New Molecular Diagnosis Approaches - From the Identification of Mutations to their Characterization
Anne Bergougnoux, Magali Taulan-Cadars, Mireille Claustres, Caroline Raynal

To cite this version:
Anne Bergougnoux, Magali Taulan-Cadars, Mireille Claustres, Caroline Raynal. New Molecular Diagnosis Approaches - From the Identification of Mutations to their Characterization. Cystic Fibrosis in the Light of New Research, InTech, 2015, 10.5772/60679 . hal-02434957

HAL Id: hal-02434957
https://hal.umontpellier.fr/hal-02434957
Submitted on 4 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
We are IntechOpen, the world’s leading publisher of Open Access books
Built by scientists, for scientists

4,500
Open access books available

119,000
International authors and editors

135M
Downloads

154
Countries delivered to

TOP 1%
Our authors are among the most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™
Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com
New Molecular Diagnosis Approaches — From the Identification of Mutations to their Characterization

Anne Bergougnoux, Magali Taulan-Cadars, Mireille Claustres and Caroline Raynal

Abstract

Molecular diagnosis of cystic fibrosis is based on the detection of mutation in the \textit{CFTR} gene, identified in 1989. During the past 20 years, thanks to evolutions of diagnostic techniques, our knowledge of mutation spectrum and pathophysiological mechanisms involved in the disease has significantly improved. Sanger sequencing and quantitative methods greatly contributed to the identification of the 2,000 sequence variations reported worldwide in \textit{CFTR}. We are now entering the new technological age with the generalisation of Next Generation Sequencing (NGS) technologies in diagnostics laboratories. These high throughput approaches allow scanning for the entire \textit{CFTR} locus, including deep intronic regions, and in parallel other candidate genes that possibly influence the clinical evolution of patients. However, this powerful technology poses new challenge in test interpretation. In this chapter, we review the current and new technologies used in molecular diagnostics of cystic fibrosis, particularly NGS approaches. We also present current and new bioinformatics tools available for the interpretation of variants and \textit{in vitro/ex vivo} and \textit{in vivo} techniques that can be used to improve the characterization of the functional impact of \textit{CFTR} variations.

\textbf{Keywords:} Next generation sequencing, \textit{CFTR}, sequence variations, interpretation, functional characterization
1. Introduction

The autosomal recessive mode of inheritance of Cystic Fibrosis (CF) was suggested by Andersen and Hodges in 1945 when they described this disease as a pathological entity.

Since the discovery of the CFTR gene (and the predicted protein cystic fibrosis transmembrane conductance regulator) in 1989 [1], close to 2,000 variations have been identified on this locus. CFTR gene studies represent one of the most frequent genetic analyses routinely performed worldwide, either to confirm the clinical diagnosis of CF or CFTR-related disorders (CFTR-RDs), or to offer carrier testing, prenatal or pre-implantation genetic diagnosis.

The most common severe mutation, p.Phe508del, is found in approximately 70% of CF alleles of European descent (therefore it is present in 49% of homozygous and 42% of compound heterozygous CF patients), with significant variations depending on ethnicities [2]. There is a clear decreasing northwest to southwest gradient in p.Phe508del frequency across Europe. Only four other mutations represent more than 1% of CF cases: p.Gly542*, p.Gly551Asp, p.Asn1303Lys, and p.Trp1282*. All other mutations are rare and many are private, only detected within a single family.

![Figure 1](image_url)

Figure 1. Two models for the classification of CFTR variants: (1) Functional: five classes of defective CFTR protein [3, 4, 5]; and (2) Clinical: four classes based on phenotypic expression in patients [6].
Therefore, a good knowledge of CFTR diseases and their molecular pathology is required when choosing tools and strategy and when interpreting results. Organization of regional or national networks of specialist clinical and molecular genetic laboratories is needed. Two models for the classification of sequence alterations have been proposed (Figure 1): (1) depending on their molecular impact on the CFTR chloride channel, which requires \textit{in vitro/in vivo} functional assays, now less used due to the difficulty in accurately classifying the new reported variants [3, 4] or (2) by their clinical consequences on patients based on detailed and prospective clinical evaluation of patients associated with molecular findings [5].

However, even if the most severe mutations were studied \textit{in vitro}, mild mutations or variants of unknown (clinical) significance (VUCs or VUS) were barely analysed. Thus, only one third of \textit{CFTR} variations are functionally classified. Finally, in CF, as in other genetic diseases, new molecular diagnosis techniques such as Next Generation Sequencing (NGS) allow the analysis of the whole \textit{CFTR} locus (including deep intronic regions) but increase the number of VUCS reported on patients.

Therefore, geneticists and clinicians have to combine CFTR epidemiological databases, functional \textit{in vitro/in vivo} analysis and exhaustive clinical data to perform pertinent genotype/phenotype correlations in patients in order to provide appropriate genetic counselling to families.

2. Technical aspects of molecular diagnosis

The \textit{CFTR} genetic testing strategy depends on the clinical and familial context and is classically performed step by step, as recommended [7]. As shown in Figure 2, tests are carried out in various situations, including confirmation of a clinical diagnosis, prenatal diagnosis (familial context or foetal suspicion of CF) and carrier screening.

A wide range of techniques is still used to identify \textit{CFTR} gene sequence variations (presented in Table 1) and there is no gold standard or preferred method for routine testing. However, laboratories should be aware of the limitations of their chosen method (e.g. some mutations are not identified). Moreover, assay performance should always be verified before diagnostic use, even though commercial kits are CE-marked \textit{in vitro} diagnostic devices (IVDD). In the perspective of accreditation (expected in France in 2020 for hospital diagnostic laboratories), all methods used in molecular biology should be validated before diagnostic use, notably by comparison with a reference method (Sanger sequencing for Single Nucleotide Variations and quantitative PCR for Copy Number Variations detection). This technical validation should be undertaken for NGS in the step 1 and 2 diagnosis.

Undetected \textit{CFTR} mutations may lie within the introns or regulatory regions, which are not routinely explored but that will be soon analysed by NGS. Comparison between NGS and reference methods for the identification of \textit{CFTR} variants on these unexplored regions will be a challenge.
2.1. Current techniques to detect CFTR variants

2.1.1. Screening techniques (Step 1)

These approaches consist of genotyping a panel of frequent CF mutations using commercial kits (Table 1) that classically cover more than 80% of CF known mutations in European populations. Additional search for mutations specific to certain regions or ethnicities (frequency higher than 1% of CF alleles in the targeted population) completes the analyses.

Data on disease and carrier frequencies or mutation frequencies in various populations are available in the WHO report [8] and should be accurately known and used by laboratories.

For many patients carrying CF-causing mutations included in commercial panels, CFTR molecular analysis generally stops at this step. There is no need for additional studies, except the confirmation of mutations by a second method, as recommended by international guidelines for genetics diagnosis.

2.1.2. Scanning techniques (Step 2)

The high heterogeneity of CFTR mutations in CF and CFTR-RD populations makes the complete molecular screening of the 27 exons and parts of the regulatory regions (5’UTR, 3’UTR and partial intronic regions) essential.

Therefore, the analysis of the CFTR locus can be performed as follows:
To detect SNV: scanning the 27 exons of the CFTR gene, intronic boundaries and four intronic regions with reported deep intronic mutations [9-12] and a part of the promoter. As an example, one can use Sanger sequencing with Single Condition Amplification Primers (SCAP) technique [13].

