Published in final edited form as:

Hum Mutat. 2014 October ; 35(10): 1249–1259. doi:10.1002/humu.22624.

Experimental Assessment of Splicing Variants Using Expression Minigenes and Comparison with In Silico Predictions

Neeraj Sharma¹, Patrick R. Sosnay^{1,2}, Anabela S. Ramalho³, Christopher Douville⁴, Arianna Franca¹, Laura B. Gottschalk¹, Jeenah Park¹, Melissa Lee¹, Briana Vecchio-Pagan¹, Karen S. Raraigh¹, Margarida D. Amaral³, Rachel Karchin², and Garry R. Cutting^{1,5,*}

¹McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland ²Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland ³Faculty of Sciences, BioFIG – Centre for Biodiversity, Functional and Integrative Genomics, University of Lisboa, Lisboa, Portugal ⁴Department of Biomedical Engineering, Institute for Computational Medicine, Johns Hopkins University, Baltimore, Maryland ⁵Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, Maryland

Abstract

Assessment of the functional consequences of variants near splice sites is a major challenge in the diagnostic laboratory. To address this issue, we created expression minigenes (EMGs) to determine the RNA and protein products generated by splice site variants ($n = 10$) implicated in cystic fibrosis (CF). Experimental results were compared with the splicing predictions of eight in silico tools. EMGs containing the full-length Cystic Fibrosis Transmembrane Conductance Regulator (*CFTR*) coding sequence and flanking intron sequences generated wild-type transcript and fully processed protein in Human Embryonic Kidney (HEK293) and CF bronchial epithelial (CFBE41o-) cells. Quantification of variant induced aberrant mRNA isoforms was concordant using fragment analysis and pyrosequencing. The splicing patterns of c.1585–1G>A and c.2657+5G>A were comparable to those reported in primary cells from individuals bearing these variants. Bioinformatics predictions were consistent with experimental results for 9/10 variants (MES), 8/10 variants (NNSplice), and 7/10 variants (SSAT and Sroogle). Programs that estimate the consequences of mis-splicing predicted 11/16 (HSF and ASSEDA) and 10/16 (Fsplice and SplicePort) experimentally observed mRNA isoforms. EMGs provide a robust experimental approach for clinical interpretation of splice site variants and refinement of in silico tools.

Keywords

expression minigene; splicing; *CFTR*; in silico tools

© 2014 Wiley Periodicals, Inc.

*Correspondence to: Garry R. Cutting, Johns Hopkins Medical Institutions, 733 North Broadway, MRB 559, Baltimore, MD 21205. gcutting@jhmi.edu.

Additional Supporting Information may be found in the online version of this article.

Disclosure Statement: The authors declare no conflict of interest.

Introduction

Pre-mRNA splicing requires introns to be correctly excised and exons to be precisely joined so as to generate mature mRNA. The nuclear ribonucleoprotein complex, known as the spliceosome, facilitates splicing by defining the native exon–intron boundaries [Hastings and Krainer, 2001]. A sizable fraction (10%–13%) of variants associated with Mendelian disorders are suspected of causing aberrant splicing as they occur in or near exon–intron boundaries [Krawczak et al., 2007]. The clinical importance of this class of variants extends beyond single gene disorders as complex disorders such as autism and psoriasis can also trace their etiology to variants that cause mis-splicing of RNA transcripts [Iossifov et al., 2012; Jordan et al., 2012]. Furthermore, variants near or within exon–intron boundaries are commonly identified in exome sequencing studies of patients with unidentified Mendelian and polygenic diseases [Sanders et al., 2012; Sankaran et al., 2012]. Finally, the introduction of next-generation sequencing in clinical laboratories is causing an explosion in the number of DNA variants identified in and around genes [Yang et al., 2013]. Unfortunately, interpreting the clinical implications of variants in or near splice sites is challenging as functional annotation of DNA variants in publically available databases is inadequate [Xue et al., 2012].

Interpretation of the clinical consequence of a splice site variant is determined by the mRNA isoforms produced and the protein synthesized from these transcripts. Aberrant splicing frequently causes a frameshift and the introduction of a premature termination codon (PTC), which in turn elicits nonsense-mediated mRNA decay (NMD) [Cartegni et al., 2002]. The end result of NMD in vivo is the loss of transcript and absence of protein [Frischmeyer and Dietz, 1999]. Alternatively, mis-splicing can alter the content but not the reading frame of the RNA transcript resulting in a stable protein that may have reduced or altered activity or an unstable protein that is degraded [Singh and Cooper, 2012]. Being able to accurately assess which, if any, protein is produced is of considerable clinical relevance. For example, complete loss of protein generally results in severe forms of recessive diseases, whereas the presence of mutated but partially functional protein can lead to moderate disease [Marden, 2008]. Mis-spliced transcripts can also encode proteins that gain function not present in the wild-type (WT) protein resulting in autosomal dominant [Petkovic et al., 2010], sporadic [Eriksson et al., 2003], or common complex disorders [Jordan et al., 2012].

The optimal method for determining whether a suspected variant impairs splicing is to analyze RNA samples directly from individuals carrying the variant [Acedo et al., 2012]. However, this approach has major limitations due to the limited accessibility of RNA samples from individuals carrying putative splicing variants. Furthermore, gene expression may be limited to tissues that are difficult to access (e.g., brain or pancreas) or if the expression is restricted to time frames when sampling cannot be performed (e.g., early development). Even if a sample can be obtained, the ubiquity of RNase enzymes requires samples to be processed using methods that preserve RNA integrity. Hybrid minigenes have been used as experimental tools for examining the effect of variants that alter RNA splicing [Cooper, 2005; MacArthur et al., 2014]. Hybrid minigenes comprising the splice site in question along with the portions of the flanking introns and upstream and downstream exon sequence can reproduce normal RNA splicing observed in vivo [Goina et al., 2011]. More

importantly, variants introduced into the splice sites faithfully replicate aberrant patterns of RNA splicing [Bonnet et al., 2008; Spurdle et al., 2008]. However, hybrid minigenes contain exons from a different gene so that the variant being tested is not in a native context [Fededa et al., 2005; Baralle et al., 2006]. Bioinformatics programs that predict the effect of DNA sequence variations upon splicing provide an alternative to experimental methods. These *in silico* tools have been used extensively in computational searches to identify exon–intron boundaries and define gene structures [Burge and Karlin, 1997; Korf et al., 2001]. Currently, a major application of these tools is to predict the impact of variations at known splice sites and their nearby regions [Jian et al., 2013]. However, *in silico* tools have limitations due to the arbitrary nature of threshold scores, challenges in prioritizing one tool over the other, and lack of reliable standard interpretation guidelines that reduce effectiveness in routine clinical practice.

To address these issues, we created expression minigenes (EMGs) composed of an entire open reading frame (i.e., cDNA) into which abridged or full length introns were incorporated. In the case of an abridged intron, inclusion of at least 200–300 bp intron sequences each from the 5′ splice site and 3′ splice site ensured that the majority of the regulatory signals necessary for constitutive and alternative splicing were present [Cooper, 2005; Yeo et al., 2007]. A major advantage of EMGs is that the effect of a variant on transcription and translation can be simultaneously assayed. Thus, EMGs can address the issue of NMD and the generation of truncated protein. In this study, we demonstrate the utility of EMGs in assessing the effect of splicing variants in the cystic fibrosis transmembrane conductance regulator gene (*CFTR*; MIM #602421; GenBank NM_000492.3) of cystic fibrosis (CF; MIM #219700) patients; and compared with *in silico* predictions.

Materials and Methods

Reference Sequence Nomenclature and Selection of *CFTR* Variants

Recommendations of the Human Genome Variation Society (<http://www.hgvs.org/mutnomen/>) were followed for exon/intron numbering and mutation name. The DNA mutation numbering system is based on cDNA using +1 as the A of the ATG translation initiation codon in the *CFTR* reference sequence (NM_000492.3) with initiation codon as codon 1 and exons are numbered 1–27. To evaluate the comparison of EMG assays with *in silico* tools for splicing defect prediction, 10 *CFTR* variants were selected as follows. Two variants (c.1585–1G>A and c.2657+5G>A) that had been previously studied using RNA from nasal epithelia cells of individuals carrying these variants were specifically selected to enable comparison of EMG results with *in vivo* results [Hull et al., 1993; Highsmith, Jr. et al., 1997]. Four variants of unknown functional effect (c.1585–8G>A, c.2657+2_2657+3insA, c.2988G>A, c.2988+1G>A) exceeding a frequency of 0.01 in the CFTR2 project [Sosnay et al., 2013] were analyzed to assist in annotation of their disease liability. Finally, four additional naturally occurring variants (c.1585–2A>G, c.1585–3T>G, c.1585–9T>A, and c.2657+3delG) from the same splice site region were chosen to assist in the interpretation of the functional effects and to provide additional experimental data for evaluation of the splicing algorithms.

