Genetics inMedicine | ORIGINAL RESEARCH ARTICLE

Small-scale high-throughput sequencing–based identification of new therapeutic tools in cystic fibrosis

Jennifer Bonini[1,2], Jessica Varilh[1,3], Caroline Raynal, PharmD, PhD[1,3], Corinne Thèze[1,3], Emmanuelle Beyne, PhD[1,3], Marie-Pierre Audrezet, PhD[4], Claude Ferec, MD, PhD[4], Thierry Bienvenu, MD, PhD[5], Emmanuelle Girodon, MD, PhD[5], Sylvie Tuffery-Giraud, PhD[1,2], Marie Des Georges, PharmD[1,3], Mireille Claustres, MD, PhD[1,2] and Magali Taulan-Cadars, PhD[1,2]

Purpose: Although 97–99% of *CFTR* mutations have been identified, great efforts must be made to detect yet-unidentified mutations.

Methods: We developed a small-scale next-generation sequencing approach for reliably and quickly scanning the entire gene, including noncoding regions, to identify new mutations. We applied this approach to 18 samples from patients suffering from cystic fibrosis (CF) in whom only one mutation had hitherto been identified.

Results: Using an in-house bioinformatics pipeline, we could rapidly identify a second disease-causing *CFTR* mutation for 16 of 18 samples. Of them, c.1680-883A>G was found in three unrelated CF patients. Analysis of minigenes and patients' transcripts showed that this mutation results in aberrantly spliced transcripts because of the inclusion of a pseudoexon. It is located only three base pairs from the c.1680-886A>G mutation (1811+1.6kbA>G), the fourth most frequent mutation in southwestern Europe. We next tested the effect of antisense oligonucleotides targeting splice sites on these two mutations on pseudoexon skipping. Oligonucleotide transfection resulted in the restoration of the full-length, in-frame *CFTR* transcript, demonstrating the effect of antisense oligonucleotide-induced pseudoexon skipping in CF.

Conclusion: Our data confirm the importance of analyzing noncoding regions to find unidentified mutations, which is essential to designing targeted therapeutic approaches.

Genet Med advance online publication 8 January 2015

Key Words: antisense oligonucleotides; cystic fibrosis; intronic mutation; next-generation sequencing; pseudoexon skipping

INTRODUCTION

Cystic fibrosis (CF) is a common autosomal recessive disorder caused by mutations in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene. This gene displays great mutational heterogeneity, depending on the ethnic and phenotypic background, with almost 2,000 *CFTR* alterations referenced (*genet.sickkids.on.ca/*). Identification of disease-causing mutations has been of particular importance, not only for diagnosis but also in the recent era of therapeutic approaches targeted to genetic defects. Of the reported *CFTR* mutations, more than 200 are classified as splicing alterations. About 60% of them affect the canonical GT/AG dinucleotides that form the consensus sequences of donor and acceptor splice sites (ss). Mutations in deep introns are poorly referenced. Among them, three (c.3718-2477C>T in intron 22 (3849+10kbC>T), c.1680-886A>G in intron 12 (1811+1.6kbA>G), and c.870-1113-870-1110delGAAT in intron 7) are severe mutations that cause cryptic exon inclusion.[1–4]

Currently, 2–5% of CF mutations remain unknown and are probably located deep in introns, inducing aberrant splicing events. Because of technical difficulties of using the classic Sanger method (a labor-intensive and time-consuming task) to sequence entire genes, intron analysis has been quite limited, and thus the identification of these remaining mutations is challenging.

The development of next-generation sequencing (NGS) technologies with an immense capacity (up to 600 Gb per run) represents major progress in human genetics. Whole-genome and exome sequencing are becoming popular approaches, and exome sequencing can be a powerful tool to identify the molecular basis of monogenic diseases.[5,6] By allowing the rapid sequencing of a considerable number of samples,[7,8] these methods have provided precious sequence data that are collected in reference data sets. Each NGS platform generates different read lengths that range from short reads (e.g., 35 bases) to reads longer than 500 bases. For a number of applications, including targeted resequencing, chromatin immunoprecipitation sequencing, and RNA sequencing, short reads are highly informative and adequate. Conversely, longer reads are more suitable for assembling de novo genomes, mapping highly-homologous regions (related gene family and pseudogenes), and sequencing repetitive DNA regions, such as introns.

The first two authors contributed equally to this work.
[1]INSERM U827, Laboratoire de Génétique de Maladies Rares, Montpellier, France; [2]Université Montpellier I, UFR de Médecine, Montpellier, France; [3]Laboratoire de Génétique Moléculaire, CHU Montpellier, Montpellier, France; [4]Laboratoire de Génétique Moléculaire et d'Histocompatibilité, CHRU, Brest, France; [5]AP-HP, Service de Biochimie et Génétique Moléculaires, Groupe Hospitalier Cochin Broca Hôtel Dieu, Paris, France. Correspondence: Magali Taulan-Cadars (magali.taulan@inserm.fr)

Submitted 9 September 2014; accepted 24 November 2014; advance online publication 8 January 2015. doi:10.1038/gim.2014.194

Despite their promises, NGS technologies have significant limitations: high error rates, enrichment of rare variants, and a large proportion of missing values, as well as the fact that most current analytical methods are designed for population-based association studies. Indeed, because NGS produces massive amounts of data, their analysis and interpretation are time consuming, not trivial, and a real challenge. Therefore, the complex and time-consuming "postsequencing" data analysis currently limits the application of high-throughput NGS technologies in most laboratories. If specific portions of a genome need to be analyzed, targeted enrichment (by hybrid capture, circularization, or polymerase chain reaction (PCR)) can be useful. For instance, two recent publications described a strategy for resequencing the *CFTR* gene in patient samples harboring previously characterized *CFTR* mutations and polymorphisms.[9,10]

Thus the first aim of this study was to develop a robust approach for fast and efficient resequencing of the whole *CFTR* gene to identify new, as-yet unidentified mutations by using any small-scale NGS sequencing platform. To this aim, we compared two target-enrichment techniques (hybrid capture and long-range PCR (LR-PCR)) and developed an "in-house" pipeline for easy postsequencing data analysis of a CF patient and his parents to identify unknown pathogenic *CFTR* variants. This pipeline then was applied to 17 samples from CF patients in whom only one mutation had hitherto been identified. The combination of bioinformatics analysis and experimental validation led to the identification of