To detect CNV (large CFTR rearrangements): scanning of exon by multiplex fluorescent quantitative PCR (Multiplex Ligation-dependent Probe Amplification or MLPA) or array-CGH [14].

These robust methods allow the detection of more than 97% of CFTR mutations involved in CF.

2.1.3. Advantages and limits

Classical screening or scanning methods used to detect mutations (in step 1 and 2) have high specificity and sensitivity (Table 1). Older techniques such as RFLP (restriction fragment length polymorphism), DGGE (denaturing gradient gel electrophoresis) or DHPLC (denaturing high performance liquid chromatography) were widely used in the past 20 years but were difficult to set up and showed inconstant performance. They have been progressively replaced by easy-to-use commercial kits, more appropriate for larger sample series. These kits allow PCR multiplexing and thus enable the detection of the most common CF mutations in one reaction. Sanger sequencing and quantitative PCR are the most effective to identify unknown mutations. However, all these classical methods are labour-intensive and time-consuming.

<table>
<thead>
<tr>
<th>Method name</th>
<th>Mutations detected</th>
<th>Advantages</th>
<th>Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reverse dot blot hybridization</td>
<td>Up to 20 mutations per multiplex</td>
<td>Partial hybridization if SNP on probe</td>
<td>Primer Design difficult</td>
</tr>
<tr>
<td>ARMS (amplification refractory mutation system)</td>
<td>Up to 50 mutations</td>
<td>Appropriate for large series</td>
<td>Need of high agarose gel quality</td>
</tr>
<tr>
<td>OLA (Oligonucleotide ligation assay)</td>
<td>32 mutations (Abbott kit)</td>
<td>High specificity</td>
<td>Partial hybridization if SNP on probe</td>
</tr>
<tr>
<td>NGS (Next Generation Sequencing)</td>
<td>139 to 166 CFTR variants</td>
<td>Cost effective</td>
<td>Need of sufficient coverage</td>
</tr>
<tr>
<td>HRM (High Resolution Melting curve analysis)</td>
<td>Heterozygous variants</td>
<td>Appropriate for large series (e.g. epidemiological studies)</td>
<td>Exon by exon</td>
</tr>
<tr>
<td>Sanger Sequencing</td>
<td>All sequenced bases</td>
<td>Close to 100% sensitivity</td>
<td>Expensive method</td>
</tr>
<tr>
<td>Quantitative fluorescent multiplex PCR</td>
<td>Deletions, duplications, insertions</td>
<td>Simple and rapid</td>
<td>False positive if SNP on probe</td>
</tr>
<tr>
<td>NGS (used for step 3)</td>
<td>All sequenced bases</td>
<td>Appropriate for large series</td>
<td>Need of sufficient coverage</td>
</tr>
</tbody>
</table>

Table 1. Overview, advantages and limits of screening and scanning techniques used in molecular genetics diagnosis (adapted from [7]).
2.2. Next Generation Sequencing (NGS)

2.2.1. Principle of NGS technologies

The development of Next Generation Sequencing (NGS) or Massively Parallel Sequencing (MPS) technologies with an immense capacity (up to 1 Terabase (Tb) of data per run with the HiSeq2500 system, Illumina) is a major technical progress in the field of human genetics. Since 2004, three principal NGS platforms have been commercially available, including (i) 454 GS FLX & GS junior from Roche (ii) Genome Analyzer, HiSeq & MiSeq from Illumina and (iii) SOLiD & Ion Torrent PGM from Life Technologies-Applied Biosystems (AB) [15]. These technologies differ in terms of sample library preparation workflow, enabling sequencing on any of the current NGS sequencers (e.g. Illumina, Life technologies, Roche). Since late 2004, three principal NGS firms developed sequencers commercially available (listed in Table 2).

454 pyrosequencing is a sequencing-by-synthesis method that measures the release of inorganic pyrophosphate upon incorporation of nucleotides, by converting it into luciferase chemiluminescent signals using a series of enzymatic reactions. Ion Torrent semiconductor technology is also based on a sequencing-by-synthesis approach but measures pH changes (instead of light) induced by the release of hydrogen ions as nucleotides are incorporated. SOLiD technology is a ligation-based sequencing system. DNA ligase is used to identify the nucleotide present at a given position in a DNA sequence; each base is read twice, which increases accuracy, even for homopolymeric regions. Base detection uses a mixture of labelled oligonucleotides, which queries the input strand with ligase. In Illumina system, clonal amplification is performed using a process termed ‘bridge amplification’ followed by two basic steps, initial priming and extending of the single-stranded, single-molecule template and bridge amplification of the immobilized template with immediately adjacent primers to form clusters. For sequencing, only dye-labelled terminators are added; then the sequence at that position is determined for all clusters; next, the dye is cleaved and another round of dye-labelled terminators is added.

<table>
<thead>
<tr>
<th>Platform</th>
<th>GAII, HiSeq &amp; MiSeq</th>
<th>Ion Torrent</th>
<th>SOLiD</th>
<th>GS FLX &amp; GS Junior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methodology</td>
<td>Sequencing by synthesis</td>
<td>Sequencing by synthesis</td>
<td>Sequencing by Ligation</td>
<td>Sequencing by synthesis</td>
</tr>
<tr>
<td></td>
<td>(Reversible termination)</td>
<td></td>
<td></td>
<td>(Pyrosequencing)</td>
</tr>
<tr>
<td>Loading</td>
<td>Adaptors on template DNA</td>
<td>Adaptors on template DNA</td>
<td>Adaptors on template DNA</td>
<td>Adaptors on template DNA</td>
</tr>
<tr>
<td></td>
<td>bind high density primers</td>
<td>bind primers on beads, one</td>
<td>bind primers on beads, one</td>
<td>bind primers on beads, one</td>
</tr>
<tr>
<td></td>
<td>across surface of slide</td>
<td>molecule per bead</td>
<td>molecule per bead</td>
<td>molecule per bead</td>
</tr>
<tr>
<td>Clonal amplification</td>
<td>Bridge PCR: Surface array on flow cell</td>
<td>Emulsion PCR: clusters on beads</td>
<td>Emulsion PCR: clusters on beads</td>
<td>Emulsion PCR: clusters on beads</td>
</tr>
<tr>
<td>Parallelisation</td>
<td>Random array on flow cell</td>
<td>Beads loaded on a chip</td>
<td>Beads bonded to high density glass slide</td>
<td>Beads loaded onto high density plate</td>
</tr>
<tr>
<td>Detection</td>
<td>Fluorescence</td>
<td>Fluorescence</td>
<td>Light (luciferase)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H+ ions (sensitive pH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>meter)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Overview of the main NGS technologies
2.2.1.1. Target enrichment

