IDENTIFICATION OF NON-TRADITIONAL MOLECULAR CONTRIBUTORS TO CYSTIC FIBROSIS

by
Melissa Lee

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, MD
June 2016

Abstract

Next generation sequencing has enabled the identification of patients with unambiguous recessive Mendelian phenotypes and incomplete genotypes, meaning zero or only one disease-causing mutation are detected by diagnostic sequencing. These patients serve as a rich source of unannotated disease variation, both within the primary disease locus and beyond. We studied two cohorts of cystic fibrosis (CF) patients with incomplete genotypes: one cohort had one CF-causing mutation in the primary locus *CFTR* after comprehensive molecular diagnostics and one cohort had zero CF-causing *CFTR* mutations.

We developed a novel computational method combining machine learning splice site models with a human-designed algorithm for the selection of high confidence splice variants to identify *CFTR* deep intronic splice variants in "one mutation" CF patients. Our method is unique in assessing the splice potential of reference sequences and comparing variant sequences as a high quality filter. Candidate deep intronic cryptic splice variants were confirmed by *in vitro* analysis with expression minigenes containing relevant introns. Additionally, we identified *CFTR* cryptic splice variants masquerading as missense mutations and experimentally validated two "deep exonic" cryptic splice variants, indicating that exonic cryptic splicing may occur more frequently than commonly appreciated.

We performed exome sequencing on CF families with zero *CFTR* mutations to identify candidate intergenic molecular contributors to CF. Putative deleterious alleles in *CA12* encoding carbonic anhydrase (CA) XII segregated with disease in two unrelated families exhibiting an atypical CF phenotype of elevated sweat chloride concentrations

and hyponatremic dehydration. Bronchiectasis was identified in an adult Caucasian proband in spite of normal pulmonary function testing, suggesting the early stages of obstructive lung disease. The adult proband carried two mutations which were shown to result in aberrant RNA splicing by analysis of patient nasal epithelia. In the second family, two affected siblings carried a homozygous missense in an essential zinc-coordinating residue of CA XII. This mutation was shown *in vitro* to cause a complete loss of CA enzymatic activity. CA XII localized to the sweat gland and airway epithelia in normal primary tissue, providing further evidence for loss of CA XII function as a cause of CF.

Advisor & 1st Reader: Garry R. Cutting, MD

Professor, Department of Pediatrics

McKusick-Nathans Institute of Genetic Medicine

Johns Hopkins University School of Medicine

2nd Reader: Michael J. Parsons, PhD

Associate Professor, Department of Surgery

McKusick-Nathans Institute of Genetic Medicine

Johns Hopkins University School of Medicine

Table of Contents

List of Tables

Chapter 3:

List of Figures

Chapter 2:

Chapter 3:

List of Figures (continued)

Chapter 1

Introduction to the genetic etiology of cystic fibrosis

Cystic fibrosis (CF) is a common autosomal recessive disorder of defective ion transport affecting approximately 70,000 individuals worldwide, including 30,000 Americans. CF is primarily caused by loss of function alleles in the *CFTR* gene, with rare cases of a milder, atypical CF phenotype attributed to defects in the epithelial sodium channel (ENaC)(*1*). *CFTR* encodes the cystic fibrosis transmembrane conductance regulator, a chloride and bicarbonate channel expressed with high specificity in secretory epithelia.

The tissue-specific expression pattern of *CFTR* is reflected in the organs affected in CF, which are primarily the eccrine sweat glands, lungs, pancreas, liver, and male reproductive tract. Salt wasting from the sweat glands causes hyponatremic dehydration in infancy and chronically elevated sweat chloride concentration throughout the lives of CF patients. Sweat chloride concentrations over 60 mEq/L are diagnostic of CF and two standard deviations over the mean sweat chloride concentration for the healthy, non-CF population(*2*). High viscosity respiratory airway surface liquid leads to persistent bacterial and fungal infections(*3*). Obstructive lung disease is the major cause of morbidity and mortality in CF patients in developed countries. Dehydrated, thickened mucosal secretions plug pancreatic ducts, eventually resulting in fibrosis of the ductal tissue and exocrine pancreatic insufficiency(*4*). Malnutrition due to malabsorption is the major cause of mortality in the developing countries where there is poor access to CF therapies(*5*). A high rate of male infertility is observed amongst CF patients due to congenital absence of the vas deferens(*4*). The most common co-morbidity amongst CF patients is CF-related diabetes (CFRD), occurring in 20% of CF adolescents and 40-50%

of CF adults. CFRD is caused by decreased secretion of *and* decreased sensitivity to insulin, making it distinct from both type I and type II diabetes(*6*).

The relatively high frequency of CF has resulted in CF being one of the longest studied Mendelian diseases and *CFTR* being among the first genes associated with disease(*7, 8*). The frequency of CF can be attributed to the F508del allele which is carried by one of every 36 people with European ancestry. 87% of CF patients carry at least one copy of F508del and approximately 50% of CF patients are homozygous for F508del (K. Raraigh, personal communication). The long molecular history of *CFTR* and high frequency of CF have lead to the establishment of CF newborn screening programs in all 50 states(*9*) (https://www.cff.org/What-is-CF/Testing/Newborn-Screening-for-CF/) and the widespread availability of comprehensive *CFTR* molecular diagnostics. All of these factors have contributed to making *CFTR* one of the most commonly interrogated genetic loci in the genome, with over 2000 variants identified (Cystic Fibrosis Mutation Database: http://www.genet.sickkids.on.ca/cftr/Home.html) across the 190 kb locus. The American Cystic Fibrosis Foundation (CFF) has established over 120 accredited CF care centers across the country and estimate that 90% of American CF patients are seen at these specialty clinics. The CFF maintains and releases a yearly patient registry to monitor changes in clinical outcomes across the patient population with the availability of new drugs and treatments, document the clinical progression of individual patients. The CFTR2 project was recently initiated to capitalize on the wealth of clinical data, both in the U.S. patient registry as well as in additional registries worldwide, to investigate genotype-phenotype correlations in patients and thus establish disease liability to *CFTR* variants(*10*). A major motivation behind this work is the rapidly advancing development

of mutation-specific small molecule therapies which can correct misfolded CFTR protein or potentiate the chloride conductance of aberrantly gated CFTR channels(*11*). To identify all patients which may qualify for these mutation-specific drugs, the Cystic Fibrosis Foundation started a program called the Mutation Analysis Program (MAP, https://www.cff.org/For-Caregivers/For-Clinicians/Mutation-Analysis-Program/) to provide CF patients with free comprehensive molecular diagnostics until their genotypes are complete (i.e. until two CF-causing mutations have been identified). The first step of the MAP is a genotyping panel including 23 CF-causing *CFTR* mutations which are responsible for disease in 49-98% of U.S. CF patients depending on their ethnic background(*12*). If a patient still has an incomplete genotype after this step, they are reflexed to sequencing of the *CFTR* coding region and flanking introns. If this does not resolve the patient's genotype, they are reflexed to an assay for the detection of deletions and duplications.

These efforts have revealed that a significant portion of CF patients have incomplete *CFTR* genotypes even after such a comprehensive gamut of diagnostics. According to the most recent release of the CFTR2 project database, approximately 9% of CF patients have only one CF-causing mutation identified and 2-3% have zero CF-causing mutations identified (B. Karczeski, personal communication). Explaining the molecular etiology of the CF phenotype in these patients is the focus of the work presented here.

<u>CF "one mutation" patients</u>

Patients with unambiguous Mendelian disease and incomplete genotypes exemplify important problems in both genetics and medicine. CF has clear diagnostic hallmarks and a relatively large number of well annotated diseause-causing mutations, making CF a strong platform upon which to test new methods by which to complete patient genotypes.

Diagnostic sequencing of the coding region and intronic flanks of *CFTR* is expected to capture the vast majority of CF-causing variation. If diagnostic sequencing is unable to complete the CF genotypes of a significant number of patients, this suggests that there are likely unaccounted disease mechanisms operating. Disease mechanisms which may not be ascertained by diagnostic sequencing include non-coding mutations which affect regulatory motifs such as the promoter or enhancers, variants which alter translational efficiency or RNA folding, and deep intronic cryptic splice variaints. Indeed, we hypothesized that a fraction of CF "one mutation" patients carried deep intronic variants which activated cryptic pseudoexons which would be undetectable by routine diagnostic genomic DNA sequencing of the coding region and intronic flanks. Deep intronic cryptic splice site activation is considered a rare event although the deep introns are infrequently interrogated. Further, there are major clinical implications for CF patients with incomplete genotypes. It is vital to know the molecular mechanism behind each CF-causing variant so that the appropriate small molecule therapies can be administered to patients.

Therefore, "one mutation" CF patients are a rich source of non-coding variants that cause substantial disruption of gene function, given that all patients selected for the study cohort truly have CF. It is possible that white U.S. carriers of F508del experiencing

any type of respiratory ailment may be incorrectly ascertained as manifesting CF and thus may not be genuine "one mutation" CF patients with a second unidentified disease-causing mutation. However, the utilization of stringent diagnostic thresholds for CF can diminish the likelihood that incorrectly ascertained individuals are selected for study. The selection criteria for this study were: (1) one identified CF-causing mutation after full molecular diagnostics of *CFTR*; (2) sweat chloride concentrations ≥ 60 mEq/L; and (3) demonstrable lung disease as reflected by pulmonary function tests and/or microbial cultures characteristic of CF(*13*). Pancreatic insufficiency was not necessary for study selection but was considered to delineate severe, "classic" CF from the milder phenotype.

42 CF "one mutation" patients were enrolled from the Johns Hopkins Twin and Sibling Study, the CFF MAP, and individual referrals. These patients included five previously studied patients who were found to have pathologically decreased transcription from the chromosome *not* carrying the one identified CF-causing mutation(*14*). These five previously studied patients underwent complete sequencing of the *CFTR* coding region and flanking introns which was unable to detect a second CF-causing mutation. An Agilent SureSelect capture for next generation sequencing was designed for the entire 190 kb *CFTR* locus plus 10 kb upstream and 5 kb downstream (B. Vecchio-Pagan, unpublished). This high density tiled capture was designed to detect with high confidence all single nucleotide variants, indels, and larger deletions and duplications. Inspection of the coding region resolved the *CFTR* genotypes of 28 "one mutation" patients, primarily through the detection of multiexon deletions or duplications and rare non-synonymous changes not identified through standard *CFTR* genotyping

panels. This left 14 “one mutation” patients who were highly likely to have non-coding unannotated CF-causing variation.

We hypothesized that these 14 patients carried novel deep intronic variants *in trans* to their identified CF-causing mutation which could pathogenically activate cryptic pseudoexons. Single nucleotide variants have been shown to be sufficient to activate deep intronic cryptic splice sites, both in CF and many other Mendelian diseases(*15-18*). The major spliceosome pathway requires a GT dinucleotide for a functional splice donor or an AG dinucleotide for a functional splice acceptor(*19*); therefore, any variant which either creates a GT or AG dinucleotide or exists within sufficient proximity to a GT or AG dinucleotide can be considered a candidate splice variant. *CFTR* itself has multiple examples of variants which activate deep intronic cryptic splice sites: 3849+10 kb C>T which occurs at a frequency of 0.74% in CF patients (CFTR2 ref) and is associated with mild pancreatic sufficient phenotypes(*20*); and 1811+1.6 kb A>G which completely obliterates normal splicing and is associated with classic CF(*21*).

The relatively high rate of intronic variation and flexible genetic code are important reasons why cryptic splice sites are not more readily identified(*22*). Variants which create novel GT or AG dinucleotides can immediately bring to mind splice mechanisms; however, variants which occur adjacent to existing GT or AG dinucleotides are far more difficult to interpret, especially if the variant is in a more distant position from the dinucleotide. For this reason, the application of artificial intelligence such as machine learning predictive models can greatly assist in determining whether a variant may activate a cryptic splice site(*19*).

"Zero mutation" patients demonstrate genetic heterogeneity of CF

CF patients with *no* disease-causing mutations in *CFTR* demonstrate the genetic heterogeneity of CF and challenge the notion that CF is a single locus disorder. In patients with diagnostic features of CF, *CFTR* can be definitely ruled out as the disease causative locus through linkage exclusion using affected and unaffected family members as demonstrated previously(*13*). In order for *CFTR* to be responsible for the CF phenotype, affected siblings must inherit the same combination of *CFTR* haplotypes and different combinations of *CFTR* haplotypes as their unaffected siblings. Conversely, to exclude *CFTR* as responsible for the observed CF phenotype, affected siblings must inherit different combinations of *CFTR* haplotypes.

We hypothesized that CF patients for whom *CFTR* had been ruled out as the causative locus carried mutations in other genes that were responsible for their disease. Non-*CFTR* genes causing CF could be identified using exome sequencing. Candidate genes could be tested using segregation analysis of non-synonymous variants of affected families, or using a patient cohort approach, both strategies having demonstrated success in revealing genetic contributors to other autosomal recessive Mendelian diseases(*23, 24*). Candidate genes were prioritized if their products were found to be part of the CFTR interactome(*25*), a network of proteins shown to interact with CFTR through ion transport, protein trafficking, and other relevant cellular processes.

Selection of CF "zero mutation" patients for exome sequencing occurred in two phases. Criteria to be considered were: (1) zero identified CF-causing mutations after full

CFTR molecular diagnostics; (2) sweat chloride concentrations ≥ 60 mEq/L; and (3) a history of respiratory illness. This resulted in a selection of patients with a broader spectrum of phenotype severity compared to patients in the "one mutation" study. Many of the selected patients manifested a milder "atypical" or "nonclassic" CF phenotype. Previous work has shown that atypical CF patients whose disease is caused by mutations in *CFTR* are indistinguishable phenotypically from those for whom *CFTR* has been definitively ruled out(*13*).

To definitively rule out *CFTR* as the disease causative locus, linkage exclusion was performed on the patient's family. If *CFTR* was ruled out, the patient and family members were subject to exome sequencing. Linkage exclusion could be inconclusive if none of the patient's siblings inherited the same combination of *CFTR* haplotypes. In these cases of inconclusive linkage exclusion results, patients and family members were subject to exome sequencing. Patients for whom family members were unavailable were subject to exome sequencing for cohort analysis if they satisfied the above listed criteria.

Exome sequencing was performed on atypical CF families with zero CF-causing mutations and non-synonymous and canonical splice variants segregating with disease were given higher priority. Predicted deleterious variants in the *CA12* gene encoding carbonic anhydrase XII were found segregating with disease in two unrelated ethnically diverse families for whom *CFTR* was ruled out by linkage exclusion. Both families exhibited atypical CF: affected individuals had diagnostically elevated sweat chloride concentrations, persistent episodes of hyponatremic dehydration, and experienced failure to thrive in infancy. *CA12* had been previously implicated in a large consanguineous Bedouin pedigree presenting this phenotype which was reported under the name

"hyperchlorhidrosis" in reference to excessive sweat gland salt wasting(*26, 27*). However, the mutation described in these previous reports appeared to be moderately hypomorphic, inconsistent with the loss of function expected with recessive disease in consanguineous pedigrees.

CA12 as a molecular contributor to atypical CF was an attractive candidate because of its role metabolizing bicarbonate to maintain proper physiological pH across epithelial membranes. Carbonic anhydrases are known to form bicarbonate transport metabolons: loose complexes of ion metabolizing enzymes and ion transport channels that facilitate the establishment and maintenance of electrochemical and electrogenic gradients(*28*). Bicarbonate transport metabolons have been described for three of the four identified transmembrane carbonic anhydrases: CA IV associates with AE1(*28, 29*), CA IX associates with AE2(*30*), and CA XIV associates with AE3(*31*). A CA XII transport metabolon has yet to be identified but it is likely that disturbing this mechanism through loss of CA XII function would have a deleterious impact on ion transport.

CA XII as a metabolizer of bicarbonate was further compelling due to persisting questions regarding the role of bicarbonate in CF. CFTR is known as an epithelial chloride channel; however, it also conducts bicarbonate and it has long been suspected that CF is as much a disease of abnormal bicarbonate transport as abnormal chloride transport(*32*). Additionally, the development of the CF pig model has revealed that the increased likelihood of respiratory infections in CF is due to pathologically low pH of the airway surface liquid, resulting in increased bacterial colonization and subsequent airway damage(*3, 33*). It was therefore considered possible that loss of CA XII function could result in a remarkably CF-like phenotype.

Chapter 2

Systematic computational identification of variants that activate exonic and intronic cryptic splice sites

Abstract

We developed and tested a novel variant annotation method that combines sequence-based machine learning classification with a context-dependent selection algorithm for the systematic identification of splice variants. Our approach is unique in that it compares the splice potential of a sequence bearing a variant to the splice potential of the reference sequence. The classifier model accurately identified 93.3% (168 of 180) canonical splice sites across five disease genes. The method identified and correctly predicted the effect of 18 of 21 (86%) known splice altering variants in *CFTR*, a well-studied gene whose loss of function alleles cause cystic fibrosis (CF). Among 1423 unannotated *CFTR* disease-associated variants, the method identified 32 novel cryptic splice variants, two of which were experimentally verified as activating "deep exonic" cryptic sites. Complete sequencing of the *CFTR* exons and introns in 14 CF patients with incomplete genotypes revealed one known (three patients) and two novel (three patients) experimentally verified intronic cryptic splice variants. Application of the method to six individuals with X-linked dyskeratosis congenita and incomplete *DKC1* genotypes identified two cryptic splice variants that were experimentally verified in patient cells. High confidence candidate splice variants selected from ClinVar "pathogenic" single nucleotide variants or "variants of uncertain significance" revealed that 28.1% (129 of 458) and 21.6% (75 of 348), respectively, were predicted to activate cryptic splice sites. Our findings indicate that cryptic splice site activation appears to be more common than previously predicted and should be routinely considered for all variants segregating with disease.

Introduction

Next generation sequencing has enabled the examination of complete gene sequences resulting in the detection of vast numbers of variants in the coding and non-coding regions of disease associated genes. Variants thought to have a deleterious effect on RNA splicing are limited primarily to the canonical splice sites. When exonic variants outside of the canonical splice sites *are* considered to impact RNA processing, disruption of splice enhancers is generally implicated, causing a significant reduction in the spliceosomeal recognition of the canonical splice sequence(*34*). However, exonic variants can activate cryptic splice sites leading to aberrant RNA splicing and loss of coding sequence(*35*). Such examples are sparse in the literature and thus the frequency of exonic cryptic splicing is unknown. It has long been suspected but not systematically shown that variants which activate splice sites have been masquerading as exonic protein altering variants or "benign" synonymous variants, leading to underappreciation of the frequency of this pathologic mechanism(*22, 36, 37*). Identifying whether exonic variation affects RNA rather than protein processing is vital as our field seeks to deploy personalized medicine initiatives such as drug therapies aimed at addressing specific disease mechanisms.

By the same token, intronic variants outside of canonical splice sites can also affect RNA processing by splice site activation. Detection of intronic cryptic splice variants is challenging as the genomic space is much larger and there is far less evolutionary constraint, resulting in a higher degree of variation. Importantly, diagnostic sequencing rarely interrogates the deep introns although deep intronic cryptic splice events have been reported when RNA transcripts are studied from patients with

incomplete Mendelian genotypes(*15-18*). Deep intronic cryptic splice activation is considered a rare and exotic event and the true frequency remains unknown. Given the higher degree of variants in the introns relative to the exons and the flexibility of splice sequences, it is possible that intronic variants may activate cryptic splice sites more often than is currently recognized. Indeed, case studies of deep intronic cryptic splice activation events have shown that a single nucleotide substitution is sufficient to cause severe disease, and that these variants do not always create novel GT or AG dinucleotides, but may also occur adjacent to existing GT or AG dinucleotides(*21, 38*). These examples suggest that single nucleotide variants in the deep introns may be as culpable as those in the exons and should be given similar scrutiny. Identifying the frequency of deep intronic cryptic splice events as well as the number of exonic variants which alter splicing is vital to the completion of Mendelian genotypes and our understanding of disease molecular mechanisms.

Materials and Methods

The computational method described here consists of two parts: splice sequence classification and splice variant selection. The foundation of classification is the application of predictive models trained on splice donor and acceptor sequence data. Selection uses the predictions provided by the classifier models to compare the splice potentials of candidate splice variant sequences to the splice potentials of the reference sequences. Filtering by distance during selection removes unlikely candidates and generates the final "high confidence" candidate splice variants.

Splice sequence classification

- Training data
    - Splice sequences from NN269(*39*) and HS3D(*40*) were utilized as training data. The NN269 and HS3D datasets contain "true" donor and acceptor sequences curated from GenBank genes annotated for canonical splice sites. "False" sequences in the training data are sequences with GT or AG dinucleotides at least 60 bp away from canonical splice sites and shown to not be recognized by the spliceosome. Donor sequences extend to seven nucleotides upstream of GT ("-7") to six nucleotides downstream of GT ("-6"). Acceptor sequences extend 68 nucleotides upstream of AG ("-68") to 20 nucleotides downstream of AG ("+20"). 1116 "true" donors from NN269 and 2796 "true" donors from HS3D were combined into a single "true" donor training data set. 1116 "true" acceptors from NN269 and

2880 “true” acceptors from HS3D were combined into a single “true” acceptor training data set. “False” sequences were randomly and proportionally selected from NN269 and HS3D to match the number of “true” sequences.

- Features
    - Features were based upon previously published features(*41*) and were chosen because of the comprehensiveness of sequence information captured. Statistical difference tables were not utilized to reduce bias and were substituted with binary coding of a feature present in the input. There were three types of features. Briefly, the “component” feature separates splice sequences into left and right segments at the consensus dinucleotide and calculates the probability of a sequence substring occurring in the left and right segments. Probabilities for all possible sequence substrings up to a length of five nucleotides were calculated. Second, the “position” (“pos”) feature describes whether a given nucleotide exists at a given position across the length of the splice sequence. Third, the “adjacent position relationship” (“apr”) feature describes whether a given dinucleotide exists at a given position across the length of the splice sequence. Feature data was extracted by custom Python scripts.
- Model selection
    - Support vector machine (SVM) models were trained using custom scripts employing the Python scikit learn machine learning library(*42*). Feature

data were transformed with linear and RBF kernels to determine if reduction in data complexity were required to achieve better model performance. Classifier models were trained with ten-fold cross-validation in which the training data were randomly split into ten equal pieces. Classifiers were trained on nine pieces and tested on the unseen tenth piece. This process was repeated ten times and averaged to determine model generalizability as measured by a distribution of accuracies. Classifiers were designed to provide probability estimates in addition to the typical binary classification of "true" or "false." Probability estimates can be considered the splice potential of a sequence, with 1 reflecting a perfect splice site and 0 indicating a sequence which does not resemble a splice site at all. Classifier models trained with RBF transformed data had the best accuracies and were utilized in all further applications.

- Creating candidate sequences
    - Candidate splice sequences were created from variants of interest and the surrounding NCBI RefSeq gene reference sequence using a custom Python script. Variants of interest were able to generate candidate splice sequences if they fulfilled either criteria: (1) the variant creates a GT or AG dinucleotide, or (2) the variant occurs within -3,+5 of an existing GT dinucleotide or -22,+3 of an existing AG dinucleotide. If either or both criteria were satisfied, sequences were extracted -7,+6 of GT dinucleotides and/or -68,+20 of AG dinucleotides. In order to provide a contextual comparison in the change of a sequence's splice potential when a variant

is introduced, reference gene sequences were subject to classification as well. Sequences were extracted -7,+6 of all GT dinucleotides and -68,+20 of all AG dinucleotides throughout the full reference sequences for all genes in this study.

- Evaluation of candidate sequences
    - All candidate splice sequences and reference sequences were subject to classification utilizing the best performing models as determined by the highest accuracies after training. Candidate sequences were given a binary classification of "true" or "false" and a probability estimate from 0 to 1.

Splice variant selection

- A selection algorithm to generate high confidence candidate splice variants was developed to compare the splice potentials of variant sequences (P(var)) to those of the reference sequence (P(ref)) and the canonical splice site (P(canon)). A variant can cause aberrant splicing in three scenarios and two metrics were utilized to test variants for these scenarios:
    - $\Delta_{variant}$: The difference in splice potential of the variant sequence P(var) and the reference sequence P(ref). Positive values indicate an increase in the splice potential with introduction of a variant and negative values indicate a weakening of the sequence's splice potential. This metric can only be calculated for variants which do *not* create a GT or AG dinucleotide. (A variant which creates a GT or AG dinucleotide cannot be compared to the reference sequence because the GT or AG dinucleotide

does not exist in the reference.) Multiple thresholds of $\Delta_{variant}$ were tested to see which captured all instances in a truth set of known *CFTR* splice variants, resulting in the default minimum threshold for this metric being set to |0.05|.