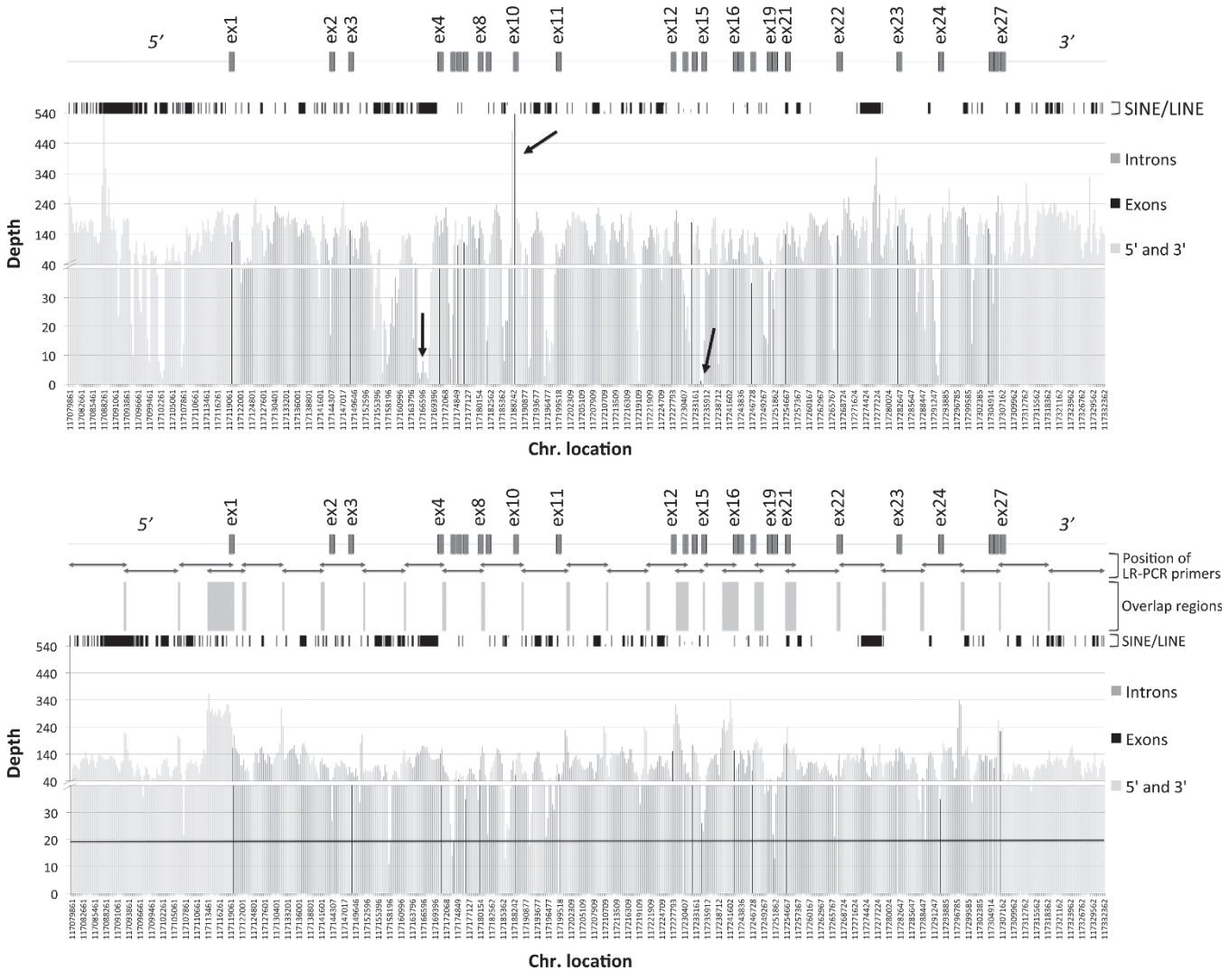


Figure 1 Graphic representation of the sequencing coverage/depth of the father's DNA sample after hybrid capture and long-range polymerase chain reaction (LR-PCR) enrichment (mapping to chromosome 7). (**a**) The whole *CFTR* locus was split into 400-bp fragments and then compared with the sequencing data. The *CFTR* flanking regions are in light gray, intronic regions in gray, and exons in black. SINEs/LINEs are shown in black at the top of the graphic representation. The three arrows highlight a region with excess coverage (exon 10, duplicated region) and two regions with low coverage (intron 3 and exon 15). (**b**) The whole *CFTR* locus sequence after LR-PCR enrichment. The position of the LR-PCR primers and overlapping regions, which often were oversequenced, are shown on top LINE, long interspersed nuclear element; SINE, short interspersed nuclear element.

a new mutation in three unrelated patients with classic CF and a single known mutation.

MATERIALS AND METHODS

Subjects

DNA was extracted from lymphocytes of peripheral blood samples from 18 unrelated patients with CF using standard protocols. All patients had classic CF (with a sweat chloride test result >60 mEq/l) and a single known mutation. All samples studied in this work had previously undergone conventional *CFTR* screening to scan exons and flanking junctions and to search for large rearrangements.

To simplify data processing we present only results from samples extracted from one CF patient and his parents (a family trio study). Written informed consent for *CFTR* studies was obtained.

Sequencing protocol

Detailed protocols on the GS Junior Sequencer (454 Life Technologies, Brandford, CT) are available in the **Supplementary Materials** online.

In silico analysis

Detailed protocols based on previous data,[11] and algorithms defined by different studies[12–16] are available in the **Supplementary Materials** online.

Functional studies: splicing reporter constructs

The impact of the newly discovered variant on splicing was tested as previously described;[17] specific information on the constructs is available in the **Supplementary Materials** online.