Targeted re-sequencing isolates genomic regions of interest in a sample library, allowing to focus efficiently and cost-effectively on a small subset of the genome, such as an exome, a particular chromosome, a set of genes or a region of interest such as a whole gene. Two main strategies can be envisioned: capture [16, 17] or amplification relevant genomic DNA [18-20] as shown in Table 3.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Capture/hybridisation-based method</th>
<th>Amplicon-based method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>HaloPlex</td>
<td>SureSelect</td>
</tr>
<tr>
<td></td>
<td>(Agilent)/SeqCap Nester (Illumina)</td>
<td>(NimbleGen)</td>
</tr>
<tr>
<td></td>
<td>Microdoplet</td>
<td>Multiplex PCR</td>
</tr>
<tr>
<td></td>
<td>(RainDance Technologies)</td>
<td>(AmpliSeq, Life</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Technology &amp; TrueSeq</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&amp; Qiagen &amp; &quot;home&quot;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>design</td>
</tr>
<tr>
<td>Main steps</td>
<td>Restriction enzyme digestion</td>
<td>Sonication/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>nebulisation</td>
</tr>
<tr>
<td></td>
<td>Tagmentation with transposon</td>
<td>PCR amplification</td>
</tr>
<tr>
<td></td>
<td>PCR amplification</td>
<td>PCR amplification</td>
</tr>
<tr>
<td></td>
<td>+ probe</td>
<td>+ probe</td>
</tr>
<tr>
<td></td>
<td>hybridisation</td>
<td>hybridisation</td>
</tr>
<tr>
<td></td>
<td>Probe hybridisation</td>
<td>Sonication/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>nebulisation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ligation of</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sequencing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>adaptors</td>
</tr>
</tbody>
</table>

Table 3. Main strategies proposed for target enrichment

2.2.1.2. Advantages and limits

The main advantages of the technology are related to its capability to process a large number of samples in parallel. NGS technologies are time saving and lower the costs per patient, of step 1 and particularly step 2 molecular analyses. But, they also have significant limitations such as high error rates, enrichment of rare variants and large proportion of missing values, as well as the fact that most current analytical methods are designed for population-based association studies. With second generation sequencing, it is necessary to clonally amplify the isolated targets in order to generate sufficient signal for detection during the sequencing run generating clusters of many thousands of identical DNA targets. In addition, each NGS platform generates different read lengths that range from short (e.g. 35 bases) to long reads (over 500 bases). For a number of applications, including targeted re-sequencing, ChIP-Seq and RNA-Seq, short reads are highly informative and adequate. Conversely, longer reads are more suitable for de novo genome assembly, mapping of high homology regions (related gene family and pseudogenes) and sequencing of repetitive DNA regions, such as introns. This is an important consideration since short read length can make accurate assembly and alignment computationally challenging.

For some clinical applications, as NGS produces massive amounts of data, their analysis and interpretation are time-consuming, not trivial and a real challenge even if specific portions of a genome is analysed.
2.2.1.3. Bioinformatics software to analyse NGS data

Then, generated data are analysed by bioinformatics tools. Quality criteria such as confidence or coverage are guarantees for optimal sensitivity and specificity of a sequencing run. Bioinformatics analyses are frequently performed by using software developed by sequencers’ companies. The three steps include (1) base calling and quality score computation, (2) assembly and alignment and (3) variant calling and annotation. Laboratories developed their own ‘in-house’ pipeline mainly to apply filters allowing an easy focus on causing-mutations. Others use free web resources to realign files including BWA (http://bio-bwa.sourceforge.net) or to perform a new variant calling such as Samtools and GATK [21-23]. Others serve as Viewer such as Integrative Genomics Viewer (http://www.broadinstitute.org/igv/). Galaxy website is an integrative platform offering the possibility to view and process files generated by NGS. Many databases, useful for variant annotation and sometimes inserted in others programs or pipelines, are publicly available including the 1000 Genomes Project [24] and the dbSNP database [25] (further described in part 3.1.1).

Many tracks are currently studied (i.e. improvement of bioinformatics tools, comparison of NGS approaches between CF laboratories, development of databases including newly detected NGS variants, publication of guidelines and definition of a diagnosis report model) to make CF molecular diagnosis by NGS suitable for different clinical and familial cases.

2.2.2. Indications and choice of CFTR analysed regions

In the case of Mendelian diseases caused by mutations in a single gene, like CFTR in Cystic Fibrosis, NGS sequencing of entire genome or exome is still useless and expensive. Collaborations between companies specialized in molecular diagnosis and academic laboratories recently led to the establishment of new molecular diagnosis tools based on NGS. Combination of (i) the enrichment of regions of interest by hybrid capture, circularization or Polymerase Chain Reaction (PCR) and (ii) high throughput sequencing now allows time-efficient and economical way to perform analyses. Use of NGS technology as a first intention is currently set up in laboratories, which will soon question the CF diagnosis strategy as a tree. As shown in Figure 2, in our actual CF molecular diagnosis strategy, NGS technology can be used instead of the ‘CFTR exome’ (step 2). Technical manipulations are similar for both, but filters can be used to focus on regions that contain mutations. Multiplicom® and Illumina® propose a locus specific design to library preparation for molecular analysis of the ‘CFTR exome’. CE-marked kits for in vitro diagnostic (CE-IVD) and efficient bioinformatics tools are commercialized. The latter could be performed by in-house pipelines or subcontract on commercial firms (Sophia Genetics®). SNV and indels are correctly detected (high specificity and sensitivity) and CNV detection is now available for some amplicon-based methods.

Furthermore, in patients with definite CF clinical diagnosis (positive sweat test, CF clinical features) and who carry no or only one CF mutation, step 3 analysis could be ultimately proposed, in combination with the analysis of potential modifiers genes (see next section)

In literature, three studies reported NGS CFTR sequencing (Table 4) on CF patients, CF carriers or controls samples [18, 26, 27]. Abou Tayoun and colleagues [26] first proposed a proof-of-
concept for a ‘CFTR exome’ analysis by NGS on 79 samples. Target enrichment was performed by PCR amplification (AmpliSeq Panel, Life technologies) and Sequencing on Ion Torrent Platform (PGM®). Their sequencing offered minimal coverage of 100X (depending on the Ion 314 or 318 chip used). Two others studies realized sequencing of the whole CFTR locus (close to 250 kb) including deep intronic CFTR regions (Figure 2, Step 3). Trujillano et al. [27] reported the CFTR re-sequencing by hybridization capture on a custom NimbleGen SeqCap EZ Choice array using HiSeq2000 (Illumina®) in a set of 92 samples. They highlighted the precise characterization of breakpoints of seven genomic rearrangements in CFTR.