- Δ_{canon}: The difference in splice potential of the variant sequence P(var) and the canonical splice site P(canon). Positive values indicate the variant sequence has a greater splice potential than the canonical splice site. Negative values indicate the variant sequence has a weaker splice potential than the canonical splice site. Multiple thresholds of Δ_{canon} were tested to see which captured all instances in a truth set of known *CFTR* splice variants, resulting in the default minimum threshold for this metric being set to |0.05|.

- Scenario 1: Weakening of the canonical splice site
  - Candidate variant sequences qualified for this scenario if P(var) ≤ 0.85 and the variant occurred within -3,+5 of a canonical splice donor or -22,+1 of a canonical splice acceptor. The Δ_{canon} had to be ≤ -0.05 to sufficiently weaken the canonical splice site in a biologically meaningful way as determined by application to a truth set of variants known to weaken *CFTR* canonical splice sites.

- Scenario 2: Deep intronic cryptic splice activation
  - Candidate variant sequences qualified for this scenario if P(var) for candidate donors was ≥ 0.7 or if P(var) for candidate acceptors was ≥ 0.6.

These thresholds were set based on the minimum probability assigned to known canonical splice sites. The variant sequence had to occur at least 100 bp into the intron and, if applicable, have $\Delta_{variant} \geq +0.05$. If a $\Delta_{variant}$ could not be calculated due to a novel GT or AG dinucleotide being created by the variant, the variant sequence was flagged if the first two criteria were satisfied.

- Scenario 3: Near canonical cryptic splice activation leading to outcompeting of the canonical splice site
    - Candidate variant sequences qualified for this scenario if P(var) for candidate donors was ≥ 0.7 or if P(var) for candidate donors was ≥ 0.6. These thresholds were set based on the minimum probability assigned to known canonical splice sites. The variant had to occur within 60 bp of the canonical splice site and, if applicable, have a $\Delta_{variant} \geq +0.05$. Based on a truth set of variants known to activate cryptic sites which outcompete the canonical, in this scenario the Δ_{canon} is softened by 0.1 to capture true examples. After this adjustment, the candidate variant sequence had to have a $\Delta_{canon} \geq +0.05$.

Sequencing of the full *CFTR* locus in CF one-mutation patients.

Genomic DNA extracted from patient whole blood using a standard phenol/chloroform protocol. A custom designed Agilent SureSelect capture was used to pull down the 215 kb region containing and surrounding *CFTR* in each sample. Samples

were were run on an Illumina HiSeq 2500. A custom designed next generation sequencing data processing pipeline was used to align reads and call variants in all samples. This pipeline included alignment by the BWA algorithm, duplicate removal by Picard, local realignment by GATK, and variant calling by four additional softwares. The intersection or merge of select variant callers was used for downstream analysis. Additionally, large indels and CNVs were called using three algorithms, but was primarily reliant upon Conifer calls.

Introduction of variants to expression minigenes.

Expression minigenes (EMGs) were developed as described previously(*43*). Candidate splice variants were introduced into WT EMGs by site-directed mutagenesis (SDM) using oligonucleotides with the candidate splice variant and flanking 20 nt sequences identical to the region of interest. SDM reactions were performed in triplicate, pooled, and digested with Dpn1 to remove plasmid template. Purified SDM products were transformed with XL10 Gold supercompetent cells and clonally expanded overnight at 37C with shaking in LB broth with ampicillin. DNA was extracted by MiniPrep and mutagenesis was confirmed by Sanger sequencing.

Transient transfection of HEK293 cells with EMGs

4 ug of EMG plasmid was diluted in 250 ul of OptiMEM and combined with 7 ul of Lipofectamine 2000 (Invitrogen) diluted in 250 ul of OptiMEM. The Lipofectamine

2000 complexes were added to confluent HEK293 cells that had been grown in antibiotic-free media for 24 hours prior to transfection. 1 ml of antibiotic-free media was added to cultures four hours post-transfection. Media was changed with antibiotic-free media 24 hours post-transfection. Cells were lysed for analysis of protein and mRNA transcripts 48 hours post-transfection.

Analysis of mRNA transcripts and protein from transiently transfected HEK293 cells

Cells were washed twice with 1X PBS 48 hours post-transfection. A standard RNA extraction using TRIzol/chloroform was performed. RNA was prepared in a DNA-free bench with dedicated plasticware and pipets. RNA preparations were treated with DNase and RT-PCR was immediately performed following RNA extraction. PCR of HEK-derived cDNAs was performed using primers lying in the exons and aberrant splice products were visualized by gel electrophoresis. Bands of interest were extracted and purified and verified by Sanger sequencing.

To analyze protein, cells were washed twice with 1X PBS 48 hours post-transfection and lysed with 250 ul of RIPA buffer containing protease inhibitors and PMSF. Cell lysates were incubated on ice for 30 minutes and vortexed for 20 seconds every 10 minutes. Cell lysates were spun at 4C for 15 minutes and the supernatant was retained for Western blotting. Loading samples were then prepared. In order to visualize CFTR clearly, the loading sample contained a minimum of 40 ug total protein. The amount of protein lysate required for 40 ug of total protein was calculated using the total protein concentrations previously determined by BCA assay. The volume of individual

protein lysates, in combination with 1X PBS, accounted for 3/4 of each total sample volume. The remaining 1/4 of the total sample volume was a dye solution containing a 1:5 dilution of DTT to 4X Laemeli Buffer. After prepared, the loading samples were further denatured by incubating at 37C for 15 minutes. Samples were loaded into a 7.5% Tris-HCL, 1.0 mm BIO RAD, Criterion Precast Gel and run with 1X Running Buffer prepared in-house. The samples were bookended with All-Blue Precision Protein Standard Ladder (BioRad) for protein size comparison. Samples were run at 150V for 2 hours. The gel was transferred to a PVDF membrane using the Trans-Blot Turbo Transfer System at 2.5A, 25V for 10 minutes. The membrane was blocked for one hour in a 5% blocking solution of non-fat dry milk reconstituted in PBST, 1X PBS containing 0.1% tween-20. The membrane was washed in PBST, and then incubated for an hour with primary anti-CFTR m570 or m596 antibody diluted 1:5000. The membrane was washed again with PBST for 30 minutes. The membrane was then incubated for one hour with an anti-mouse secondary antibody (GE Healthcare) diluted 1:150,000. The secondary anti-mouse antibody was removed and the membrane was washed thoroughly with PBST for 45 minutes prior to imaging.

Analysis of mRNA transcripts from CF patients

Human nasal epithelial (HNE) cells were collected by nasal cytology brush by brushing the inferior surface of the inferior nasal turbinate of each nostril from each patient as previously described(*44, 45*). HNE cells were collected under IRB approval from the University of Alabama at Birmingham (IRB # F090916001). RNA was

extracted from HNE cells using Qiagen RNeasyPlus Mini Kit following the manufacturer's protocol. RNA was eluted in 30 ml RNase-free water. The quantity and quality of RNA was determined by OD260 and OD260/OD280, respectively, using the NanoDrop ND-1000.

RT-PCR of patient-derived RNA was performed using the BioRad iScript kit and a total of 250 ng of RNA. Patient cDNA was amplified as-is as well as in a 1:10 dilution. Products were visualized by gel electrophoresis, extracted, and purified, and subject to Sanger sequencing to verify aberrantly spliced transcripts.

Analysis of mRNA transcripts from DKC patients

Patient lymphocytes from whole blood were used to establish lymphoblastoid cell lines as previously described(*46*). RNA from patient lymphoblastoid cell lines was isolated using the Qiagen RNeasy following the manufacturer's protocol. RT-PCR of RNA from lymphoblastoid cell lines established from patient-derived lymphocytes was performed using the Invitrogen Super Script First Strand cDNA synthesis kit. Patient cDNA was amplified using primers lying in distant exons as well as primers laying across the intron 2-exon 3 boundary to select transcripts with retained intron 2 sequence. Products were separated, extracted, and purified by gel electrophoresis and cloned using the Thermo Fisher TOPO Cloning kit. DNA isolated from clones was subject to Sanger sequencing to verify aberrantly spliced transcripts.

All genomic coordinates correspond to build hg19/GRCh37 and all *CFTR* variants and exons are identified by HGVS nomenclature, and *not* with legacy names unless otherwise indicated.

Results

Training and classification of candidate splice donor and acceptor sequences

The foundation of our method was the application of machine learning models to agnostically evaluate any sequence containing a variant as a splice sequence so long as the sequence contained the requisite GT or AG dinucleotide. Support vector machine (SVM) donor and acceptor classifier models were trained using canonical "true" and "false" splice sequences from annotated gene sequences in GenBank(*39, 40*). Classifier models were trained on feature data that reflected the essential distinguishing characteristics of a genuine splice sequence; therefore, feature data were entirely sequence-based and included nucleotide content and substring frequency(*41*) (see **Materials and Methods**). Classifier models were internally validated using ten-fold cross validation and the highest performing models were selected for classification of candidate splice sequences. Classifier models provide a binary classification ("true" or "false") and a probability estimate from 0 to 1. Probability estimates can be considered approximations of a sequence's splice potential, where a sequence with a score of 1 most closely resembles a genuine splice sequence.

To predict the impact of a variant on the splice potential of a given sequence, variants were introduced into NCBI reference gene sequences to create candidate splice sequences which were subject to classification. Variants created candidate splice sequences only if they satisfied these criteria: (1) the variant created a novel GT or AG, or occurred -3 to +5 of an existing GT (relative to GT at position 00) and/or -22 to +1 of an existing AG (relative to AG at position 00); and (2) the variant did not occur outside

the exons and introns of the gene of interest (i.e. in the UTRs) (**Figure 1A**). If these criteria were met, sequence windows were extracted -7 to +6 of the GT dinucleotide and/or -68 to +20 of the AG dinucleotide. These larger window sizes were chosen to maximize inclusion of variants likely to impact splicing based on numerous splice site studies(*47*). Variants often occurred in sufficient proximity to multiple GT or AG dinucleotides to result in multiple overlapping sequences. Thus, multiple candidate splice sequences were assessed for some variants.

A unique and novel feature of our method is the context-dependent determination of how a variant alters the splice potential of its native context. Due to the inactivation of ancient exons over evolutionary time, the genome is littered with many sequences with high splice potential even in the absence of variation. These cryptic splice sites are especially vulnerable to activation. The accurate identification of cryptic splice sites therefore requires comparison of the splice potential of the variant splice sequence to the splice potential of the reference sequence. To assess the splice potential of native reference sequences, full reference gene sequences were also subject to classification.

Validation of the splice sequence classifier models using canonical splice sites of five disease-associated genes

To evaluate the sensitivity of the classifier models, classification was performed on the entire gene sequences for and five disease-associated genes: *BRCA2*, *CFTR*, *DKC1*, *HEXB*, and *LMNA*. 88 of 90 (97.8% sensitivity) canonical donors across these five genes were accurately given a “true” classification, and 80 of 90 (88.9% sensitivity)

of canonical acceptors were accurately given a "true" classification. The combined sensitivity of the donor and acceptor classifier models was 93.3%. The probability estimates of these 90 canonical splice sites were used to set the minimum thresholds for the P(var) of selected candidate variant sequences as described above (canonical donors had probabilities over 0.7 and canonical acceptors had probabilities over 0.6). (See **Materials and Methods**.)

Selection of high confidence splice variants

The premise of this work was that variants could activate or inactivate splice sites; therefore, any method which accurately detected cryptic splice activation should also be able to detect disruption of canonical splicing. A selection algorithm was designed to assess each variant for its possible effect on splicing by comparing the splice probabilities of candidate splice sequences bearing variants of interest (P(var)) to the splice probabilities of reference sequences (P(ref)) and/or nearby canonical splice sites (P(canon)) (**Figure 1B**). The Δ_{canon} metric is defined as the splice potential of the variant sequence (P(var)) minus the splice potential of the canonical site (P(canon)). The $\Delta_{variant}$ metric is defined as the splice potential of the variant sequence (P(var)) minus the splice potential of the reference sequence (P(ref)). (Variants which create novel GT or AG dinucleotides could not be compared to the corresponding reference sequences due to the lack of the necessary dinucleotide.) The thresholds for $\Delta_{variant}$ and Δ_{canon} were established using known standards in variant annotation and calling (i.e. truth sets) of canonical splice sequences and known splice variants. (See **Materials and Methods**.) The $\Delta_{variant}$

and Δ_{canon} metrics enabled selection of high confidence splice variants based on three possible scenarios of aberrant splicing: (1) weakening of the canonical splice site (**Figure 1B, *left***); (2) activation of a cryptic splice site leading to outcompeting of the canonical splice site (**Figure 1B, *center***); and (3) activation of a deep intronic cryptic splice site leading to pseudoexon inclusion (**Figure 1B, *right***).

Validation of the context-specific selection algorithm on known *CFTR* splice altering variants.

The sensitivity of the splice variant selection algorithm was tested using 21 *CFTR* splice altering variants manually curated from the literature(*20, 21, 43, 48-50*). Eight of the 21 *CFTR* splice altering variants cause cryptic splice site activation either in the deep intron or near enough to a canonical splice site to outcompete it. The remaining 13 known *CFTR* splice altering variants weaken the canonical splice sites in which they occur. All but three of the 21 known splice variants were correctly predicted to alter splicing by their previously reported mechanisms using default thresholds set to $\Delta_{variant} = 0.05$ and $\Delta_{canon} = 0.05$, yielding a sensitivity of 85.7%. Two variants were classified by the predictive model to allow splicing as they generated high probability estimates: c.2988 G>A (rs121908797, P(var) = 0.889) and c.1766+3 A>G (rs397508298, P(var) = 0.940); however, *in vitro* experimental studies indicate that these variants disrupt splicing(*43, 50*). Upon further inspection, it became apparent that the P(var) for the candidate splice sequences with either c.2988 G>A or c.1766+3 A>G were statistical outliers in the distribution of P(var) for variants reported to abolish *CFTR* canonical splice sites, lying at

least three standard deviations above the mean (0.263 ± 0.313). The third variant, c.2816 A>G (rs397508440), was excluded during splice variant selection due to the specification that cryptic splice sites which outcompete the canonical splice site lie within 60 bp of the canonical splice site. This 60 bp threshold was established because exonic cryptic splice sites are unlikely to be utilized if the remaining exon is of insufficient length(*51*). c.2816 A>G activates a cryptic donor which outcompetes the canonical splice site (see experimental data below), but is 97 bp upstream of the canonical donor. This validation exercise using known *CFTR* splice altering variants gave the selection algorithm a sensitivity of 85.7%. Most importantly, the 18 selected known splice variants were assigned the correct mechanism by which they altered splicing according to previous reports (i.e. c.3700 A>G activates an exonic cryptic site(*35*)).

Evaluation of CF-associated *CFTR* variants for effect on RNA splicing

To test the ability of the classification and selection method to identify variants that affect splicing, we evaluated *CFTR* variants that had been detected in individuals with cystic fibrosis (CF, OMIM #219700). 1477 variants were obtained from the CFTR2 database, an extensive catalog of *CFTR* variants (cftr2.org(*10*)). After removing variants with nonstandard HGVS nomenclature cDNA naming, variants in the promoter and UTRs, and variants with indeterminate breakpoints, 1423 variants from CFTR2 were left to be evaluated. 308 of the 1423 variants either created a GT or AG dinucleotide or occurred within sufficient proximity of an existing GT or AG dinucleotide to create 3315 candidate donor and acceptor sequences (**Figure 2**). 98 of the 308 (31.8%) variants were

selected as having a high likelihood of altering native splice patterns (**Table S1**). All candidate variants were assessed for the manner in which they were predicted to disrupt RNA processing, as well as predicted protein impact. 18 (18.4%) of the 98 selected predicted splice variants were in an exonic nucleotide position *other* than the first or last nucleotide of the exon, locations that are known to affect splicing when altered. 32 (32.7%) of the 98 selected variants were predicted to activate cryptic splice sites,14 of which were exonic. Five of the 32 cryptic splice variants were predicted to weaken the canonical acceptor while simultaneously activating a cryptic acceptor. Two of these five variants (c.1585-8 G>A, rs193922503 and c.1585-9 T>A, rs397508234) have been previously reported to activate cryptic acceptors that outcompete the canonical acceptor(*43*). Notably, 33 (33.7%) of the 98 variants were predicted to be "missense" variants, suggesting that a significant portion of variants with a high likelihood of causing aberrant splicing could be mistaken as protein altering, rather than RNA altering.

Experimental evaluation of predicted exonic cryptic splice variants in *CFTR*

Two of the "missense" exonic variants predicted to activate cryptic splice sites were selected for *in vitro* investigation using expression minigenes (EMGs). EMGs containing multiple entire or abridged introns and the entire coding region of *CFTR* enable simultaneous analysis of the consequence of variants upon RNA splicing and protein translation in a near native context(*34, 43*). c.454 A>G (p.Met152Val, rs397508721) was predicted to create a novel GT dinucleotide, activating a cryptic donor (P(var) = 0.996) in exon 4 that could outcompete the canonical donor (Δ_{canon} = +0.162)

resulting in a 36 bp in-frame deletion (**Figure 3A, *top panel***). To test this prediction, c.454 A>G was introduced into exon 4 of a *CFTR* EMG with sequences from flanking introns (2, 3, 4, and 5). RNA extracted from HEK293 cells transiently transfected with the resulting EMG was reverse transcribed and PCR amplified (RT-PCR). Sanger sequencing of this cDNA confirmed the predicted 36 bp deletion (**Figure 3A, *middle panel***). As expected, mature protein was not generated by the EMG bearing the c.454 A>G variant compared to the WT *CFTR* EMG (**Figure 3A, *lower panel, lane 4***). A faint signal corresponding to the molecular mass of immature protein was observed, consistent with the presence of incompletely glycosylated protein misfolded due to an in-frame deletion of 12 amino acids. To demonstrate the importance of studying variants in a near native context including relevant introns, c.454 A>G was introduced into plasmids bearing *CFTR* cDNA only and *no* introns. Western blotting showed mature, properly folded protein as seen in WT controls. The observation of stable and normally processed CFTR protein is consistent with the *in silico* prediction that c.454 A>G predicts a benign amino acid substitution (methionine to valine) by PolyPhen and SIFT.

c.2816 A>G (p.His939Arg, rs397508440) occurs in the +3 position of an existing GT in the middle of exon 17 and was predicted to activate a cryptic donor ($\Delta_{variant}$ = +0.783) 97 bp upstream of the canonical donor (P(canon) = 0.915) resulting in a frameshift starting at codon 939 and a premature termination codon (PTC) nine codons downstream (**Figure 3B, *top panel***). To test this prediction, c.2816 A>G was introduced into exon 17 of an EMG with sequences from introns 14 through 18. Sequencing of RT-PCR products from transiently transfected HEK293 cells confirmed the predicted 97 bp deletion **(Figure 3B, *middle panel***). Mature protein was produced by the normally

spliced WT EMG (**Figure 3B, *lower panel*, *lane 4***) but not by the EMG bearing c.2816 A>G, consistent with the expectation of nonsense-mediated RNA decay (NMD) of aberrantly spliced transcript bearing a PTC. Very low amounts of immature protein were observed for c.2816 A>G, likely due to the generation of some normally spliced transcript bearing the predicted histidine to arginine amino acid substitution. These data show that c.2816 A>G should be annotated as disease causing due to missplicing of RNA leading to pathogenic decrease in RNA transcript. Together these results verify that c.454 A>G and c.2816 A>G, although annotated as missense variants, activate exonic cryptic splice sites. These experiments further highlight the importance of studying exonic variants in a near native context incorporating relevant introns in order to accurately determine the underlying disease molecular mechanism.

Identification of novel intronic cryptic splice variants in *CFTR*

The entire *CFTR* gene was sequenced in 14 individuals with diagnostically elevated sweat chloride concentrations and phenotypic features of CF but only one disease-causing mutation identified after analysis of all exons and flanking introns. As these individuals had compelling clinical evidence of CF, it was likely that they carried a second deleterious variant outside of the regions analyzed, such as an intronic variant that affected *CFTR* splicing(*18*). Sequencing identified 41 candidate intronic *CFTR* variants in these 14 subjects after excluding 388 intronic variants that were found *in cis* with CF-causing variants in 31 CF patients with two known CF-causing alleles. The 41 intronic variants created 27 candidate donor sequences and 48 candidate acceptor sequences.

Given the small number of candidate sequences generated and the assumption of deep intronic cryptic splice activation, all candidate sequences with a "true" classification were evaluated manually. Five different intronic variants created sequences that were classified as "true" splice sequences. One variant, c.3140-26 A>G (rs76151804), was identified in three patients and had been previously characterized as CF-causing(*49*). This variant creates an AG dinucleotide and a novel acceptor with a probability of P(var) = 0.896. The remaining four variants were predicted to activate deep intronic cryptic donors: c.1680-877 G>T (two subjects; rs397508261, $\Delta_{variant}$ = +0.228); c.3717+40 A>G (rs397508595, $\Delta_{variant}$ = +0.071); c.1584+689 G>A (chr7:117200398, $\Delta_{variant}$ = +0.361); and c.2491-1190 C>T (chr7:117233794, P(var) = 0.994).

Experimental evaluation of predicted deep intronic cryptic splice variants in *CFTR*

c.1680-877 G>T was identified in two of the 14 subjects. Both were of Hispanic ancestry and carried F508del on their other *CFTR* gene. The variant is in the +4 position of an existing GT dinucleotide and was predicted to create a novel deep intronic donor ($\Delta_{variant}$ = +0.228). An intronic acceptor (chr7:117229470, P(ref) = 0.870) was predicted 60 bp upstream of the variant. Spliceosomal recognition of the intronic acceptor at chr7:117229470 and the cryptic donor activated by c.1680-877 G>T would result in the inclusion of a 53 bp pseudoexon between exons 12 and 13 in the final processed mRNA (**Figure 4A, *upper panel***). Notably, the variant is nine bp downstream from a different, previously identified CF-causing deep intronic cryptic donor created by c.1680-886 A>G (*CFTR* legacy name 1811+1.6kb A>G, rs397508266(*21*). c.1680-886 A>G activates a

cryptic exon using the same acceptor predicted for c.1680-877 G>T. EMGs bearing c.1680-877 G>T and abridged intron 12 generated *CFTR* transcript containing the predicted 53 bp cryptic pseudoexon between exons 12 and 13 (**Figure 4A, *middle panel***). Western blotting of EMGs bearing c.1680-877 G>T showed a complete loss of normal mature CFTR protein compared to WT EMG i12 using two different antibodies with epitopes on opposite termini of CFTR (**Figure 4A, *lower panel***). Protein expression from WT EMG i12 was consistent across multiple transfections (n=4) but, for unknown reasons, lower than that observed from other EMGs.

c.3717+40 A>G was identified in an adult patient carrying F508del, exhibiting features consistent with a mild CF phenotype (mean sweat chloride concentration of 83 mEq/L and pancreatic sufficiency). c.3717+40 A>G was predicted to create a novel donor sequence ($\Delta_{variant}$ = +0.071) 40 bp from a canonical splice donor. Furthermore, the novel donor created by c.3717+40 A>G was predicted to outcompete the nearby canonical donor (Δ_{canon} = +0.261) (**Figure 4B, *upper panel***). The EMG bearing c.3717+40 A>G and abridged introns 21, 22, and 23 generated *CFTR* transcript including the first 40 nucleotides of intron 22 immediately followed by exon 23, consistent with splicing occurring at c.3717+40 A>G (**Figure 4B, *middle panel***). Inclusion of the 40 intronic nucleotides resulted in a frameshift and introduction of a PTC at codon 1248. Western blotting of cell lysates showed very low amounts of truncated CFTR protein with a molecular mass of 142 kDa consistent with translation of aberrantly spliced transcript (**Figure 4B, *bottom panel***). The amount of CFTR protein generated by the EMG bearing c.3717+40 A>G was greatly reduced compared to controls consistent with the reduction in misspliced transcript due to NMD. (Very low amounts of protein

migrating at a size consistent with mature protein were also observed, suggesting that a degree of normal splicing occurs in the presence of c.3717+40 A>G, resulting in translation of minimal WT CFTR protein, consistent with the mild CF phenotype observed in this patient).

c.1584+689 G>A was predicted to create a novel donor ($\Delta_{variant}$ = +0.361) that would activate a deep intronic cryptic pseudoexon, leading to loss of transcript due to NMD. This variant was identified in a patient whose one identified CF-causing variant was c.489+1 G>T (legacy name 621+1 G>T). Previous studies of RNA from this patient's nasal epithelia did not show inclusion of a pseudoexon; however, RNA transcription from the chromosome *not* bearing c.489+1 G>T was significantly diminished, consistent with the prediction of NMD(*14*). The fourth variant, c.2491-1190 C>T was predicted to create a novel GT (P(var) = 0.994) that would activate a deep intronic cryptic donor. However, it was not considered a strong candidate to activate cryptic exon splicing due to the absence of a suitable upstream acceptor. To test this prediction, *CFTR* RNA transcripts from the nasal epithelial cells of a pair of dizygotic CF twins carrying this variant *in trans* with F508del were analyzed. No aberrant splicing of transcript from the non-F508del allele was observed in either subject (data not shown).