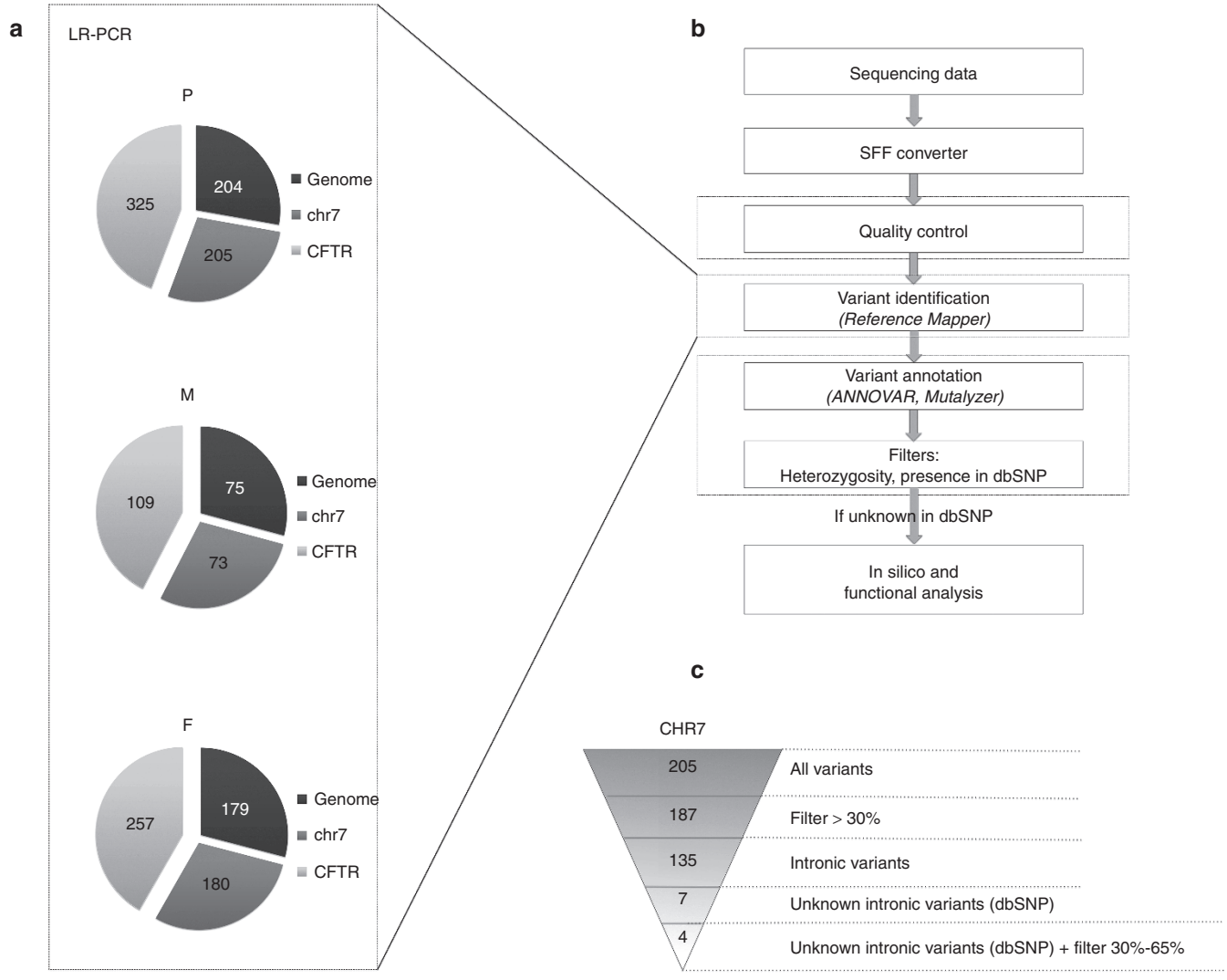
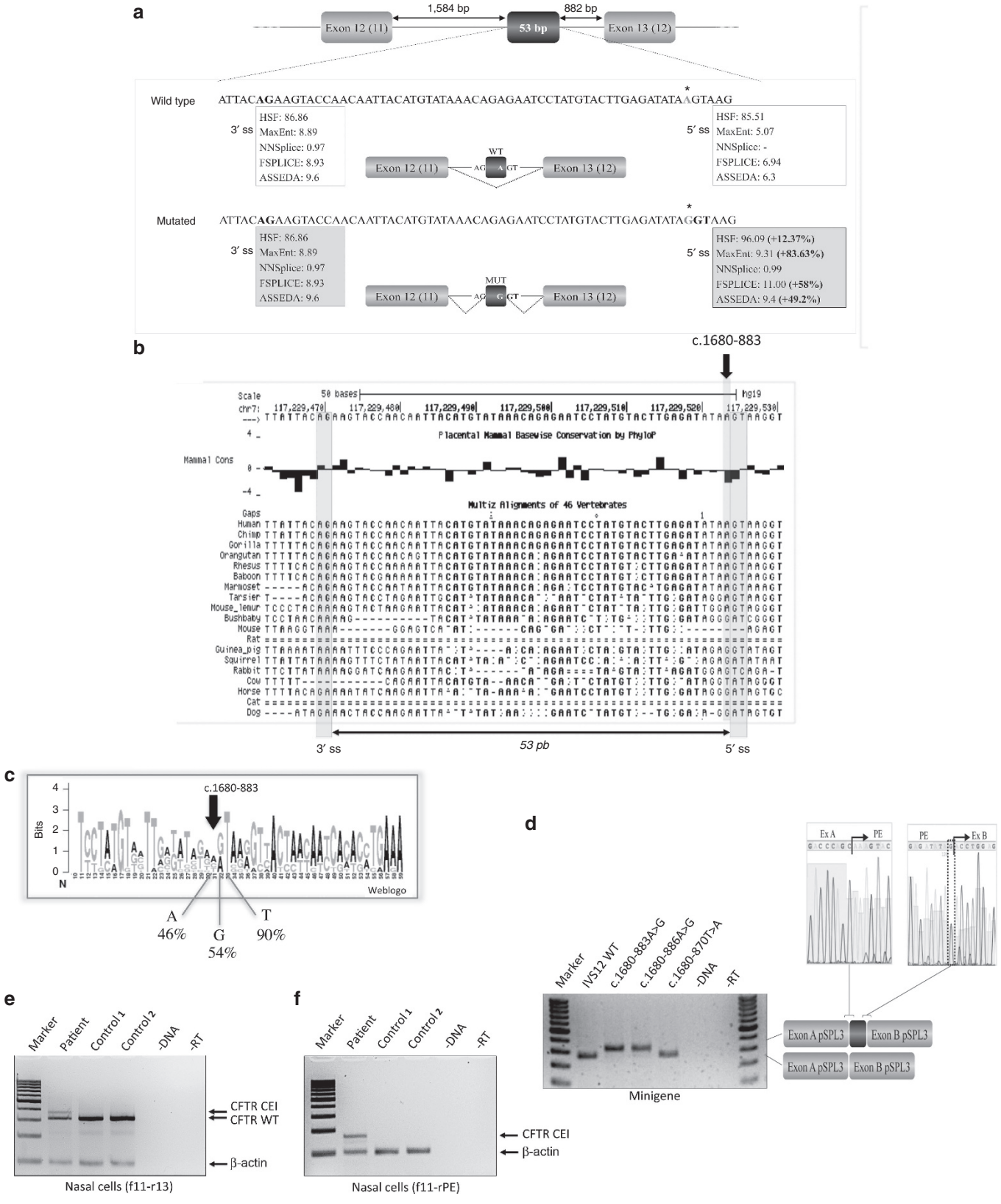


Figure 2 Variant mapping, in-house pipeline, and variant identification. (**a**) Circular diagrams showing the number of variants found after alignment (hg19) to the entire human genome (genome), chromosome 7 (chr7), or *CFTR* as the reference. The number of variants depends on the mapping reference. F, father; M, mother; P, patient. (**b**) In-house pipeline for variant identification based on the local Web-based Galaxy platform (dotted box). Variant annotation is implemented with in-house Perl modules. Heterozygosity (threshold fixed at 30–65%) and dbSNP filtering were applied to all variants extracted after annotation using Annovar and Mutalyzer. (**c**) The pipeline was used to analyze the variants identified in the patient's sample. The results shown in **a** and **c** correspond to data generated by long-range polymerase chain reaction enrichment.

Cell culture, transfection, and target site blocker treatment. Human bronchial BEAS-2B cells were cultured as previously described.[18] Twenty-four hours before transfection, cells were plated in six-well plates and transiently transfected once, at about 80% confluence, with 1.5 μg of each minigene construct using the PolyFect transfection reagent (Qiagen, Courtaboeuf,



France). Cells were harvested after 48 h for transcript analysis. For target site blocker (TSB) treatment (European Patent application number EP13306250.5), cells were cotransfected with the *CFTR* minigene constructs and 25, 50, and 100 nmol/l or 1 µmol/l TSB using the Interferin transfection reagent (Polyplus; Ozyme, Illkirch, France).