Creation of EMGs

EMGs were created by inserting either partial 5' and 3' intron sequences or a complete intron into a plasmid harboring full length *CFTR* cDNA as described by Sosnay et al. (2013). Details are provided for introduction of sequence from intron 11 as the process is similar for introduction of additional intron sequence (Fig. 1). The first 326 bp of 5' and the last 320 bp of 3' intron 11 along with flanking exons were PCR amplified from 200 ng of genomic DNA using 1.5 mM MgSO₄, 0.2 mM dNTP (each), 0.3 μM forward and reverse primer, and 1 U “KOD hot start” DNA polymerase (Novagen, Darmstadt, Germany) (Fig. 1, Step A). The PCR conditions were polymerase activation 95°C/2 min, 25 cycles of denaturation 95°C/20 sec, annealing 50°C/10 sec, and extension 70°C/10 sec. Both 5' splice donor and 3' splice acceptor site amplification products were gel extracted. In the next step, fusion PCR was performed using the exonic primers on the amplicons (50 ng each) generated in the first step to create an “abridged” intron 11 along with respective exons on either side (Fig. 1, Step B). PCR conditions were the same as above. The resulting “abridged intron 11” PCR product was gel extracted. *CFTR* cDNA (4,630 nt) was obtained from pcDNA5/FRT/GFP*CFTR* [Krasnov et al., 2008] digested with *Aat*II, *Sac*II, *Kpn*I, and *Eco*RV and subcloned into *Kpn*I and *Eco*RV restriction sites of pcDNA5/FRT vector (Invitrogen, Carlsbad, CA, USA) to create pcDNA5/FRT/*CFTR* (hereafter called pcDNACFTR). Sticky feet mutagenesis of pcDNACFTR was performed using “abridged” intron 11 as a mega-primer to create *CFTR* a.i11 EMG (Fig. 1, Step C). Mega-primer annealed to the pcDNACFTR template with “sticky feet” that paired with complementary exonic sequences. During subsequent rounds of thermal cycling, abridged intron got incorporated into the cDNA at the correct exon–exon junction. PCR conditions were activation at 95°C/2 min, 25 cycles of denaturation 95°C/20 sec, annealing 50°C/10 sec, and extension 70°C/6 to 8 min. XL10 Gold-ultracompetent cells (Agilent Technologies, Santa Clara, CA, USA) were transformed with *Dpn*I digested PCR product (Fig. 1, Step D). Plasmid was extracted and sequenced to verify the correct insertion of intron into the *CFTR* cDNA sequence. The process of amplification, fusion PCR and sticky feet mutagenesis was repeated to incorporate additional abridged introns. When full-length introns were introduced, the amplification step was followed by sticky feet mutagenesis (i.e., Step A to Step C; Fig. 1). *CFTR* a.i11 a.i12 EMG contained abridged intron 11 (a.i11; 326 bp of 5' and 320 bp of 3') and abridged intron 12 (a.i12; 224 bp of 5' and 226 bp of 3'). *CFTR*_i14_a.i18_EMG contained full-length intron 14 (i14; 2,272 bp), abridged intron 15 (a.i15; 259 bp of 5' and 350 bp of 3'), full-length intron 16 (i16; 668 bp), abridged intron 17 (a.i17; 300 bp of 5' and 320 bp of 3'), and abridged intron 18 (a.i18; 333 bp of 5' and 339 bp of 3'). Primer sequences are provided in Supp. Table S1.

Site-Directed Mutagenesis

Intron 11 variants (c.1585–1G>A, c.1585–2A>G, c.1585–3T>G, c.1585–8G>A, and c.1585–9T>A) were introduced individually into *CFTR*_a.i11_a.i12_EMG. Intron 16 variants (c.2657+2_2657+3insA, c.2657+3delG, c.2657+5G>A), exon 18 variant (c.2988G>A), and intron 18 variant (c.2988+1G>A) were introduced into *CFTR*_i14_a.i18_EMG. Primer sequences are provided in Supp. Table S2. Briefly, the mutagenesis involved: thermal cycling (1× buffer, 1.5 mM MgSO₄, 0.25 mM dNTP each, 20 ng WT EMG, 125 ng primers,

1 U “KOD hot start” DNA polymerase), *DpnI* (NEB) digestion of the PCR products, transformation of XL10-Gold ultracompetent cells (Agilent, Santa Clara, CA, USA), selection of the colonies on LB-Ampicillin plates (Quality Biologicals, Gaithersburg, MD, USA), plasmid extraction (Qiagen, Valencia, CA, USA), and sequencing. Finally, the sequence verified full-length WT and mutant EMGs were restriction digested and recloned into the *KpnI* and *EcoRV* restrictions sites in the pcDNA5FRT vector to omit chances of nucleotide errors introduced in the plasmid backbone during mutagenesis steps. Sequence confirmation of the WT and mutant EMGs was performed by the Synthesis and Sequencing Facility, Johns Hopkins University School of Medicine.

Transfection of HEK293 and CFBE41o- Cells

Human Embryonic Kidney (HEK293) and CF Bronchial Epithelial (CFBE41o-) cells (a generous gift from Prof.D. Gruenert, University of California-San Francisco, San Francisco, CA) were grown to approximately 95% confluency on six-well plates containing growth medium (MEM supplemented with 10% fetal bovine serum and 1% penicillin–streptomycin) kept at 37°C in a 5%CO₂ incubator. The plates were fibrinogen–collagen coated for the growth of CFBE41o cells. The cells were transfected with 4 µg of WT or mutant EMG plasmids and 6 µl of Lipofectamine2000 (Invitrogen, Carlsbad, CA, USA) according to the manufacturer’s instructions. Plasmid containing *CFTR* cDNA (pcDNACFTR) was used as a positive control while plasmid without insert (pcDNA5FRT) was used as a negative control. Forty eight hours post-transfection, the cells were prepared for RNA and protein analysis.

RT-PCR and Quantification of mRNA Isoforms

Total RNA was prepared using RNeasy Mini Kit (Qiagen, Valencia, CA, USA) as per manufacturer’s instructions and stored at –80°C. Reverse transcription was carried out with 1 µg total RNA using i-Script cDNA synthesis kit (BioRad, Hercules, CA, USA). The reaction mix was incubated for 5 min at 25°C, 30 min at 42°C and 5 min at 85°C. The resulting cDNA product was diluted 10-fold and stored at –20°C. RT-PCR was performed using 2 µl cDNA in a standard 50 µl reaction set up: 10× buffer, 25 mM MgSO₄, 2 mM each dNTPs mix, 10 µM each forward and reverse primers, and 1 U “KOD Hot Start” DNA polymerase (Novagen, Darmstadt, Germany). The forward primers were fluorescently labeled with 6 – FAM (carboxyfluorescein) at the 5’ end (See Supp. Table S3 for primer sequences and annealing temperatures). PCR conditions were 2 min at 95°C, followed by 35 cycles of 20 sec at 95°C, 10 sec at the annealing temperature, and 15 sec at 70°C. The RT-PCR products (2 µl, 100-fold dilution) were mixed with 18 µl of Hi-Di Formamide (Applied Biosystems, Grand Island, NY, USA) and 0.25 µl of an internal size standard (GeneScan-500 Rox; Applied Biosystems, Grand Island, NY, USA). Products were separated by capillary electrophoreses on an ABI 3100 Genetic Analyser using POP4 polymer (Applied Biosystems, Grand Island, NY, USA) and analyzed with the Gene Mapper Software Version 3.7 (Applied Biosystems, Grand Island, NY, USA). Additionally, RT-PCR products were pyrosequenced to determine the percentages of the different isoforms [Mereau et al., 2009]. The RT-PCR products were obtained as described above using biotin labeled reverse primers (See Supp. Table S3 for primer sequences). The products were sequenced according to the protocol of the PyroMark Q24 system (Qiagen, Valencia, CA,

USA) with 0.4 μ M of specific pyrosequencing primers. The Pyromark assay file was adapted to quantitate each splice product, according to the expected sequence. Pyrograms were analyzed with the software PyroMark Q24 V.2.0.6 (Qiagen, Valencia, CA, USA).