Identification and characterization of novel *DKC1* intronic splice variants in dyskeratosis congenita patients

To determine the generalizability of the splice variant classification and prioritization method to other genes, we analyzed variants discovered in individuals with

dyskeratosis congenita (DKC, OMIM #305000). DKC is a genetically heterogeneous multisystem disease of shortened telomeres resulting in a variable phenotype including skin abnormalities and mortality from bone marrow failure. An X-linked recessively inherited form of DKC has been identified and 50% of the cases can be attributed to mutations in *DKC1*(*46*), leaving a significant number of DKC patients with incomplete genotypes(*52*). A variant call file extracted from a 3 Mb region surrounding the 15 kb *DKC1* locus from genome sequencing of five DKC patients and a single variant identified by Sanger sequencing in the exon 3 canonical splice acceptor of a sixth DKC patient were subject to classification. 4685 unique variants were found in the six patients, resulting in five candidate donor sequences and 12 candidate acceptor sequences. After splice variant selection, two variants generated candidate splice sequences predicted to alter native splice patterns. The first variant, *DKC1* c.16+592 C>G was predicted to create a novel GT (chrX:153991848, $\Delta_{variant}$ = +0.709) that would activate a deep intronic cryptic pseudoexon (**Figure 5A, *top panel***). This variant was found in a patient manifesting mild DKC(*46*). An acceptor of moderate splice potential (chrX:153991611, P(ref) = 0.579) was identified 249 bp upstream of c.16+592. Recognition of the cryptic donor activated by c.16+592 C>G and the acceptor at chrX:153991611 would result in activation of a novel pseudoexon of 234 bp. Sequencing of patient-derived RNA transcripts confirmed activation of the predicted cryptic donor and inclusion of a pseudoexon (**Figure 5A, *lower panel***). Interestingly, the acceptor used by this pseudoexon was not the one identified by the classifier model but a sequence 14 bp upstream that had a low P(ref) = 0.134. The second variant, *DKC1* c.85-5 C>G was predicted to moderately decrease the splice potential of the exon 3 canonical acceptor

(chrX:153993717, Δ_{canon} = -0.152) (**Figure 5B, *top panel***). Given the relatively high P(var) of this candidate splice sequence compared to the P(var) of validation sequences bearing variants known experimentally to completely abolish normal splicing, two transcripts were predicted: normally spliced WT transcript and transcript with exon 3 skipped. An acceptor of moderate splice potential (P(ref) = 0.684) was identified 82 bp upstream of the exon 3 canonical acceptor, yielding the possibility of a third transcript utilizing the upstream cryptic acceptor resulting in extension of exon 3, frameshift, premature termination, and NMD. Sequencing of patient-derived cDNA showed four products: the three predicted transcripts plus one transcript with intron 2 retained (**Figure 5B, *lower panel***). The predominant isoforms amplified from patient cDNA were the normally spliced WT transcript and exon 3 skipped transcript (data not shown). The transcript retaining intron 2, which is 477 bp, had very low expression relative to the WT and exon 3 skipped transcripts. Retention of exon 2 and skipping of exon 3 are both consistent with the prediction that c.85-5 C>G would disrupt the splicing of exon 3.

Identification of candidate splice variants among ClinVar SNVs

The identification of splice variants from amongst CF-associated *CFTR* variants demonstrated that splice variants may be under-ascertained in disease variant databases. To test this hypothesis, splice site classification and selection were applied to 24,787 SNVs labeled “pathogenic” or “probably pathogenic” in the October 2015 ClinVar FTP download (URL). Over 15,000 unique SNVs generated nearly 29,000 candidate splice sequences by creating a new GT or AG dinucleotide or existing in sufficient proximity to

an existing GT or AG dinucleotide (**Figure S1**). 459 candidate splice variants were selected as having a high likelihood of altering native splice patterns (**Table S2**) and were evaluated for their effect on RNA processing. 129 (28.1%) of the 459 high confidence splice variants were predicted to activate cryptic splice sites. Interestingly, while most of the selected variants involving splice donors were predicted to weaken nearby canonical donors (315 of 379, 83%), most of the selected variants involving splice acceptors were predicted to activate cryptic acceptors (65 of 80, 81.3%). 293 (63.8%) of the 459 selected splice variants were exonic, and 205 (44.7%) of all selected variants were annotated as missense mutations in ClinVar.

To explore whether variants designated in ClinVar as "variants of uncertain significance" (VUS) might affect splicing, classification and splice variant selection were applied to 28,147 VUS in the October 2015 ClinVar FTP download. Over 18,500 unique VUS created novel GT or AG dinucleotides or occurred within sufficient proximity to existing GT or AG dinucleotides to generate over 35,000 candidate donor and acceptor sequences (**Figure S2**). 348 candidate splice variants were selected as having a high likelihood of altering native splice patterns (**Table S3**), with the majority (273, 78.4%) of VUS predicted to weaken nearby canonical donors or acceptors. 75 (21.6%) of 348 selected VUS were predicted to result in cryptic splice site activation either near canonical donors or in the deep intron. 137 of the selected VUS were exonic and 113 (32.5%) of all selected VUS were annotated as missense mutations in ClinVar. Overall, 28.1% ClinVar "pathogenic" variants and 21.6% of ClinVar VUS selected with high confidence to affect splicing were predicted to activate cryptic splice sites (**Figure S1, S2**), a similar rate to that observed with CFTR2 variants (32.7%).

Discussion

The activation of cryptic splice sites by single nucleotide variants is a long understood disease-causing molecular mechanism, with one of the earliest disease variants ever identified causing cryptic activation of an acceptor 19 bp upstream of the canonical exon 2 acceptor of the β-globin gene causing β^+ thalassemia(*53*). Using a stringent method for the *in silico* identification of variants that activate cryptic splice sites, we discovered that this mechanism is not rare. Consequently, we propose that cryptic splice activation should be considered when evaluating pathogenicity of exonic and intronic variants. While we present Mendelian recessive examples where cryptic splice site activation caused substantial reductions in normally spliced transcripts, it is reasonable to posit that the same mechanism could account for dominant and complex patterns of inheritance. Variants that activate splice sites leading to loss of protein or aberrant protein could underlie dominant disorders caused by haploinsufficiency, dominant negative, or gain of function mechanisms. Variability in splicing efficiency shown here was associated with mild phenotypes suggesting that some variants could have subtle effects on protein expression. Thus, intronic variants associated with complex traits should be considered for splice site activation.

Exonic SNVs are often assumed to primarily impact protein, especially if the variant is non-synonymous. For example, there are efforts to determine the impact of every exonic variant on protein processing and function. However, these massively parallel mutagenesis methods presume exonic variants will not impact mRNA processing and thus interrogate variants in cDNA in the absence of introns(*54, 55*). The importance of considering whether variants affect splicing is illustrated by the dramatic differences in

the effect of *CFTR* c.454 A>G upon protein synthesis when cDNA (without introns) or expression minigenes (with introns) were employed (see **Figure 3A**). Correct assessment of variant effect upon gene function is essential to ensure that the appropriate molecular therapeutics may be administered to patients(*10, 56-58*). Of particular note, a recent clinical trial evaluated the efficacy of the small molecule ivacaftor in CF patients carrying variants that premitted residual protein function(*59*) as determined by *in vitro* chloride conductance measurements of mutant CFTR(*60*). At the completion of the clinical trial, all patients had positive clinical responses to ivacaftor except for those carrying p.Gly970Arg. The variant which causes p.Gly970Arg (c.2908 G>C) alters last nucleotide of the exon and was predicted to cause a complete loss of normal splicing at the canonical donor (**Table S1**). Thus, the lack of clinical response to ivacaftor in CF patients carrying c.2908 G>C is explained if the underlying molecular defect impacts mRNA processing instead of protein function, as we propose here. The importance of elucidating molecular mechanisms will be highly relevant to disease research communities that seek to deploy small molecule therapies.

Second, deep intronic variants are largely dismissed as disease-causing given the high degree of intronic variation, the distance of these variants from coding or other highly conserved regions, and the practical issues inherent to obtaining and working with RNA from patient specimens(*61*). Experimental detection of aberrantly spliced transcripts with premature termination codons can be difficult, especially in primary tissues, due to RNA degradation caused by nonsense mediated RNA decay. For example, the deep intronic variant, *CFTR* c.1584+689 G>A was predicted to result in activation of a cryptic pseudoexon. The patient carrying this variant had been shown in previous

studies(*14*) to have severely decreased RNA transcription from the chromosome bearing c.1584+689 G>A. Amplification of cDNA derived from this patient's nasal epithelial RNA did not show pseudoexon inclusion. However, it is reasonable to posit that transcripts containing this predicted pseudoexon were sufficiently degraded by NMD to preclude detection by PCR. Of note, we were able to find intronic variants that caused aberrant RNA transcripts in a substantial fraction of CF (7 of 14) and DKC (2 of 6) patients with incomplete genotypes. These findings suggest that deleterious splice variants are likely to be present in the introns of other genes responsible for loss of function Mendelian disorders; therefore, intronic variants in the suspected disease associated genes should be inspected in cases where exome sequencing is unable to complete the genotype of a patient with unambiguous Mendelian disease.

Variants that affect nucleotides outside of highly conserved position in acceptor sites and the impact of these variants are known to be difficult to assess. We observed a lower detection rate for canonical acceptors compared to canonical donors in our validation using the canonical sites of five disease genes, with 88.9% sensitivity for the canonical acceptors compared to 97.8% sensitivity for the canonical donors (**Materials and Methods**). Further, we were unable to computationally identify the acceptor that paired with the donor cryptically activated by *DKC1* c.16+592 C>G (**Figure 4A**). Multiple factors likely contribute to the lower ascertainment of acceptors relative to donors using our method. The longer length of acceptors and the flexibility of the polypyrimidine tract allow for a greater diversity of nucleotide combinations. Further, the splice variant selection algorithm is designed to find variants with a high likelihood of altering native splice patterns through comparisons of splice potentials ($\Delta_{variant}$ and Δ_{canon},

where applicable). $\Delta_{variant}$ and Δ_{canon} metrics rely on the probability estimates provided by the splice sequence classifiers and therefore do not take into consideration non-sequence contributors to splice site recognition. The role of splice enhancers and inhibitors on the maintenance of normal splicing patterns is well established(*22*) and the incorporation of predictive tools which model the functions of splice enhancers and inhibitors(*62, 63*) into the method described here could possibly lead to better ascertainment of cryptic splice sites, particularly acceptors. However, to reduce the complexity of our method and enrich for high confidence candidates, we focused on spliceosome recognition and therefore splice *sequence* only. Indeed, in our validation using known *CFTR* splice variants, all five acceptor variants were correctly identified. While using such stringent criteria could increase the number of false negatives, it also decreases the number of false positives, making each selected candidate splice variant a high confidence candidate. Conversely, the flexibility of the $\Delta_{variant}$ and Δ_{canon} thresholds allow for variability in the permissiveness of the selection algorithm, enabling the user to accept a more or less stringent sensitivity when generating candidate splice variants.

Our findings reveal that the deep introns remain an untapped reservoir of cryptic splice variants sufficient to cause severe, life-limiting disease as demonstrated in both cystic fibrosis and dyskeratosis congenita patients. These variants may create novel GT or AG dinucleotides that may immediately bring splicing mechanisms to mind; however, the majority of these variants lie adjacent to existing GT or AG dinucleotides and are thus easier to assess accurately with the aid of artificial intelligence tools such as the classification-based method described here(*19*). Although existing *in silico* splice prediction tools have demonstrated success, other classifiers are limited by simple

human-delineated rules that do not capture the complexity of sequence information required for spliceosome recognition. NNSplice, for example, uses neural networks based on the frequencies of dinucleotides within a sequence(*64*); in the training of the splice sequence classifiers described here, dinucleotide frequency is but one of three sequence features used. Other commonly used splice prediction tools utilize thermodynamics and conservation data to assess splice potential, both of which are descriptive of spliceosome recognition but not deterministic (as is nucleotide sequence). Further, the utilization of the selection algorithm refines the classifier predictions into intelligible findings based on the unique genomic context of each candidate variant. The comparison of a variant sequence's splice potential to the splice potential of the reference provides a powerful filter with demonstrable accuracy, a context-dependency which is still relatively new to variant annotation tools but will enable the generation of stronger candidate disease variants. Our findings emphasize the need to consider all variants as splice variants, even if they appear to cause amino acid substitutions predicted to have deleterious consequences on the resultant protein.

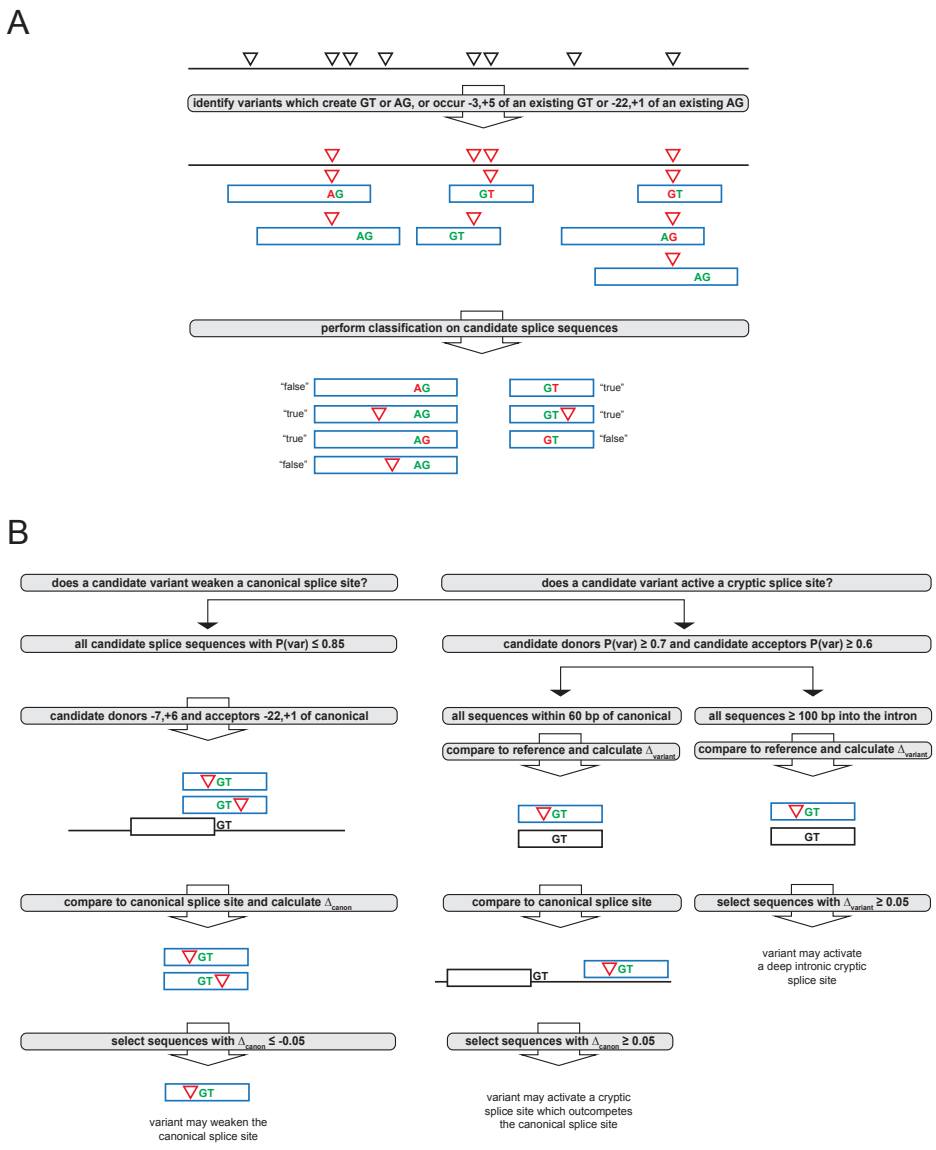


Figure 2.1. Overview of splice sequence classification and selection of high confidence candidate splice variants. (A) Variants (black triangles) are chosen for splice sequence classification if they either create a novel GT or AG dinucleotide or occur within the specified ranges of an existing GT or AG dinucleotide. Candidate splice sequences containing chosen variants are indicated by the blue rectangles where candidate variants are represented by red triangles or, in the event that the variant creates

a novel GT or AG, red text. Candidate splice sequences are subject to classification and are given binary classifications ("true" or "false") and probability estimates of splice potential (P(var) for variant sequences). **(B)** A custom algorithm for selecting high confidence candidate splice sequences (blue rectangles) utilizing variant distance from canonical splice sites (black rectangles) and changes in splice potential ($\Delta_{variant}$ and Δ_{canon}) was used to determine which variants were likely to: (*left*) weaken the canonical splice site, (*center*) activate a cryptic splice site which could outcompete the canonical, or (*right*) activate a cryptic splice site in the deep intron. See Materials and Methods for more information.

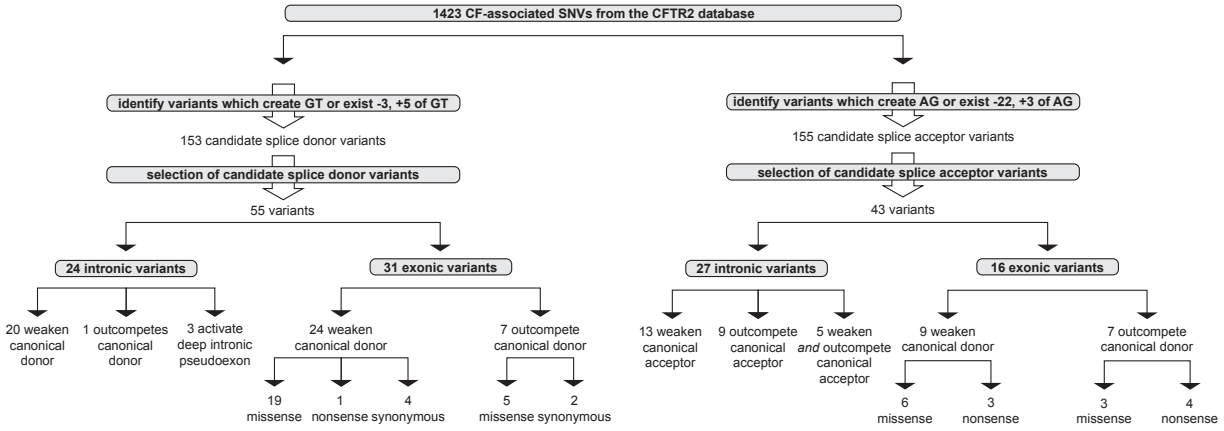


Figure 2.2. Selection of high confidence candidate splice variants from the CFTR2 database. Out of 1477 variants in the CFTR2 database, 1423 variants had standard HGVS nomenclature cDNA names, were not in the promoter or UTRs, and did not have indeterminate breakpoints, and were thus subject to splice sequence classification and splice variant selection as described in **Figure 2.1**. Out of 153 candidate splice donor variants (*left*), 55 were selected as high confidence candidates. Out of 155 candidate splice acceptor variants (*right*), 43 were selected as high confidence candidates.

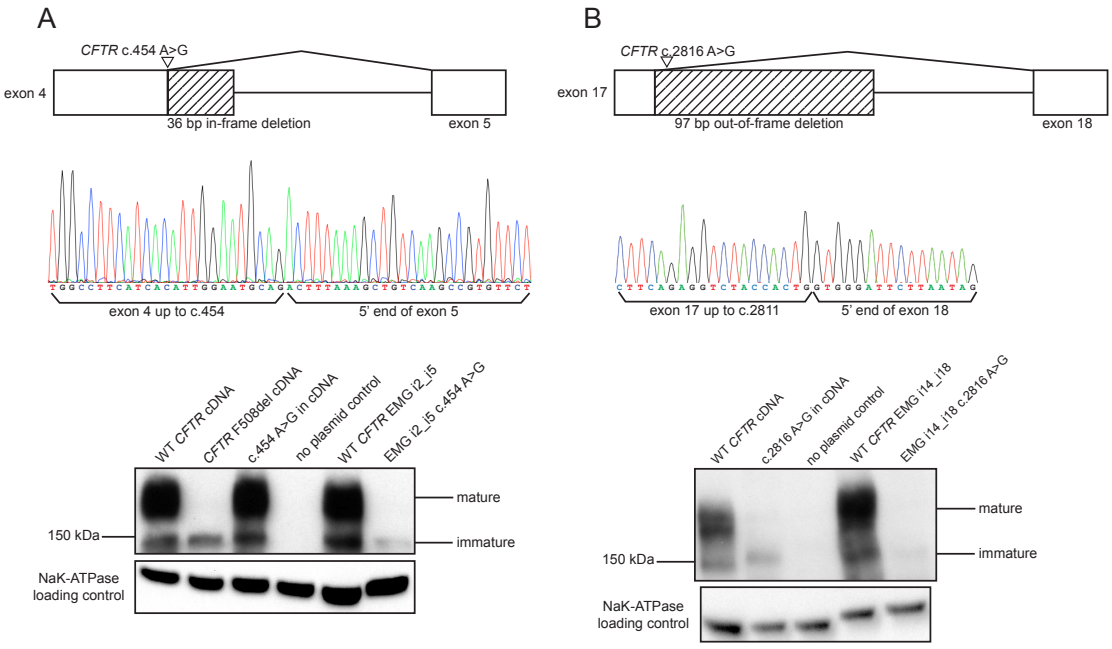


Figure 2.3. Experimental validation of exonic *CFTR* cryptic splice variants. (A) c.454 A>G (predicted missense: p.Met152Val) was predicted to activate a cryptic donor upstream of the exon 4 canonical donor, resulting in an in-frame deletion of 36 bp that would result in a 12 amino acid deletion as indicated by the diagonally hashed rectangle. Sequence analysis of RNA transcripts from HEK293 cells transfected with EMGs bearing c.454 A>G and flanking introns showed deletion of the last 36 exonic nucleotides from the final processed transcript due to utilization of the cryptic donor at c.454 over the canonical exon 4 donor. Western blotting of EMGs with c.454 A>G showed a drastic loss of normal CFTR protein and a very faint amount of immature protein. By contrast, c.454 A>G in a plasmid bearing only cDNA with *no* intronic sequences showed CFTR protein identical to that seen in WT transfections. **(B)** c.2816 A>G (predicted missense: p.His939Arg) was predicted to activate a cryptic donor upstream of the exon 17 canonical donor, resulting in a 97 bp deletion as indicated by the diagonally hashed rectangle.

Sequencing of transcripts from HEK293 cells transfected with EMGs bearing c.2816 A>G and flanking introns showed deletion of the last 97 exonic nucleotides due to utilization of the cryptic donor at c.2811. Western blotting of EMGs with c.2816 A>G showed a near complete loss of normal protein and a very faint amount of immature protein consistent with translation of a small amount of normally spliced transcript bearing a deleterious missense mutation leading to misfolded protein.