Transcript analysis. Total RNA was extracted from BEAS-2B cells using the RNeasy Plus kit (Qiagen). At least two independent transfections were carried out for all experimental conditions. The impact on splicing was tested, as previously reported.[17] The reverse transcriptase PCR products also were sequenced using the Big Dye Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems/Life Technologies, St Aubin, France) on an ABI-3130XL Genetic Analyzer. The relative amount of each *CFTR* splicing product was determined by measuring the peak area (evaluated by GeneMapper software) and dividing it by the sum of all peak areas detected in the same PCR reaction.

Total RNA was extracted from the patient's nasal cells and from two controls without CF using the RNeasy Plus kit (Qiagen). Reverse transcription was produced from 500 ng of total RNA with the M-MLV Reverse Transcriptase kit (Invitrogen, Villebon sur Yvette, France). One microliter of each reverse transcriptase product was used for PCR amplification with primers encompassing intron 12 (f11-r13) and specific primer pairs amplifying a pseudoexon (PE) (f11-rPE). Details are available in the **Supplementary Materials** online.

Nomenclature

We used the international nomenclature recommended by the Human Genome Variation Society (http://www.hgvs.org/mutnomen/), in accordance with the *CFTR* gene numbering in which nucleotide +1 in the coding DNA reference sequence (GenBank NM_00492.3) corresponds to the *A* of the ATG translation initiation codon. For convenience, the legacy name of previously described mutations was added in parentheses.

RESULTS

Sequencing and coverage

To test our approach for small-scale NGS projects, after enrichment by hybrid capture or LR-PCR, we sequenced the entire *CFTR* locus on chromosome 7q32 in 18 CF samples. In all samples and on average, we found 95% of coverage for the capture and 98% for the LR-PCR, with depth up to 10×. Under 10×, we considered the reads as artifacts, that is, problems of sequencing. For the LR-PCR, the remaining 2% are probably due to problems in sequencing, as no uncovered region is common between the patients. The details of the runs for one CF patient and his parents (in family trio studies, the parents serve as controls to filter out benign variants and establish the pipeline) are summarized in **Supplementary Table S1** online. On average, on the complete gene, the mean depth of coverage varied from 50× to 112× (44–113× for exons, 54–118× for introns, and 74–152× for gene regions, including 5′ untranslated region (UTR) and 3′ UTR sequences) with both methods (**Supplementary Table S2** online). Because the father's DNA from the family trio study was associated with the lowest number of reads, we present his DNA to show that this amount of reads was sufficient to obtain informative sequence data.

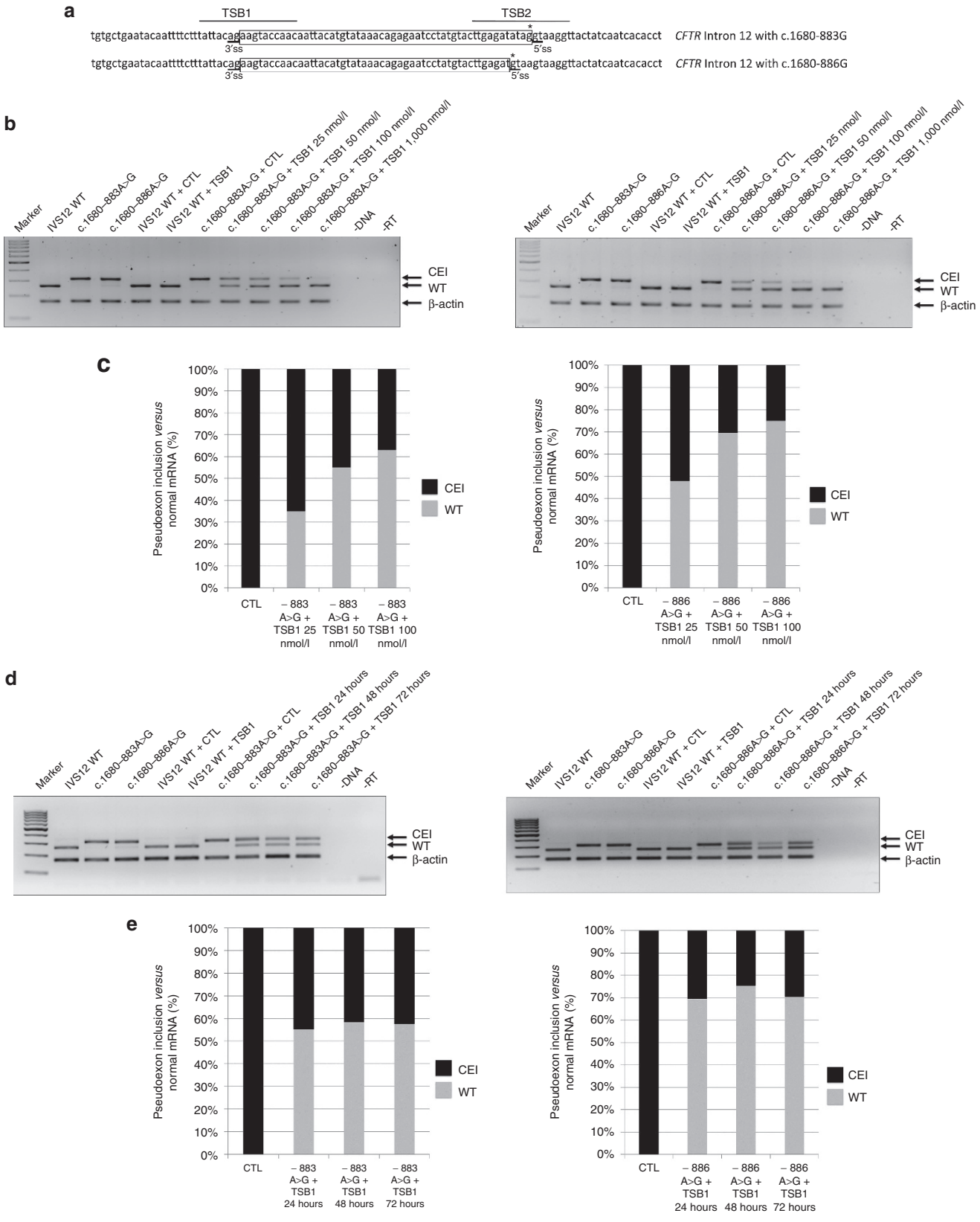
Hybrid capture versus LR-PCR enrichment

To compare the *CFTR* gene coverage generated using the two target enrichment approaches, data were processed using the in-house pipeline (in Galaxy) to visualize all reads for the *CFTR* gene. Globally, these data helped us to verify whether the absence of variants in some regions in this patient was due to sequencing problems (**Figure 1**). First, we observed the presence of large sequencing gaps, particularly with the hybrid capture method, shown in **Figure 1a** (composition of the entire *CFTR* gene is also shown in **Supplementary Figure S1** online). Second, we observed a high peak encompassing exon 10 (arrow in **Figure 1a**) following hybrid capture, but not after LR-PCR enrichment (**Figure 1b**). Analysis of human *CFTR* exon 10 and its flanking regions using the Basic Local Alignment Search Tool revealed a duplication of the region 117,188,495–117,189,477, as previously reported in a study that focused on exon 10 (PEs located on chromosomes 9, 12, and 20).[19] Alignment of reads against the entire genome as reference, rather than against *CFTR* or chromosome 7, improved the problem of alignment of pseudogenes.