SDS-PAGE and Western Blot

The whole cell lysates from HEK293 and CFBE41o- cells were prepared in a lysis buffer (50 mM Tris, 150 mM NaCl, 1% NP40, 1% deoxycholate, sodium azide, sodium orthovanadate pH 7.4, PMSF, and protease inhibitor cocktail [Sigma, St. Louis, MO, USA]). Protein concentration was measured using the bicinchoninic acid (BCA) assay (Pierce Thermo Scientific, Rockford, IL). Total protein (20 μ g) was loaded on 7.5% SDS-PAGE and subjected to Western blot analysis to evaluate the amount of complex glycosylated form, called CFTR Band C [Cheng et al., 1990]. Mouse monoclonal antibody “MM13–4” that recognizes amino acid residues 25–36 in exon 2 (Millipore, Billerica, MA, USA) was used to detect CFTR generated from WT EMG a.i11 a.i12 and corresponding mutant EMGs. Mouse monoclonal antibody “570” that recognizes amino acid residues 731–742 in exon 14 (UNC antibody distribution program sponsored by Cystic Fibrosis Foundation Therapeutics) was used to detect CFTR generated from WT EMG_i14_a.i18 and corresponding mutant EMGs. GAPDH used as a loading control was detected using rabbit polyclonal anti-GAPDH antibody (Sigma #G9545, St. Louis, MO, USA), which recognized approximately 36 kDa protein. The blots were quantified using Image J software (NIH) to determine the amount of processed CFTR (Band C) for each variant relative to pcDNACFTR.

Splicing Predictions

In total, eight in silico prediction tools were used in this study to assess the effect of DNA sequence variations on splicing. Four tools were used to assess the effect of variants upon splicing strength and four were used to predict the splicing isoforms. To assess whether a variant caused aberrant splicing, following four tools were used to record the output scores as a numerical measure of the splicing signal strength: MaxEntScan (MES) using Maximum entropy score (http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html), Splice-Site Analyser Tool (SSAT) using position weight matrix score (<http://ibis.tau.ac.il/ssat/SpliceSiteFrame.html>), NNSplice using Neural network (NN) score (http://www.fruitfly.org/seq_tools/splice.html), and Sroogle using position specific scoring matrix score for 3' splice site and Shapiro and Senapathy score for 5' splice site (<http://sroogle.tau.ac.il/>) [Shapiro and Senapathy, 1987; Reese et al., 1997; Yeo and Burge, 2004; Schwartz et al., 2009]. Next, to facilitate the comparison between different methods, percent score variation (%) was calculated by comparing the mutant score with the reference score ($[(WT \text{ score} - Mut \text{ score})/WT \text{ score}] \times 100$). Since this value is arbitrary across different tools [Bonnet et al., 2008; Houdayer et al., 2008], we used a 15% cut-off for the scores from MES and 5% cut-off for SSAT, NNSplice and Sroogle to predict aberrant splicing [Houdayer et al., 2012; Jian et al., 2013]. To predict the splice isoforms generated by variants, we used following four programs:

Fsplice (<http://linux1.softberry.com/berry.phtml>),

SplicePort (<http://spliceport.cbcb.umd.edu/SplicingAnalyser.html>), Human Splicing Finder (HSF, <http://www.umd.be/HSF/>), and

Automated Splice Site and Exon Definition Analysis (ASSEDA, <http://splice.uwo.ca/>) [Dogan et al., 2007; Desmet et al., 2009; Mucaki et al., 2013].

Creation of Flp-In 293 Stable Cells

Flp-In 293 cell line (Invitrogen, Carlsbad, CA, USA), which contains the Flp Recombinase Target (FRT) but expresses no endogenous CFTR, was grown on 60 mm dishes and cotransfected using Lipofectamine2000 (Invitrogen, Carlsbad, CA, USA) with either WT EMG/EMGs with intron 16 variants/pcDNACFTR and pOG44 recombinase (Flp-In System Kit; Invitrogen, Carlsbad, CA, USA) at ratio of 1:9 that facilitated site-specific recombination. Hygromycin-resistant clones were PCR amplified using *CFTR* exon spanning primers to confirm *CFTR* integration and selected as described previously [Krasnov et al., 2008].

Results

Phenotypic Details of *CFTR* variants

The phenotype associated with each *CFTR* variant was ascribed by virtue of its presence on either both chromosomes or in trans with second CF-causing variant previously shown to have minimal residual function [Castellani et al., 2009]. For a given variant, the clinical outcome variables such as sweat chloride, lung function, infection, and pancreatic status were averaged across homozygous and compound heterozygous genotypes (Table 1). Among the clinical features associated with each variant, the rate of severe pancreatic disease (pancreatic insufficiency) was notably lower in patients carrying variants in intron 16 compared with the overall CF population (~85%–90%) [Sosnay et al., 2013]. As *CFTR* function has been shown to correlate with pancreatic status, the lower rate of PI associated with the intron 16 variants suggests that they might allow some degree of normal splicing and production of some full-length *CFTR* protein.

Functional Assessment of the 3' Splice Site Variants of *CFTR*

To study the variants in the 3' splice site of intron 11, an EMG was created which composed of the *CFTR* cDNA into which abridged intron 11 sequence flanking 3' of exon 11 and 5' of exon 12 (646 bp) and abridged intron 12 sequence flanking 3' of exon 12 and 5' of exon 13 (450 bp) were introduced (*CFTR*_a.i11_a.i12_EMG; Fig. 2A). The five variants in intron 11 listed in Table 1 were introduced to the *CFTR*_a.i11_a.i12_EMG using site-directed mutagenesis. Splicing of *CFTR* pre-mRNA transcribed from WT and mutant EMGs in HEK293 cells was assessed by RT-PCR followed by fragment analysis using a forward primer from exon 11 and a reverse primer from exon 13. Each mutant EMG generated RT products of different sizes compared with WT EMG (Fig. 2B). The WT EMG yielded a 259 bp product as expected from the correct splicing of exons 11, 12, and 13. However, c.1585–1G>A, c.1585–2A>G, and c.1585–3T>G variants resulted in a shorter isoform of 164 bp, in addition to 258, 265, and 261 bp isoforms, respectively. Other two variants, c.1585–8G>A and c.1585–9T>A, yielded isoforms of 265 and 266 bp, respectively (Fig. 2B). Sequencing confirmed that none of the mutant EMGs generated normally spliced RNA transcripts (Fig. 2C). The mutant EMGs synthesized RNA transcripts that are consistent with the use of an alternative 3' splice site (shown in larger bold font) that lead to

frameshifts (c.1585–1G>A, c.1585–3T>G, and c.1585–9T>A) or retention of six nucleotides that added termination codons to the RNA transcript (c.1585–2A>G and c.1585–8G>A). With the exception of c.1585–3T>G variant, each novel splice site conformed to the consensus “AG” 3' splice site sequence. The c.1585–2A>G variant led to use of a “GG” dinucleotide in the 3' splice site, which has been observed previously [Burset et al., 2000] (Fig. 2C). Sequencing of the shorter amplification product (164 bp) revealed skipping of exon 12 due to splicing of exon 11 to exon 13. After sequence confirmation, RT-PCR products were quantified using two approaches: measuring the peak area of the fluorescent signal generated by each isoform (Fig. 2 B) and calculating the relative proportion of the different isoforms by pyrosequencing (Fig. 2D). Comparison of two methods demonstrated high concordance (Supp. Table S4).

It is important to note that all of the RNA transcripts generated by the mutant EMGs contained PTCs. As noted earlier, PTCs generally induce NMD leading to loss of protein synthesis [Chang et al., 2007]. In rare cases, RNA transcripts with PTCs are stable and can translate truncated forms of protein [Silva et al., 2008]. To determine between these possibilities, lysates from HEK293 cells transfected with WT and mutant EMGs were analyzed by Western blot. The WT *CFTR* EMG expressed similar amounts of mature CFTR as observed for the intron-less *CFTR* cDNA (Fig. 2E). Note that CFTR protein is expressed primarily as mature, complex glycosylated form, called Band C (~170 kDa) [Cheng et al., 1990]. A minor fraction of CFTR protein was present in an immature form as the results of partial glycosylation and is designated as Band B (~150 kDa). None of the *CFTR* mRNA generated by the mutant EMGs was able to synthesize a stable protein product (Fig. 2E). As *CFTR* splicing and expression can vary due to different cell types, we repeated these experiments in human bronchial epithelial cells (CFBE41o-), homozygous for p.Phe508delmutation. These are highly differentiated cells that form tight junctions and represent the native pulmonary airway cell type in which CFTR is expressed. CFTR protein generated in CFBE41o- cells corroborated results obtained from HEK293 cells (Fig. 2F).