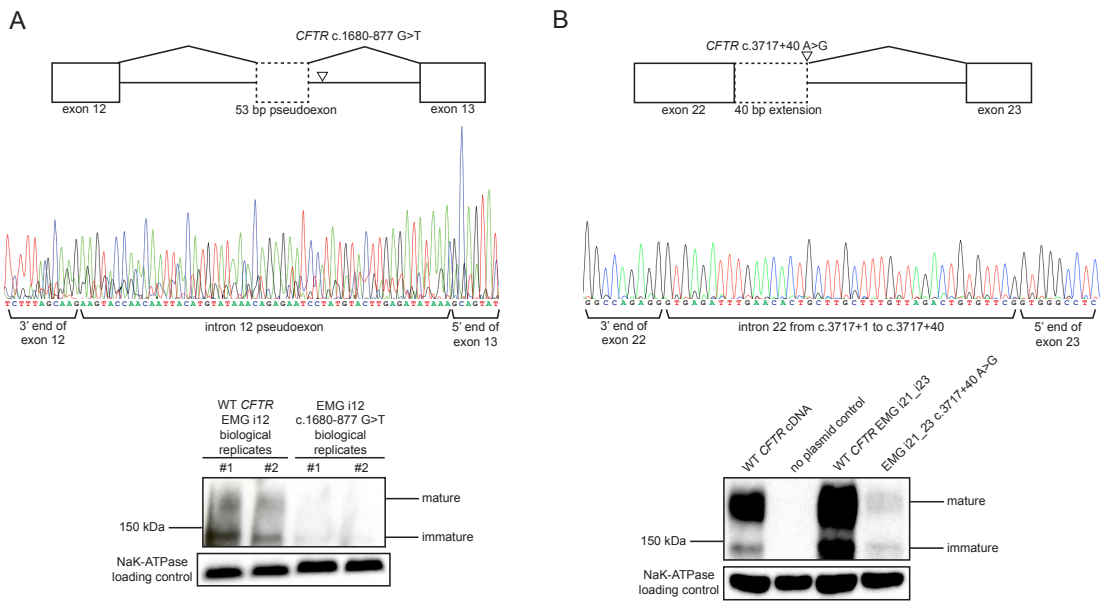


Figure 2.4. Experimental validation of intronic *CFTR* cryptic splice variants. (A) c.1680-877 G>T was predicted to activate a deep intronic cryptic donor and a strong latent acceptor 55 bp upstream. Sanger sequencing of RNA transcripts from HEK293 cells transiently transfected with EMGs bearing c.1680-877 G>T and abridged intron 12 showed activation of the expected 53 bp cryptic pseudoexon as indicated by the dotted edge rectangle. Western blotting revealed loss of normal CFTR protein in EMGs bearing c.1680-877 G>T **(B)** c.3717+40 A>G was predicted to activate a cryptic donor 40 nucleotides downstream of the exon 22 canonical donor, resulting in exon extension (as indicated by the dotted edge rectangle), frameshift, and a premature termination. Sequence analysis of RNA transcripts from HEK293 cells transfected with EMGs with c.3717+40 A>G and flanking introns showed retention of the first 40 nucleotides of intron 22 in the final processed transcript due to activation of the cryptic donor at c.3717+40. Western blotting showed a severe loss of normal protein and truncated CFTR

protein of a molecular weight consistent with translation of a transcript with a premature termination due to cryptic pseudoexon activation.

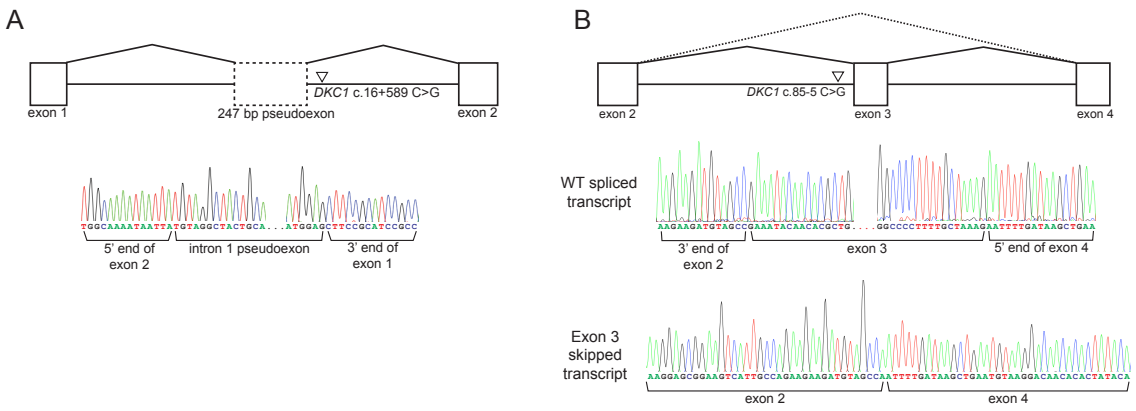


Figure 2.5. Experimental validation of *DKC1* cryptic splice variants. (A) *DKC1* c.16+589 C>G was predicted to activate a deep intronic cryptic donor resulting in inclusion of a pseudoexon in the final processed mRNA as indicated by the dotted edge rectangle. Sanger sequencing of transcripts from patient-derived lymphoblastoid cell lines confirmed cryptic pseudoexon inclusion. **(B)** *DKC1* c.85-5 C>G was predicted to moderately weaken the canonical exon 3 acceptor, leading to two predicted transcripts: skipping of exon 3 and normally spliced transcript as indicated by the solid and dotted lines joining the exons. Amplification and Sanger sequencing of RNA transcripts from patient-derived lymphoblastoid cell lines confirmed the two predicted transcripts and a third transcript with full retention of intron 2.

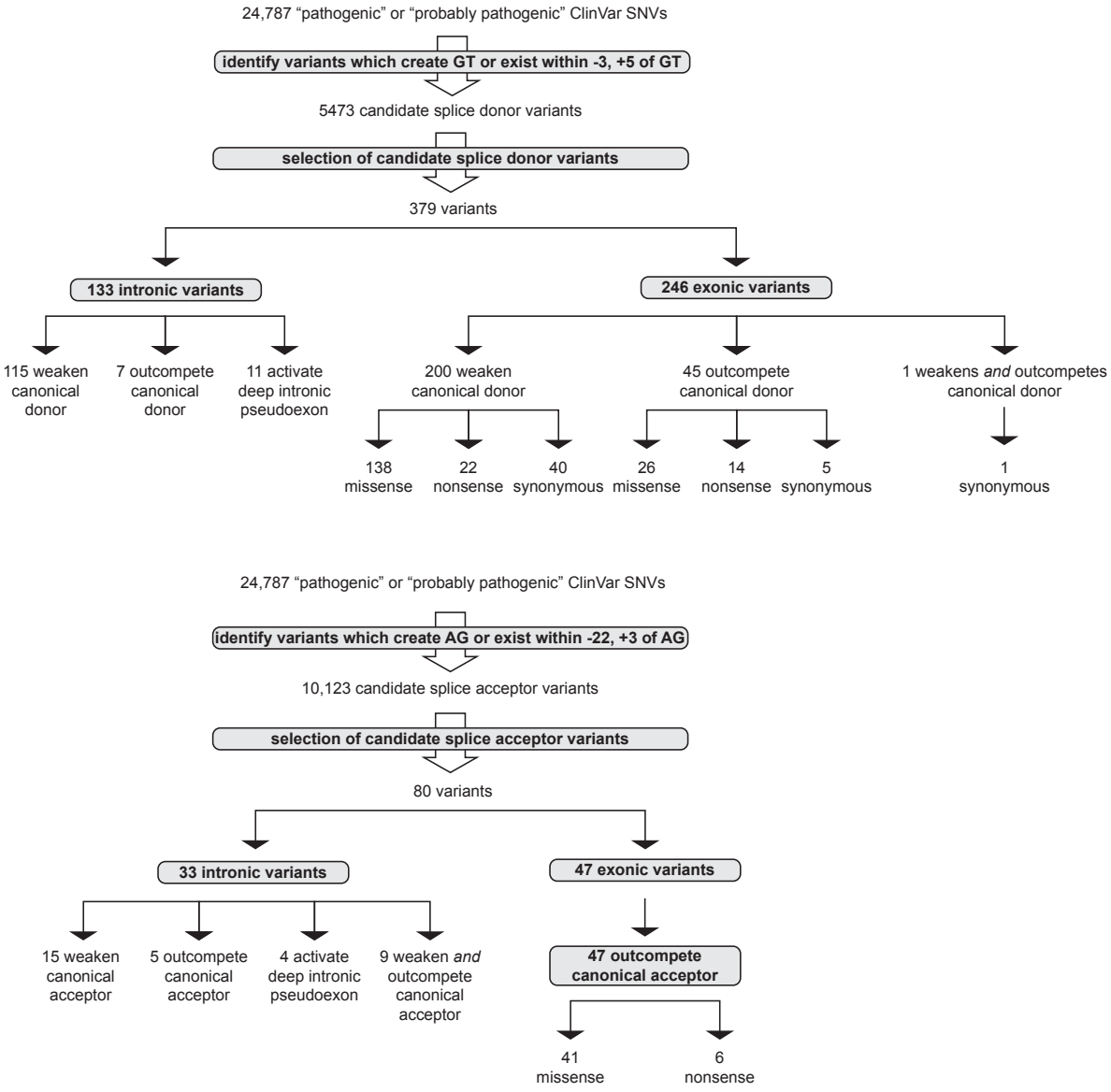


Figure S2.1. Flowcharts of selected high confidence splice variants from ClinVar "pathogenic" or "probably pathogenic" variants. Predicted high confidence splice variants broken down by impact on splicing and predicted impact on protein. Flowcharts are split by predicted impact on splice donors or splice acceptors.

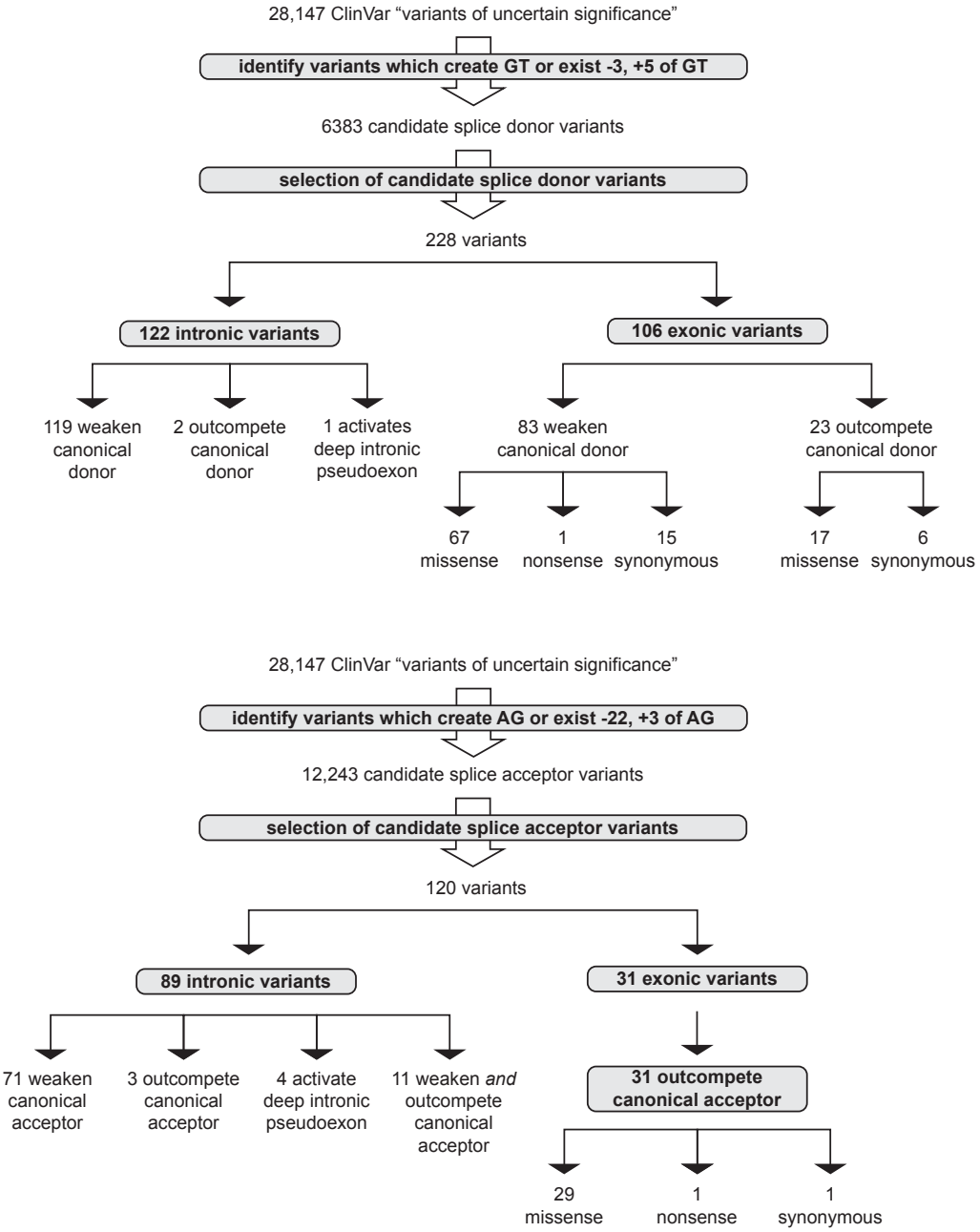


Figure S2.2. Flowcharts of selected high confidence splice variants from ClinVar "variants of uncertain clinical significance." Predicted high confidence splice variants broken down by impact on splicing and predicted impact on protein. Flowcharts are split by predicted impact on splice donors or splice acceptors.

References

1. Soukarieh, O., Gaildrat, P., Hamieh, M., Drouet, A., Baert-Desurmont, S., Frebourg, T., Tosi, M., and Martins, A. (2016). Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico Tools. PLoS Genet 12, e1005756.

2. Molinski, S.V., Gonska, T., Huan, L.J., Baskin, B., Janahi, I.A., Ray, P.N., and Bear, C.E. (2014). Genetic, cell biological, and clinical interrogation of the CFTR mutation c.3700 A>G (p.Ile1234Val) informs strategies for future medical intervention. Genet Med 16, 625-632.

3. Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3, 285-298.

4. Wang, G.S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet 8, 749-761.

5. Singh, R.K., and Cooper, T.A. (2012). Pre-mRNA splicing in disease and therapeutics. Trends Mol Med 18, 472-482.

6. Bonini, J., Varilh, J., Raynal, C., Theze, C., Beyne, E., Audrezet, M.P., Ferec, C., Bienvenu, T., Girodon, E., Tuffery-Giraud, S., et al. (2015). Small-scale high-throughput sequencing-based identification of new therapeutic tools in cystic fibrosis. Genet Med 17, 796-806.

7. Bonifert, T., Karle, K.N., Tonagel, F., Batra, M., Wilhelm, C., Theurer, Y., Schoenfeld, C., Kluba, T., Kamenisch, Y., Carelli, V., et al. (2014). Pure and syndromic

optic atrophy explained by deep intronic OPA1 mutations and an intralocus modifier. Brain 137, 2164-2177.

8. Pezeshkpoor, B., Zimmer, N., Marquardt, N., Nanda, I., Haaf, T., Budde, U., Oldenburg, J., and El-Maarri, O. (2013). Deep intronic 'mutations' cause hemophilia A: application of next generation sequencing in patients without detectable mutation in F8 cDNA. J Thromb Haemost 11, 1679-1687.

9. den Hollander, A.I., Koenekoop, R.K., Yzer, S., Lopez, I., Arends, M.L., Voesenek, K.E., Zonneveld, M.N., Strom, T.M., Meitinger, T., Brunner, H.G., et al. (2006). Mutations in the CEP290 (NPHP6) gene are a frequent cause of Leber congenital amaurosis. Am J Hum Genet 79, 556-561.

10. Chillon, M., Dork, T., Casals, T., Gimenez, J., Fonknechten, N., Will, K., Ramos, D., Nunes, V., and Estivill, X. (1995). A novel donor splice site in intron 11 of the CFTR gene, created by mutation 1811+1.6kbA-->G, produces a new exon: high frequency in Spanish cystic fibrosis chromosomes and association with severe phenotype. Am J Hum Genet 56, 623-629.

11. Reboul, M.P., Bieth, E., Fayon, M., Biteau, N., Barbier, R., Dromer, C., Desgeorges, M., Claustres, M., Bremont, F., Lacombe, D., et al. (2002). Splice mutation 1811+1.6kbA>G causes severe cystic fibrosis with pancreatic insufficiency: report of 11 compound heterozygous and two homozygous patients. J Med Genet 39, e73.

12. NN269.

13. HS3D: Homo Sapiens Splice Sequence Database.

14. Li, J.L., Wang, L.F., Wang, H.Y., Bai, L.Y., and Yuan, Z.M. (2012). High-accuracy splice site prediction based on sequence component and position features. Genet Mol Res 11, 3432-3451.

15. Python scikit learn Machine Learning Library.

16. Sharma, N., Sosnay, P.R., Ramalho, A.S., Douville, C., Franca, A., Gottschalk, L.B., Park, J., Lee, M., Vecchio-Pagan, B., Raraigh, K.S., et al. (2014). Experimental assessment of splicing variants using expression minigenes and comparison with in silico predictions. Hum Mutat 35, 1249-1259.

17. Reynolds, S.D., Rios, C., Wesolowska-Andersen, A., Zhuang, Y., Pinter, M., Happoldt, C., Hill, C.L., Lallier, S.W., Cosgrove, G.P., Solomon, G.M., et al. (2016). Airway Progenitor Clone Formation is Enhanced by Y-27632-dependent Changes in the Transcriptome. Am J Respir Cell Mol Biol.

18. Poole, A., Urbanek, C., Eng, C., Schageman, J., Jacobson, S., O'Connor, B.P., Galanter, J.M., Gignoux, C.R., Roth, L.A., Kumar, R., et al. (2014). Dissecting childhood asthma with nasal transcriptomics distinguishes subphenotypes of disease. J Allergy Clin Immunol 133, 670-678 e612.

19. Parry, E.M., Alder, J.K., Lee, S.S., Phillips, J.A., 3rd, Loyd, J.E., Duggal, P., and Armanios, M. (2011). Decreased dyskerin levels as a mechanism of telomere shortening in X-linked dyskeratosis congenita. J Med Genet 48, 327-333.

20. Pagani, F., and Baralle, F.E. (2004). Genomic variants in exons and introns: identifying the splicing spoilers. Nat Rev Genet 5, 389-396.

21. Highsmith, W.E., Burch, L.H., Zhou, Z., Olsen, J.C., Boat, T.E., Spock, A., Gorvoy, J.D., Quittel, L., Friedman, K.J., Silverman, L.M., et al. (1994). A novel mutation in the cystic fibrosis gene in patients with pulmonary disease but normal sweat chloride concentrations. N Engl J Med 331, 974-980.

22. Tzetis, M., Efthymiadou, A., Doudounakis, S., and Kanavakis, E. (2001). Qualitative and quantitative analysis of mRNA associated with four putative splicing mutations (621+3A-->G, 2751+2T-->A, 296+1G-->C, 1717-9T-->C-D565G) and one nonsense mutation (E822X) in the CFTR gene. Hum Genet 109, 592-601.

23. Amaral, M.D., Pacheco, P., Beck, S., Farinha, C.M., Penque, D., Nogueira, P., Barreto, C., Lopes, B., Casals, T., Dapena, J., et al. (2001). Cystic fibrosis patients with the 3272-26A>G splicing mutation have milder disease than F508del homozygotes: a large European study. J Med Genet 38, 777-783.

24. Dujardin, G., Commandeur, D., Le Jossic-Corcos, C., Ferec, C., and Corcos, L. (2011). Splicing defects in the CFTR gene: minigene analysis of two mutations, 1811+1G>C and 1898+3A>G. J Cyst Fibros 10, 212-216.

25. Yu, J., Yang, Z., Kibukawa, M., Paddock, M., Passey, D.A., and Wong, G.K. (2002). Minimal introns are not "junk". Genome Res 12, 1185-1189.

26. Sosnay, P.R., Siklosi, K.R., Van Goor, F., Kaniecki, K., Yu, H., Sharma, N., Ramalho, A.S., Amaral, M.D., Dorfman, R., Zielenski, J., et al. (2013). Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. Nat Genet 45, 1160-1167.

27. Sheridan, M.B., Hefferon, T.W., Wang, N., Merlo, C., Milla, C., Borowitz, D., Green, E.D., Mogayzel, P.J., Jr., and Cutting, G.R. (2011). CFTR transcription defects in pancreatic sufficient cystic fibrosis patients with only one mutation in the coding region of CFTR. J Med Genet 48, 235-241.

28. Walter, J.E., Armanios, M., Shah, U., Friedmann, A.M., Spitzer, T., Sharatz, S.M., and Hagen, C. (2015). CASE RECORDS of the MASSACHUSETTS GENERAL HOSPITAL. Case 41-2015. A 14-Year-Old Boy with Immune and Liver Abnormalities. N Engl J Med 373, 2664-2676.

29. Busslinger, M., Moschonas, N., and Flavell, R.A. (1981). b^+ thalassemia: Aberrant splicing results from a single point mutation in an intron. Cell 27, 289-298.

30. Kitzman, J.O., Starita, L.M., Lo, R.S., Fields, S., and Shendure, J. (2015). Massively parallel single-amino-acid mutagenesis. Nat Methods 12, 203-206, 204 p following 206.

31. Starita, L.M., Young, D.L., Islam, M., Kitzman, J.O., Gullingsrud, J., Hause, R.J., Fowler, D.M., Parvin, J.D., Shendure, J., and Fields, S. (2015). Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. Genetics 200, 413-422.

32. Dietz, H.C. (2010). New therapeutic approaches to mendelian disorders. N Engl J Med 363, 852-863.

33. Finkel, R.S., Flanigan, K.M., Wong, B., Bonnemann, C., Sampson, J., Sweeney, H.L., Reha, A., Northcutt, V.J., Elfring, G., Barth, J., et al. (2013). Phase 2a study of ataluren-mediated dystrophin production in patients with nonsense mutation Duchenne muscular dystrophy. PLoS One 8, e81302.

34. Lynch, T.J., Bell, D.W., Sordella, R., Gurubhagavatula, S., Okimoto, R.A., Brannigan, B.W., Harris, P.L., Haserlat, S.M., Supko, J.G., Haluska, F.G., et al. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med 350, 2129-2139.

35. Committee, F.P.-A.D.A. (2014). Ivacaftor for the Treatment of Cystic Fibrosis in Patients Age 6 Years and Older with an R117H-CFTR Mutation in the CFTR Gene.

36. Yu, H., Burton, B., Huang, C.J., Worley, J., Cao, D., Johnson, J.P., Jr., Urrutia, A., Joubran, J., Seepersaud, S., Sussky, K., et al. (2012). Ivacaftor potentiation of multiple CFTR channels with gating mutations. J Cyst Fibros 11, 237-245.

37. Baralle, D., and Baralle, M. (2005). Splicing in action: assessing disease causing sequence changes. J Med Genet 42, 737-748.

38. Fairbrother, W.G., Yeh, R.F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science 297, 1007-1013.

39. Smith, P.J., Zhang, C., Wang, J., Chew, S.L., Zhang, M.Q., and Krainer, A.R. (2006). An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Hum Mol Genet 15, 2490-2508.

40. Libbrecht, M.W., and Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nat Rev Genet 16, 321-332.

41. Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. J Comput Biol 4, 311-323.

Chapter 3

Loss of carbonic anhydrase XII function in individuals with elevated sweat chloride concentration and pulmonary airway disease

Melissa Lee; Briana Vecchio-Pagán; Neeraj Sharma; Abdul Waheed; Xiaopeng Li; Karen S. Raraigh; Sarah Robbins; Sangwoo T. Han; Arianna L. Franca; Matthew J. Pellicore; Taylor A. Evans; Kristin M. Arcara; Hien Nguyen; Shan Luan; Deborah Belchis; Jozef Hertecant; Joseph Zabner; William S. Sly; Garry R. Cutting. *Human Molecular Genetics* 2016. doi: 10.1093/hmg/ddw065**.**

Abstract

Elevated sweat chloride levels, failure to thrive (FTT), and lung disease are characteristic features of cystic fibrosis (CF, OMIM #219700). Here we describe variants in *CA12* encoding carbonic anhydrase XII in two pedigrees exhibiting CF-like phenotypes. Exome sequencing of a white American adult diagnosed with CF due to elevated sweat chloride, recurrent hyponatremia, infantile FTT and lung disease identified deleterious variants in each *CA12* gene: c.908-1 G>A in a splice acceptor and a novel frameshift insertion c.859_860insACCT. In an unrelated consanguineous Omani family, two children with elevated sweat chloride, infantile FTT, and recurrent hyponatremia were homozygous for a novel missense variant (p.His121Gln). Deleterious *CFTR* variants were absent in both pedigrees. CA XII protein was localized apically in human bronchiolar epithelia and basolaterally in the reabsorptive duct of human sweat glands. Respiratory epithelial cell RNA from the adult proband revealed only aberrant *CA12* transcripts and in vitro analysis showed greatly reduced CA XII protein. Studies of ion transport across respiratory epithelial cells in vivo and in culture revealed intact *CFTR*-mediated chloride transport in the adult proband. CA XII protein bearing either p.His121Gln or a previously identified p.Glu143Lys missense variant localized to the basolateral membranes of polarized MDCK cells, but enzyme activity was severely diminished when assayed at physiologic concentrations of extracellular chloride. Our findings indicate that loss of CA XII function should be considered in individuals without *CFTR* mutations who exhibit CF-like features in the sweat gland and lung.

