ref_ MaxEnt Score SS3	Alt_ MaxEnt Score SS3	Diff_ MaxEnt Score SS3	Kind	Possibility	Decision
chr1:156100609T>G	-2.91	7.29	10.20	-11.31	-13.67	-2.36	5SS_Gain	High	YES
chr1:156105681C>G	-23.47	-999	-999	9.23	2.91	-6.32	3SS_Loss	High	YES
chr1:156106824G>C	11.45	5.33	-6.12	-99	-99	0	5SS_Loss	High	YES
chr1:156096389A>G	-26.80	-17.82	8.98	8.46	9.08	0.62	3SS_Gain	High	YES

Table 3

gBlocks and other primers used in laboratory protocols

Item Name	Sequence	Comments
CMVp-N500	TAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTC CGCGTTACATACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCC CGCCCATTTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGAC TTTCCATTGACGTCAATGGGTGGAGTATTACGGTAAACTGCCCACTTGGCAGTA CATCAAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAA TGGCCCGCCTGGCATTATGCCAGTACATGACCTTATGGGACTTTCCTACTT GGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTGG CAGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCGAAGTCTCCA CCCCATTGACGTCAATGGGAGTTTGTGTTTGGCACCAAAATCAACGGGACTTCCAAAAT GTCGTAACAACCTCCGCCCATTTGACGCAAATGGGCGGTAGGCGGTGACGGTGGGAGGT CTATATAAGCAGAGCTGGTTTGTAGTGAACCGTCAGTTACGCCAAGTTATTTAGGTGACA	Order as IDT gBlock
SV40PA-N500	CATAATCAGCCATACCACATTTGTAGAGGTTTTACTTGCTTTAAAAA CCTCCACACCTCCCCCTGAACCTGAAACATAAAATGAATGCAATTGTTGTTGT TAACCTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGCATCACAA ATTTACAAATAAAGCATTTTTTCACTGCAITCTAGTTGTGGTTTGTCCAAACT CATCAATGTATCTTA	Order as IDT gBlock
CMVp-N500F	TAGTAATCAATTACGGGGTCATTAGTTCATA	Primer used in PCR #2
CMVp-N500R	TGTCACCTAAATAACTTGGCGTAACTG	Primer used in troubleshooting
SV40PA-N500R	TAAGATACATTGATGAGTTTGGACAAACC	Primer used in PCR #1 and PCR #2
N500-RTPCR-F2	ACGCCAAGTTATTTAGGTGACA	Primer used in PCR #1 and in library preparation
N500-RTPCR-R1	CAAATGTGGTATGGCTGATTATG	Primer used in cDNA synthesis and library preparation
Multiplexing Adapters	5' P - GATCGGAAGAGCACACGTCT 5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT	To anneal: Dilute each oligonucleotide to 200uM conc. Take the same volumes from each and mix. Then heat to 95C for 3min followed by gradually cooling down to 4C. This makes a 100 uM stock solution.
Multiplexing Primer 1.0	5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'	Order as primer for use in final amplification before MiSeq
Indexing Primer (Sample)	5' CAAGCAGAAGACGGCATACGAGATNNNNNGTGAAGTTCAGACGTGTGCTCTTCCGATCT 3'	"N" indicates barcode used to index each library.

Table 4

PCR #1 set up. Repeat for as many gBlocks as are required.

Reagent	Concentration	Amount	Notes
Test sequence gBlock construct	10ng/ul	1ul	Elute using Tris-EDTA
SV40PA-N500	10ng/ul	1ul	Elute using Tris-EDTA
SV40PA-N500R	10uM	2.5ul	
N500RTPCRF2	10uM	2.5ul	
Phusion2xMix (GC buffer)		10ul	Must be kept on ice. Do not vortex.
dH2O		3ul	
TOTAL		20 ul	

Table 5

PCR #2 set up. Repeat for as many gBlocks as are required.