Functional Assessment of the 5' Splice Site Variants of *CFTR*

To determine the effect of variants at the 5' splice site of introns 16 and 18, we created an EMG with full-length introns 14 (2,272 nt) and 16 (668 nt) and abridged versions of introns 15 (609 nt), 17 (620 nt), and 18 (672 nt) each cloned in the correct orientation into the *CFTR* cDNA (*CFTR* i14_a.i18_ EMG; Fig. 3A). Intron 16 variants (c.2657+2_2657+3insA, c.2657+3delG, and c.2657+5G>A), exon 18 variant (c.2988G>A), and intron 18 variant (c.2988+1G>A) were introduced into wild-type_i14_a.i18_ EMG. To study the effect on splicing, RT-PCR followed by fragment analysis was performed using forward primer from exon 15 and reverse primer from exon 17 for intron 16 variants; and forward primer from exon 17 and reverse primer from exon 19 for exon/intron 18 variants. RT-PCR products of 395 and 484 bp revealed correct splicing of exons 15, 16, and 17; and exons 17, 18, and 19, respectively, from the WT EMG_i14_a.i18 (Fig. 3B). The c.2657+2_2657+3insA, c.2657+3delG, and c.2657+5G>A variants generated normal spliced (395 bp) and exon 12 skipped (357 bp) isoforms, whereas c.2988G>A and c.2988+1G>A variants yielded exon 18 skipped (404 bp) isoform only (Fig. 3B). As before, fragment analysis and pyrosequencing were used to quantify the RT-PCR products. Concordance was observed among both methods

(Supp. Table S5). To exclude cell type differences, spliced isoforms generated by each variant were determined in both HEK293 and CFBE41o- cells by pyrosequencing. Comparable levels of correctly spliced isoforms were observed in both cell lines (Fig. 3C).

To assess whether CFTR protein was generated by EMGs bearing any of the intron 16 or 18 variants, we performed western blotting. Variable levels of fully processed CFTR were observed with EMGs bearing intron 16 variants (Fig. 3D), whereas the EMGs bearing c.2988G>A or c.2988+1G>A variants generated no detectable CFTR protein (Fig. 3E). The WT EMG with introns 14–18 generated similar amounts of mature CFTR as observed for the pcDNACFTR (Fig. 3D and E). Interestingly, the EMG containing c.2657+2_2657+3insA variant synthesized mature CFTR protein at level comparable to the pcDNACFTR ($89 \pm 9.3\%$ and $81.8 \pm 10.21\%$ in CFBE41o- and HEK293 cells, respectively) (Fig. 3F). However, the EMG with c.2657+3delG variant generated low amounts of CFTR protein ($8.3 \pm 2.4\%$ and $7.3 \pm 2.9\%$), whereas the EMG with c.2657+5G>A variant generated $18.2 \pm 5.3\%$ and $15 \pm 6.1\%$ of protein in CFBE41o- and HEK293 cells, respectively (Fig. 3F).

As variable levels of expression were found with intron 16 mutant EMGs, we wanted to verify that these variants did not dramatically alter CFTR protein level due to unidentified off-target effects of site-directed mutagenesis or contamination with the WT EMG. The variant, c.2657+2_2657+3insA*, was introduced again into WT EMG_i14_a.i18 and sequence verified (“re-created”) in a separate experiment. The recreation of the variant ensured that the observed splicing patterns were due to the introduced variant rather than an unrecognized sequence change elsewhere in the plasmid construct (Fig. 3G). Similarly, the c.2657+5G>A variant was mutated back to WT (called reverted and annotated as c.2657+5A>G**). The recovery of normal splicing indicated that the variant was the cause of the aberrant splicing (Fig. 3G).

Assessing protein levels translated from episomal plasmids can be problematic as multiple copies of plasmids can be present per cellular genome leading to the synthesis of higher protein levels than observed in vivo. Therefore, we assessed the amount of CFTR protein expressed from a single copy of an EMG integrated into the cellular genome of Flp-In 293 cells. Cell lines were created with stable integrations of WT EMG, EMGs with each of the intron 16 variants, and pcDNACFTR into the single FRT site in Flp-In 293 cells. This approach minimized variation observed in RNA and protein levels due to the random integration or integration of multiple copies of the EMG. The amount of CFTR Band C generated from the EMG bearing c.2657+2_2657+3insA ($85.62 \pm 8.54\%$) was similar to that translated from the WT EMG or CFTR cDNA, whereas very low levels of CFTR were expressed in c.2657+3delG ($2.83 \pm 1.1\%$) and c.2657+5G>A ($1.43 \pm 0.78\%$) mutant cell lines (Fig. 3H).

Comparisons of EMG Results with In Silico Predictions

To determine the accuracy of splicing algorithms, we tested the ability of publically available programs to predict whether variants would cause mis-splicing and to predict which RNA transcripts are generated. To predict whether alterations in or near splice sites affect splicing, four programs were used that calculate the strength of the splicing efficiency

at the consensus native site (MES, SSAT, NNSplice, and Sroogle). The percent change in splicing strength at the native site was calculated from the scores generated for sequence containing each individual variant and the corresponding WT sequence. The magnitude of the change is expected to correlate with decrease in efficiency and increase in the likelihood of mis-splicing. A variant was predicted to affect splicing using thresholds suggested in the previous studies (MES – 15%, and NNSplice, SSAT or Sroogle – 5%) [Houdayer et al., 2012; Jian et al., 2013]. Predictions were concordant with experimental results for nine out of the 10 variants using MES and eight out of the 10 variants using NNSplice and seven out of the 10 variants using Sroogle and SSAT (Fig. 4 and Supp. Table S6). SSAT, NNSplice, and Sroogle incorrectly predicted normal splicing for the c.1585–8G>A variant. Similarly, SSAT and Sroogle incorrectly predicted normal splicing for the c.1585–9T>A variant. All four programs predicted mis-splicing for the c.2657+2_2657+3insA variant, however, a correctly spliced product was observed (Fig. 4).

To evaluate existing algorithms that predict RNA isoforms resulting from splice site variants, we tested Fsplice, SplicePort, and HSF predictions against transcript isoforms generated by each variant. Among the 16 isoforms generated by 10 variants, Fsplice and SplicePort each predicted 10 isoforms, whereas HSF and ASSEDA each predicted 11 isoforms correctly (Table 2). However, some notable discrepancies were observed between in silico and experimental data. First, the EMG containing c.1585–1G>A variant generated two isoforms: exon 12 skipped and one nucleotide exon deletion. Whereas Fsplice predicted the exon 12 skipped isoform and SplicePort predicted one nucleotide exon deletion isoform. Second, the c.1585–2A>G variant caused both retention of six intronic nucleotides and skipping of exon 12, but exon 12 skipping was predicted by all three programs. Third, the c.1585–3T>G variant caused retention of two intronic nucleotides and skipping of exon 12. Fsplice predicted exon 12 skipping alone, whereas SplicePort and HSF predicted two nucleotide intron retention isoform only. Finally, the WT isoform observed with c.2657+2_2657+3insA variant was not predicted by any of the three programs (Table 2).

A fourth prediction tool was also evaluated. ASSEDA attempts not only to predict mutant mRNA isoforms but also their relative abundance by Information theory-based Exon definition [Mucaki et al., 2013]. In our study, it correctly predicted 10 of the 16 isoforms generated by the 10 variants (Table 2). With regard to relative abundance, a striking difference was observed with c.1585–8G>A, where full-length transcript is predicted to be the major isoform by ASSEDA. However, c.1585–8G>A generated a single isoform with a six nucleotide intron retention with no WT transcript. In case of other variants, less abundant isoforms predicted by ASSEDA (third and fourth isoforms for c.1585–1G>A, third isoform for c.1585–2A>G, third isoform for c.1585–3T>G, and second and third isoforms for c.2988+1G>A) were not observed in EMG experiments. It was also prone to predict exon skipping events as WT. For instance, exon 16 and 18 skipping events observed with the variants c.2657+5G>A and c.2988G>A, respectively, in our experiments were predicted to generate full-length transcript by ASSEDA. Lastly, six nucleotide intron retention observed with c.1585–2A>G was incorrectly predicted as a seven nucleotide intron retention isoform. On the other hand, ASSEDA predicted a WT transcript for c.2657+2_2657+3insA, contrary to all the bioinformatics tools used in this study. Interestingly, the EMG harboring c.