Data comparison (**Figure 1a**,**b**) showed that the LR-PCR enrichment method improved the coverage, with more homogeneous coverage despite higher depth at overlapping regions. The poorly covered 4.3-kbp region in intron 3 (chr7:117,165,300–117,169,577; arrow in **Figure 1a**), which was not amplified by the capture method, was correctly amplified by LR-PCR (**Figure 1b**). In addition, exon 15, which

Figure 3 Identification of a putative intronic mutation. (**a**) Schematic representation of the *CFTR* region that contains the c.1680-883A>G mutation showing the pseudoexon (53 bp) included in the patient's sequence, as suggested by in silico analysis. Asterisks indicate the position of the wild-type and mutant nucleotides. The numbers in the white boxes indicate the scores for the corresponding 5′ and 3′ splice sites obtained using different bioinformatics tools. Theoretically, the mutation strengthens a 5′ donor splice site (+12% with Human Splicing Finder, +83.6% with MaxEnt, +58% with Fsplice, and +49% with Automated Splice Site and Exon Definition Analysis), leading to the inclusion of a cryptic exon. (**b**) Phylogenetic analysis of the potential exon cryptic region by Multiz Alignment (UCSC Genome Browser). Nucleotide c.1680-883 is conserved in primates. (**c**) Sequence logo of the mutated region. Among the various species, nucleotide c.1680-883 is less conserved as compared with the juxtaposed donor site. (**d**) Aberrant splicing induced by the newly identified intronic mutation by splicing minigene assay. Wild-type (intron 12 wt) or mutant minigenes (c.1680-883A>G, c.1680-886A>G as positive control, and c.1680-870T>A as negative control) were transiently transfected in bronchial BEAS-2B cells. RNA was isolated and analyzed by reverse transcriptase (RT) polymerase chain reaction (PCR) using minigene-specific primers. PCR products were then analyzed by agarose gel electrophoresis (left panel). The lower band represents the correctly spliced exons, whereas the upper band represents the pseudoexon inserted between the minigene exons. The right panels show the sequence of the splicing products (Sanger method). (**e** and **f**) Aberrant splicing induced by the newly identified intronic mutation by RT-PCR on nasal cells from CF patients and two control samples. PCR amplifications were done with f11-r13, which amplifies (**e**) the normal transcript, or with f11-rPE, which specifically amplified (**f**) the cryptic exon. CEI, cryptic exon inclusion.

was poorly covered by the hybrid capture system (arrow in **Figure 1a**), could be sequenced following LR-PCR enrichment (20× coverage). Using this method, among the five fragments that remained poorly sequenced, regions located in short interspersed nuclear element/long interspersed nuclear element sequences, repeats, or AT-rich motifs were covered by



11–14 reads. For instance, sequencing of region 117,173,740–117,174,800, which contains only 24.5% GC (**Supplementary Figure S1b** online, panel 4), remained difficult, whatever the enrichment method used. In addition, 454 programs failed to correctly align homopolymers, including the well-known IVS9-poly(TG)mT(n) at the 5′ end of exon 10 (5′ end of exon 9 with legacy nomenclature) on the *CFTR* gene. Because we obtained 35,000 reads, at least for the father, using LR-PCR, we deduced that this read number per run is the cutoff to cover the majority of the 250 kb of the *CFTR* locus with 40× coverage; this was based on the family trio study and confirmed in the 17 other CF samples.

Influence of mapping choice in sequence variant identification

To facilitate data processing, we showed only the data from the trio previously studied. Using the sequencing data obtained after LR-PCR amplification, we next checked whether the number of *CFTR* variants depends on the reference used (the *CFTR* gene, chromosome 7, or the entire human genome) (**Figure 2a**). When reads were aligned against *CFTR* as reference, we identified, on average, 325 (patient), 109 (mother), and 257 (father) variants (**Figure 2a**). When we compared the alignments using chromosome 7 or the entire human genome as the reference, the number of identified variants varied. By analyzing the DNA composition of all identified variants, we noted that 82.4% of the variants in the patient's sequence that had been found only when using the *CFTR* reference (not chromosome 7 or the entire human genome) were located in short interspersed nuclear element/long interspersed nuclear element repeats. This suggests that using the *CFTR* gene sequence as the reference forced the alignment of reads located in short interspersed nuclear element/long interspersed nuclear element regions, thus explaining the difference in the variant number found with the three references (**Figure 2a**). We thus decided to use chromosome 7 as the reference for mapping with LR-PCR enrichment. When using a hybrid capture system, human genome mapping must be considered to avoid the alignment of potential PEs. In addition, because most (96%) of the common variants (both after hybrid capture and LR-PCR enrichment) found using alignment against the different references had 30% read support (i.e., for a given variant, 30 of 100 reads included the nucleotide change), we fixed this threshold as a filter of heterozygosity. Thus, variants were considered as heterozygous when the read support was 30–65% and as homozygous when it was >70%. These filters have been validated on the 17 other CF samples.

Sequence variant identification

To facilitate the data and in silico analyses, we generated a workflow based on the family trio studies (**Figure 2b**). This pipeline is partially implemented in our local Galaxy Web-based platform with in-house Perl scripts. Using LR-PCR (and chromosome 7 as the reference for mapping), the detected variants included point mutations and small insertions/deletions. Among the 205 identified variants (**Figure 2c**), 42 were located in the 5′ UTR (117,079,912–117,120,149), 10 in the 3′ UTR (117,307,162–117,332,742), 2 in exons (previously identified by Sanger sequencing as non–disease causing variants), and 151 in introns. The previously reported variants detected with the Sanger method (four variants including p.Phe508del and five variants of repeat sequences corresponding to one polypyrimidine tract and four microsatellites) were all validated, with the exception of two microsatellites and the polypyrimidine tract. Indeed, one significant limitation of pyrosequencing is its apparent inability to correctly determine the number of bases within a homopolymeric stretch.[20] Among the 151 intronic variants identified, 135 had a read support higher than 30%. Seven of these 135 intronic variants were not referenced in the dbSNP site (http://www.ncbi.nlm.nih.gov/SNP), and only four had a read support between 30 and 65% and were thus heterozygous (the heterozygosity filter was fixed between 30 and 65%). In silico analysis was used next to predict the deleterious effect on splicing of each variant unreferenced in the dbSNP site. We then applied this workflow to the other CF samples, identifying only one mutation.