We proposed a complete CFTR gene sequencing of DNA samples from patients with a confirmed CF clinical diagnosis but with an incomplete genotype [18]. Although large unexplored intronic regions might contain few mutations (about 1%–3% of CF mutations), we identified a new pathogenic mutation, which creates a pseudo-exon (Table 4). Moreover, we compared hybridization capture and Long-Range PCR to target enrichment and used a small-scale NGS platform for sequencing (GS Junior Sequencer, 454 Life Sciences®). Some promising variants were then confirmed as deleterious by in vitro/ex vivo functional assays. However, for most detected intronic variants, classification will be a long and difficult way. This approach is currently under development for CF diagnosis.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Study objecif</td>
<td>Proof of concept of CFTR</td>
<td>Resequencing mutations</td>
<td>Identification of the second mutation</td>
</tr>
<tr>
<td></td>
<td>assay by NGS</td>
<td>Validation technique</td>
<td>Comparison of two target enrichment approaches</td>
</tr>
<tr>
<td>Sequecing protocol</td>
<td>Custom AmpliSeq panel (Life technologies)</td>
<td>Hybrid capture: Custom NimbleGen SeqCap EZ</td>
<td>Hybrid capture: Custom NimbleGen SeqCap EZ</td>
</tr>
<tr>
<td>Regions analyzed</td>
<td>All CFTR exons and 20bp of</td>
<td>All CFTR locus (208kb → 181kb)</td>
<td>All CFTR locus (250kb)</td>
</tr>
<tr>
<td></td>
<td>all intron/splice sites (10</td>
<td>after masking repetitive DNA elements)</td>
<td>2.1million probes (60-90bp)</td>
</tr>
<tr>
<td></td>
<td>348bp)</td>
<td></td>
<td>36 overlapping fragments (7.3-14.5kb)</td>
</tr>
<tr>
<td>Target enrichment</td>
<td>Custom AmpliSeq panel (Life technologies)</td>
<td>Hybrid capture: Custom NimbleGen SeqCap EZ</td>
<td>Hybrid capture: Custom NimbleGen SeqCap EZ</td>
</tr>
<tr>
<td>Initial DNA quantity per sample</td>
<td>20ng</td>
<td>500ng</td>
<td>750ng</td>
</tr>
<tr>
<td>Library preparation</td>
<td>2 separate primer pools (2x36 amplicons), end-repair, ligation to adaptors, quantification of Library pooled bar-coded libraries Clonal PCR</td>
<td>Sonication/Covaris</td>
<td>Nebulization; End-repair, ligation to adapters (TrueSeq protocol CR), PCR amplification, quantification of Library (bioanalyzer), in-solution hybridization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>End-repair, A-tailing, ligation to index adapters (TrueSeq protocol)</td>
<td>Nebulization; End-repair, ligation to adapters, quantification of Library, Clonal emulsion PCR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PCR, amplification, quantification of Library</td>
<td></td>
</tr>
<tr>
<td>Sample Multiplexing and cost</td>
<td>5 to 35 samples</td>
<td>8 to 24 samples</td>
<td>No multiplexing 4 samples</td>
</tr>
<tr>
<td>NGS technology</td>
<td>200bp single reads</td>
<td>2x100bp paired-end reads</td>
<td>Single or Paired-end reads, up to 450 bp</td>
</tr>
<tr>
<td></td>
<td>Ion 314TM (10Mb) or Ion 318 (1Gb)</td>
<td>HiSeq 2000 instrument (Illumina)</td>
<td>GS Junior Sequencer (454 Life Sciences)</td>
</tr>
</tbody>
</table>
### Mapping

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Human genome</th>
<th>Human genome (GATK pipeline and picard-tools)</th>
<th>Human Genome (454 Roche GS Junior data analysis pipeline)</th>
<th>Chromosome 7 (454 Roche GS Junior data analysis pipeline)</th>
</tr>
</thead>
</table>

### Minimal coverage

| Minimal coverage | Mean: 231X | 99.7% covered by minimum 5 reads | 40X |

### CFTR Sequencing statistics

| Specificity | 97% (22 on 23 mutations) | 100% | ND |
| Sensitivity | 100% (23 on 23 mutations) | 100% (122 on 122 variants) | ND |
| Reproducibility | 100% (36 on 56 variants in 3 independent runs) | 100% (8 on 8 variants) | ND |

### Bioinformatics tools

| CNV | Depth-Of-Coverage tool from the Genome Analysis ToolKit (GATK) -> Log2 ratios < -0.7: deletion | Pindel; Conifer; PelSV-Fisher 5 large deletions, 1 duplication and 1 genomic rearrangement | ND |
| SNV | 7 variants per sample on average | 113 SNV + 28 InDels on average per sample | 197 variants on average per sample (118 in introns) |
| Variant identification and calling | Not notified | Variant prediction tools: GATK Unified Genotyper, samtools mpileup, SHORE | Reference Mapper Mutalyzer SeqNext (SS medical systems) |
| Variant filter | Min coverage: 20 reads Threshold: 25% reads Both strands | GATK Variant filtration (MQ<30.0; QUAL=25.0; QD=4.0; DP=5; DP=2000; GQ=15) GATK Combine Variant | Threshold 30% of reads Heterozygous: Read support: 30-65% Homozygous >70% |
| Databases | Variant frequency > 1% | dbSNP Exome Variant Server CFTR2 | 1000 Genomes project dbSNP Exome Variant Server CFTR2 |
| In silico analysis | SIFT Mutation Taster | Annovar SIFT, PolyPhen2, PhyloP, Mutation Taster, phastCons UCSC Genome Browser | Annovar, HSF, MaxEnt, NNSplice, SIFT, Polyphen2, PhyloP, UCSC Genome Browser, Ensembl, MutationTaster |
| Performance | Diagnostic rate | ND | 98.91% | 88.9% (16 / 18) |
| Particular cases | False positive c.2052del (2184delA) Sanger sequencing to determine the c.1210-34TG(11-13);-12T(5-9) haplotype | In-house script to determine the c.1210-34TG(11-13);T(5-9) haplotype | False positive c.2052del (2184delA) Sanger sequencing to determine the c.1210-34TG(11-13);-12T(5-9) haplotype |
| Advantages | Robust, specific, limited number of VUCs Breakpoints of large genomic rearrangement accurately determined | LR-PCR more adapted for complete gene resequencing New intronic mutations identified |
| Limits | SINE/LINE repeats (amplification and mapping) Homopolymer stretches (Sequencing) | GC rich genomic segments (target enrichment) Errors in index tags (sequencing) SINE/LINE repeats (amplification and mapping) Homopolymer stretches (Sequencing) Uncaptured region in intron 3 |

Table 4. Comparison of three NGS strategies for CFTR sequencing
2.3. Other loci

With the perspectives of high throughput molecular diagnosis in genetics laboratories, CF, CFTR-RDs and CF-like diseases could be simultaneously explored in patients using NGS. Various gene panels could be investigated according to patients’ phenotypes. Therefore, NGS approaches could contribute to (i) identify and confirm the implication of modifiers genes and (ii) improve molecular diagnosis of atypical Cystic Fibrosis, CFTR-RDs or CF-like diseases. Several sequence changes located in the so-called ‘modifier genes’ have been associated with progression of lung disease in CF patients. A decrease of pulmonary function measured by FEV1 (the forced expired volume after 1 s of blowing out) was associated with SNVs in EDNRA, ACER, IFRD1, IL8, MUC5AC and TGF-β1 genes. Haplotype 8.1 and variants in MBL2 gene were related to Pseudomonas aeruginosa colonization [28, 29]. SNVs in SNAP23, PPP2R4, PPP2R1A and KRT19 were recently associated with a decrease of lung function. Interactions between CFTR and these altered proteins may modify CFTR trafficking and membrane stability and therefore modify phenotype of CF patients [30]. Moreover, changes in other modifier genes were suspected to have an effect on intestinal obstruction (DCTN4, ADIPOR2 and MSRA genes), CF-associated diabetes (TCF7L2 gene) or liver disease (SERPIN1A1) [31].

CFTR-Related disorders comprise congenital bilateral absence of vas deferens (CBAVD), pancreatitis, diffuse bronchiectasis and nasal polyposis. Classically, two CFTR mutations (a severe and a mild mutation or two mild mutations) are found in well-characterized CBAVD patients [32]. In addition, Sharma et al. reported 2 SNVs on TGF-β1 and ENDRA genes associated with urogenital anomalies [33]. Chronic pancreatitis is caused by CFTR variations in some cases, and mutations in CTRC, PRSS1, PRSS2 or SPINK1 are also involved and must be analysed.