2657+2_2657+3insA variant generated majority of correctly spliced WT transcript and abundant CFTR Band C in the experiments (Table 2).

Discussion

This study illustrates the use of EMG technology to efficiently predict the impact of splicing variants by analysis of RNA and protein products. Our confidence that the EMGs generated splicing isoforms that represent those found in vivo is based on four observations. First, the WT EMGs containing either introns 11 and 12 or introns 14, 15, 16, 17, and 18 used in our study spliced normally and generated full-length CFTR protein. Second, the EMGs bearing c.1585-1G>A and c.2657+5G>A variants generated CFTR RNA transcripts that were comparable to the transcripts from the nasal epithelium of the individuals carrying these variants [Hull et al., 1993; Highsmith, Jr. et al., 1997; Masvidal et al., 2014]. Third, the effect of variants upon RNA splicing and protein translation of the EMGs was studied in two human cell lines of different origin. One of two cell lines, CFBE41o-, has been generated by immortalizing airway cells isolated from a CF (p.Phe508del/p.Phe508del) individual and is suggested to retain properties of bronchial epithelial cells known to express CFTR [Bruscia et al., 2002; Ehrhardt et al., 2006]. Results of RNA splicing and protein translation were highly concordant between the two cell lines. Lastly, three variants in intron 16 that caused partial disruption of splicing were re-analyzed after integration of a single copy of the EMG bearing each variant into the genome of HEK293 cells. Integration ensured that the copy number of EMGs was constant among cells and between cell lines bearing EMGs with different variants.

Comparing functional results with clinical features can improve our ability to differentiate deleterious variants from variants of uncertain significance [Jacob et al., 2013; Sosnay et al., 2013]. When the results of the functional study were combined with clinical data, nine of the 10 variants could be assigned as disease-causing (<http://www.cftr2.org/>). Notably, two of the three variants in intron 16 had reduced, but not absent full-length *CFTR* transcript consistent with their more moderate CF phenotypes (Table 1). The remaining variant, c.2657+2_2657+3insA, showed a consistent minimal effect on *CFTR* splicing and no biologically relevant decrease in CFTR protein level. While patients bearing this variant exhibited a lower rate of severe pancreatic disease (i.e., pancreatic insufficiency) and lower sweat chloride concentration than observed in CF patients (<http://www.cftr2.org/>), the normal levels of CFTR Band C generated by the EMG bearing c.2657+2_2657+3insA is incongruous with a disease phenotype [Sosnay et al., 2013]. There is a possibility that an undetected variant in *cis* with c.2657+2_2657+3insA is responsible for the severe phenotype, as has been noted for other variants in *CFTR* and other disease-associated genes [Pagani et al., 2000; Hefferon et al., 2004; Cooper et al., 2013].

Algorithms that accurately predict the likelihood that a variant causes mis-splicing and the splicing isoforms produced are important tools in the genome sequencing era. The functional effect of splicing variants in the *BRCA1* and the *BRCA2* genes has been evaluated systematically by testing both patient RNA and hybrid minigenes to optimize cut-off values for splicing defect prediction tools [Bonnet et al., 2008; Spurdle et al., 2008; Houdayer et al., 2012]. More recently, *CFTR* splice site variants studied using different hybrid minigene

systems (exon trap vector pSPL3; exon trap vector pET01, and splice vector pDEST) suggested that both bioinformatics tools and hybrid minigenes can be used for predicting the pathogenicity of variants of unknown clinical significance [Scott et al., 2012; Aissat et al., 2013; Raynal et al., 2013]. In this study, we show that splicing algorithms are quite accurate in estimating the likelihood that mis-splicing will occur.

There were notable differences between EMG assays and bioinformatics predictions for some of the variants. Most strikingly, c.2657+2_2657+3insA was predicted to affect splicing by all programs except ASSEDA, but our experimental data showed that the majority of the transcript was normally spliced (>80%) and abundant mature CFTR protein was produced (>80%). The discrepancy between prediction and experimental observation for c.2657+2_2657+3insA could be explained by the absence of deep intronic sequences. Most of the prediction programs take into consideration only minimal sequence surrounding the core dinucleotide (GT) splicing signal at 5' splice donor site. However, the EMG used to evaluate c.2657+2_2657+3insA contained full-length intron 16. Second, the adenine insertion occurs at the +3 position, which is more preferred nucleotide at this location (62.1%) compared with guanine (32.8%), cytosine (2.2%), and thymine (3%) [Zhang et al., 2005]. In silico substitution of guanine at +3 position with adenine (c.2657+3G>A) is predicted to be tolerant by all eight programs, whereas cytosine or thymine substitutions are deleterious. A prior study evaluated 24 mutations at +3 positions and revealed heterogeneous predictions ranging from 20.83% to 100% with overall prediction efficiency of 68.06% [Desmet et al., 2010]. These results suggested that position +3 is dependent on sequence context; therefore, splice site prediction at this site can be challenging [Desmet et al., 2010]. Finally, exon 16 is composed of 38 nucleotides and is the smallest among the 27 exons in *CFTR*. Previously, it has been reported that mini-exons less than 12 nucleotides in length are not incorporated in processed mRNA unless accompanied by the upstream exons [Stern and Berget, 1993]. Moreover, constitutive exons that are shortened to less than 51 nucleotides are skipped unless the strength of splicing signals is increased [Dominski and Kole, 1991; Stern and Berget, 1993]. ASSEDA was the only program used in our study that made predictions based on exon definition [Mucaki et al., 2013] and was the only tool that correctly predicted c.2657+2_2657+3insA to be normally spliced.

Another challenge in bioinformatics prediction is the creation of a new splice site as a consequence of sequence variation. If the new splice site is some distance from the original site then sequence information may not be analyzed by the prediction program [Spurdle et al., 2008]. This phenomenon was observed in our study where no deleterious effect was predicted for c.1585-8G>A and c.1585-9T>A by three and two programs, respectively. However, in each case, novel aberrant splice sites were created 6 and 7 nucleotides upstream of the 3' splice acceptor site, respectively, resulting in intron retention and addition of a PTC. The discrepancies between experimental results and in silico predictions in this study support the recommendation that use of a single prediction model is not sufficient, as recommended for the *BRCA1* and the *BRCA2* genes [Spurdle et al., 2008; Acedo et al., 2012; Houdayer et al., 2012]. In summary, we show that EMGs provide a robust method to evaluate variants that affect pre-mRNA splicing with the ability to simultaneously assess translation from processed mRNA isoforms. In addition to testing variants near splice sites,

EMGs could interrogate deep exonic and intronic variants that influence splicing by altering silencer and enhancer binding sites. Thus, EMGs should enable the generation of comprehensive splicing data to train and improve algorithms that simplify interpretation of DNA variants.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank contributors of the CFTR2 patient data, Roxann Ashworth at Genetic Resource Core Facility, Johns Hopkins University School of Medicine, for pyrosequencing assay design, and Patricia Cornwall for her administrative assistance.

Contract grant sponsors: NIDDK (5R37DK044003); US CF Foundation (CUTTING08A, CUTTING09A, CUTTING10A); US CF Foundation (SOSNAY10Q); PEstOE/BIA/UI4046/2011 (PIC/IC/83103/2007).