Introduction

Persistently elevated sweat chloride concentration caused by loss of function mutations in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene is the diagnostic hallmark of cystic fibrosis (CF). Individuals with features of CF who do not carry any disease-causing *CFTR* alleles have been reported. These patients were phenotypically indistinguishable from CF patients carrying two known CF-causing mutations (*13*). Some individuals presenting a milder, atypical CF were found to carry variants that altered the function of subunits that form the epithelial sodium channel (ENaC) (*1*),(*65*).

CA12, the gene encoding carbonic anhydrase (CA) XII, has been implicated as a cause of elevated sweat chloride concentration, failure to thrive in infancy, and recurrent hyponatremia in two consanguineous Bedouin kindreds (*26*),(*27*). The same missense variant was identified in both pedigrees. However, the variant caused only a modest reduction (~30%) in enzymatic activity (*27*), which was unexpected as autosomal recessive disorders are generally associated with severe loss of function variants. The authors speculated that the minimal reduction in CA XII function produced a phenotype limited to the sweat gland (OMIM #143860) (*27*),(*66*).

In this study, we report the discovery and analysis of loss of function variants in *CA12* that associate with elevated sweat chloride concentrations in two unrelated pedigrees. An adult proband in one pedigree also displayed pulmonary features that overlap with CF; namely recurrent pulmonary exacerbations, Pseudomonas in sputum

cultures, and mild but distinct bronchiectasis upon high resolution chest CT scanning. These findings indicate that loss of CA XII activity is uncompensated in certain epithelia and that CA XII may play a key role in the function of the pulmonary airways as well as the sweat gland.

Results

Identification of *CA12* variants segregating in two unrelated pedigrees.

The proband in pedigree A (II:1, **Figure 3.1A**) presented with failure to thrive at 2.5 months of age and sweat chloride concentrations ranging from 82 to 88 mEq/L. She was diagnosed with cystic fibrosis (CF) and prescribed pancreatic enzymes to improve growth. At 7 months of age she had an episode of hyponatremic dehydration requiring hospitalization (plasma sodium 120 mm/L upon admission). Spirometry from ages 7-9 years indicate three episodes of airway obstruction, with forced expiratory volumes (FEV1) and forced expiratory flows (FEF25-75%) falling below 80%. At age nine, repeat sweat chloride testing revealed elevated levels (range=112-116 mEq/L) and serum IRT levels were within normal range, resulting in the discontinuation of pancreatic enzymes. Nasal potential difference (NPD) testing performed at this time reported aberrant chloride transport consistent with CF. It should be noted that NPD testing can have considerable technical variability and was standardized after this test was administered to proband A. Available clinical records between the ages of 19-22 revealed a persistent cough and cultures of bacteria common to CF patients, including Pseudomonas aeruginosa in the throat (age 19), Stenotrophamonas maltophilia in the throat (age 20), and Pseudomonas fluorescens in sputum (age 22). The proband's pulmonary exacerbations were often

treated with a regiment consistent with her CF diagnosis, including bronchodilators, antibiotics, and steroids. At age 25, NPD testing repeated at the same clinical facility was not consistent with CF (response to low chloride and isoproterenol: -24 mV on right and -16 mV on left). The proband continued to be seen regularly at an accredited CF care center and reported compliance with daily respiratory treatments including aerosolized albuterol, acetylcysteine, hypertonic saline, and chest physiotherapy. High resolution chest CT scanning revealed mild bronchiectasis without scarring, inflammation, or mucus plugging (**Figure 3.2**). Assessment of airway dilatation was confirmed by two adult CF pulmonologists and an additional interpretation by a radiologist who was masked to the clinical status of proband A. The proband has also been treated by a dermatologist for axillary hyperhidrosis. Exome sequencing was performed on the proband, her unaffected sister, and both parents. Average depth of coverage was 87X, and 93.2% of the targeted regions were covered at a depth ≥ 10X. No deleterious variants were found in *CFTR* or the three genes (*SCNN1A, SCNN1B,* and *SCNN1G*) that encode the epithelial sodium channel (ENaC). Loss of ENaC function can cause pseudohypoaldosteronism, a secondary and rare cause of elevated sweat chloride concentration (*1*),(*67*). Two variants within *CA12* were discovered in trans (compound heterozygosity) in the proband (II:1): a variant inherited from her father (I:1) in the canonical splice acceptor site of exon 10, c.908-1 G>A (rs148438059, chr15:63,619,433, ClinVar accession# SVC000255965) and an insertion variant of four nucleotides inherited from her mother (I:2), c.859_860insACCT (chr15:63,631,029-63,631,030, ClinVar accession# SVC000255963) in exon 8. The splice acceptor variant is found in heterozygosity in 53 individuals in the Exome Aggregation Consortium (ExAC) variant browser (*68*) with a global MAF of

0.00471%. It is the most common predicted deleterious *CA12* variant found in ExAC. The proband's insertion mutation was not found in ExAC. *CA12* variants were confirmed in all family members via Sanger sequencing.

In a second unrelated family, a six year old Omani boy (proband B, II:3, **Figure 3.1B**) presented with a history of hyponatremic dehydration, elevated sweat chloride, and bilateral hyperkeratosis of the heels. Hyponatremic dehydration was alleviated with administration of Pedialyte and unrestricted access to dietary salt. Four sweat chloride measurements ranged from 90 to 110 mEq/L. Pulmonary function tests and fecal elastase measurements were within the normal ranges, ruling out chronic pulmonary and exocrine pancreatic insufficiency associated with CF. Aldosterone measurements excluded pseudohypoaldosteronism. Clinical diagnostic sequencing of the coding and intron flanking regions of *CFTR* and *SCNN1A* encoding the α subunit of ENaC in proband B did not detect sequence variations predicted to be deleterious. An 11 year old sister (II:1) of proband B initially considered to be asymptomatic was discovered to have a sweat chloride of 130 mEq/L. At a two year follow up, this sister was found to have developed a phenotype concordant with that of proband B, reporting episodes of hyponatremic dehydration as well as mild bilateral hyperkeratosis of the heels. Hyperkeratosis of the heels was not observed in proband A or the previously reported patients (*66*) and could be due to unrelated deleterious recessive alleles that may be present in this consanguineous pedigree. Unaffected siblings in pedigree B had sweat chloride measurements within the normal range for the referring laboratory (<50 mEq/L, personal communication Jozef Hertecant). No evidence of respiratory disease was reported in either patient; however, we were unable to obtain high resolution chest CT scans.

Proband B was reported to have a normal chest X-ray at age 11 and his affected sister had no pulmonary testing of any kind. *CFTR*, *SCNN1B*, and *SCNN1G* were excluded via genetic linkage analysis of all individuals in pedigree B with the assumption of recessive inheritance. Exome sequencing was conducted on proband B, his affected sister, and all unaffected siblings in pedigree B. The average depth of coverage was 34X, and 62.5% of the targeted regions were covered at a depth ≥ 8X. A previously unreported variant c.363 C>A (chr15:63,637,742, ClinVar accession# SVC000255964) in exon 4 of *CA12* was found in homozygosity only in the two affected individuals (**Figure 3.1B**). It is predicted to cause a substitution of His at codon 121 with Gln (p.His121Gln). Segregation of the *CA12* c.363 C>A variant in an autosomal recessive inheritance pattern was confirmed in all family members in pedigree B via Sanger sequencing.

CA XII is localized to the basolateral membrane of ductal epithelia in sweat gland and apical

membrane in airway epithelia.

To ascertain whether CA XII was expressed in the organs affected in the two probands, namely the sweat gland and the airways, immunohistochemistry (IHC) of normal skin and lung sections was performed. IHC of whole skin tissue showed robust sweat gland expression of CA XII in the basolateral compartment of the reabsorptive ductal cells (**Figure 3.3B, 3.3C**). Basolateral CA XII staining in the reabsorptive sweat duct was distinguishable from apically localized CA II, a ubiquitously expressed cytosolic carbonic anhydrase (**Figure 3.3D, 3.3E**). To determine if CA XII is expressed in the airways, IHC of lung was performed and showed robust apical localization of CA

XII in bronchiolar epithelia (**Figure 3.3F, 3.3G**). Of note, IHC of CA XII in the lung was performed using the same anti-CA XII antibody (ProteinTech #15180-1-AP) that was utilized for IHC of the sweat gland. The varying subcellular localization of CA XII (basolateral within the sweat gland; apical within bronchioles), may indicate alternative roles for this protein in reabsorptive or secretory epithelial membranes.

The c.908-1 G>A and c.859_860insACCT variants found in proband A generate aberrant RNA transcripts.

The two variants identified in proband A were predicted to affect RNA processing. To evaluate this supposition, respiratory epithelia from the inferior nasal turbinate was obtained for RNA and functional studies. CA XII is expressed in nasal epithelial cells (**Supplemental Table 3.1**) and respiratory epithelia of the nasal turbinates have been used as a proxy for respiratory epithelia of the airways (*69*). The *CA12* gene is composed of 11 exons that constitute its primary mRNA transcript (**Figure 3.4A**). Alternative splicing of *CA12* has been observed in native and cancerous tissues by both RT-PCR and RNA sequencing (*70*). The most common alternative isoform of *CA12* (CCDS# 10186) removes exon 9, a small exon composed of 33 bp, allowing the downstream transcript to retain the same reading frame. A review of publicly available RNA-sequencing splicing data from the Human Protein Atlas (*71*) reveals this isoform is predominately expressed in brain, and select other tissues (**Supplemental Table 3.1**). Two additional rare isoforms of *CA12* result from skipping of exons 9 and 10, or exon 10 only. These alternative *CA12* transcripts are of very low abundance in nasal and bronchial

epithelial cells compared to the full-length transcript with 11 exons. PCR of cDNA derived from nasal epithelial cell RNA of proband A generated DNA products of 1069 bp, 980 bp, and 947 bp. Each product was gel purified and subject to Sanger sequencing. The 1069 bp product corresponded to full-length *CA12* transcript bearing the insertion c.859_860insACCT. This variant introduces a frameshift that is predicted to lead to the incorporation of 49 novel residues following codon 287 and a premature termination codon (PTC) in exon 11 (predicted size 336 residues; **Figure 3.4B**). Despite the presence of a PTC, the transcript was stable due to the location of the PTC in the last exon of *CA12*, thereby allowing the transcript to evade nonsense mediated RNA decay (*72*). The splice site variant found in proband A, c.908-1 G>A, was predicted to cause misplicing of *CA12* exon 10 as it alters an invariant nucleotide of the canonical 3' splice acceptor site. Indeed, the 980 bp amplicon was *CA12* transcript missing exon 10 and the 947 bp product was an alternatively spliced *CA12* transcript missing exons 9 and 10 (**Figure 3.4C**). Loss of exon 10 was predicted to result in a frameshift beginning at codon 302 and translational read-through of the native termination codon. A novel termination codon in the 3' UTR occurs at amino acid position 413. The resulting protein is predicted to be composed of the first 302 amino acids of CA XII followed by 111 novel residues, 89 of which are translated from the 3' UTR. Skipping of exons 9 and 10 would add the same 111 novel residues but the frameshift would start at codon 291 (predicted size 402 residues). Finally, amplification from exon 8 to exon 10 and Sanger sequencing verified that all transcripts bearing exons 9 and 10 contained the c.859_860insACCT insertion (data not shown). In summary, all *CA12* mRNA transcripts in the nasal epithelial RNA of proband A were abnormal and each was predicted to generate aberrant CA XII protein.

CA12 variants found in proband A generate unstable CA XII protein .

Expression vectors with *CA12* cDNA modified to correspond to each of the three transcripts observed in the nasal epithelial cells of proband A, the missense variant p.His121Gln (c.363 C>A) observed in proband B, and a previously described Bedouin missense variant p.Glu143Lys (c.427 G>A) were transfected into HEK293 cells and lysates were subjected to analysis by Western blot. Wild-type (WT) CA XII was present in both unglycosylated (39 kDa) and fully glycosylated (43 kDa) forms (*73*) (**Figure 3.5, lane 2**). CA XII protein was severely reduced in the lysate of cells transfected with *CA12* expression vectors bearing the insertion variant c.859_860insACCT (7.2%-13.3% of WT) and only a single band of the predicted mass of the unglycosylated protein (37.9 kDa) was observed (**Figure 3.5**, lane 4). CA XII lacking residues encoded by exon 10 and exons 9 and 10 were barely visible (**Figure 3.5**, lanes 5, and 6). CA XII bearing the missense variants p.His121Gln, and p.Glu143Lys generated protein of a molecular mass comparable to WT and unglycosylated and glycosylated forms were observed (**Figure 3.5**, lanes 7 and 8). CA XII with p.Glu143Lys had a higher fraction of unglycosylated protein, suggesting a possible effect of the amino acid substitution on processing and post-translation modifications. These results indicate that each of the changes in amino acid composition due to the *CA12* variants found in proband A cause substantial instability in CA XII.

Nasal respiratory epithelial cells from proband A demonstrate CFTR-mediated chloride transport.

Cultured epithelial cells from proband A and controls were mounted in Ussing chambers for short circuit current measurements. To increase the driving force for chloride secretion through *CFTR*, the apical membrane was hyperpolarized by administration of amiloride that inhibits sodium current conducted by epithelial sodium channels. To specifically examine chloride currents mediated by *CFTR*, calcium-activated chloride channels were inhibited by DIDS (4, 4'-diisothiocyanoto-stilbene -2, 2'-disulfonic acid). Application of DIDS did not result in a significant change in current in any sample. *CFTR* was activated by elevating cellular levels of cAMP with forskolin and 3-isobutyl-1-methylxanthine (IBMX). Upon treatment with forskolin and IBMX ("F+I"), the change in *CFTR*-mediated chloride transport in nasal epithelia from proband A (9.34 and 9.15 uA/cm2) were higher than that observed in nasal epithelial from a CF subject tested concurrently (2.1 uA/cm2). The values in proband A are consistent with short circuit measures of nasal epithelia from other non-CF and CF subjects (*74*). Substantial reduction in the current of cells from proband A upon addition of GlyH-101 (-18.8 and -21.5 uA/cm2) is consistent with the chloride secretion being mediated by *CFTR*. Together, these findings suggest that loss of CA XII does not ablate *CFTR*-mediated chloride secretion across nasal respiratory epithelia.

p.His121Gln and p.Glu143Lys mutations cause near complete loss of enzyme activity of CA XII.

As the missense variants permitted the generation of stable full-length protein, immunocytochemistry and confocal microscopy was utilized to test whether either variant affected the subcellular localization of CA XII. Expression in polarized epithelial Madin-Darby canine kidney (MDCK) cells revealed that WT CA XII localized to basolateral membranes (green, n=7, **Figure 3.6**, left panel). CA XII bearing p.His121Gln (green, n=8, **Figure 3.6**, middle panel) and p.Glu143Lys (green, n=13, **Figure 3.6**, right panel) showed basolateral localization indistinguishable from that of WT CA XII. Comparable staining patterns were observed when immunocytochemistry was performed with alternative CA XII antibodies (rabbit: Sigma Prestige; mouse: Novus) that detected different extracellular CA XII epitopes (data not shown). Since the p.His121Gln and p.Glu143Lys mutants were localized to plasma membranes, we tested the effect of each variant on CA function by measuring carbonic anhydrase enzyme activity. Carbonic anhydrase activity is the reversible rate of CO2 hydration. When assayed in a 2 mM NaCl solution, the activity of CA XII bearing p.His121Gln is reduced by 84.6 ± 3.6% compared to WT (n=11) while the enzymatic activity of CA XII with p.Glu143Lys is reduced by 24.4 ± 4.9%, consistent with the approximate 30% reduction in activity previously reported under the same conditions (*27*) (**Figure 3.7**). When assayed at physiological salt concentrations, the enzyme activity of p.His121Gln was reduced by 99.2 ± 0.5% compared to WT, and the activity of the p.Glu143Lys mutant was reduced by 97.1 ± 1.2% compared to WT (n=9) (**Figure 3.7**). The enzyme activities of p.His121Gln and p.Glu143Lys were not statistically different ($p = 0.12$; t test) when assayed in the presence of 100 mM NaCl. These findings reveal a chloride-sensitive

abolition of carbonic anhydrase enzyme activity for CA XII p.His121Gln and p.Glu143Lys mutants compared to WT.

Discussion

Each of the *CA12* variants reported in the three individuals described here with elevated sweat chloride, recurrent hyponatremia and failure to thrive in infancy cause severe loss of CA XII activity. The two variants found in proband A generate mRNA transcripts that are missing nucleotides that form transmembrane domains and enable dimerization via a key glycine zipper motif (*73*). Translation of cDNA that replicate the mRNA transcripts identified in proband A generated unstable protein products in heterologous cells. On the other hand, CA XII proteins bearing the missense variant found in this study and the previously reported missense variant were stable and, in each case, localized to the basolateral membrane of MDCK cells, consistent with native CA XII location in kidney cells (*75*). Point mutation energy modeling by FoldX suggested that both missense variants should affect CA XII structure and catalytic function. Modeling indicated that the secondary amine of histidine 121 is essential for tetrahedral coordination of the zinc ion within the catalytic domain. In the WT conformation, the H121–zinc bond distance is ~2.1Å (**Supplemental Figure 3.1**). When mutated to glutamine as in p.His121Gln, the distance from zinc to the hydroxyl of the carboxyl group is 3.042Å. This increased distance likely precludes formation of a coordinating bond, and the zinc ion is instead coordinated by a free hydroxide ion (red sphere). In the extracellular aqueous environment, this coordination would only be transient, potentially leading to poor catalytic activity. The distances of other residues, such as the highly conserved second shell glutamic acid 143, are also altered by p.His121Gln by a

magnitude similar to that reported in previous studies of the Bedouin p.Glu143Lys mutation (*26*). CA XII bearing the p.Glu143Lys mutation was particularly sensitive to inhibition by chloride, as previously reported (*27*). Given that the active site of CA XII lies on the extracellular face of the basolateral compartment in the sweat duct, enzymatic activity of both mutants was assayed in the presence of increasing NaCl concentrations. The concentration of NaCl which most closely mimics the enzyme's native physiological environment is 100 mM NaCl (*76*). At this concentration, catalytic activity of CA XII bearing each missense mutation was reduced to less than 3% of WT activity. As the affected individuals are homozygous for the *CA12* missense mutations, it is reasonable to conclude that the sweat gland dysfunction observed in each is due to near complete loss of CA XII activity. This conjecture is supported by the studies of proband A, where mutations in each *CA12* gene lead to severe instability of CA XII protein.

Robust expression of CA XII in lung epithelia and the observation of bronchiectasis in proband A suggest a role for this protein in the maintenance of airways. Elevated sweat chloride concentration indicates aberrant chloride transport in the sweat duct and is a consistent feature of the three individuals carrying the loss of function *CA12* variants reported here, as well as in 11 individuals homozygous for the p.Glu143Lys mutation reported previously (*66*). However, *CFTR*-mediated chloride transport appeared to be intact in the nasal respiratory epithelia as determined by in vitro and in vivo methods, suggesting that other pathways of ion transport in the airways might be disrupted by the loss of CA XII. Given the importance of CAs in the maintenance of pH via bicarbonate metabolism, the mechanism underlying airway damage could be related to aberrantly low pH of airway surface liquid (ASL) due to loss of bicarbonate production

or transport. The pH of the ASL has been shown to be integral to the proper expansion and processing of mucins which play a key role in CF-related bronchiectasis (*32*). Alternatively, the bronchiectasis observed in proband A might be unrelated to the loss of CA XII function. Spirometry measurements in all tested individuals homozygous for the p.Glu143Lys mutation were reported as normal (*66*) and chest X-rays of proband B were reported to be normal at age 11; however, lung function measures and chest X-rays were also normal throughout the life of proband A up to her current age of 25. Detection of abnormal airway dilatation in proband A required high resolution CT scanning; therefore, it is possible that bronchiectasis remains undetected in the previously reported patients with loss of CA XII function and the affected individuals in pedigree B. Without comparison chest CTs from the affected individuals in pedigree B, it cannot be determined if proband B and his affected sister manifest bronchiectasis similar to that observed in proband A. Further, if loss of CA XII function does lead to bronchiectasis, the possibly ameliorating impact of CF-specific airway treatments is an important question. Proband A has undergone routine airway clearance, therapies, and antibiotic courses since being diagnosed with CF as an infant. It is possible that a lifetime of diligent pulmonary monitoring and treatments managed by an accredited CF care center minimized the effect of bronchiectasis upon pulmonary function. However, estimating the impact of respiratory therapy on degree of airway dilatation is difficult without study of additional CA XII-deficient individuals manifesting pulmonary disease. Indeed, the identification of additional individuals with loss of CA XII function and high resolution chest CT scanning will clarify the role of CA XII in the airways. Given the strong causative connection between aberrant chloride transport and bronchiectasis in CF, CA

XII loss of function should be considered as a potential explanation for non-CF bronchiectasis.

Loss of CA XII function in a patient with respiratory disease in humans suggests a previously unsuspected role for this isozyme in the lung. Although each individual CA isozyme follows a tissue-specific expression pattern (*77, 78*), RNA expression studies show that multiple isozymes can be expressed in certain tissues (*79*). Loss of function of one isozyme may be compensated for by other isozymes in certain tissues. This explanation is offered for the lack of a phenotype in patients with loss of function mutations in CA I (*80*). Our findings show that isozyme redundancy does not compensate for loss of CA XII function in the sweat gland. Further, although other transmembrane CA isozymes such as CA IV have been localized to the plasma face of lung microcapillaries, our results suggest that CA XII may also play an important role in the maintenance of the airways.

IHC of the sweat gland revealed CA XII to be highly expressed in the resorptive duct in basolateral distribution consistent with its location in other epithelia, namely endometrium, kidney, and large intestine (*81*), (*82*), (*75*). Faint staining was observed at the apical membrane which may be non-specific signal or evidence of dual CA XII localization. However, the pattern of CA XII staining was distinctly different for that of CA II, which was discretely localized near apical membranes of the ductal cells. CA XII was only found on basolateral membranes of polarized MDCK cells. In contrast, CA XII was discretely localized to the apical regions of normal airway. The distribution of CA XII does not appear to be due to non-specific signal as two antibodies that detect different antigenic regions of CA XII revealed the same immunolocalization pattern. Localization

to the terminal bar is consistent with CA XII location in other tissues, including the bronchus and fallopian tubes (*71*). A possible factor in the different localization of CA XII could be its involvement in one of the transport metabolon complexes formed between CA isozymes and bicarbonate transporters that facilitate the exchange of bicarbonate across membranes (*28*). Bicarbonate transport metabolons comprised of CA isozymes and anion exchangers have so far been described for three of the transmembrane isozymes: CA IV associates with AE1 (*28, 29*), CA IX associates with AE2 (*30*), and CA XIV associates with AE3 (*31*). AE1 and CA IV have been localized to both basolateral and apical membranes in different cell types in the kidney (*29*). Since the kidney can absorb and secrete ions including bicarbonate, it is possible that the different localization of the AE1/CA IV metabolon may be related to the direction of ion flow. However, to date, no metabolon interaction has been reported for the remaining transmembrane isozyme, CA XII.