Application of the pipeline to the 18 CF samples

Considering all the CF DNA and before filtering, 197 variants in the *CFTR* gene per sample, on average, were detected, with 118 variants (filter >30%) found in introns. On average, we thus detected 4.6 variants per intron, with lengths ranging from 512 to 28,084 bp (in other words, 0.65 variants/1,000 bp). The pipeline allowed the detection of a potential second CF mutation in 16 samples and was inconclusive for 2. For 2 of 16 samples, the second mutation had not been previously identified because the DNA samples are related to old cases not processed by all conventional strategies, c.2875delG (3007delG, exon 17) and c.870-1113_1110delGAAT (a deep intronic mutation previously

Figure 4 Correction of aberrant splicing in bronchial cells by target site blockers (TSBs). (**a**) *CFTR* pre–messenger RNA fragment that contains the 53-bp pseudoexon harboring the newly identified c.1680-883A>G mutation (top) and the 49-bp pseudoexon harboring the c.1680-886A>G mutation (bottom). TSBs that target the splicing acceptor site (3′ splice site) and splicing donor site (5′ splice site) are above the sequence. (**b** and **d**) Reverse transcriptase polymerase chain reaction (RT-PCR) analysis of total RNA from bronchial BEAS-2B cells was performed using specific primers to analyze splicing after cotransfection of wild-type (IVS12 wt) or mutant (c.1680-883A>G, c.1680-886A>G) minigenes and TSBs. CEI, cryptic exon inclusion; WT, wild-type transcript. (**b**) Effect of TSB concentration on aberrant splicing. Cells were transfected with different TSB1 concentrations (25, 50, 100, and 1,000 nmol/l) for 24 h. (**d**) Effect of incubation time (24, 48, and 72 h) on splicing correction using 50 nmol/l TSB1. TSB1 specificity was confirmed using a control TSB (CTL) at the different concentrations tested and in combination with wild-type and mutant minigenes; to avoid overloading the figure, only the assays at 50 nmol/l or 24 h are shown. (**c** and **e**) Quantification of CEI and WT transcripts after transfection of TSBs at (**c**) different concentrations or at (**e**) different time points after transfection. RT-PCR was performed using a fluorescein amidite (FAM)–labeled forward primer located within the splice donor exon and a reverse primer within the splice acceptor exon of the pSPL3 plasmid. Quantification (noted as a percentage) was performed by dividing the area of the CEI peak by the area of all peaks (wild-type + CEI). Data correspond to the mean value of at least two independent experiments.

described[2]). In addition to the workflow, in silico tests predicted aberrant splicing for an additional 14 samples. A single, new, deep intronic mutation—c.1680-883A>G—was found in three unrelated patients. For one of them, the parents' DNA samples were also available; a familial segregation study was performed, revealing that the c.1680-883A>G intronic variant found in the patient (61%, 129×) was also present in the mother (44%, 108×). In the same family, allelic segregation thus confirmed that the new mutation was in trans of c.1521_1523del (p.Phe508del with 54% of the aligned sequences in the index case carrying this mutation; depth of 24×), as inherited from the father (p.Phe508del found at 35%, 41×).

Identification of a new, putative disease-causing mutation

To explore the effect of c.1680-883A>G (located in intron 12; chromosome location: 117,229,524; hg19), donor (5′ ss) and acceptor site (3′ ss) in silico predictions were generated for the mutated sequence (**Figure 3a**). The mutation generated a new, high-score 5′ ss in intron 12, suggesting that this site could be used for alternative splicing. The newly identified, putative disease-causing mutation c.1680-883A>G is three nucleotides away from a well-known splice mutation (c.1680-886A>G (1811+1,6kbA>G)) that creates a donor site, causing the inclusion of a PE in mature transcripts.

To assess the putative functional importance of the new variant, a large-scale comparison of the orthologous *CFTR* intron 12 region from several mammalian species representing the Primate, Artiodactylia, and Lagomorpha orders (including 46 vertebrates and 9 primates; **Figure 3b**) was carried out using the UCSC Multiz alignment tool. This comparison revealed that nucleotide A at position c.1680-883 was remarkably conserved among primates and poorly conserved in the other species, including placental mammals (mouse, dog, rat, rabbit, pig, cow, sheep, and squirrel) (**Figure 3b,c**). Then the same locus was compared in selected orthologous genes using ClustalW2. The T nucleotide of the cryptic 5′ ss was conserved in all studied species (90%); the G nucleotides of the cryptic 5′ ss and at position c.1680-883 (black arrow in **Figure 3c**) were conserved (54 and 46%, respectively). Based on the results of the in silico analysis, this nucleotide variant could be considered as a disease-causing mutation through the inclusion of a PE. In other respects, the mutation was tested in 200 control chromosomes by Sanger sequencing analysis and was not found.

Confirmation of the c.1680-883A>G intronic mutation using a splicing reporter assay and in nasal cells of a patient

We next used a splicing reporter assay to test the impact of the c.1680-883A>G variant on splicing (**Figure 3d**). When BEAS-2B cells were transfected with the minigene carrying the c.1680-886A>G variant (used as a positive control), which causes the inclusion of a 49-bp intronic sequence (cryptic exon inclusion), aberrantly spliced transcripts were ~90–95% of the total *CFTR* products (wild type and aberrantly spliced). Conversely, transfection of the minigene carrying the neutral variant c.1680-870T>A (negative control) did not have any effect on splicing. Finally, transfection of the minigene carrying the newly identified c.1680-883A>G mutation led to activation of a PE, resulting in the inclusion of an additional 53-bp sequence, as shown by Sanger sequencing, and a complete loss of the wild-type CFTR products.

We next checked whether this mutation retained the sequence in nasal cells from the CF patient included in the family trio analysis. Thus, we confirmed that, compared with controls, the patient harbored a 53-bp PE inclusion in intron 12 by PCR amplification using nonspecific and specific primers pairs to the PE (f11-r13 and f11-rPE, respectively) (**Figure 3e,f**).

Correction of CFTR aberrant splicing by using TSBs

We designed antisense oligonucleotides (TSB1 and TSB2) that block access to the 3′ ss (acceptor site) and 5′ ss (donor site), respectively (**Figure 4a**), in order to correct aberrant splicing caused by the c.1680-883A>G and c.1680-886A>G mutations. To determine the effect of TSB concentration on aberrant splicing, human bronchial BEAS-2B cells were cotransfected with the minigenes harboring the two mutations and four different TSB concentrations (25, 50, and 100 nmol/l, and 1 μmol/l) for 24 h. At a low concentration (50 nmol/l), TSB1, which targets the 3′ ss (acceptor site), had a marked corrective effect on aberrant splicing caused by the c.1680-883A>G and

Table 1 In silico predictions for the 10 new putative disease-causing mutations

	Software				
Variants	MaxEnt	HSF	NNSplice	ASSEDA	Fsplice
c.53+3158A>G					
c.274-2354A>C					
c.1209+2330A>G					
c.1393-2734G>A					
c.1393-2883G>A					
c.1393-2921G>A					
c.1680-883A>G					
c.2989-313A>T					
c.3874-4522A>G					
c.3469-1304C>G					

Gray cells indicate noted variants with deleterious impact predicted by different programs.