References

- Acedo A, Sanz DJ, Duran M, Infante M, Perez-Cabornero L, Miner C, Velasco EA. Comprehensive splicing functional analysis of DNA variants of the BRCA2 gene by hybrid minigenes. *Breast Cancer Res.* 2012; 14:R87. [PubMed: 22632462]
- Aissat A, de Becdelievre A, Golmard L, Vasseur C, Costa C, Chaoui A, Martin N, Costes B, Goossens M, Girodon E, Fanen P, Hinzpeter A. Combined computational-experimental analyses of CFTR exon strength uncover predictability of exon-skipping level. *Hum Mutat.* 2013; 34:873–881. [PubMed: 23420618]
- Baralle M, Skoko N, Knezevich A, de Conti L, Motti D, Bhuvanagiri M, Baralle D, Buratti E, Baralle FE. NF1 mRNA biogenesis: effect of the genomic milieu in splicing regulation of the NF1 exon 37 region. *FEBS Lett.* 2006; 580:4449–4456. [PubMed: 16870183]
- Bonnet C, Krieger S, Vezain M, Rousselin A, Tournier I, Martins A, Berthet P, Chevrier A, Dugast C, Layet V, Rossi A, Lidereau R, et al. Screening BRCA1 and BRCA2 unclassified variants for splicing mutations using reverse transcription PCR on patient RNA and an ex vivo assay based on a splicing reporter minigene. *J Med Genet.* 2008; 45:438–446. [PubMed: 18424508]
- Bruscia E, Sangiuolo F, Sinibaldi P, Goncz KK, Novelli G, Gruenert DC. Isolation of CF cell lines corrected at DeltaF508-CFTR locus by SFHR-mediated targeting. *Gene Ther.* 2002; 9:683–685. [PubMed: 12032687]
- Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* 1997; 268:78–94. [PubMed: 9149143]
- Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* 2000; 28:4364–4375. [PubMed: 11058137]
- Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet.* 2002; 3:285–298. [PubMed: 11967553]
- Castellani C, Southern KW, Brownlee K, Dankert RJ, Duff A, Farrell M, Mehta A, Munck A, Pollitt R, Sermet-Gaudelus I, Wilcken B, Ballmann M, et al. European best practice guidelines for cystic fibrosis neonatal screening. *J Cyst Fibros.* 2009; 8:153–173. [PubMed: 19246252]
- Chang YF, Imam JS, Wilkinson MF. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem.* 2007; 76:51–74. [PubMed: 17352659]
- Cheng SH, Gregory RJ, Marshall J, Paul S, Souza DW, White GA, O’Riordan CR, Smith AE. Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis. *Cell.* 1990; 63:827–834. [PubMed: 1699669]
- Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet.* 2013; 132:1077–1130. [PubMed: 23820649]

- Cooper TA. Use of minigene systems to dissect alternative splicing elements. *Methods*. 2005; 37:331–340. [PubMed: 16314262]
- Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res*. 2009; 37:e67. [PubMed: 19339519]
- Desmet, FO.; Hamroun, D.; Collod-Beroud, G.; Claustres, M.; Beroud, C. Bioinformatics identification of splice site signals and prediction of mutation effects. In: Mohan, RM., editor. Recent advances in nucleic acid research. Kerala, India: Global Research Network; 2010. p. 1-14.
- Dogan RI, Getoor L, Wilbur WJ, Mount SM. SplicePort—an interactive splice-site analysis tool. *Nucleic Acids Res*. 2007; 35:W285–W291. [PubMed: 17576680]
- Dominski Z, Kole R. Selection of splice sites in pre-mRNAs with short internal exons. *Mol Cell Biol*. 1991; 11:6075–6083. [PubMed: 1944277]
- Ehrhardt C, Collnot EM, Baldes C, Becker U, Laue M, Kim KJ, Lehr CM. Towards an in vitro model of cystic fibrosis small airway epithelium: characterisation of the human bronchial epithelial cell line CFBE41o- *Cell Tissue Res*. 2006; 323:405–415. [PubMed: 16249874]
- Eriksson M, Brown WT, Gordon LB, Glynn MW, Singer J, Scott L, Erdos MR, Robbins CM, Moses TY, Berglund P, Dutra A, Pak E, et al. Recurrent de novo point mutations in lamin A cause Hutchinson–Gilford progeria syndrome. *Nature*. 2003; 423:293–298. [PubMed: 12714972]
- Fededa JP, Petrillo E, Gelfand MS, Neverov AD, Kadener S, Nogues G, Pelisch F, Baralle FE, Muro AF, Kornblihtt AR. A polar mechanism coordinates different regions of alternative splicing within a single gene. *Mol Cell*. 2005; 19:393–404. [PubMed: 16061185]
- Frischmeyer PA, Dietz HC. Nonsense-mediated mRNA decay in health and disease. *Hum Mol Genet*. 1999; 8:1893–1900. [PubMed: 10469842]
- Goina E, Fernandez-Alanis E, Pagani F. Approaches to study CFTR pre-mRNA splicing defects. *Methods Mol Biol*. 2011; 741:155–169. [PubMed: 21594784]
- Hastings ML, Krainer AR. Pre-mRNA splicing in the new millennium. *Curr Opin Cell Biol*. 2001; 13:302–309. [PubMed: 11343900]
- Hefferon TW, Groman JD, Yurk CE, Cutting GR. A variable dinucleotide repeat in the *CFTR* gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proc Natl Acad Sci USA*. 2004; 101:3504–3509. [PubMed: 14993601]
- Highsmith WE Jr, Luranell BH, Zhaoqing Z, Olsen JC, Strong TV, Smith T, Friedman KJ, Silverman LM, Boucher RC, Collins FS, Knowles MR. Identification of a splice site mutation (2789+5G>A) associated with small amounts of normal CFTR mRNA and mild cystic fibrosis. *Hum Mutat*. 1997; 9:332–338. [PubMed: 9101293]
- Houdayer C, Caux-Moncoutier V, Krieger S, Barrois M, Bonnet F, Bourdon V, Bronner M, Buisson M, Coulet F, Gaildrat P, Lefol C, Leone M, et al. 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; 33:1228–1238. [PubMed: 22505045]
- Houdayer C, Dehainault C, Mattler C, Michaux D, Caux-Moncoutier V, Pages-Berhouet S, d'Enghien CD, Lauge A, Castera L, Gauthier-Villars M, Stoppa-Lyonnet D. Evaluation of in silico splice tools for decision-making in molecular diagnosis. *Hum Mutat*. 2008; 29:975–982. [PubMed: 18449911]
- Hull J, Shackleton S, Harris A. Abnormal mRNA splicing resulting from three different mutations in the CFTR gene. *Hum Mol Genet*. 1993; 2:689–692. [PubMed: 7689009]
- Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A, Kendall J, Grabowska E, et al. De novo gene disruptions in children on the autistic spectrum. *Neuron*. 2012; 74:285–299. [PubMed: 22542183]
- Jacob HJ, Abrams K, Bick DP, Brodie K, Dimmock DP, Farrell M, Geurts J, Harris J, Helbling D, Joers BJ, Kliegman R, Kowalski G, et al. Genomics in clinical practice: lessons from the front lines. *Sci Transl Med*. 2013; 5:194cm5.
- Jian X, Boerwinkle E, Liu X. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet Med*. 2013