Our study suggests a mechanism for the well-established salt wasting complication of topiramate, a CA inhibitor and anticonvulsant commonly prescribed to people suffering from epilepsy. Elevated sweat chloride concentration is an established phenomenon observed in epileptic children being treated with topiramate (*83*). In these children, CF was clinically and/or molecularly excluded as being the cause of this increase in sweat chloride value, and the effect disappeared when topiramate treatment ended. Investigations into the inhibition potency of topiramate across the α family of CAs have shown that topiramate is a strong inhibitor of CA XII, and not the other transmembrane isozymes CA IV, IX, and XIV (*84*),(*85*). These pharmacologic observations are consistent with our hypothesis regarding the role of CA XII in the

maintenance of proper ion composition in the sweat gland. Topiramate is also a strong inhibitor of the ubiquitous and highly active cytoplasmic isozyme CA II. Correspondingly, individuals being treated with topiramate have been observed to develop renal tubular acidosis (*86*), a hallmark feature of CA II deficiency syndrome.

In summary, the individuals studied here demonstrate that severe loss of function mutations in *CA12* cause an autosomal recessive disorder affecting chloride and sodium resorption in the sweat duct. The observation of airway dilatation in proband A suggests a possible molecular etiology for some forms of non-CF bronchiectasis, a disease that affects over 110,000 individuals in the U.S. (*87*).

Materials and Methods

Recruitment

Family A was recruited and consented into the study Genome-wide Sequencing to Identify the Genes Responsible for Mendelian Disorders at Johns Hopkins University (IRB# NA_00045758). Family B was referred to Johns Hopkins University via private communication of Dr. Jozef Hertecant at Tawam Hospital, United Arab Emirates. All members of family A were consented into the Molecular Genetics of Cystic Fibrosis (IRB# NA_00050260).

Linkage exclusion assays

Three highly polymorphic deCODE STRs were selected for each locus to be excluded (*CA12*, and *SCNN1B* and *SCNN1G*, which lie in close proximity to one another) from the STS Marker track on the UCSC Genome Browser. Each marker was no more than 1Mb from either end of the locus to be excluded. Primer sequences from the STS Marker track were run through BLAT to verify specificity. Oligonucleotides were synthesized by IDT. Forward primers were fluorescently labeled with 6-FAM. STRs were PCR amplified from genomic DNA and amplification products were separated on an ABI Prism 3100 Genetic Analyzer by automated capillary electrophoresis. Fragment sizing and visualization were performed using ABI GeneMapper software. Haplotype phasing was performed by manual inspection. If the proband and at least one unaffected sibling were found to be IBD2 for a locus, it was deemed to be excluded, given an assumption that the disease phenotype follows an autosomal recessive Mendelian inheritance pattern.

DNA sample acquisition, exome sequencing, and *CA12* genotyping

Peripheral blood was obtained from all consented individuals in pedigree A and genomic DNA was extracted by a phenol/chloroform protocol. Exome capture was performed on all siblings within the pedigree using the Agilent SureSelect Human All Exon (51 Mb), and 100 bp paired-end reads were subsequently obtained from an Illumina HiSeq 2500 system as part of a study within the Baylor-Hopkins Center for Mendelian Genomics conducted by the Center for Inherited Disease Research. Reads were aligned to the hg19 reference genome using Burrows-Wheeler Aligner software (*88*) and subsequent

alignment processing was completed using SAMtools (*89*), PicardTools, and GATK softwares (*90, 91*) in a manner similar to (*92*). Filtering of variants was conducted via custom scripts, and variant prioritization was conducted using Enlis Genomic Research software. To verify mutations, a 654 bp region encompassing c.908-1 G>A and c.859_860insACCT was amplified from genomic DNA from the proband A and her mother by PCR using the following primers (IDT): 5'F GCCCTGTACTGCACACACAT and 3'R AGGATGATGCCCAGACTCAG. PCR products were purified using QIAquick PCR purification kit (Qiagen), and then sequenced using the Applied Biosystems 3730xl DNA Analyzer. The resulting sequences were analyzed via the Sequencher analysis suite (Gene Codes). Genomic DNA extracted from peripheral blood was obtained from all consented individuals in pedigree B. Exome capture was performed on all siblings using the Illumina Truseq Exome Enrichment kit (62 Mb), and 90 bp paired-end reads were subsequently obtained from an Illumina HiSeq 2000 system (Otogenetics). Exome sequencing data analysis was conducted in a manner similar to pedigree A, however, variant prioritization was conducted using VAAST software (*93*). *CA12* (RefSeq# NM_001218.4) mutations and mode of inheritance were verified via Sanger sequencing. A 475 bp region encompassing c.363 C>A was amplified from genomic DNA of all siblings in pedigree B by PCRs using the following primers (IDT): 5'F GTCCCATGCTCTGGTGTTATC and 5'R CTTTCCAAGGTGAACCAAGAA. PCR products were purified, sequenced, and analyzed as described for pedigree A. The resulting sequences were analyzed via the Sequencher analysis suite (Gene Codes).

It should be noted that all variant nomenclature is specific to *CA12* nucleotide and CA XII amino acid numbering and represent the minus strand sequence unless otherwise

specified. All genomic coordinates are specific to hg19. All variants discovered in this study have been submitted to ClinVar.

IHC of CA XII in human sweat duct and lung

Frozen discarded unidentified skin and lung obtained from the Division of Surgical Pathology, Johns Hopkins Hospital, Baltimore, MD, were embedded in Optimal Cutting Temperature (OCT) compound and held at -70°C prior to sectioning. Six μm cryo-sections were mounted onto uncoated microscope slides. Staining with H&E (Sigma, St Louis, MO, USA) for 1 min was performed for morphological evaluation. The rest of the slides were stored at -70°C until use. Sections were fixed for 10 min in pre-cooled acetone followed by 5 min peroxidase block at room temperature to quench the endogenous peroxidase activity. Sections were further incubated in serum-free protein block (Dako # X0909) for 20 min at room temperature and incubated overnight at 4oC with the following primary antibodies: anti-rabbit *CA12* (ProteinTech # 15180-1-AP) and anti-rabbit CA2 (LSBio # C138796). Relevant universal negative control antibodies: mouse (Dako # N1698) and rabbit (Dako # IR600) were used to ascertain nonspecific staining. After washing, staining was performed using EnVision+System-HRP (AEC) kit from Dako (# K4008). Sections were covered with peroxidase-labeled polymer for 30 mins. For visualization of the reaction, sections were developed in AEC+substrate-chromogen for 5-20 mins. After washing, the sections were counterstained with hematoxylin (Dako # S3309) for 30 seconds, cleared, and mounted on Faramount

aqueous mounting medium (Dako # S3025). Samples were analyzed under an Olympus BX51 microscope.

Nasal epithelial culture and Ussing chamber studies

Nasal epithelia from the proband of pedigree A were expanded and cultured using previous described method (*94*). Cells were mounted in Ussing chambers and studied as previously described (*95*). Apical and basolateral chambers contained the same bathing solution with symmetrical Cl- concentrations. *CFTR*-mediated Cl- current were measured using a previously described protocol (*95*).

Identification of *CA12* mutant transcripts

RNA was isolated from expanded nasal brushings of proband A by standard Trizol-chloroform method. cDNA was made using RT-PCR (Qiagen iScript) and served as template for *CA12* amplification. PCR products of *CA12* transcript isoforms were separated by 3% agarose gel electrophoresis and subject to Sanger sequencing.

Development of mutant *CA12* expression vectors

Full length wild-type (WT) *CA12* cDNA in bacterial pBS II expression vector was generously provided by Dr. William Sly. The proband B variant c.363 C>A (p.His121Gln) was introduced into full length WT *CA12* cDNA in bacterial pBSII

expression vector using the QuikChange II XL Site-Directed Mutagenesis kit reagents and protocol (Agilent). Mutagenesis products were confirmed by Sanger sequencing. Each cDNA was removed from the bacterial expression vector using restriction enzymes KpnI and BamHI, purified by gel electrophoresis, and ligated into the eukaryotic pcDNA5 FRT expression vector. Subcloning was confirmed by Sanger sequencing. Proband A variant c.859_860insACCT was introduced into pcDNA5 FRTexpression vector bearing *CA12* cDNA using the QuikChange II XL Site-Directed Mutagenesis kit reagents and protocol (Agilent). A previously described *CA12* missense variant, c.427 G>A (p.Glu143Lys), deemed a moderate hypomorph and reported in a large consanguineous Bedouin kindred manifesting a highly similar phenotype (*27*), (*26*) was also introduced into the full length *CA12* cDNA. To replicate the consequences of the splice acceptor variant c.908-1 G>A found in proband A, *CA12* cDNAs lacking exon 10 and lacking exons 9 and 10 were custom synthesized (GeneWiz, South Plainfield, NJ). Since exon 10 skipping was predicted to result in a frameshift and subsequent readthrough of the natural termination codon, 287 bp of the *CA12* 3' UTR were included in each construct. Both constructs were removed from the bacterial pUC57 expression vector using restriction enzymes KpnI and EcoRV, purified by gel electrophoresis, and ligated into eukaryotic pcDNA5 FRT expression vector. Subcloning was confirmed by Sanger sequencing.

Expression of WT CA XII and mutant proteins in HEK293 cells

HEK293 cells were transiently transfected with 500 ng of WT and mutant *CA12* vectors using Lipofectamine 2000 Reagent and standard protocol (Invitrogen). Cells were lysed 24 hours post-transfection. Western blotting of cell lysates was performed using anti-CA XII antibody (Novus #NBP1-81668), and loading control GAPDH antibody (Sigma #G9545).

Immunocytochemistry of CA XII in a polarized epithelial cell line

Madin-Darby canine kidney (MDCK) cells were transiently transfected with 1.6 µg of *CA12* cDNA using 3.2 µl Lipofectamine 2000 Transfection Reagent and protocol (Invitrogen). Cells were fixed one day post-transfection with 4% paraformaldehyde for 20 mins and rinsed with 1X PBS. Cells were permeabilized with 0.5% Triton X-100 for 5 mins, then rinsed with 1X PBS, and blocked overnight at 4C with 2.5% goat serum. Cells were immunostained using rabbit anti-CA XII primary antibody (ProteinTech #15180-1-AP) diluted 1:200 and anti-ZO1 primary antibody (Invitrogen) with a conjugated anti-mouse red fluorophore diluted 1:200, followed by incubation in anti-rabbit secondary antibody diluted 1:50. Findings were validated by staining with different primary anti-CA XII antibodies from Sigma Prestige (mouse) and Novus (rabbit). All antibodies were diluted in 2.5% goat serum blocking solution. Cells were washed in 1X PBS three times for 10 mins following the 90 mins primary antibody incubation. Cells were washed in 1X PBS four times for 15 mins following the 30 mins secondary antibody incubation. Cells were mounted on microscope slides with Molecular Probes ProLong Gold Antifade

Reagent with DAPI and viewed with the Zeiss LSM510-Meta single-point confocal laser-scanning microscope and Zen imaging software.

Carbonic anhydrase enzyme activity assay

Cell pellets were lysed by sonication in 300 µl lysis buffer (PBS containing protease inhibitors, 1 mM each of PMSF, o-phenanthroline, EDTA, benzamidine-hydrochloride and iodoacetamide plus 1% NP-40) and left on ice. The media were centrifuged to remove dead cells. The protein concentration of cell lysates was determined by microLowry's procedure using bovine serum albumin as a standard (*96*). The carbonic anhydrase activity was determined using Maren's procedure (*97*) as described (*98*). To reflect extracellular physiological chloride concentration (*76*), 100 mM NaCl was utilized.

Acknowledgements

This study would not be possible without the participation of the patients and families described in this manuscript. The authors would like to acknowledge Diane Acquazzino for acquisition of patient records. This work was supported by the National Institutes of Health [R01 DK044003]; and the Cystic Fibrosis Foundation [CUTTIN13A2].

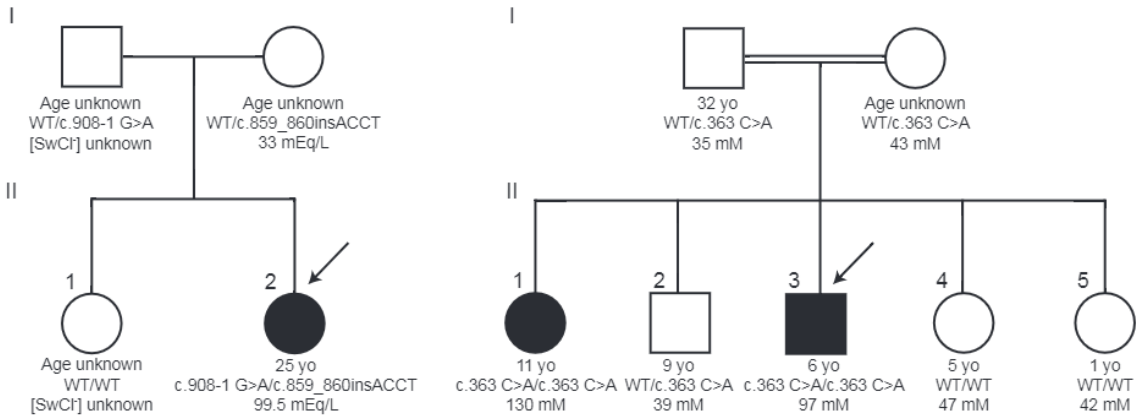


Figure 3.1. Segregation of putative deleterious *CA12* variants in two unrelated families. Filled shapes indicate status as affected and arrows indicate the proband in each family. The ages indicate the age of the individual at the time of exome sequencing. **(A)** Pedigree A: A white American family in which the proband exhibits consistently elevated sweat chloride concentration and bronchiectasis. The proband carries a splice acceptor variant c.908-1 G>A and an insertion frameshift variant c.859_860insACCT. **(B)** Pedigree B: An Omani family with first-cousin parents as indicated by the double horizontal line. The proband and affected sister exhibit elevated sweat chloride concentrations and have experienced multiple episodes of hyponatremic dehydration. Only the proband and affected sister are homozygous for c.363 C>A (p.His121Gln).

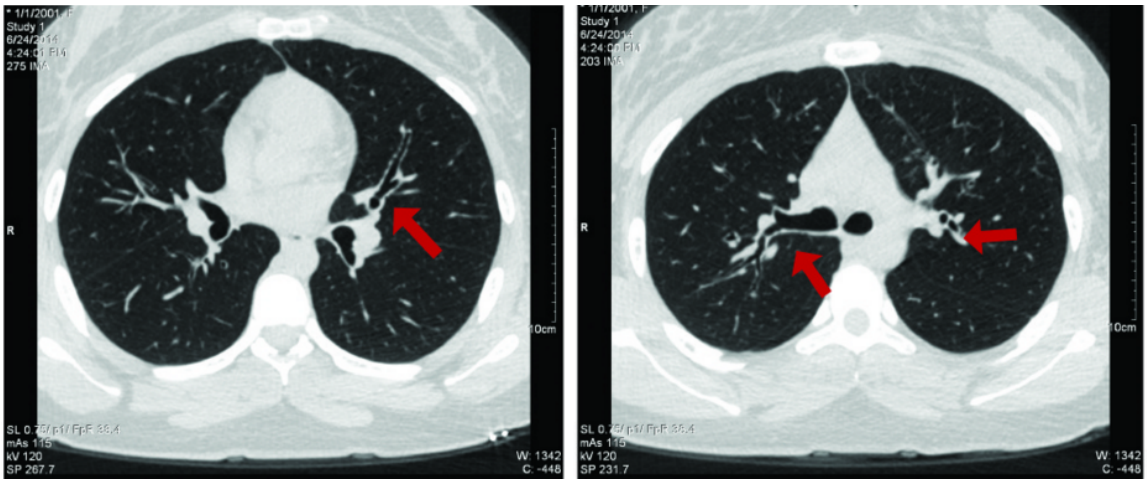


Figure 3.2. Axial plane high resolution CT images of proband A. Examples of enlargement of the airways (tram-tracking and signet rings) are indicated by red arrows. This bronchiectasis is seen in the absence of mucus plugging, scar tissue, or surrounding inflammation. It is unknown as to whether this mild pulmonary phenotype would be more exacerbated had the proband not been undergoing daily preventative lung therapies (aerosolized albuterol, hypertonic saline, chest physiotherapy, etc.) due to her original CF diagnosis.

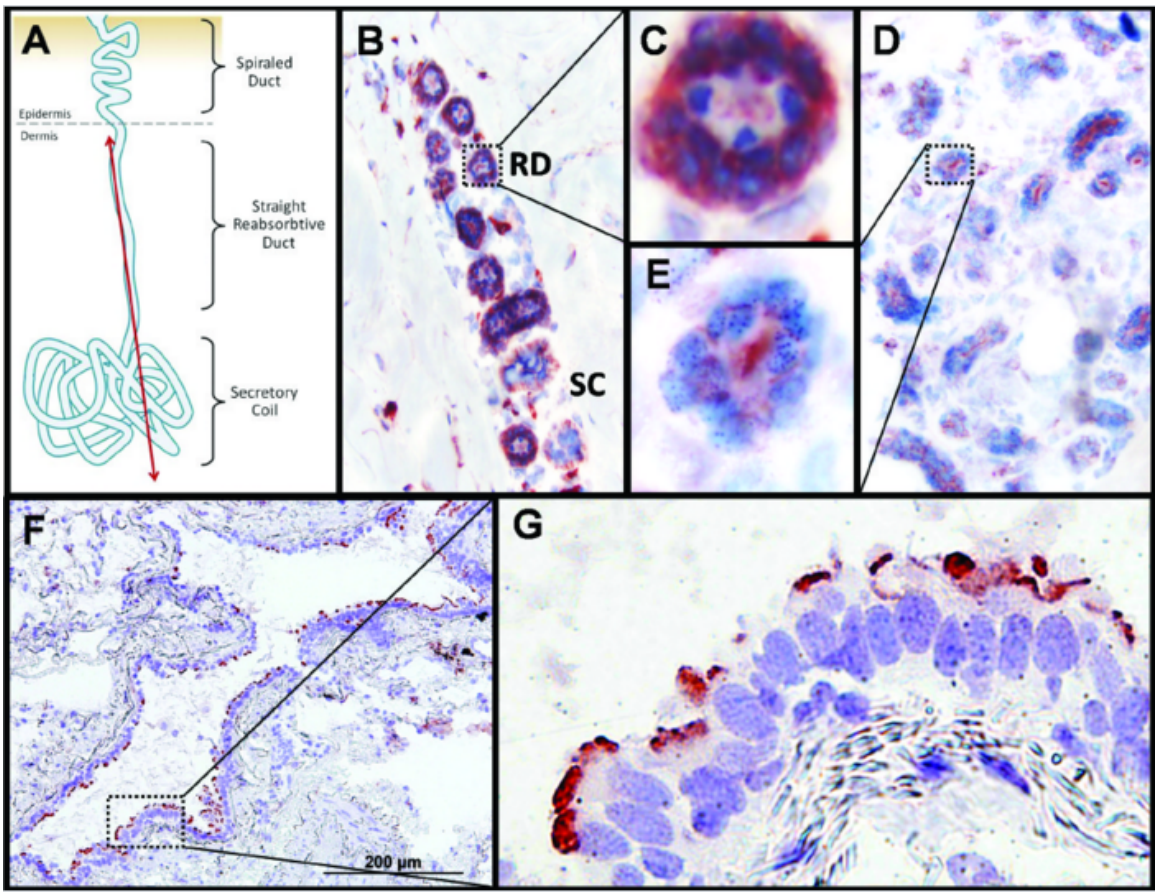


Figure 3.3. Immunohistochemical staining of CA XII and CA II in human skin and lung. (A) Diagram depicting longitudinal view of sweat gland components with red line indicating a hypothetical plane used to generate slices for the micrographs shown in panels B, C, D, E. **(B)** Two populations of sweat gland resorptive ducts ("RD") and secretory coils ("SC") immunostained for CA XII and counter-stained with hematoxylin and eosin. The resorptive ducts (n=9 different cross-sections captured in this panel) show positive staining for CA XII in a two cell thick layer of cuboidal epithelia cells. The myoepithelial cells surrounding the secretory coils (n=3 different cross-sections captured in this panel) show light staining of CA XII that may be non-specific. Magnification of this micrograph is 100x. **(C)** Enlargement of CA XII positive staining of the basolateral membrane in resorptive ductal cells from panel B. **(D)** Positive staining of apically

localized control protein CA II in sweat ducts and secretory coils. Magnification of this micrograph is 100x. **(E)** Enlargement of resorptive ducts from panel E showing apical localization of CA II. **(F)** CA XII positive staining of the luminal edge of a bronchiole cross-section. Magnification of this micrograph is 100x. **(G)** Positive staining of apically localized CA XII in the terminal bar of bronchial epithelia. Magnification of this micrograph is 400x.

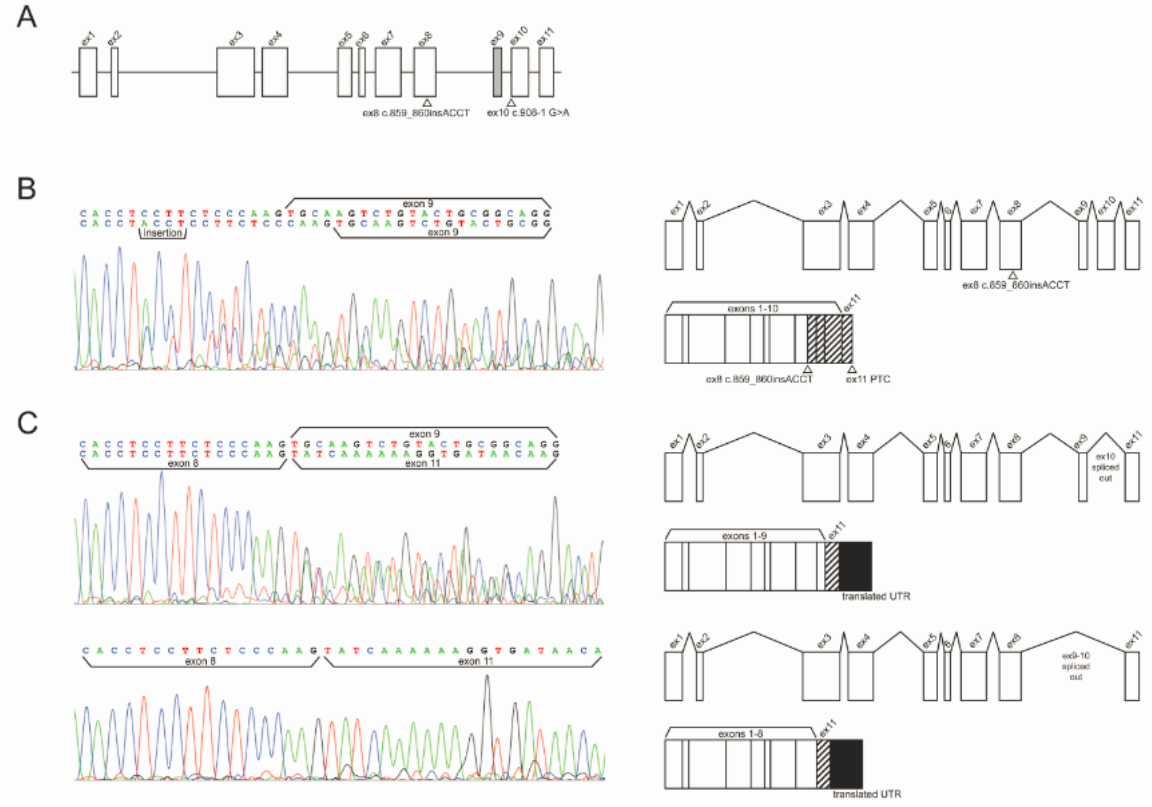


Figure 3.4. Effect of *CA12* variants upon RNA processing in nasal epithelial cells from proband A. (A) Exon and intron structure of *CA12* with locations of proband A variants identified by NGS. Rectangles represent exons and the lines through the center of the rectangles represent the genomic axis. Variants are indicated with triangles and HGVS cDNA names. The gray rectangle indicating exon 9 is spliced out in an alternative *CA12* transcript whose function and tissue distribution is unknown. **(B)** (*Left*) Electropherogram of Sanger sequencing of cDNA reverse transcribed from RNA extracted from proband A cultured nasal epithelial. Sequencing detected a heterozygous insertion variant c.859_860insACCT on a transcript retaining alternatively spliced exon 9. A second transcript detected by sequencing does not bear the insertion, retains exon 9, and is consistent with transcript lacking exon 10 as a result of the in trans variant c.908-1 G>A. Presence of this second transcript in this sequencing reaction is likely due to

imperfect isolation of the similarly sized transcripts by gel purification as transcript bearing the insertion is only 89 bp longer than transcripts lacking exon 10 only. (*Right*) Gene models depict RNA processing of the insertion variant and the predicted gene product. The insertion variant causes a frameshift starting in exon 8, a premature termination codon in exon 11, and was predicted to generate a misfolded protein targeted by ERAD. Hashed rectangles indicate an altered exonic reading frame. **(C)** (*Left*) Electropherogram of Sanger sequencing detecting transcript missing exons 9 and 10, and a second transcript missing exon 10 only, due to heterozygous splice acceptor variant c.908-1 G>A. Imperfect isolation of transcripts in this reaction is due to alternatively spliced exon 9 which is only 33 bp long. (*Right*) Gene models depict the two processed RNAs and protein products lacking exon 10 observed by RT-PCR: one with exon 9 alternatively spliced out and one retaining exon 9. This variant is predicted to cause skipping of exon 10, a frameshift resulting in readthrough of the native stop codon, translation of 267 nucleotides from the 3' UTR, and a misfolded protein targeted by ERAD. Hashed rectangles indicate an altered exonic reading frame and filled rectangles indicate translation of the 3' UTR.