Reagent	Concentration	Amount	Notes
Individual PCR #1 Product	1ng/ul	3 ul	
CMVp-N500	1ng/ul	3 ul	Elute using Tris-EDTA
CMVp-N500F	10 uM	1.25 ul	
SV40PA-N500R	10uM	1.25 ul	
Phusion2xMix (GC buffer)		12.5 ul	Must be kept on ice. Do not vortex.
dH2O		4 ul	
Total		25 ul	

Table 6

qPCR set up for enrichment of test sequences. SYBR Green is used for detection of fluorescence.

Reagent	Concentration	Amount	Notes
Purified cDNA Product	Any	1 ul	
N500-RTPCR-F2	10uM	1.25 ul	
N500-RTPCR-R1	10uM	1.25 ul	
Phusion2xMix (GC Buffer)		12.5 ul	Must be kept on ice. Do not vortex.
dH ₂ O		8 ul	
DMSO		0.75 ul	
1:1,000 SYBR Green		0.25 ul	
Total		25 ul	

Table 7

PCR set up for enrichment of test sequences. This is an identical set up to Table 6, however, no SYBR Green is used and the equivalent volume is replaced by water.

Reagent	Concentration	Amount	Notes
Purified cDNA Product	Any	1 ul	
N500-RTPCR-F2	10uM	1.25 ul	
N500-RTPCR-R1	10uM	1.25 ul	
Phusion2xMix (GC Buffer)		12.5 ul	Must be kept on ice. Do not vortex.
dH2O		8.25 ul	
DMSO		0.75 ul	
Total		25 ul	

Table 8

End Repair of Amplified cDNA Product

Reagent	Concentration	Amount
Amplified cDNA Product		100 ul
End Repair 10x Buffer		15 ul
dNTPs	25 mM	15 ul
ATP	10 mM	15 ul
End Repair Enzyme Mix		3 ul
Total		148 ul

Table 9

A-Tailing of End Repaired cDNA

Reagent	Concentration	Amount (ul)
End Repaired cDNA Product		40 ul
NEBuffer 2 (10X)		5 ul
dATP	5 mM	2 ul
Klenow Fragment (3'-5' exo-)		3 ul
Total		50 ul

Table 10

Adapter Ligation of A-Tailed cDNA

Reagent	Concentration	Amount (ul)
A-Tailed cDNA Product		20 ul
Quick Ligase Reaction Buffer (2×)		14 ul
Multiplex Adapter (Illumina)	100 uM	2 ul
Quick Ligase Enzyme		2 ul
Total		38 ul

Table 11

qPCR for Final Indexing Amplification

Reagent	Concentration	Amount (ul)
Adapter-Ligated Product	Any	1
Multiplexing Primer (1.0)	25 uM	0.5
Indexing Primer *	25 uM	0.5
Phusion2xMix (GC Buffer)		12.5
dH ₂ O		10.25
1:1,000 SYBR Green		0.25
Total		25 ul

* This primer must vary for each sample if indexing is desired.

Table 12

Final Indexing Amplification PCR

Reagent	Concentration	Amount (ul)
Adapter-Ligated Product	Any	1
Multiplexing Primer (1.0)	25 uM	0.5
Indexing Primer *	25 uM	0.5
Phusion2xMix (GC Buffer)		12.5
dH2O		10.5
Total		25 ul

* This primer must vary for each sample if indexing is desired.

Table 13

Essential Parameters for Illumina MiSeq Order Form

Input Parameter	Selection
Number of Sequencing Reads	Determined by depth of coverage required. We prefer an average of 50,000 reads per test sequence contained within the final library pool. In our experience, a MiSeq V2 (10-20 million reads) works best.
Library Type	Amplicon/PCR Product
Library Category	DNA-seq
Organism	Homo Sapiens
Adapter-Type	Custom Paired End Adaptors
Min Size	200
Max Size	700
Read Length	Long (1-523 cycles)
Paired End vs. Single End	Paired end single index
Phi-X Spike	Yes (15%)
Read 1 Length	250 bp
Read 2 Length	250 bp
Index Length	6 bp