- Jordan CT, Cao L, Roberson ED, Pierson KC, Yang CF, Joyce CE, Ryan C, Duan S, Helms CA, Liu Y, Chen Y, McBride AA, et al. PSORS2 is due to mutations in CARD14. *Am J Hum Genet.* 2012; 90:784–795. [PubMed: 22521418]
- Korf I, Flicek P, Duan D, Brent MR. Integrating genomic homology into gene structure prediction. *Bioinformatics.* 2001; 17(Suppl 1):S140–S148. [PubMed: 11473003]
- Krasnov KV, Tzetis M, Cheng J, Guggino WB, Cutting GR. Localization studies of rare missense mutations in cystic fibrosis transmembrane conductance regulator (CFTR) facilitate interpretation of genotype–phenotype relationships. *Hum Mutat.* 2008; 29:1364–1372. [PubMed: 18951463]
- 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; 28:150–158. [PubMed: 17001642]
- MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, Barrett JC, Biesecker LG, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature.* 2014; 508:469–476. [PubMed: 24759409]
- Marden JH. Quantitative and evolutionary biology of alternative splicing: how changing the mix of alternative transcripts affects phenotypic plasticity and reaction norms. *Heredity (Edinb).* 2008; 100:111–120. [PubMed: 17006532]
- Masvidal L, Igreja S, Ramos MD, Alvarez A, de Gracia J, Ramalho A, Amaral MD, Larriba S, Casals T. Assessing the residual CFTR gene expression in human nasal epithelium cells bearing CFTR splicing mutations causing cystic fibrosis. *Eur J Hum Genet.* 2014; 22:784–791. [PubMed: 24129438]
- Mereau A, Anquetil V, Cibois M, Noiret M, Primot A, Vallee A, Paillard L. Analysis of splicing patterns by pyrosequencing. *Nucleic Acids Res.* 2009; 37:e126. [PubMed: 19671523]
- Mucaki EJ, Shirley BC, Rogan PK. Prediction of mutant mRNA splice isoforms by information theory-based exon definition. *Hum Mutat.* 2013; 34:557–565. [PubMed: 23348723]
- Pagani F, Buratti E, Stuardi C, Romano M, Zuccato E, Niksic M, Giglio L, Faraguna D, Baralle FE. Splicing factors induce cystic fibrosis transmembrane regulator exon 9 skipping through a nonevolutionary conserved intronic element. *J Biol Chem.* 2000; 275:21041–21047. [PubMed: 10766763]
- Petkovic V, Godi M, Lochmatter D, Eble A, Fluck CE, Robinson IC, Mullis PE. Growth hormone (GH)-releasing hormone increases the expression of the dominant-negative GH isoform in cases of isolated GH deficiency due to GH splice-site mutations. *Endocrinology.* 2010; 151:2650–2658. [PubMed: 20351314]
- Raynal C, Baux D, Theze C, Bareil C, Taulan M, Roux AF, Claustres M, Tuffery-Giraud S, des Georges M. A classification model relative to splicing for variants of unknown clinical significance: application to the CFTR gene. *Hum Mutat.* 2013; 34:774–874. [PubMed: 23381846]
- Reese MG, Eeckman FH, Kulp D, Haussler D. Improved splice site detection in Genie. *J Comput Biol.* 1997; 4:311–323. [PubMed: 9278062]
- Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature.* 2012; 485:237–241. [PubMed: 22495306]
- Sankaran VG, Ghazvinian R, Do R, Thiru P, Vergilio JA, Beggs AH, Sieff CA, Orkin SH, Nathan DG, Lander ES, Gazda HT. Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. *J Clin Invest.* 2012; 122:2439–2443. [PubMed: 22706301]
- Schwartz S, Hall E, Ast G. SROOGLE: webserver for integrative, user-friendly visualization of splicing signals. *Nucleic Acids Res.* 2009; 37:W189–W192. [PubMed: 19429896]
- Scott A, Petrykowska HM, Hefferon T, Gotea V, Elnitski L. Functional analysis of synonymous substitutions predicted to affect splicing of the CFTR gene. *J Cyst Fibros.* 2012; 11:511–517. [PubMed: 22591852]
- Shapiro MB, Senapathy P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* 1987; 15:7155–7174. [PubMed: 3658675]

- Silva AL, Ribeiro P, Inacio A, Liebhaber SA, Romao L. Proximity of the poly(A)-binding protein to a premature termination codon inhibits mammalian nonsense-mediated mRNA decay. *RNA*. 2008; 14:563–576. [PubMed: 18230761]
- Singh RK, Cooper TA. Pre-mRNA splicing in disease and therapeutics. *Trends Mol Med*. 2012; 18:472–482. [PubMed: 22819011]
- Sosnay PR, Siklosi KR, Van Goor F, Kaniecki K, Yu H, Sharma N, Ramalho AS, Amaral MD, Dorfman R, Zielenski J, Masica DL, Karchin R, et al. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat Genet*. 2013; 45:1160–1167. [PubMed: 23974870]
- Spurdle AB, Couch FJ, Hogervorst FB, Radice P, Sinilnikova OM. Prediction and assessment of splicing alterations: implications for clinical testing. *Hum Mutat*. 2008; 29:1304–1313. [PubMed: 18951448]
- Sterner DA, Berget SM. In vivo recognition of a vertebrate mini-exon as an exon-intron-exon unit. *Mol Cell Biol*. 1993; 13:2677–2687. [PubMed: 7682652]
- Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, Phillips AD, Shaw K, Stenson PD, Cooper DN, Tyler-Smith C. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am J Hum Genet*. 2012; 91:1022–1032. [PubMed: 23217326]
- Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z, Hardison M, Person R, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med*. 2013; 369:1502–1511. [PubMed: 24088041]
- Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. 2004; 11:377–394.
- Yeo GW, Van Nostrand EL, Liang TY. Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genet*. 2007; 3:e85. [PubMed: 17530930]
- Zhang XH, Leslie CS, Chasin LA. Computational searches for splicing signals. *Methods*. 2005; 37:292–305. [PubMed: 16314258]

Figure 1.

Construction of an expression minigene to test splicing variants in *CFTR*. **A:** Amplification of 5' fragment (first 326 bp) and 3' fragment (last 320 bp) of intron 11 along with exon 11 and 12, respectively, from genomic DNA. Overhang on 5'R primer (black dotted line) was reverse complement of the first sixteen nucleotides of the 3' fragment. Overhang on the 3'F primer (gray dotted line) was identical to the last seventeen nucleotides of the 5' fragment. **B:** Fusion PCR resulted in an "abridged" intron 11 (a.i11) by using exonic primers (5'F and 3'R) only. **C:** Sticky feet mutagenesis of the pcDNA5FRT plasmid with full-length *CFTR*

cDNA (pcDNACFTR) incorporated “abridged” intron 11 (a.i11) sequence. The fusion PCR product served as a megaprimer. **D:** Transformation of XL10 Gold-ultracompetent cells with *DpnI* digested mutagenesis PCR product. *DpnI* digested the parental plasmid, and transformation nick repaired and circularized the plasmid to generate expression minigene containing abridged intron 11 (*CFTR_a.i11_EMG*). An arrow indicates that repeated rounds of amplification of subsequent introns, fusion PCR, sticky feet mutagenesis, and transformation are required to incorporate additional introns in the EMG.

Figure 2.

EMG analysis of the variants in 3' splice acceptor site of *CFTR*. **A:** Plasmid map of the EMG containing abridged intron 11 (a.i11) and abridged intron 12 (a.i12). Open arrow indicates cytomegalovirus promoter, open arc indicates *CFTR* cDNA and dark arcs in-between indicate a.i11 (827 bp) and a.i12 (617 bp) introduced after exon 11 and 12, respectively. **B:** Fragment analysis of the splice isoforms obtained by RT-PCRs of the total RNA extracted from HEK293 cells transiently transfected with WT EMG_a.i11_a.i12 (WT EMG) and mutant EMGs. RT-PCR products labeled with 6-FAM (shadowed blue peaks)

were electrophoresed on an ABI 3100 sequencer with Genescan ROX 500 (red peaks) as size standard. WT and aberrant isoforms are indicated by arrows. RFU refers to Relative Fluorescence Units. **C:** Diagram of the sequences around 3' splice acceptor site of *CFTR* intron 11 shows aberrant splicing caused by each of the variants. The underlined nucleotide denotes the variant. Dinucleotides in bold are the new 3' splice acceptor sites compared to the WT sequence. Boxes indicate sequence incorporated into the RNA transcript. **D:** Quantification of the mRNA isoforms by pyrosequencing. Data are from three independent experiments performed in HEK293 cells and presented as mean $\pm$ SD. **E:** Western blot of lysates from HEK293 cells transiently transfected with WT EMG_a.i11_a.i12, mutant minigenes, empty plasmid (pcDNA5FRT) and plasmid with *CFTR* cDNA (pcDNACFTR). Upper panel shows mature (Band **C**) and immature (Band **B**) forms of CFTR detected by anti-CFTR mouse monoclonal antibody (MM13–4; Millipore, Billerica, MA, USA). Lower panel shows GAPDH was used as loading control. **F:** Graph indicates the amount of mature CFTR protein, Band **C**, generated by WT EMG and mutant EMGs relative to pcDNACFTR in CFBE41o- and HEK293 cells. Protein was quantified using Image J software (NIH). Results are expressed as% relative to mature CFTR from pcDNACFTR (mean $\pm$ SD, $n = 3$ independent experiments).

Figure 3.