ASSEDA, Automated Splice Site and Exon Definition Analysis; HSF, Human Splicing Finder.

c.1680-886A>G mutations (**Figure 4b**, upper and lower panel, respectively). TSB1 specificity was confirmed using a TSB control. The efficiency of wild-type splicing restoration was quantified by fragment analysis PCR (**Figure 4c**). TSB2 required a higher concentration to act on splicing (**Supplementary Figure S2a** online). We next performed time course experiments by transfecting 50 nmol/l TSB1 and harvesting cells after 24, 48, and 72 h. A marked effect was evident after 24 h (**Figure 4d**). Specifically, quantification showed that the percentage of aberrantly spliced transcripts (containing the cryptic exon) was reduced to 45% (c.1680-883A>G) and to 30% (c.1680-886A>G) of the total *CFTR* messenger RNA (mRNA) (wild-type and aberrantly spliced transcripts). Thus, transfection of 50 nmol/l TSB1 for 24 h induced a restoration of 55 and 70% of normal *CFTR* mRNA, respectively (**Figure 4e**). Finally, we assessed the duration of action of both TSBs in BEAS-2B cells and found that they had a strong effect on splicing up to 72 h after washing off the transfection medium (16 h of incubation) (**Supplementary Figure S2b** online). Partial restoration of correctly spliced *CFTR* mRNA induced by TSB1 (24 h at 100 nmol/l) was confirmed in primary nasal cultures obtained from a control individual (**Supplementary Figure S2c** online).

DISCUSSION

Although the majority of disease-causing mutations are typically found in the coding region or in canonical ss, a number of mutations also occur in noncoding regions. The value of NGS for whole-genome or exome sequencing for the identification of mutations in common and rare disease is now recognized.[5,6] In the case of monogenic Mendelian diseases caused by mutations in a single, well-defined gene, however, sequencing this single gene, rather than the entire genome or exome, could be sufficient and also less expensive. Indeed, the combination of target enrichment and NGS now offers the possibility to perform such analyses in a time-efficient and economical way. In CF, great efforts to identify new modifier genes[21–23] that influence disease severity have been made, but analysis of *CFTR* intronic sequences has been neglected.

Comprehensive sequence information on the entire gene locus can help identify a disease-causing mutation and avoid interpretation errors due to repeated sequences, PEs,[24] and pseudo-mutations.[19] Recent publications reported *CFTR* resequencing by targeting exon and intron flanking sequences using the IonTorrent sequencer or by exploring the complete gene (custom NimbleGen SeqCap EZ Choice array) using HiSeq2000 in a selected set of 92 DNA samples; however, these studies focused only on previously characterized *CFTR* mutations and polymorphisms.[9] No data about the robustness of *CFTR* resequencing for finding new mutations in yet unexplored regions have been published.

As proof of concept, we sequenced 18 samples from CF patients with a single known CF-causing mutation. We blindly applied the automated "in-house" pipeline implemented in Galaxy (prioritization strategy) that allowed the rapid identification of previously detected mutations (by conventional approaches). This prioritization strategy is based on a comparison of the data obtained by hybrid capture or LR-PCR enrichment, and the read mapping against *CFTR* and chromosome 7, and application of a heterozygosity filter (30–65%), familial segregation (when possible), and dbSNP interrogation. The pipeline first maps the sequence being studied against chromosome 7 as a reference. Sequence analysis indicates that enrichment by multiple LR-PCR amplification (when possible) was better than hybrid capture for resequencing the complete gene. We found a putative second mutation for 16 of 18 tested DNA samples, including the new c.1680-883A>G mutation, identified in 3 unrelated patients. For the remaining samples, functional studies to confirm the deleterious effect of the variants predicted by bioinformatics tools (**Table 1**) are still in process. When no new mutation is found, the pipeline offers the possibility of obtaining more variants by mapping against the *CFTR* reference. The next step could then be the search for variants in the 5′ and 3′ UTR regions, which contain regulatory elements,[25] among the unreferenced variants in dbSNP. Indeed, a recent work described a 3′ UTR variant associated with CFTR-related disorders.[26] A remaining challenge is to define the real impact of variants in noncoding regions. Minigene splicing assay is the classic tool for determining the product generated by splicing variants. When nasal cells are not available, expression minigenes represent a real opportunity to confirm the RNA and protein products generated by splicing variants.[16] Overall, creating a database compiling information obtained from high-throughput sequencing can improve the gap in our ability to interpret the clinical relevance of genomic variations, as has previously been done for well-known mutations.[27]

Here, functional analysis, minigene splicing assay, and PCR on nasal cells from a CF patient carrying c.1680-883A>G showed that this deep intronic mutation generated a new, high-score 5′ ss (donor site) in intron 12 that is involved in PE inclusion. Among the 1,976 reported *CFTR* mutations, 228 (11.54%) are believed to affect pre-mRNA splicing (http://www.genet.sickkids.on.ca/). Most splicing mutations disrupt the canonical ss sequences, completely abolishing exon recognition and/or leading to a nearly complete absence of correctly spliced transcripts. The new c.1680-883A>G mutation strengthens the use of a cryptic donor site. Despite the abundance in the genome of potential PEs (intronic sequences between 50 and 200),[1,3,28,29] their inclusion does not seem to be a frequent event during normal pre-mRNA processing. The c.1680-883A>G mutation neither creates a splice donor site nor is well conserved, so only functional tests can confirm its deleterious impact. Functional assessment of deep intronic variants may improve our knowledge on the usage of poorly conserved cryptic 5′ or 3′ ss in eukaryote evolution and/or may improve the accuracy of splicing algorithms for these intronic variants. Surprisingly, this mutation is close to another previously identified deep intronic mutation, c.1680-886A>G, which also induces PE inclusion, suggesting that this intronic region may be prone to mutation. The c.1680-886A>G mutation occurs with a frequency

of 0.2% in France.[30] Conversely, c.1680-883A>G has never been described before, although here it was identified in three unrelated patients.

The final objective of this work was the design of antisense oligonucleotides for CF treatment. Antisense oligonucleotides for inherited diseases, including Duchenne muscular dystrophy, have been used for several years.[31] Indeed, PE exclusion by antisense modification of pre-mRNA splicing represents a type of personalized genetic medicine. The development of oligonucleotides that block access to a target site (TSBs) offers new treatment opportunities for other genetic disorders.[32,33] Here, we used this approach to correct the aberrant splicing caused by deep intronic mutations in the *CFTR* gene (c.1680-883A>G and c.1680-886A>G). The effect of TSBs on aberrant splicing correction in bronchial BEAS-2B cells was rapid and maintained over time, suggesting that TSBs could be a therapeutic tool in patients with CF who have deep intronic mutations in the *CFTR* gene because TSBs restore normal transcripts. These data are particularly interesting for patients with CF because the c.1680-886A>G mutation is the fourth most frequent in southwestern Europe (3.4%), and the threshold of functional mRNA and, subsequently, of CFTR protein required for normal functions is very low, having been estimated at 5%.[34] It would be interesting to test, if possible, these TSBs in airway cells from patients with CF harboring both mutations tested in this work and to assess TSBs for other intronic splicing mutations in *CFTR*.