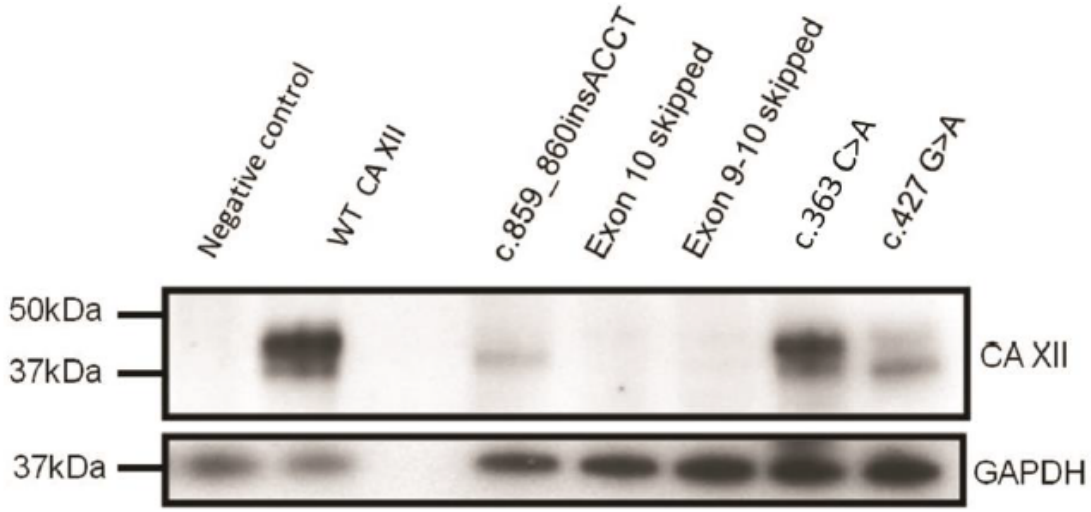


Figure 3.5. Expression of transiently transfected wild-type and mutant CA XII protein in HEK293 cells. Western blot of cell lysates extracted from HEK 293 cells (*top*) following transfection with CA XII expression vectors. Probing with anti-CA XII antibody (Novus) shows unglycosylated (39 kDa) and glycosylated (43 kDa) protein generated from transfections with WT CA12 (lane 2), c.363 C>A (lane 7) and c.427 G>A (lane 8). The insertion variant in proband A (c.859_860insACCT) produced a faint band of approximately 38 kDa while CA XII cDNA missing exon 10 and exons 9 and 10 sequence generated only faint bands. The negative control in the first lane is a mock transfection. The third lane has no lysate. GAPDH loading control (*bottom*) shows loading of cell lysates.

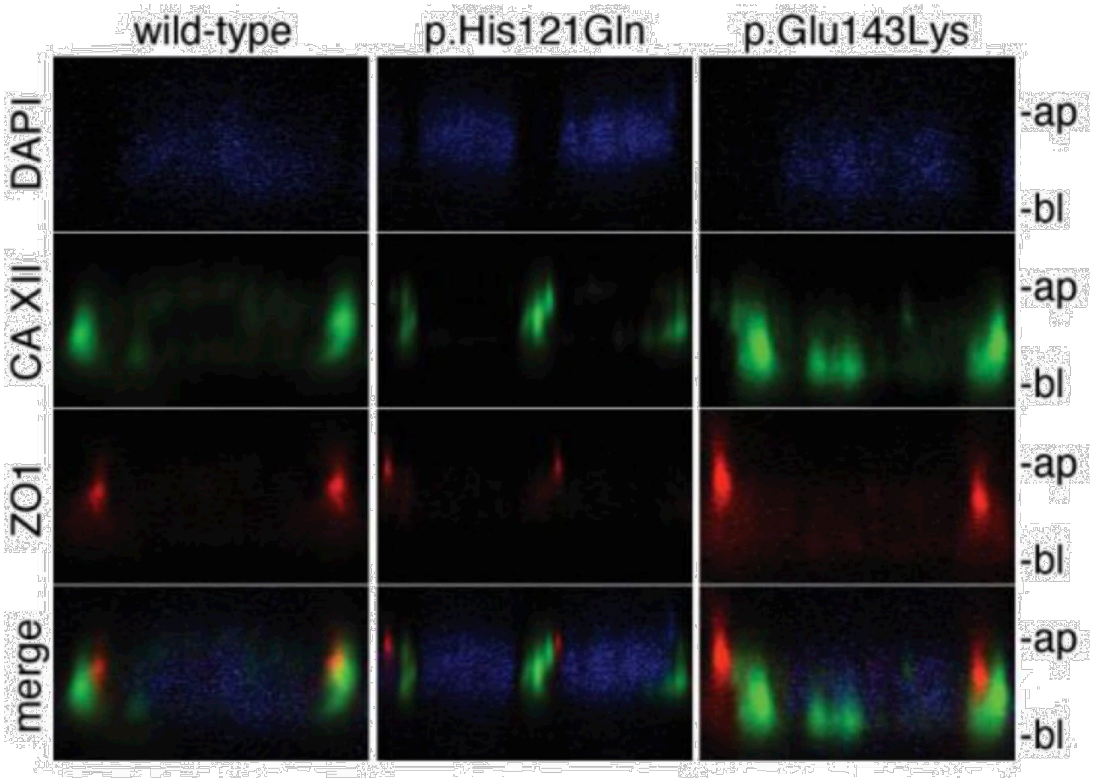


Figure 3.6. Subcellular localization of WT and mutant CA XII in polarized MDCK cells. Fluorescent co-staining of (*left*) WT CA XII (green), (*center*) p.His121Gln (green), and (*right*) p.Glu143Lys (green) with endogenous tight junction protein ZO1 (red) and nuclear stain DAPI (blue) in polarized MDCK cells imaged in the xz-plane. This micrograph reveals primarily lateral staining of CA XII; however, basal and lateral staining were observed for WT (n=7 different micrographs), p.His121Gln (n=8 different micrographs), and p.Glu143Lys (n=13 different micrographs). The apical membrane is indicated by "ap" and the basal membrane is indicated by "bl."

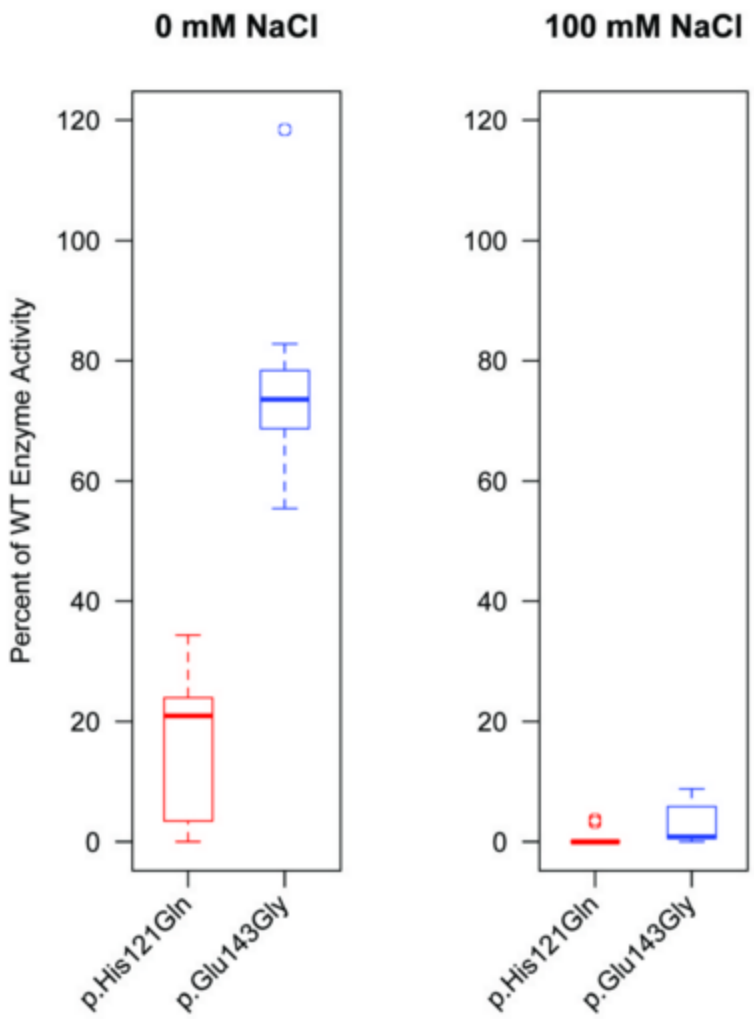
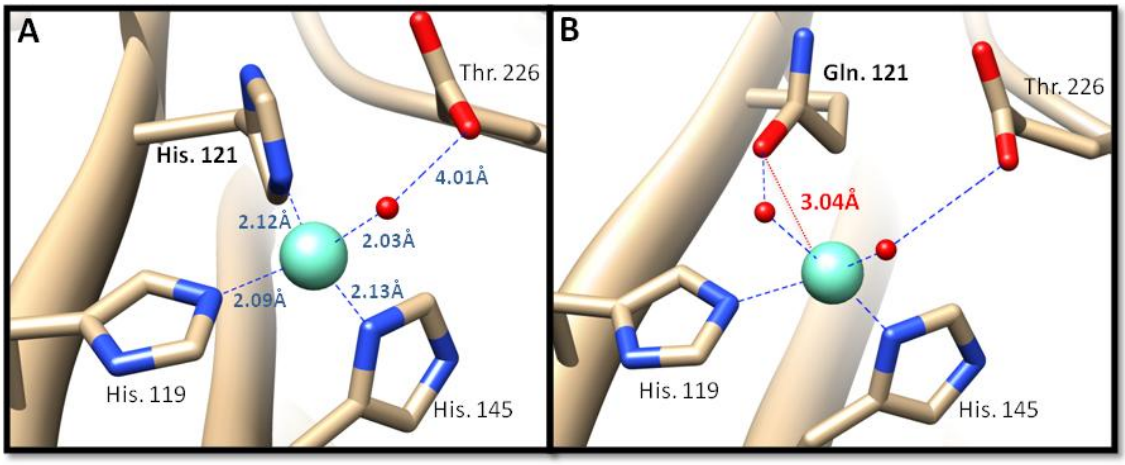


Figure 3.7. Enzymatic activity of CA XII proteins bearing p.His121Gln or p.Glu143Gly substitutions. Enzymatic activity of CA XII mutants p.His121Gln (red boxplot series) and p.Glu143Gly (blue boxplot series) was determined by assaying the reversible rate of hydration of CO2 as previously described in the absence (*left*) or presence (*right*) of physiological concentration of NaCl and normalizing to wild-type. Boxes represent the interquartile range (IQR) and the horizontal bars within the boxplots represent the median. The top whisker represents the 75th percentile plus 1.5 times the IQR and the bottom whisker represents the 25th percentile minus 1.5 times the IQR. Circles represent statistical outliers. (Boxplot statistics calculated in R.)



Supplemental Figure 3.1. Computational modeling of CA XII active site. (A) Wild-type active site showing zinc ion (teal) coordinated by three histidines (119, 121, 145) and hydroxide group (red sphere) stabilized by threonine 226. **(B)** Mutation of amino acid 121 to glutamine leading to transient coordination of zinc by hydroxide ion and an increased theoretical bond distance of 3.04Å when compared to native histidine. Bond lengths and zinc distance shown by red and blue lines respectively.

	WT	Ex 9 Skip	Ex 9 and 10 skip	Ex 10 skip	Total Depth	% WT	% Ex 9 skip	% Ex 9 & 10 skip	% Ex 10 skip
Kidney 1	592	0	0	0	592	100.00%	0.00%	0.00%	0.00%
Kidney 2	454	0	2	0	456	99.56%	0.00%	0.44%	0.00%
Bronchial Epithelia 3	196	0	13	0	209	93.78%	0.00%	6.22%	0.00%
Skin 1	166	23	11	8	208	79.81%	11.06%	5.29%	3.85%
Fallopian Tube	27	153	2	0	182	14.84%	84.07%	1.10%	0.00%
Bronchial Epithelia 2	153	0	12	0	165	92.73%	0.00%	7.27%	0.00%
Colon 1	111	9	1	0	121	91.74%	7.44%	0.83%	0.00%
Panc 2	92	0	0	0	92	100.00%	0.00%	0.00%	0.00%
Skin 2	45	0	4	0	49	91.84%	0.00%	8.16%	0.00%
Nasal Epithelia 1	42	0	1	0	43	97.67%	0.00%	2.33%	0.00%
Colon 2	36	0	0	0	36	100.00%	0.00%	0.00%	0.00%
Thyroid	32	0	0	1	33	96.97%	0.00%	0.00%	3.03%
Nasal Epithelia 2	32	0	0	0	32	100.00%	0.00%	0.00%	0.00%
Panc 1	27	0	0	0	27	100.00%	0.00%	0.00%	0.00%
Small Intestine	24	0	0	0	24	100.00%	0.00%	0.00%	0.00%
Brain 1	0	12	0	0	12	0.00%	100.00%	0.00%	0.00%
Stomach	11	0	0	0	11	100.00%	0.00%	0.00%	0.00%
Cervix	2	5	0	0	7	28.57%	71.43%	0.00%	0.00%
Testes 1	3	3	0	0	6	50.00%	50.00%	0.00%	0.00%
Brain 2	0	5	0	0	5	0.00%	100.00%	0.00%	0.00%
Lung 1	2	0	0	1	3	66.67%	0.00%	0.00%	33.33%
Ovary	0	2	0	0	2	0.00%	100.00%	0.00%	0.00%
Bronchial Epithelia 1	1	1	0	0	2	50.00%	50.00%	0.00%	0.00%
Heart 1	0	1	0	0	1	0.00%	100.00%	0.00%	0.00%
Lymph Node	0	1	0	0	1	0.00%	100.00%	0.00%	0.00%
Lung 2	1	0	0	0	1	100.00%	0.00%	0.00%	0.00%
Testes 2	1	0	0	0	1	100.00%	0.00%	0.00%	0.00%

Supplemental Table 2.1. RNA-sequencing splice junction data shows distribution of CA XII isoforms in various tissues. Raw reads obtained from various tissues through the sequence read archive were aligned to the hg19 reference genome and splice junctions were detected using TopHat software. Depth indicates number of reads supporting each splice junction, with lower depths indicative of limited expression of *CA12* in that tissue. Percentages relative to all splice junctions detected for this region are

also included. All general tissue samples (non-epithelia) were obtained from the Human Protein Atlas through the SRA (ERP003613). Non-epithelia tissues are presumed to be of mixed cell types. Epithelia specific data were obtained from SRA studies: SRP018883, SRP044906, and SRP058237.

Chapter 4

Discussion and Conclusions

The work presented here focuses on explaining the genetic etiology of cystic fibrosis (CF) in two patient cohorts with incomplete genotypes. One patient cohort had one identified CF-causing mutation after full molecular diagnostics of the coding region and flanking intronic sequence of *CFTR*. This cohort was considered a source of rare, unannotated, non-coding CF-causing variation within *CFTR*. The second patient cohort had no identified CF-causing mutations and had *CFTR* definitively ruled out by linkage exclusion. This cohort was considered a source of rare, unannotated CF-causing variation outside of *CFTR*, demonstrating that CF is a genetically heterogeneous disease.

Chapter 2 presents a novel systematic computational method for identifying with high confidence variants that cause both loss and gain of splicing. This method was applied to *CFTR* deep intronic variants identified in "one mutation" CF patients. Four deep intronic cryptic splice variants were identified in six of 14 "one mutation" patients. One variant, *CFTR* c.3140-26 A>G had been previously identified as activating a novel cryptic acceptor in the flanking intron(*49*). Two variants, c.1680-877 G>T and c.3717+40 A>G, were shown to activate intronic cryptic donors *in vitro* using expression minigenes. The fourth variant, c.1584+689 G>A, was identified in a previously studied patient who had been shown to have severely decreased transcript from the same chromosome bearing c.1584+689 G>A(*14*). PCR of cDNA from this previously studied patient's nasal epithelia with overlapping primers did not identify aberrant splicing of the *CFTR* cDNA. However, given a 143 predicted cryptic pseudoexon and premature termination codon (PTC), nonsense-mediated decay could be expected to eradicate PTC-bearing transcripts sufficiently to avoid detection by PCR amplification. *CFTR* has demonstrated low expression in upper airways such as nasal epithelia and can be difficult to detect from this

primary tissue. This warrants further *in vitro* investigation of this variant's potential to pathologically alter splicing.

The efficacy of this method was further demonstrated in a cohort of six patients with X-linked dyskeratosis congenita (DKC) for whom no DKC-causing mutations had been identified after exhaustive searches of the coding region of the primary disease locus *DKC1*. Incomplete genotypes are a documented obstacle for DKC, with over 50% of patients showing unambiguous clinical features of DKC yet missing DKC-causing mutations(*46*). Two *DKC1* intronic splice variants were identified in the six DKC patients with incomplete genotypes. One variant, c.16+592 C>G, activated a 234 bp long deep intronic cryptic pseudoexon. A second variant in the *DKC1* exon 3 canonical acceptor, c.85-5 C>G, was predicted to result in three transcripts, one of which is the product of activation of an upstream cryptic acceptor.

These findings show the deep introns are latent reservoirs of disease-causing variation and that deep intronic cryptic splicing may not be rare and exotic, but more frequent than currently appreciated. Due to the inactivation of ancient exons over evolutionary time, the introns contain regions with remarkably high splice potential. These regions are particularly vulnerable to cryptic splice activation by single nucleotide variants as demonstrated by our findings.

The deep introns of patients with unambiguous Mendelian disease and incomplete genotypes specifically should be scrutinized for cryptic splice variants. Historically, deep intronic splice variants are identified when patient-derived RNA is analyzed *in vitro*. The introns contain far more variation than the exons due to lack of evolutionary constraint;

therefore, determining the disease liability of intronic variants is more difficult than exonic variants. This highlights the importance and urgency for the development of the next generation of variant annotation tools for context-dependent assessment of loss of function variants. The method described here compares the splice potential of the reference sequence to the splice potential of the same sequence bearing a variant of interest, providing a high quality filter to select variants with the highest chance of pathologically altering native splice patterns.

We show that a significant number of non-synonymous *CFTR* variants identified in CF patients are in fact splice variants. Some of these missense mutations are "deep exonic" and have highly deleterious predicted amino acid substitutions, emphasizing the need for an unbiased computational method for ascertaining splice variants that can be fit into a high volume variant annotation pipeline. The replacement of diagnostic genotyping panels with next generation sequencing panels has resulted in the detection of thousands of variants of uncertain clinical significance (VUS) in disease-associated genes by clinical genetics laboratories(*99*). Misannotation of a VUS recently precipitated a lawsuit against Quest Diagnostics, one of the largest American clinical genetics testing entity (https://www.genomeweb.com/molecular-diagnostics/mothers-negligence-suit-against-quests-athena-could-broadly-impact-genetic). There is a pressing need for the accurate annotation of VUS for disease liability. Given that any single nucleotide variant may impact splicing due to the flexibility of splice sequences, a computational tool with an agnostic decision-making foundation such as machine learning predictive models is ideal.

Chapter 3 evaluates *CA12* as a candidate gene for atypical CF, a phenotype consisting of elevated sweat chloride concentration and persistent hyponatremic

dehydration indistinguishable from mild CF caused by hypomorphic *CFTR* alleles. Three *CA12* variants identified in two unrelated and ethnically diverse families and segregating with disease were studied *in vitro* to show that loss of CA XII function is responsible for the atypical CF observed in these families.

Indeed, each proband was clinically evaluated for CF due to their diagnostic sweat chloride concentrations. All affected individuals performed within normal ranges on pulmonary function tests, suggesting that CA XII deficiency may be distinct from CF in having no respiratory involvement. However, upon closer inspection of the adult proband using high resolution chest CT scanning, moderate bronchiectasis highly similar to that observed in CF patients was detected. Our findings show CA XII is expressed in the bronchiolar epithelia. The *in vivo* studies of CFTR-mediated chloride conductance (both nasal potential difference in the adult proband and short circuit studies of cultured nasal epithelial biopsies) did not reveal the loss of CFTR-mediated chloride conductance in the absence of CA XII function that is observed in patients with two loss of function *CFTR* alleles. Given the strong resemblance of the adult proband's bronchiectasis to that observed in adult CF patients with two loss of function *CFTR* alleles, it is important that further studies determine the role of CA XII in the lung. High resolution CT scanning of additional individuals carrying loss of function *CA12* alleles will reveal whether the bronchiectasis observed in the adult proband is unrelated.

Searches for *CA12* loss of function alleles in other atypical CF patients have so far been unsuccessful. This indicates that there are likely other genes which are responsible, suggesting CF may be even more genetically heterogeneous. Up until now, *CA12* had not

been linked to CF so it is possible that there are other genes which contribute to CF which have not been captured by studies of the CFTR interactome(*100*).

Overall, this work emphasizes the need for careful phenotyping and genotyping of patients with Mendelian disease. Even for a highly interrogated disease locus such as *CFTR*, there remain many severe loss of function variants left to be identified. Patients with incomplete genotypes face multiple obstacles, including denial of insurance coverage, barring of access to treatments and therapeutics, lack of information for family planning, and the psychological challenges of a seemingly unknown disease etiology. Further, the underlying disease mechanisms of known variants need to be determined as the field of genetics works to develop personalized medicine initiatives such small molecules to correct protein folding or function and antisense oligonucleotides to correct aberrant mRNA splicing. Given the high cost of therapeutics to treat genetic disorders(*101*) (http://www.nytimes.com/2012/02/01/business/fda-approves-cystic-fibrosis-drug.html), it is vital that patients are administered the appropriate therapeutics with respect to the underlying disease mechanism of their mutations.

Additionally, *CA12* was found to be responsible for disease in two unrelated families, demonstrating that CF is not a single locus disorder but is genetically heterogeneous. The importance of CA XII in a CF-like phenotype provides additional understanding of the importance of pH regulation in the airways, corroborating previous findings in the porcine CF model(*3*).

It is important that CF care centers be aware that patients without *CFTR* mutations may in fact suffer from CA XII deficiency. The impact of CF treatments on the

prognosis of patients with loss of CA XII function remains to be determined; however, it is likely that the course of their disease is milder than the average patient with severe, pancreatic insufficient CF. Increased awareness and understanding of how CA XII deficiency causes a CF-like phenotype will lead to better clinical care of these patients. Given that patients suffering from CA XII deficiency may greatly resemble patients with loss of CFTR function, it is important that they are distinguished to avoid being needlessly administered expensive small molecule therapies which are only proven to operate on CFTR.