EMG analysis of the variants in 5' splice donor sites of *CFTR*. **A:** Plasmid map of the EMG containing full-length intron 14 (i14), abridged intron 15 (a.i15), full-length 16 (i16), abridged intron 17 (a.i17) and abridged intron 18 (a.i18). Open arrow indicates cytomegalovirus promoter, open arc indicates *CFTR* cDNA and dark arcs in-between indicate i14 to a.i18 introduced at the correct exon/exon junction. **B:** Fragment analysis of the splice isoforms obtained by RT-PCRs of the total RNA extracted from HEK293 cells transiently transfected with WT EMG_i14_a.i18 (WT EMG) and mutant EMGs. RT-PCR

products labeled with 6-FAM (shadowed blue peaks) were electrophoresed on an ABI 3100 sequencer with Genescan ROX 500 (red peaks) as size standard. WT and aberrant isoforms are indicated by arrows. RFU refers to Relative Fluorescence Units. **C:** Quantification of the mRNA isoforms by pyrosequencing. Data are from three independent experiments performed in CFBE41o- and HEK293 cells and presented as mean $\pm$ SD. **D and E:** Representative Western blot of lysates from CFBE41o- cells transiently transfected with WT EMG, intron 16 mutant EMGs (**D**), exon/intron 18 mutant EMGs (**E**), empty plasmid (pcDNA5FRT), and plasmid with *CFTR* cDNA (pcDNACFTR). **F:** Quantification of the mature CFTR (Band C) levels from CFBE41o- and HEK293 cells transiently transfected with WT EMG and mutant EMGs. Protein was detected using anti-CFTR mouse monoclonal antibody 570 (UNC antibody distribution program sponsored by Cystic Fibrosis Foundation Therapeutics, USA). The blots were quantified using Image J software (NIH). Results are expressed as % relative to mature CFTR from pcDNACFTR (mean $\pm$ SD, $n = 3$ independent experiments). **G:** Western blot of lysates from CFBE41o- cells transiently transfected with WT EMG, c.2657+2_2657+3insA*(* refers to mutation was introduced again), c.2657+3delG, c.2657+5G>A, c.2657+5A>G**(** refers to c.2657+5G>A mutated back to WT). **H:** Western blot of lysates from Flp-In 293 cells stably expressing WT EMG, intron 16 mutant EMGs and pcDNACFTR. Upper panels in **D**, **E**, **G**, and **H** shows mature (Band C) and immature (Band B) forms of CFTR translated from RNA transcripts generated by each EMG. The “pcDNA5FRT” and “parental” lanes show that no endogenous CFTR is present in CFBE41o- and Flp-In 293 cells, respectively. Lower panels show GAPDH was used as a loading control.

Figure 4.

Concordance among in silico tools and expression minigenes results in predicting splicing aberration caused by the variations in *CFTR*. Splicing predictions using MaxEntScan (MES), Neural network splice (NN splice), Splice site analyzer tool (SSAT) and Sroogle that are concordant with experimental results are shown inside the gray circle while discordant predictions are outside the gray circle.

Table 1Clinical Characteristics of Patients with *CFTR* Splice Site Variants

Location in <i>CFTR</i>	Variant	Alleles (Frequency in CFTR2)	Sweat [CF] in mmol/L (N)	% PI (N)	FEV1 % predicted (N)	% Pa Infection (N)	Source
3' Splice site Intron 11	c.1585-1G>A	635 (1.01)	102 (336)	97.3 (398)	73.7 (301)	49.7 (179)	CFTR2
	c.1585-2A>G	1	–	–	–	–	CFMD
	c.1585-3T>G	1	–	–	–	–	CFMD
5' Splice site Intron 16	c.1585-8G>A	9 (0.01)	106 (3)	100 (3)	82.4 (3)	0 (0)	CFTR2
	c.1585-9T>A	1	74.6 (1)	–	–	–	CFMD
5' Splice site Intron 18	c.2657+2_2657+3insA	25 (0.04)	76 (18)	33.3 (4)	85.3 (8)	0 (0)	CFTR2
	c.2657+3delG	3 (0.005)	86 (3)	33.3 (3)	109.5 (2)	0 (1)	CFTR2
	c.2657+5G>A	538 (0.86)	96 (291)	47.3 (138)	77.9 (241)	38.4 (114)	CFTR2
5' Splice site Intron 18	c.2988G>A	40 (0.063)	95 (24)	59.3 (16)	70.4 (29)	65.5 (19)	CFTR2
	c.2988+1G>A	266 (0.42)	102 (115)	95.8 (114)	71.7 (92)	51.5 (70)	CFTR2

The DNA mutation numbering system is based on cDNA using +1 as the A of the ATG translation initiation codon in the *CFTR* reference sequence (NM_000492.3) with initiation codon as codon 1 and exons numbered 1–27. Alleles refer to the number of chromosomes on which a particular variant was observed. Sweat chloride concentration is represented as mmol/L. The sweat chloride concentration from a single subject is averaged if performed more than once. PI refers to pancreatic insufficiency, FEV1 % predicted refers to lung function and Pa refers to *Pseudomonas aeruginosa* colonization. The c.1585–3T>G variant is identified in a CF subject with congenital absence of vas deferens (CBAVD). There are no details available on c.1585–2A>G except that it was seen only once in over 100 non-p.Phe508del CF chromosomes screened. Only sweat chloride data is available on c.1585–9T>A variant. CFMD refers to the Cystic Fibrosis Mutation Database (<http://www.genet.sickkids.on.ca>). CFTR2 refers to Clinical and Functional Translation of CFTR (<http://www.cftr2.org/>). Gray highlight indicates variants of unknown functional consequence selected from the CFTR2 project.

Table 2

Comparison of Experimental Results with in Silico Isoform Predictions

Variant	Location	EMG experimental results		In silico isoform predictions									
		Isoform1 (Effect on protein)	Isoform2 (Effect on protein)	Spliceport		HSF		ASSEDA					
				Isoform1	Isoform2	Isoform1	Isoform2	Isoform 1	Isoform 2	Isoform 3	Isoform 4		
c.1585-1G>A	Intron 11	1 nt exon deletion (p.Asp529SerfsX7)	Exon 12 skipped (p.Asp529SerfsX7)	1 nt exon deletion	Exon 12 skipped	1 nt exon deletion	Exon 12 skipped	1 nt exon deletion	Exon 12 skipped	7 nt intron retention	19 nt exon deletion		
c.1585-2A>G	Intron 11	6 nt intron retention (p.Glu528_Asp529insXX)	Exon 12 skipped (p.Asp529SerfsX7)	Exon 12 skipped	Exon 12 skipped	Exon 12 skipped	Exon 12 skipped	7 nt intron retention	Exon 12 skipped	19 nt exon deletion	19 nt exon deletion		
c.1585-3T>G	Intron 11	2 nt intron retention (p.Asp529ArgfsX12)	Exon 12 skipped (p.Asp529SerfsX7)	2 nt intron retention	Exon 12 skipped	2 nt intron retention	Exon 12 skipped	2 nt intron retention	Exon 12 skipped	7 nt intron retention	7 nt intron retention		
c.1585-8G>A	Intron 11	6 nt intron retention (p.Glu528_Asp529insXX)	Exon 12 skipped (p.Asp529SerfsX7)	6 nt intron retention	Exon 12 skipped	6 nt intron retention	Exon 12 skipped	6 nt intron retention	Exon 12 skipped	7 nt intron retention	7 nt intron retention		
c.1585-9T>A	Intron 11	7 nt intron retention (p.Asp529ValfsX41)	Exon 12 skipped (p.Asp529SerfsX7)	7 nt intron retention	Exon 12 skipped	7 nt intron retention	Exon 12 skipped	7 nt intron retention	Exon 12 skipped	7 nt intron retention	7 nt intron retention		
c.2657+2,2657-3insA	Intron 16	WT	Exon 16 skipped (p.Val874HisfsX9)	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped		
c.2657+3delG	Intron 16	Exon 16 skipped (p.Val874HisfsX9)	Normal spliced	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped		
c.2657+5G>A	Intron 16	Exon 1 skipped (p.Val874HisfsX9)	Normal spliced	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped	Exon 16 skipped		
c.2988G>A	Exon 18	Exon 18 skipped (p.Glu970Val506X)	Exon 18 skipped	Exon 18 skipped	Exon 18 skipped	Exon 18 skipped	Exon 18 skipped	Exon 18 skipped	Exon 18 skipped	Exon 18 skipped	Exon 18 skipped		
c.2988+1G>A	Intron 18	Exon 18 skipped (p.Gly970Val506X)	Exon 18 skipped	Exon 18 skipped	Exon 18 skipped	Exon 18 skipped	Exon 18 skipped	Exon 18 skipped	Exon 18 skipped	Exon 18 skipped	Exon 18 skipped		

HSF refers to Human Splicing Finder tool and ASSEDA refers to Automated Splice Site and Exon Definition Analysis tool. Note that the maximum number of isoforms observed in EMG is two. Major transcript is assigned as "Isoform 1" and minor transcript as "Isoform 2". For each predicted isoform, table cell in green indicates agreement; red indicates disagreement and yellow indicates partial agreement.