Table 14

Explanation of Column Headers in Final Output File

Column Header	Interpretation
Construct_Name	This is the name of the test sequence, as originally created in the gBlock order file in Basic Protocol 2.
ID	This 10 base pair sequence contains the 4 base pairs upstream of the barcode, the 2 bp barcode, and the 4 base pairs downstream of the barcode. It serves as a unique identifier for each construct.
Allele	This sequence contains the allele (reference base or alternate base) and the surrounding sequence context. The presence of this sequence in the fastq file indicates a read containing the reference base or variant base.
NoSplicing	This sequence contains the region surrounding the 3' acceptor splice site. Note that introns are lowercase and exons are uppercase. The presence of this sequence in the fastq file indicates a read that was not spliced at the normal position.
NormalSplicing	This sequence contains the end of the first exon and the start of the second exon, pasted together into a single sequence. The presence of this sequence in the fastq file indicates a read that was normally spliced.
AberrantSplicing	This column contains the sequence that corresponds to an aberrant splice product. If the variant was predicted to lead to a novel splice site, this contains the specific sequence that results when the predicted aberrant splice occurs.
N_AnalyzableReads	The number of reads that contained the 10 bp ID <u>and</u> also contained a sequence consistent with no splicing, normal splicing, or aberrant splicing. This is also the sum of N_NoSplice, N_NormalSplice, and N_AberrantSplice.
N_NoSplice	The number of reads that contained the 10 bp ID <u>and</u> also contained the sequence listed in the NoSplicing column.
N_NormalSplice	The number of reads that contained the 10 bp ID <u>and</u> also contained the sequence listed in the NormalSplicing column.
N_AberrantSplice	The number of reads that contained the 10 bp ID <u>and</u> also contained the sequence listed in the AberrantSplicing column.
Percent_NoSplice	The percent of reads which demonstrated no splicing.
Percent_NormalSplice	The percent of reads which demonstrated normal splicing.
Percent_AberrantSplice	The percent of reads which demonstrated aberrant splicing.
p_value	The p-value comparing the distribution of reads between the REF and ALT constructs. The p-value is calculated after normalizing all counts to 100 (i.e. using percentages) and then calculating a two-sided Fisher's exact test.
Var_Name	The initial variant for which the test sequence was designed.
REF_ALT	Identifies whether the data refers to the reference (REF) or alternate (ALT) construct.
Kind	Indicates whether the variant was predicted to lead to splice site loss (SS_Broken) or splice site gain (SS_Create)
TranscriptID	Indicates the ENSEMBL transcript used to design the test sequence.
Ref_MaxEntScore	The MaxEnt score of the reference sequence as first determined in Basic Protocol 2.
Alt_MaxEntScore	The MaxEnt score of the alternate sequence as first determined in Basic Protocol 2.
Diff_MaxEntScore	The difference in MaxEnt scores as a result of the variant.

Table 15

Interpreting the Final Output of the Minigene Experiment

Construct_Name	N_Analyzable Reads	N_No Splice	N_Normal Splice	N_Aberrant Splice	%_No Splice	%_Normal Splice	%_Aberrant Splice	p_value
REF_chr1_156105681C_G	34833	6085	27684	1064	17.5	79.5	3.1	1
ALT_chr1_156105681C_G	10067	9602	187	278	95.4	1.9	2.8	1.44E-34
REF_chr1_156096389A_G	273090	2972	270116	2	1.1	98.9	0	1
ALT_chr1_156096389A_G	127598	1387	126210	1	1.1	98.9	0	1
REF_chr1_156100609T_G	4698	680	4011	7	14.5	85.4	0.1	1
ALT_chr1_156100609T_G	10262	746	546	8970	7.3	5.3	87.4	5.35E-45
REF_chr1_156106824G_C	148947	2369	144384	2194	1.6	96.9	1.5	1
ALT_chr1_156106824G_C	34491	8103	6009	20379	23.5	17.4	59.1	5.53E-34

Table 16

Amplification of CMVp-N500

Reagent	Concentration	Amount (ul)	Notes
CMVp-N500	Any	1	
CMVp-N500F	10uM	1.5	See Table 2 for sequence
CMVp-N500R	10uM	1.5	
Phusion2xMix (GC)		12.5	
dH2O		9	
Total		25 ul	