To conclude, these data provide proof that small-scale NGS approaches are suitable for the identification of new mutations in noncoding regions and that TSBs can be envisaged for personalized therapies for CF patients.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at http://www.nature.com/gim

ACKNOWLEDGMENTS

This work was supported by grants from the French association Vaincre La Mucoviscidose Agence de la Biomedecine, the CHU, and INSERM. We thank Fanny Verneau and Jean-Pierre Altieri (Montpellier, France) for technical assistance.

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

1. Faà V, Incani F, Meloni A, et al. Characterization of a disease-associated mutation affecting a putative splicing regulatory element in intron 6b of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. *J Biol Chem* 2009;284:30024–30031.
2. Costa C, Pruliere-Escabasse V, de Becdelievre A, et al. A recurrent deep-intronic splicing CF mutation emphasizes the importance of mRNA studies in clinical practice. *J Cyst Fibros* 2011;10:479–482.
3. Chillón M, Dörk T, Casals T, 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* 1995;56:623–629.
4. Highsmith WE, Burch LH, Zhou Z, et al. A novel mutation in the cystic fibrosis gene in patients with pulmonary disease but normal sweat chloride concentrations. *N Engl J Med* 1994;331:974–980.
5. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009;461:272–276.
6. Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci USA* 2009;106:19096–19101.
7. Mardis ER. A decade's perspective on DNA sequencing technology. *Nature* 2011;470:198–203.
8. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010;11:31–46.
9. Trujillano D, Ramos MD, González J, et al. Next generation diagnostics of cystic fibrosis and CFTR-related disorders by targeted multiplex high-coverage resequencing of CFTR. *J Med Genet* 2013;50:455–462.
10. Abou Tayoun AN, Tunkey CD, Pugh TJ, et al. A comprehensive assay for CFTR mutational analysis using next-generation sequencing. *Clin Chem* 2013;59:1481–1488.
11. Raynal C, Baux D, Theze C, et al. A classification model relative to splicing for variants of unknown clinical significance: application to the CFTR gene. *Hum Mutat* 2013;34:774–784.
12. Desmet FO, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 2009;37:e67.
13. Mucaki EJ, Shirley BC, Rogan PK. Prediction of mutant mRNA splice isoforms by information theory-based exon definition. *Hum Mutat* 2013;34:557–565.
14. Houdayer C, Dehainault C, Mattler C, et al. Evaluation of in silico splice tools for decision-making in molecular diagnosis. *Hum Mutat* 2008;29:975–982.
15. Le Guédard-Méreuze S, Vaché C, Baux D, et al. Ex vivo splicing assays of mutations at noncanonical positions of splice sites in USHER genes. *Hum Mutat* 2010;31:347–355.
16. Sharma N, Sosnay PR, Ramalho AS, et al. Experimental assessment of splicing variants using expression minigenes and comparison with in silico predictions. *Hum Mutat* 2014;35:1249–1259.
17. Lopez E, Viart V, Guittard C, et al. Variants in CFTR untranslated regions are associated with congenital bilateral absence of the vas deferens. *J Med Genet* 2011;48:152–159.
18. René C, Lopez E, Claustres M, Taulan M, Romey-Chatelain MC. NF-E2-related factor 2, a key inducer of antioxidant defenses, negatively regulates the CFTR transcription. *Cell Mol Life Sci* 2010;67:2297–2309.
19. El-Seedy A, Dudognon T, Bilan F, et al. Influence of the duplication of CFTR exon 9 and its flanking sequences on diagnosis of cystic fibrosis mutations. *J Mol Diagn* 2009;11:488–493.
20. Brockman W, Alvarez P, Young S, et al. Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res* 2008;18:763–770.
21. Wright FA, Strug LJ, Doshi VK, et al. Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13.2. *Nat Genet* 2011;43:539–546.
22. Gu Y, Harley IT, Henderson LB, et al. Identification of IFRD1 as a modifier gene for cystic fibrosis lung disease. *Nature* 2009;458:1039–1042.
23. Gallati S. Disease-modifying genes and monogenic disorders: experience in cystic fibrosis. *Appl Clin Genet* 2014;7:133–146.
24. Pagani F, Buratti E, Stuani C, Bendix R, Dörk T, Baralle FE. A new type of mutation causes a splicing defect in ATM. *Nat Genet* 2002;30:426–429.
25. Zhang Z, Ott CJ, Lewandowska MA, Leir SH, Harris A. Molecular mechanisms controlling CFTR gene expression in the airway. *J Cell Mol Med* 2012;16:1321–1330.
26. Amato F, Seia M, Giordano S, et al. Gene mutation in microRNA target sites of CFTR gene: a novel pathogenetic mechanism in cystic fibrosis? *PLoS One* 2013;8:e60448.
27. Sosnay PR, Siklosi KR, Van Goor F, et al. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat Genet* 2013;45:1160–1167.

28. Vaché C, Besnard T, le Berre P, et al. Usher syndrome type 2 caused by activation of an USH2A pseudoexon: implications for diagnosis and therapy. *Hum Mutat* 2012;33:104–108.
29. Cavalieri S, Pozzi E, Gatti RA, Brusco A. Deep-intronic ATM mutation detected by genomic resequencing and corrected in vitro by antisense morpholino oligonucleotide (AMO). *Eur J Hum Genet* 2013;21:774–778.
30. Federici S, Iron A, Reboul MP, et al. [CFTR gene analyis in 207 patients with cystic fibrosis in southwest France: high frequency of N1303K and 1811+1.6bA>G mutations]. *Arch Pediatr* 2001;8:150–157.
31. Douglas AG, Wood MJ. Splicing therapy for neuromuscular disease. *Mol Cell Neurosci* 2013;56:169–185.
32. Webb TR, Parfitt DA, Gardner JC, et al. Deep intronic mutation in OFD1, identified by targeted genomic next-generation sequencing, causes a severe form of X-linked retinitis pigmentosa (RP23). *Hum Mol Genet* 2012;21: 3647–3654.
33. Nuzzo F, Radu C, Baralle M, et al. Antisense-based RNA therapy of factor V deficiency: in vitro and ex vivo rescue of a F5 deep-intronic splicing mutation. *Blood* 2013;122:3825–3831.
34. Ramalho AS, Beck S, Meyer M, Penque D, Cutting GR, Amaral MD. Five percent of normal cystic fibrosis transmembrane conductance regulator mRNA ameliorates the severity of pulmonary disease in cystic fibrosis. *Am J Respir Cell Mol Biol* 2002;27:619–627.