The implication of CFTR in diffuse bronchiectasis or nasal polyposis is more controversial [34]; variants in other genes were previously reported as possibly causative. However, the identification of mutations in other genes with sufficient significance remains difficult and needs large patient cohorts. In fact, in airway tract diseases, the influence of environment (pollutants, drugs/therapy and way of life) complicated the achievement of unbiased studies. Nevertheless, the hypothesis of oligogenism is supported by a study that reported mutations in ENaC channel genes (SCNN1A, SCNN1B, SCNN1G) or SERPIN1A1 in CF-like patients (borderline sweat-test and suggestive CF clinical features without two CFTR mutations) [35].

Finally, a pleiotropic effect of SLC26A9 on meconium ileus, pancreatic damage and lung disease has been identified [36], as well as SLC9A3 for meconium ileus and lung disease and SLC6A14 for meconium ileus and both lung disease and age at first P. aeruginosa infection [37]. Thus, the existence of pleiotropic effect of modifier genes on CF evolution may encourage the development of new therapeutic targets with multi-organ benefits.

3. Functional characterization of CFTR sequence variations

To provide appropriate diagnosis and prognosis to CF patients and also appropriate genetic counselling to families, the impact of variants identified by the techniques detailed above has to be functionally characterized.
Some variants such as frequent mutations (found in more than 1% of CF mutated alleles), nonsense or frameshift mutations are readily classified as pathogenic mutations. However, the frequent identification of rare sequence alterations of unknown pathogenicity (VUCS, VUS) substantially complicates test interpretation. Moreover, their number will increase with the diffusion of NGS technologies. To facilitate classification of these variants, CF laboratories have to combine several tools like central mutation databases or CFTR locus specific databases, in silico prediction tools and ex vivo/in vivo functional analyses [38, 39].

3.1. Epidemiological data and locus specific databases dedicated to CFTR

3.1.1. Core or central mutation databases

Their goal is to collect all sequence variations detected in all genes and to describe each mutation briefly. These databases are used to assess the frequency of a variation (minor allele frequency (MAF) lower or higher than 1%) in the general population (i.e. unaffected individuals). Since a MAF higher than 1% reflects a low probability for the variant to be pathogenic, such data may be highly informative for the interpretation of variants.

3.1.1.1. The National Center for Biotechnology Information (NCBI) short genetic variations database dbSNP

dbSNP [25] is the most comprehensive directory of single nucleotide variations. It catalogues short variations in nucleotide sequences from a wide range of organisms. Genetic Variations may be common, thus representing true polymorphisms, or they may be rare. Some of these rare human entries have additional information associated with them, including disease associations, genotype information and allele origin, as some variations are somatic rather than germ-line events. Genotypes and allele frequencies information for various populations from different studies, including data form the HapMap project, are also available.

3.1.1.2. Databases specifically collecting data from NGS projects

1000Genomes [24] aimed to find most genetic variants that have frequencies of at least 1% in samples from five populations: East Asian, South Asian, African, European and American ancestries. As in dbSNP, genotypes and allele frequencies information are available for a large number of variants [40].

Exome Variant Server [41] is a database that collects data of the NHLBI GO Exome Sequencing Project (ESP). This project aimed to discover novel genes and mechanisms contributing to various disorders by sequencing the protein coding regions of the human genome (i.e. exome) using NGS technology. As the CFTR gene is widely studied, this tool would not be of added value compared to dbSNP and 1000 Genomes.

3.1.2. Locus Specific Databases (LSDBs) dedicated to CFTR

LSDBs are now recognized as the best mode of collecting and curating lists of mutations related to human genetic diseases [42]. They compile in a single bioinformatics tool disease-causing
and non-disease-causing sequence variations identified by genetics laboratories in families with a history of a given Mendelian disease. The most sophisticated ones integrate clinical and biological data, information on the geographic and/or ethnic origin, frequency of variations in the general population, mutation hot spots and all useful information for diagnosis, prognosis and the evaluation of genotype/phenotype relationships [43].

Here we choose to detail three LSDB dedicated to CFTR that provide complementary information for the interpretation and the characterization of variants identified in diagnostics practice.

3.1.2.1. Cystic fibrosis mutation database

The Cystic Fibrosis Mutation Database (CFMDB) [44] also called ‘CFTR1’ is an open access database dedicated to the collection of sequence variations in the CFTR gene for the international CF genetics research community. It was initiated by the Cystic Fibrosis Genetic Analysis Consortium (CFGAC) in 1989 and is maintained by the Cystic Fibrosis Centre at the Hospital for Sick Children in Toronto. CFMDB allows the direct submission of new variants by laboratories, by filling out an on-line standardized form with the possibility to detail phenotypic data, genotype (i.e. other variants identified in patient) or epidemiological data. The key point of this database is to collect the largest number of CFTR sequence variations identified in patients, relatives and partners. On the other side, because the submission procedure applies only to the initial report of each variant, CFMDB does not provide frequency data, available with the two other databases described below. Finally, contributors do not always follow HGVS recommendations and a same variant can be reported by several laboratories under different names, possibly leading to misinterpretation or misreporting in diagnosis reports.

3.1.2.2. Clinical and functional translation of CFTR database CFTR2

CFTR2 [45] is a website designed to provide information about specific CF mutations to patients, researchers and the general public. For each mutation included in the database, it provides information about whether a given mutation causes cystic fibrosis when combined with another CF-causing mutation and clinical and biological information (sweat chloride, lung function, pancreatic status and pseudomonas infection rates) in patients carrying the mutation. A specific section for health practitioners and scientists provides more in-depth and research-related information.

The goal of the CFTR2 project is to categorize all mutations seen in CF patients as disease-causing (always resulting in CF when combined with another CF-causing mutation), neutral or mutation of varying clinical consequences (CF and CFTR-RD). Mutations that have not been fully analysed are considered of unknown clinical significance.

The major advantages of CFTR2 are (i) the collection of detailed clinical characteristics on large cohorts of individuals [46] that provide useful information related to a given genotype, and (ii) results of functional testing that are key arguments for their final interpretation [47].

However, this database only collects clinical and genetic data of CF patients (from national registers) that can lead to a bias of phenotypic spectrum assessment of several mutations.
considered as CF-causing mutations while they were also reported in CFTR-RD patients in trans of other CF-causing mutations.

3.1.2.3. The French molecular database CFTR-France

CFTR-France [48] has been developed since 2012 with the aim to collect, store and process any category of variants identified in the CFTR gene, thanks to the collaboration of nine French laboratories with high expertise in the molecular analyses of this gene. Its specificity is to compile and annotate any category of variations (disease-causing, non-disease-causing and variants of unknown clinical significance) that have been identified by collaborators in patients affected with CF or CFTR-RD, in foetuses with abnormal ultrasonography (e.g. echogenic bowel), newborns with pending or inconclusive diagnosis and asymptomatic individuals carrying at least one sequence variation on each CFTR gene (i.e. carrying two variations in trans). The database includes the main clinical data of these individuals, genetic information from familial segregation studies and various variant annotations (frequency in patients and controls populations, sequence homology, predicted or experimentally assessed functional impact, etc.), allowing the analysis of genotype/phenotype relationships.

Thus, CFTR-France, by collecting all phenotypes, reflects the phenotypic spectrum of a large number of mutations. It also reports mutations in complex alleles with association frequencies (related to all individuals recorded in the database), and gives the up-to-date HGVS nomenclature of mutations.

Data collected in CFTR-France are provided by level 2 (specialised) and reference laboratories, so that patients analysed only by level 1 laboratories (searching for the most common mutations) are not included in the database.