References

1. M. B. Sheridan *et al.*, Mutations in the beta subunit of the epithelial Na+ channel in patients with a cystic fibrosis-like syndrome. *Hum. Mol. Genet* **14**, 3493-3498 (2005).
2. B. J. Rosenstein, G. R. Cutting, The diagnosis of cystic fibrosis: A consensus statement. *J. Pediatr* **132**, 589-595 (1998).
3. A. A. Pezzulo *et al.*, Reduced airway surface pH impairs bacterial killing in the porcine cystic fibrosis lung. *Nature* **487**, 109-113 (2012).
4. M. J. Welsh, B. W. Ramsey, F. J. Accurso, G. R. Cutting, in *The Metabolic and Molecular Bases of Inherited Disease,* C. R. Scriver, A. L. Beaudet, D. Valle, W. S. Sly, Eds. (McGraw-Hill, Inc., New York, 2001), vol. III, chap. 201, pp. 5121-5188.
5. C. Stewart, M. S. Pepper, Cystic fibrosis on the African continent. *Genet Med*, (2015).
6. K. L. Ode *et al.*, Oral glucose tolerance testing in children with cystic fibrosis. *Pediatr. Diabetes* **11**, 487-492 (2010).
7. B. Kerem *et al.*, Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073-1080 (1989).
8. J. R. Riordan *et al.*, Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066-1073 (1989).
9. J. S. Wagener, E. T. Zemanick, M. K. Sontag, Newborn screening for cystic fibrosis. *Curr Opin Pediatr* **24**, 329-335 (2012).

10. P. R. Sosnay *et al.*, Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat. Genet* **45**, 1160-1167 (2013).

11. F. Van Goor, H. Yu, B. Burton, B. J. Hoffman, Effect of ivacaftor on CFTR forms with missense mutations associated with defects in protein processing or function. *J. Cyst. Fibros* **13**, 29-36 (2014).

12. A. M. Elliott, J. Radecki, B. Moghis, X. Li, A. Kammesheidt, Rapid detection of the ACMG/ACOG-recommended 23 CFTR disease-causing mutations using ion torrent semiconductor sequencing. *J Biomol Tech* **23**, 24-30 (2012).

13. J. D. Groman, M. E. Meyer, R. W. Wilmott, P. L. Zeitlin, G. R. Cutting, Variant cystic fibrosis phenotypes in the absence of CFTR mutations. *N. Engl. J. Med* **347**, 401-407 (2002).

14. M. B. Sheridan *et al.*, CFTR transcription defects in pancreatic sufficient cystic fibrosis patients with only one mutation in the coding region of CFTR. *J. Med. Genet* **48**, 235-241 (2011).

15. A. I. den Hollander *et al.*, Mutations in the CEP290 (NPHP6) gene are a frequent cause of Leber congenital amaurosis. *Am J Hum Genet* **79**, 556-561 (2006).

16. B. Pezeshkpoor *et al.*, Deep intronic 'mutations' cause hemophilia A: application of next generation sequencing in patients without detectable mutation in F8 cDNA. *J Thromb Haemost* **11**, 1679-1687 (2013).

17. T. Bonifert *et al.*, Pure and syndromic optic atrophy explained by deep intronic OPA1 mutations and an intralocus modifier. *Brain* **137**, 2164-2177 (2014).

18. J. Bonini *et al.*, Small-scale high-throughput sequencing-based identification of new therapeutic tools in cystic fibrosis. *Genet Med* **17**, 796-806 (2015).

19. M. W. Libbrecht, W. S. Noble, Machine learning applications in genetics and genomics. *Nat Rev Genet* **16**, 321-332 (2015).

20. W. E. Highsmith *et al.*, A novel mutation in the cystic fibrosis gene in patients with pulmonary disease but normal sweat chloride concentrations. *N Engl J Med* **331**, 974-980 (1994).

21. M. Chillon *et al.*, A novel donor splice site in intron 11 of the CFTR gene, created by mutation 1811+1.6kbA-->G, produces a new exon: high frequency in Spanish cystic fibrosis chromosomes and association with severe phenotype. *Am J Hum Genet* **56**, 623-629 (1995).

22. R. K. Singh, T. A. Cooper, Pre-mRNA splicing in disease and therapeutics. *Trends Mol Med* **18**, 472-482 (2012).

23. S. B. Ng *et al.*, Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet* **42**, 30-35 (2010).

24. J. C. Roach *et al.*, Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636-639 (2010).

25. X. Wang *et al.*, Hsp90 cochaperone Aha1 downregulation rescues misfolding of CFTR in cystic fibrosis. *Cell* **127**, 803-815 (2006).

26. E. Muhammad *et al.*, Autosomal recessive hyponatremia due to isolated salt wasting in sweat associated with a mutation in the active site of Carbonic Anhydrase 12. *Hum. Genet* **129**, 397-405 (2011).

27. M. Feldshtein *et al.*, Hyperchlorhidrosis caused by homozygous mutation in CA12, encoding carbonic anhydrase XII. *Am J Hum. Genet* **87**, 713-720 (2010).

28. H. McMurtrie *et al.*, The bicarbonate transport metabolon. *Journal of enzyme inhibition and medicinal chemistry* **19**, 231-236 (2004).

29. D. Sterling, B. V. Alvarez, J. R. Casey, The Extracellular Component of a Transport Metabolon EXTRACELLULAR LOOP 4 OF THE HUMAN AE1 Cl−/HCO EXCHANGER BINDS CARBONIC ANHYDRASE IV. *Journal of Biological Chemistry* **277**, 25239-25246 (2002).

30. P. E. Morgan, S. Pastorekova, A. K. Stuart-Tilley, S. L. Alper, J. R. Casey, Interactions of transmembrane carbonic anhydrase, CAIX, with bicarbonate transporters. *Am. J. Physiol Cell Physiol* **293**, C738-C748 (2007).

31. J. R. Casey, W. S. Sly, G. N. Shah, B. V. Alvarez, Bicarbonate homeostasis in excitable tissues: role of AE3 Cl−/HCO3− exchanger and carbonic anhydrase XIV interaction. *American Journal of Physiology-Cell Physiology* **297**, 1091-1102 (2009).

32. P. M. Quinton, Cystic fibrosis: impaired bicarbonate secretion and mucoviscidosis. *The Lancet* **372**, 415-417 (2008).

33. D. A. Stoltz *et al.*, Cystic fibrosis pigs develop lung disease and exhibit defective bacterial eradication at birth. *Sci. Transl. Med* **2**, 29ra31 (2010).

34. O. Soukarieh *et al.*, Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico Tools. *PLoS Genet* **12**, e1005756 (2016).

35. S. V. Molinski *et al.*, Genetic, cell biological, and clinical interrogation of the CFTR mutation c.3700 A>G (p.Ile1234Val) informs strategies for future medical intervention. *Genet Med* **16**, 625-632 (2014).

36. L. Cartegni, S. L. Chew, A. R. Krainer, Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* **3**, 285-298 (2002).

37. G. S. Wang, T. A. Cooper, Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8**, 749-761 (2007).

38. M. P. Reboul *et al.*, Splice mutation 1811+1.6kbA>G causes severe cystic fibrosis with pancreatic insufficiency: report of 11 compound heterozygous and two homozygous patients. *J Med Genet* **39**, e73 (2002).

39. NN269.

40. HS3D: Homo Sapiens Splice Sequence Database.

41. J. L. Li, L. F. Wang, H. Y. Wang, L. Y. Bai, Z. M. Yuan, High-accuracy splice site prediction based on sequence component and position features. *Genet Mol Res* **11**, 3432-3451 (2012).

42. Python scikit learn Machine Learning Library.

43. N. Sharma *et al.*, Experimental assessment of splicing variants using expression minigenes and comparison with in silico predictions. *Hum. Mutat* **35**, 1249-1259 (2014).

44. S. D. Reynolds *et al.*, Airway Progenitor Clone Formation is Enhanced by Y-27632-dependent Changes in the Transcriptome. *Am J Respir Cell Mol Biol*, (2016).

45. A. Poole *et al.*, Dissecting childhood asthma with nasal transcriptomics distinguishes subphenotypes of disease. *J Allergy Clin Immunol* **133**, 670-678 e612 (2014).

46. E. M. Parry *et al.*, Decreased dyskerin levels as a mechanism of telomere shortening in X-linked dyskeratosis congenita. *J Med Genet* **48**, 327-333 (2011).

47. F. Pagani, F. E. Baralle, Genomic variants in exons and introns: identifying the splicing spoilers. *Nat Rev Genet* **5**, 389-396 (2004).

48. M. Tzetis, A. Efthymiadou, S. Doudounakis, E. Kanavakis, Qualitative and quantitative analysis of mRNA associated with four putative splicing mutations (621+3A-->G, 2751+2T-->A, 296+1G-->C, 1717-9T-->C-D565G) and one nonsense mutation (E822X) in the CFTR gene. *Hum Genet* **109**, 592-601 (2001).

49. M. D. Amaral *et al.*, Cystic fibrosis patients with the 3272-26A>G splicing mutation have milder disease than F508del homozygotes: a large European study. *J Med Genet* **38**, 777-783 (2001).

50. G. Dujardin, D. Commandeur, C. Le Jossic-Corcos, C. Ferec, L. Corcos, Splicing defects in the CFTR gene: minigene analysis of two mutations, 1811+1G>C and 1898+3A>G. *J Cyst Fibros* **10**, 212-216 (2011).

51. J. Yu *et al.*, Minimal introns are not "junk". *Genome Res* **12**, 1185-1189 (2002).

52. J. E. Walter *et al.*, CASE RECORDS of the MASSACHUSETTS GENERAL HOSPITAL. Case 41-2015. A 14-Year-Old Boy with Immune and Liver Abnormalities. *N Engl J Med* **373**, 2664-2676 (2015).

53. M. Busslinger, N. Moschonas, R. A. Flavell, b^+ thalassemia: Aberrant splicing results from a single point mutation in an intron. *Cell* **27**, 289-298 (1981).

54. J. O. Kitzman, L. M. Starita, R. S. Lo, S. Fields, J. Shendure, Massively parallel single-amino-acid mutagenesis. *Nat Methods* **12**, 203-206, 204 p following 206 (2015).

55. L. M. Starita *et al.*, Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* **200**, 413-422 (2015).

56. H. C. Dietz, New therapeutic approaches to mendelian disorders. *N. Engl. J Med* **363**, 852-863 (2010).

57. R. S. Finkel *et al.*, Phase 2a study of ataluren-mediated dystrophin production in patients with nonsense mutation Duchenne muscular dystrophy. *PLoS One* **8**, e81302 (2013).

58. T. J. Lynch *et al.*, Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med* **350**, 2129-2139 (2004).

59. F. P.-A. D. A. Committee, Ivacaftor for the Treatment of Cystic Fibrosis in Patients Age 6 Years and Older with an R117H-CFTR Mutation in the CFTR Gene. (2014).

60. H. Yu *et al.*, Ivacaftor potentiation of multiple CFTR channels with gating mutations. *J. Cyst. Fibros* **11**, 237-245 (2012).

61. D. Baralle, M. Baralle, Splicing in action: assessing disease causing sequence changes. *J Med Genet* **42**, 737-748 (2005).

62. W. G. Fairbrother, R. F. Yeh, P. A. Sharp, C. B. Burge, Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007-1013 (2002).

63. P. J. Smith *et al.*, An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum Mol Genet* **15**, 2490-2508 (2006).

64. M. G. Reese, F. H. Eeckman, D. Kulp, D. Haussler, Improved splice site detection in Genie. *J Comput Biol* **4**, 311-323 (1997).

65. J. D. Groman *et al.*, Phenotypic and genetic characterization of patients with features of "nonclassic" forms of cystic fibrosis. *J Pediatr* **146**, 675-680 (2005).

66. Y. Feinstein *et al.*, Natural history and clinical manifestations of hyponatremia and hyperchlorhidrosis due to carbonic anhydrase XII deficiency. *Hormone Research in Paediatrics* **81**, 336-342 (2014).

67. F. G. Riepe *et al.*, Revealing a subclinical salt-losing phenotype in heterozygous carriers of the novel S562P mutation in the α subunit of the epithelial sodium channel. *Clinical endocrinology* **70**, 252-258 (2009).

68. M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*, (2015).

69. K. Mosler *et al.*, Feasibility of nasal epithelial brushing for the study of airway epithelial functions in CF infants. *Journal of Cystic Fibrosis* **7**, 44-53 (2008).

70. J. Haapasalo *et al.*, Identification of an alternatively spliced isoform of carbonic anhydrase XII in diffusely infiltrating astrocytic gliomas. *Neuro-oncology* **10**, 131-138 (2008).

71. M. Uhlen *et al.*, Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

72. L. E. Maquat, Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol* **5**, 89-99 (2004).

73. D. A. Whittington *et al.*, Crystal structure of the dimeric extracellular domain of human carbonic anhydrase XII, a bitopic membrane protein overexpressed in certain cancer tumor cells. *Proc Natl Acad Sci U S A* **98**, 9545-9550 (2001).

74. M. Mall *et al.*, Effect of genistein on native epithelial tissue from normal individuals and CF patients and on ion channels expressed in Xenopus oocytes. *British journal of pharmacology* **130**, 1884-1892 (2000).

75. S. Parkkila *et al.*, Expression of the membrane-associated carbonic anhydrase isozyme XII in the human kidney and renal tumors. *Journal of Histochemistry & Cytochemistry* **48**, 1601-1608 (2000).

76. M. Lee, *Basic skills in interpreting laboratory data*. (ASHP, 2009).

77. C. R. Scriver, S. Kaufman, in *Metabolic and Molecular Bases of Inherited Disease,* C. R. Scriver, A. L. Beaudet, D. Valle, W. S. Sly, Eds. (McGraw-Hill, Inc., New York, 2001), chap. 77, pp. 1667-1724.

78. W. S. Sly, G. Shah, in *The Metabolic and Molecular Bases of Inherited Disease,* C. R. Scriver, A. L. Beaudet, W. S. Sly, D. Valle, Eds. (McGraw-Hill, Inc., New York, 2001), vol. IV, chap. 208, pp. 5331-5343.

79. K. Gilmour, Perspectives on carbonic anhydrase. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* **157**, 193-197 (2010).

80. R. E. Tashian, D. HEWETT-EMMETT, S. J. Dodgson, R. E. Forster, W. S. Sly, The Value of Inherited Deficiencies of Human Carbonic Anhydrase Isozymes in Understanding Their Cellular Rolesa. *Annals of the New York Academy of Sciences* **429**, 262-275 (1984).

81. A. Kivelä *et al.*, Expression of a novel transmembrane carbonic anhydrase isozyme XII in normal human gut and colorectal tumors. *The American journal of pathology* **156**, 577-584 (2000).

82. P. Karhumaa *et al.*, Identification of carbonic anhydrase XII as the membrane isozyme expressed in the normal human endometrial epithelium. *Molecular human reproduction* **6**, 68-74 (2000).

83. L. Guglani, B. Sitwat, D. Lower, G. Kurland, D. J. Weiner, Elevated sweat chloride concentration in children without cystic fibrosis who are receiving topiramate therapy. *Pediatr. Pulmonol* **47**, 429-433 (2012).

84. J. Y. Winum, S. A. Poulsen, C. T. Supuran, Therapeutic applications of glycosidic carbonic anhydrase inhibitors. *Medicinal research reviews* **29**, 419-435 (2009).

85. C. T. Supuran, Carbonic anhydrases: novel therapeutic applications for inhibitors and activators. *Nature reviews Drug discovery* **7**, 168-181 (2008).

86. N. Mirza, A. G. Marson, M. Pirmohamed, Effect of topiramate on acid–base balance: extent, mechanism and effects. *British journal of clinical pharmacology* **68**, 655-661 (2009).

87. D. Weycker, J. Edelsberg, G. Oster, G. Tino, Prevalence and economic burden of bronchiectasis. *Clinical Pulmonary Medicine* **12**, 205-209 (2005).

88. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

89. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

90. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010).

91. M. A. Depristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet* **43**, 491-498 (2011).

92. S. B. Ng *et al.*, Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-276 (2009).

93. M. Yandell *et al.*, A probabilistic disease-gene finder for personal genomes. *Genome Res*, (2011).

94. X. Li *et al.*, Integrin alpha6beta4 identifies human distal lung epithelial progenitor cells with potential as a cell-based therapy for cystic fibrosis lung disease. *PLoS One* **8**, e83624 (2013).

95. X. Li *et al.*, CFTR is required for maximal transepithelial liquid transport in pig alveolar epithelia. *American journal of physiology. Lung cellular and molecular physiology* **303**, L152-160 (2012).

96. G. L. Peterson, Review of the Folin phenol protein quantitation method of Lowry, Rosebrough, Farr and Randall. *Analytical biochemistry* **100**, 201-220 (1979).

97. T. H. Maren, A simplified micromethod for the determination of carbonic anhydrase and its inhibitors. *Journal of Pharmacology and Experimental Therapeutics* **130**, 26-29 (1960).

98. V. Sundaram, P. Rumbolo, J. Grubb, P. Strisciuglio, W. S. Sly, Carbonic anhydrase II deficiency: diagnosis and carrier detection using differential enzyme inhibition and inactivation. *American journal of human genetics* **38**, 125 (1986).

99. J. Y. Cheon, J. Mozersky, R. Cook-Deegan, Variants of uncertain significance in BRCA: a harbinger of ethical and policy issues to come? *Genome Med* **6**, 121 (2014).

100. C. Li, A. P. Naren, Analysis of CFTR interactome in the macromolecular complexes. *Methods Mol Biol* **741**, 255-270 (2011).

101. M. Schlander, M. Beck, Expensive drugs for rare disorders: to treat or not to treat? The case of enzyme replacement therapy for mucopolysaccharidosis VI. *Curr Med Res Opin* **25**, 1285-1293 (2009).

Melissa Lee

melissalee0@gmail.com - DOB 7/13/1984

EDUCATION

Johns Hopkins University, Baltimore, MD June 2016
Ph.D., Human Genetics

University of Georgia, Athens, GA 2006
B.S., Cellular Biology

EXECUTIVE SUMMARY

- Meticulous and resourceful computational biologist specializing in the genetic etiology of disease
- Fluent at scripting and implementing machine learning in Python and intermediate statistical analysis in R
- Deep understanding of genetic principles from 10 years of academic research in human genetics and 4 years of experience in clinical diagnostic genetics
- Strong communication skills for facilitating collaborative work as demonstrated by the selection of a bioinformatics platform presentation at a large national clinical conference and highly interdisciplinary publications

TECHNICAL SKILLS

Programming

- Python including scikit learn, pandas, Numpy
- R
- Command line utilities
- HTML, CSS

Bioinformatics software

- Alignment: bowtie2, BWA
- Variant calling: GATK, SAMtools, Breakdancer
- Analysis: Plink, bedtools
- Visualization: IGV, R including ggplot, Haploview

Databases

- 1000 Genomes
- Exome Aggregation Consortium (ExAC)
- UCSC Genome Browser
- dbSNP
- ClinVar
- Ensembl
- HGMD
- OMIM
- The Cancer Genome Atlas (TCGA)

RESEARCH EXPERIENCE

Doctoral Research, Advisor: Garry Cutting, M.D. 2010-present
Johns Hopkins University, Institute of Genetic Medicine, Baltimore, MD

- Used machine learning to develop SVM models for the prediction of cryptic splice sites from NGS data for the diagnosis of cystic fibrosis (CF) patients resulting in a first author manuscript (in preparation)
- Performed qualitative and statistical analyses of genetic and phenotypic data in large public online databases to assess disease variant annotations resulting in a platform presentation
- Developed pipeline for joint variant calling and annotation of NGS data and haplotype phasing for a patient cohort utilizing GATK and Plink
- Evaluated candidate genes responsible for atypical CF utilizing exome and whole genome sequencing and a novel variant filtering and prioritization strategy resulting in a first author manuscript (under review)
- Coordinated physician consultants and biochemistry collaborators at outside institutions for experimental validation of atypical CF candidate genes

Undergraduate Research, Advisor: Mary Bedell, Ph.D. 2005-2006
University of Georgia, Department of Genetics, Athens, GA

- Investigated differential tissue and developmental expression of Kit ligand mRNA and protein

Undergraduate Research, Advisor: Stephanie Sherman, Ph.D. 2004-2005
Emory University, Department of Human Genetics, Atlanta, GA

- Investigated collagen 6A genes and *PDE9A* spliceforms as candidates for congenital heart defects in Down syndrome
- Evaluated the *FMR1* CGG repeat array for correlations with disease severity in Fragile X syndrome
- Investigated correlation of skewed X chromosome inactivation with disease severity in Fragile X syndrome

PROFESSIONAL EXPERIENCE

Clinical Genetics Technologist 2007-2010
Emory Genetics Laboratory, Emory University Department of Human Genetics, Decatur, GA

- Performed a wide variety of molecular diagnostic tests in accordance with standards specified by CLIA and the College of American Pathologists (CAP)
- Lead trainer for new personnel on molecular genetics techniques, proper use and maintenance of instrumentation, and QA/QC compliance
- Developed and performed testing for the first clinical diagnostic gene-targeted CGH array, in addition to R&D of other new clinical assays
- Responsible for procurement of laboratory consumables, equipment, software, and maintenance contracts with a purchase value of approximately $1M per year

Freelance Website Design and Development 2002-present

- Developed and implemented web and graphic design solutions for a variety of clients, including small local and international businesses, large online fan communities, and cosplayer Monika Lee (averaging 2k unique visits per month)

PUBLICATIONS

M. Lee, P. Roos, N. Sharma, T. A. Evans, M. J. Pellicore, S. Stanley, S. Khalil, G. M. Solomon, A. N. Lam, K. S. Raraigh, B. Vecchio-Pagan, and G. R. Cutting. Systematic computational identification of variants that activate exonic and intronic cryptic splice sites. 2016. (in preparation)

M. Lee, B. Vecchio-Pagan, N. Sharma, A. Waheed, X. Li, K. S. Raraigh, S. Robbins, S. T. Han, A. L. Franca, M. J. Pellicore, T. A. Evans, H. Nguyen, S. Luan, D. Belchis, J. Hertecant, J. Zabner, W. S. Sly, and G. R. Cutting. Loss of carbonic anhydrase XII function in individuals with elevated sweat chloride concentration and pulmonary airway disease. *Hum Mol Genet*. 2016 Feb 23. pii: ddw065. [Epub ahead of print]

N. Sharma, P. R. Sosnay, A. S. Ramalho, C. Douville, A. Franca, L. B. Gottschalk, J. Park, M. Lee, B. Vecchio-Pagan, K. S. Raraigh, M. D. Amaral, R. Karchin, and G. R. Cutting. Experimental assessment of splicing variants using expression minigenes and comparison with in silico predictions. *Hum Mutat*. 2014 Oct;35(10):1249-59.

PRESENTATIONS

North American Cystic Fibrosis Conference, Atlanta, GA 2014

M. Lee, K. S. Raraigh, and G. R. Cutting. Evaluation and navigation of online resources for the clinical interpretation of *CFTR* variants. (Platform)

American Society of Human Genetics, San Diego, CA 2014

M. Lee, B. Vecchio-Pagàn, and G. R. Cutting. Assessment of the variant annotation interpretive gap among major variant databases. (Poster)

North American Cystic Fibrosis Conference, Salt Lake City, UT 2013

M. Lee, B. Vecchio-Pagàn, N. Sharma, A. Waheed, J. Hertecant, W. Sly, and G. R. Cutting. Recurrent hyponatremic dehydration and elevated sweat chloride concentration due to a mutation in carbonic anhydrase XII. (Poster)

American Society of Human Genetics, Boston, MA 2013

B. Vecchio-Pagàn, M. Lee, N. Sharma, A. Waheed, D. Belchis, J. Hertecant, W. Sly, and G. R. Cutting. Loss of function mutations in Carbonic Anhydrase XII result in hyponatremic dehydration and elevated sweat chloride concentration. (Poster)

Association for Genetic Technologists, Jacksonville, FL 2009

M. Lee, E. L. H. Chin, B. H. Billotte, B. Coffee, L. J. H. Bean, and M. Hegde. Combining sequencing and CGH array to detect mutations in Sandhoff disease. (Poster)

Foundation for Genetic Technology, Atlanta, GA 2008

M. Lee, E. L. H. Chin, B. H. Billotte, C. R. Alexander, B. Coffee, L. J. H. Bean, and M. Hegde. Gene conversion resulting in multiple Gaucher disease-causing mutations. (Poster)

American Society of Human Genetics, Toronto, ON 2004

R. E. Pyatt, M. Gupte, M. Lee, C. Torfs, C. Capone, K. Dooley, S. B. Freeman, and S. L. Sherman. Genetic variation in *COL6A1* in Down syndrome individuals with congenital heart defects and their parents. (Poster)

CERTIFICATIONS

American Society of Clinical Pathologists, Board of Certification 2009-2012

Certified Technologist in Molecular Biology, MB(ASCP)