Note: Access to CFTR-France is currently restricted to collaborators. A public access program is in progress for the medical and scientific community and for patients and families.

3.2. In silico prediction analyses

3.2.1. Variants located in exons and exon-intron boundaries

3.2.1.1. Prediction tools for the assessment of the impact on protein

Prediction methods of amino acid substitutions use protein sequence, structure and/or annotation. Disease-causing mutations that affect protein function tend to occur at evolutionarily conserved sites and/or at key positions in protein structure. Multiple sequence alignment of orthologous sequences reveal what positions have been conserved through evolution, and these positions are supposed to be important for protein function. Annotation can enhance prediction for variants located in structurally and functionally important domains, but this information is often sparse.

The issue of the efficiency of prediction tools in assessing possible pathogenicity of missense variants in the CFTR gene is of major interest, since they constitute the vast majority of VUS...
identified in patients. Diagnostics laboratories frequently use those tools and particularly in problematic situations. Unfortunately their performance has not been clearly established and results (i.e. score of pathogenicity) may be discordant for a given variant.

Predictions of the impact of non-synonymous substitutions in CFTR are mainly based on multiple sequence alignment of orthologous sequences. Indeed, even if a partial 3D model of the CFTR protein has been established [49, 50], prediction tools do not take into account these elements in the final ‘score of pathogenicity’. It is classically recommended to use several prediction tools to obtain concordant predictions that could be considered for variant interpretation.

Table 5 summarizes bioinformatics programs classically used by diagnostics laboratories [51-53] and the new software SuSPect [54].

<table>
<thead>
<tr>
<th>Method and Web site</th>
<th>Algorithm</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>PolyPhen 2 [54]</td>
<td>Sequence conservation, structure to model position of amino acid substitution, and SWISS-PROT annotation</td>
<td>Score ranges from 0 to a positive number, where 0 is neutral, and a high positive number is damaging</td>
</tr>
<tr>
<td>SIFT [55]</td>
<td>Sequence homology: scores are calculated using position-specific scoring matrices with Dirichlet priors</td>
<td>Score ranges from 0 to 1, where 0 is damaging and 1 is neutral</td>
</tr>
<tr>
<td>Align GVGD [56]</td>
<td>Multiple sequence alignments to characterise the biochemical properties (composition, polarity and volume) of the observed amino acids at each position</td>
<td>The prediction classes form a spectrum (C0, C15, C25, C35, C45, C55, C65) with C65 most likely to interfere with function and C0 least likely.</td>
</tr>
<tr>
<td>SuSPect [57]</td>
<td>Sequence-, structure*- and systems biology-based features to predict the phenotypic effects of missense mutations. Algorithm trained using Variflank (contains information for experimentally verified effects) *Structure-based comparison with other ABC transporters</td>
<td>Table of scores from 0-100, colour-coded according to predicted deleteriousness (blue=neutral, red=disease-causing). A score of 50 is recommended as a cut-off between neutral and disease-causing variants</td>
</tr>
<tr>
<td>Alamut© [58]</td>
<td>Commercial software (Interactive Biosoftware) Databases: RefSeq, dbSNP, Uniprot, InterPro, UCSC Genome Browser Database, PubMed, Ensembl. And Automated access to on-line prediction tools: PolyPhen2, SIFT, Align GVGD Access to splicing prediction tools (detailed in Section III.2.1.2.) &quot;Alamut Batch&quot; High-throughput annotation software for NGS analysis</td>
<td>Summarized report of the results of different prediction tools, according to their output format Designed for intensive variant analysis workflows</td>
</tr>
</tbody>
</table>

Table 5. Bioinformatics tools for the prediction of amino acid changes: websites, characteristics and output format [55-59]

A recent work has emphasised the importance of sequence alignments on the performance of prediction tools [60]. The authors constructed custom multiple sequence alignments called phenotype-optimized sequence ensembles (POSEs) that was tested on a training set of CFTR mutations.
A previous work already suggested that providing SIFT or PolyPhen-2 with custom alignments increased their performance relative to the default alignments employed by the algorithms [61]. This could explain that, if Alamut® is a highly interesting tool with its ease of use, some predictions obtained by using each tool separately with a custom algorithm could differ from Alamut® results (obtained with default alignments).

3.2.1.2. Prediction tools for the assessment of the impact on pre-mRNA splicing

Splicing mechanisms comprise exon recognition within large pre-mRNA molecules and the precise removal of flanking introns. Three elements constitute the core splicing signals: the intronic branch point, the acceptor site (or 3’splice site), including an inconstant upstream polypyrimidine tract (PPT), and the donor site (or 5’ splice site). These core human splice site motifs contain only a part of the information that defines exons, whereas the rest corresponds to less conserved splicing regulatory elements. The latter are located within the exon or flanking introns, promoting or inhibiting exon recognition through exonic/intronic splicing enhancers (ESE or ISE) or silencers (ESS or ISS), respectively (Figure 3).

Figure 3. A schematic of key splicing motifs and regulatory elements. Adapted by Le Guédard-Méreuze S. from Wang and Burge, 2008 [62, 63].

Many bioinformatics tools have been developed to predict which splicing modification is the most probable for a given sequence variation — exon skipping, cryptic splice sites activation, use of de novo splice sites — or if the variant may be considered as neutral regarding its impact on splicing. Most algorithms were developed based on biostatistical and experimental analyses of information contained in the genomic sequence. They provide a score depending on the strength of the considered splice site. Indeed, the strength of splicing motifs is a key parameter to predict the impact of a sequence variation. Performance of these tools has been widely studied by comparing the results of predictions with experimental assays for various genes including CFTR [39, 63-66]. In 2012, Houdayer and collaborators performed a large-scale study of VUCS in BRCA genes in order to assess the performance of six prediction tools [67]. This work provided guidelines for the proper use of these tools and for the interpretation of prediction results.

Table 6 summarizes principle and main characteristics of the most ‘popular’ bioinformatics programs and ASSEDA, a recently developed program [68-73].
<table>
<thead>
<tr>
<th>Method and Web site</th>
<th>Algorithm</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaxEntScan [72]</td>
<td>Based on a Maximum Entropy Model (MEM): Modelled the sequences of short sequence motifs such as those involved in RNA splicing which simultaneously accounts for non-adjacent as well as adjacent dependencies between positions.</td>
<td></td>
</tr>
<tr>
<td>Human Splicing Finder (HSF)</td>
<td>Associates weight matrix model (WMM) and Maximum Entropy Model (MEM)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HSF (WMM): Quantification of the relative likelihood of candidate splice site sequence to show coincidence with consensus sequence</td>
<td>Numerical Score (0-100)</td>
</tr>
<tr>
<td></td>
<td>MaxEnt (MEM): Statistical approach representing the least biased approximation for the distribution of sequence motifs, from available data (real and decoy splice sites). Incorporates local adjacent and non-adjacent position dependencies</td>
<td>Numerical Score (0-10)</td>
</tr>
<tr>
<td></td>
<td>Consensus splice sites</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>ESEfinder</em></td>
<td>See Prediction tools for cis-regulatory elements</td>
</tr>
<tr>
<td></td>
<td>RESCUE-ESE*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Predicted PESE AND PESS Octamers from Zhang &amp; Chasin</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cis-regulatory elements</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ESEfinder*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RESCUE-ESE*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>*See Prediction tools for cis-regulatory elements</td>
<td></td>
</tr>
<tr>
<td>NNSplice [74]</td>
<td>Machine learning approach that recognizes sequence patterns once it is trained with sets of DNA sequences encompassing authentic and decoy splice sites. Based on a hidden Markov Model (HMM): Incorporates local adjacent position dependencies</td>
<td>Score between 0 and 1</td>
</tr>
<tr>
<td>Automated Splice Site and Exon Definition Analysis (ASSEDA) [75]</td>
<td>Analyses splicing mutations according to changes in total exon information:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>∆Ri_{total}, values (differences in probability of relative inclusion or exclusion of the wild type and mutated exon in mature mRNA)</td>
<td></td>
</tr>
<tr>
<td>Prediction tools for cis-regulatory elements</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ESEfinder [76]</td>
<td>ESE motif score</td>
</tr>
<tr>
<td></td>
<td>Prediction of SR protein specific putative ESE, based on an in vitro SELEX approach dependent on addition of individual SR proteins</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RESCUE-ESE [77]</td>
<td>Z-score</td>
</tr>
<tr>
<td></td>
<td>(Relative enhancer and silencer classification by unanimous enrichment)</td>
<td>(Highlights extremely over- or under-represented hexamers)</td>
</tr>
<tr>
<td></td>
<td>Statistical approach based upon different distribution of hexamers in exons and introns with different properties (e.g. weak and strong splice sites)</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Main characteristics of several Splicing prediction tools [74-79]
It is important to note that consequences on splicing of exonic synonymous and non-synonymous CFTR variants must be assessed, as suggested by recent experimental studies [39, 80].

3.2.2. Deep intronic variants

The examples of insertion of intronic sequences called pseudo-exons (or cryptic exons) in mature transcripts of various genes are becoming ever more numerous and their role in human diseases has been largely demonstrated. We saw in section 2.2 that NGS strategies currently allow scanning of CFTR deep intronic regions [18], resulting in a growing number of new identified deep intronic variants.

Bioinformatics tools described above, which assess the impact of variants on splicing, can also be used to evaluate deep intronic mutations. We tested these algorithms on mutations identified in CF patients after NGS sequencing of the entire CFTR locus and they showed satisfactory results [18]. Indeed, prediction tools allowed the selection of possible disease-causing mutations (i.e., predicted impact on splicing by inclusion of pseudo-exons) and predictions were confirmed by in vitro functional studies using minigene constructs (see section 3.3.1.1) and by direct analysis of aberrant transcripts from nasal epithelial cells of patients (see 3.3.2.1).

3.3. In vitro/ex vivo functional analyses

3.3.1. Cell lines transfection experiments

The type of cells used for transfection depends on the tissue that is studied and the clinical context. Pulmonary (BEAS-2B, A549, Calu-3) or intestinal/colic (CACO-2, T84) immortalized cells (by SV40 or carcinoma) contain an appropriate concentration of transcriptional and splicing factors for CFTR protein synthesis. Cells stably transfected with mutated CFTR can also be used (CFBe41o-, CFPAC-1). Stable expression is usually obtained by lentivirus transduction and transient transfection by chemical agent (Polyfect, interferin). In this case, the endogenous CF molecular and cellular context (inflammation) should also be considered.

3.3.1.1. Splicing assessment

Minigenes are autonomic cyclic entity containing promoter and exons and are produced by clonal amplification in bacteria [81, 82]. They contain a genomic segment from the gene of interest (here CFTR) that includes exon and flanking intronic regions (length can range from ten to thousands of nucleotides, an average of 300 bp) or only intronic regions in the case of evaluation of potential creation of a pseudo-exon. To determine whether a mutation is responsible for altered splicing, minigenes can also include cis-regulatory elements if affected (ESE, ESS, ISE or and ISS) [83]. These regions of interest are framed by two invariable exons, which are part of the system. Every assay of transfection in cell lines compares the wild-type and mutated (through directed mutagenesis) constructs [84]. All CFTR exons are needed to produce a mature and functional protein. Thus, a modification of transcript in the in vitro system suggests that the assessed CFTR change has a deleterious effect on exon splicing. An
ever-increasing number of mini-gene studies have been performed to assess the pathogenicity of CFTR variants [39, 60, 80, 84]. This strategy, despite its limitations, is of high interest in the overall strategy for the characterization of rare sequence variations.

3.3.1.2. Expression vectors for the quantification of mRNA, protein or CFTR-specific chloride conductance

Full-length CFTR cDNA is classically inserted in expression vector system (e.g. pcDNA3 or p-Tracer) upon a promoter that may be drug-activated (G418, tetracycline or doxycycline-activation). To assess point variants or small indels, directed mutagenesis is carried out (usually QuikChange Mutagenesis kits®, Agilent Technologies). To assess molecular consequences of large rearrangements concerning one or more exons, a truncated CFTR cDNA can be inserted in the expression vector [85-87]. Transient or stable transfection can be performed in eukaryotic cells (describe in III.3.1. section). 3-HA tag (in the fourth loop of CFTR) can be introduced to easily visualize protein expression. Then, measurement of mRNA expression and evaluation of function and localization of the CFTR protein can be performed for each alternative transcript construct, compared to wild-type.

Automated real-time RT-PCR allows the relative straightforward quantification of mRNA transcripts with specific primers and appropriate reference genes for normalization. mRNA level informs about future protein quality and quantity.

Protein assessment consists in the implementation of complementary experiments for protein quantification, evaluation of its maturation or its cellular localization. Main techniques are detailed below. The effect of variants on CFTR expression and maturation is assessed based on the detection of immature (core-glycosylated, B band, ~150-kD) and/or mature (additional glycosylation in the Golgi, C band ~170-190 kDa) CFTR forms by immunoblotting. Long-term pulse-chase experiments can provide additional information on the lifetime of CFTR on cellular compartments [88]. Immunocytochemical assays (Immunofluorescence (IF) based) can highlight the cellular localization of the CFTR protein. However, most difficulties noted in IF experiments relate to non-specific antibody staining and the effect of sample processing on characteristics of cell development. Moreover, confusion between cell surface (where CFTR is active) and subsurface (where it would not) may occur. Therefore, more sensitive and specific antibodies as well as co-localization assays with other cell surface markers (such as β-tubulin or WGA) are needed. Finally, this remains a qualitative or semi-quantitative method.

CFTR function and activity, i.e. CFTR-specific chloride conductance, can be determined by patch-clamp electrophysiology, halide selective electrode technique, radioisotope efflux assays and by fluorescence-based halide efflux measurement. The use of a CFTR-activating appropriate drug (such as forskolin, isobutylmethylxanthine (IBMX), isoproterenol, terbutaline, genistein, adenosine, etc.) or ATP followed by specific CFTR inhibitor CFTRinh-172 permits CFTR-dependant or independent chloride transport, respectively. To date, the easiest approach developed consists in Iodide efflux based on fluorescence measurement. YFP fluorescence is dependent on YFP expression levels and iodide concentration. Compared with conventional plate-bound CFTR functional assays, the flow cytometric approach can be used to study CFTR function in cell suspension. It may be further adjusted to study CFTR function
in heterologous cell populations using cell surface markers and selection of cells that display high CFTR function. Technical limitations include the need to perform this assay in specialized centres (using expensive imaging equipment).

All these methods offer the possibility to evaluate the functional consequences of molecular abnormalities on CFTR and finally improve the classification of variants.

3.3.2. Ex vivo CF patients’ cells assays

3.3.2.1. Analyses on primary airway cells and tissues

Characterization of CFTR molecular defect can partially be performed directly after biopsies of tissues that show sufficient CFTR expression. Nasal or lung airway epithelia are optimal. These tissues are accessible by minimally invasive interventions and display an endogenous expression of CFTR transcripts and protein. Moreover, nasal and bronchial epithelia show the same cellular composition (ciliated, goblet, columnar and immune cells) although ratios differ slightly [89]. Since quantification and detection of aberrant splicing and quantification or localization of proteins are possible in human tissues, information that they bring is crucial to assess the effect of variants and to propose a functional classification. However, this approach has its limitations and requires other functional assays to perform large-scale genotype-phenotype correlation studies. In addition, especially for nasal tissue, the low quantity of cells collected (out of 500,000 cells per brushing) only allows ‘one-shot’ tests and hinders mechanistic assays. Moreover, highly variable CFTR expression in heterogeneous cell types, in healthy individuals and in p.Phe508del homozygous patients has been described, varying from 0 to 100% [90]. Other genetic (see below) or environmental parameters could also modify CFTR expression levels.

3.3.2.2. Ex vivo culture of primary airway cells

Culture of primary cells from CF patients can be performed with brushed nasal or bronchial cells after biopsies. Wild-type endogenous CFTR protein is expressed at the apical membrane of polarized cells. Therefore, in vitro monolayer culture seems no longer adapted. Obtaining polarized cells is promoted by air-liquid interface culture (ALI), proposed since the 2000s, by an ex vivo system of collagen-coated porous membrane on which cells are plated after a phase of monolayer amplification or directly after nasal brushing. Basal adherent cells differentiate in all airways epithelial cell types, which organize into a pseudo-stratified epithelium [91].

This model offers the opportunity to perform functional assays described above to determine CFTR dysfunction. Molecular defect induced by a specific mutation can be qualitatively determined if the cell donor is homozygous for this mutation.

Technical limitations such as bacterial or fungi contaminations or absence of adherence complicate culture of cells obtained from CF patients. Moreover, further studies are needed to determine if extrapolation is possible between observations in primary cells directly after
brushing and after several weeks in culture media, particularly for quantitative level assessment.

3.3.2.3. Intestinal current measurement (ICM)

ICM was developed as a research tool to assess CFTR function in the 1990s and has been used as a diagnostic test since the early 2000s [92, 93]. At least four superficial rectal biopsies per patient, obtained by suction, are needed and mounted on adequate tissue sliders. ICM consists in blocking epithelial sodium channels by amiloride and stimulates cAMP-mediated CFTR-dependent chloride transport in a chloride-free solution with forskolin or IBMX. Cholinergic chloride transport and histaminic reaction were also evaluated and the sum of the response ∆Isc_{forskolin/IBMX+carbachol+histamine} appears to be the best parameter to evaluate CFTR function by ICM, but reference values and ranges have not been established. This combination of ionic responses discriminates patients with CF from healthy subjects but not CF patients with pancreatic insufficiency (PI) or sufficiency (PS). ICM can detect a loss of CFTR function above 80%, therefore CF carriers and CFTR-RD patients may not be identified by this method. Moreover, mild mutations could result in a false-negative ICM.

ICM is not altered by secondary damage on tissue and thus better reveals the primary CFTR dysfunction (compared with nasal potential difference, see below). A new functional CFTR assay using primary CF intestinal organoids derived from patients and cultured in vitro may offer new tools to screen for therapy [94].

Finally, setup and maintenance of dedicated equipment by experienced and trained staff limits its use.

3.3.3. In vivo biomarkers assays of CFTR function

Sweat chloride Test (ST) and Nasal potential difference (NPD) measurement are used as diagnostic tests for CF. In atypical clinical context these in vivo tests can give additional arguments to further explore the CFTR locus (cf. Figure 2 that describes the molecular diagnostics step-by-step strategy). Furthermore, they provide complementary information for the interpretation of CFTR variations.

ST higher than 60 mmol/L is the ‘gold standard’ and discriminates between healthy and typical CF. Sweat electrolytes are higher in the most severely affected and are lower in those with mild mutation, who have partial rescue of channel function [95-97]. There are two major advantages for the use of sweat test in the evaluation of the CFTR mutation severity: (i) stability of the measure throughout life and (ii) non-invasive way of measurement which maintains skin integrity. Indeed, sweat electrolytes levels reflect the primary defect on the CFTR protein and do not highlight secondary consequences of its absence or dysfunction on affected organs. Furthermore, a new means for assessing the secretory function of CFTR has been recently developed based on β-adrenergic-mediated sweating. This method provides a unique evaluation of the purely secretory function of CFTR in vivo. It has been shown to be more sensitive in individuals carrying mutations that commonly exhibit normal or borderline sweat chloride rates, such as c.3718-2477C>T (legacy name: 3849+10kbC>T) [98].
The transepithelial NPD measurement estimates the net ion conductance across the nasal airway epithelium and indirectly measure CFTR activity. Sodium conductance is determined after amiloride perfusion (Na+). NPD parameters discriminate CF and CFTR-RD phenotypes and a correlation between NPD and disease severity has been shown [99], although there is an overlap in subjects with mild phenotype [97]. However, other sources of phenotypic variability (modifiers genes or environment) and technical limitations (i.e. maintenance of dedicated equipment and trained personnel) contribute to the inconsistencies of NPD across genotype-phenotype correlation studies. CFTR expression and NPD response can also be modified if nasal epithelium is affected by rhinosinusitis, polyposis or exposure to smoking [100]. However, despite the difficulty of setting up this technique, NPD assessment could be used as a complementary step to support CFTR dysfunction in inconclusive clinical cases and then to confirm the deleterious effect of CFTR variants identified in these patients.

4. Conclusion

New issues are emerging from the use of NGS technology in CF molecular diagnosis. On one hand, NGS approaches offer new possibilities by multiplexing samples and provide a wider coverage of the CFTR locus including deep intronic regions. NGS assay design can also include additional modifiers genes [31]. On the other hand, molecular diagnoses in emergency contexts challenge the possibility of sample multiplexing, and the increased number of VUCS will require complex functional analyses. However, as tools described above are constantly improving, the knowledge about CFTR variations is rapidly expanding, allowing geneticists and clinicians to provide patients with high quality information and adequate genetic counselling. Finally, the functional characterization of CFTR variations will provide rationale for a personalised medicine strategy driven by patients’ genotype in the very near future.

Acknowledgements

We acknowledge all members of the technical team for CF molecular diagnosis: Jean-Pierre Altieri, Fanny Verneau and Lise Larrieu. We also thank the coordination team of the CFTR-France genetic database: Corinne Bareil and Corinne Thèze and all the collaborating members of this project. We are also grateful to the French Network of laboratories specialized in molecular diagnosis of CF (GenMucoFrance). Finally, we thank the French Association against CF (Vaincre la Mucoviscidose) for its constant support.

Author details

Anne Bergougnoux1*, Magali Taulan-Cadars2, Mireille Claustres2 and Caroline Raynal1

*Address all correspondence to: anne.bergougnoux@inserm.fr
1 Laboratory of Molecular Genetics, Montpellier Hospital, France

2 Montpellier University, France

References


[53] Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A. Comprehensive statistical study of 452 BRCA1 missense sub-


