SHORT THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

MOLECULAR EPIDEMIOLOGIC AND DIAGNOSTIC INVESTIGATIONS IN CYSTIC FIBROSIS

by Gergely Ivády

Supervisor: István Balogh

UNIVERSITY OF DEBRECEN
DOCTORAL SCHOOL OF MOLECULAR CELLULAR AND IMMUNE BIOLOGY

DEBRECEN, 2019

MOLECULAR EPIDEMIOLOGIC AND DIAGNOSTIC INVESTIGATIONS IN CYSTIC FIBROSIS

By Gergely Ivády, MD

Supervisor: István Balogh, PhD

Doctoral School of Molecular Cellular and Immune Biology,
University of Debrecen

Head of the **Examination Committee**:	Prof. Gábor Szabó, MD, PhD, DSc
Members of the Examination Committee:	Prof. Márta Széll, MD, PhD, DSc
	Prof. Gábor Méhes, MD, PhD, DSc

The Examination takes place at the Library of the Department of Laboratory Medicine, Faculty of Medicine, University of Debrecen, at 11:00 am, 2nd of April, 2019

Head of the **Defense Committee**:	Prof. Gábor Szabó, MD, PhD, DSc
Reviewers:	Olga Török, MD, PhD
	Prof. István Raskó, MD, PhD, DSc
Members of the Defense Committee:	Prof. Márta Széll, MD, PhD, DSc
	Prof. Gábor Méhes, MD, PhD, DSc

The PhD Defense takes place at the Lecture Hall of Bldg. A, Department of Internal Medicine, Faculty of Medicine, University of Debrecen at 12:30, 2nd of April, 2019

1. INTRODUCTION

Cystic fibrosis

Cystic fibrosis (CF), also known as mucoviscidosis is one of the most common monogenic hereditary diseases in the Caucasian population. Approximately one out of 27 individuals carries a pathogenic mutation in the gene in heterozygous form. The disease develops as a consequence to a chloride channel dysfunction, caused by mutations in the cystic fibrosis transmembrane conductance regulator gene (*CFTR*). It has an autosomal recessive inheritance. Since the exchange of sodium, chloride and bicarbonate ions is hindered during the epithelial fluid transport, thick and viscous secretions are going to form in the efferent ducts of the involved exocrine glands. The clinical picture is extremely diverse, progressive and characteristic of the given patient. The course of the disease depends on which organs and to what extent are concerned.

Epidemiology

CF affects ca. 70.000 people worldwide. According to WHO data, a sick child is born from every 2.000 – 3.000 pregnancies in Europe. Mortality varies widely in the context of the country's health care and the living standard of the population. Life expectancy has increased considerably over the last four decades, and is currently the highest in Canada (47.7 years). In the past, only scarce information was available about the prevalence of the disease in Hungary, the most recent dated to 1996. According to data published by the European Cystic Fibrosis Society (ECFS), 558 CF patients were registered in our country in 2017. In the near future with the spread and advancement of neonatal screening, and with more frequent recognition of the milder forms of the disease, incidence data is likely to increase, which will decrease gradually because of the improved prenatal and pre-implantation diagnosis. However prevalence will increase over the long term due to the more complex therapy.

Structure and function of CFTR

CFTR protein is a cAMP-dependent chloride (Cl^-) / bicarbonate (HCO_3^-) channel in the ABC (ATP Binding Cassette) protein family. It is located in the apical plasma membrane of the cells in the airways, intestinal tract, pancreas, liver and reproductive organs. The

macromolecule consists of two homologous parts, which are built up of one hexahelical transmembrane domain (TMD1 and 2) and one intracellular nucleotide binding domain (NBD1 and 2). The two sections are connected by an intracellular regulatory (R) domain. The protein's biosynthesis and maturation follows the traditional endoplasmic reticulum - Golgi pathway, finally it completes its physiological role in the cell membrane. After internalization, the protein eventually enters the endosomes, where it is mostly recycled. The half life of the molecule is 12-14 hours on average. CFTR-controlled ion transport in the epithelial tissues counts as the rate-determining step of anion secretion, thus determining the degree of transepithelial fluid movement, which affects the degree of hydration and pH of the epithelial lumen surfaces. Overall, this channel regulates the final concentration and volume of pancreatic fluid, sweat and airway fluids, so it plays a key role in digestion, body temperature regulation and in the natural immune defense of the lungs, so the reduced volume and altered composition of these secretions lead to severe complications very early.

Pathogenesis of CF

The first symptoms of the disease most commonly appear in infants and toddlers. Tissues expressing CFTR (pancreatitis, lung, gastrointestinal and hepatobiliary system as well as reproductive tract) undergo a cystic, connective tissue type degeneration accompanied by functional deterioration. Functional disorder is primarily caused by organ damage due to obstruction by the viscous mucus. The *CFTR* gene is located on the long arm of chromosome 7 (7q31.2), has a size of 230 kb and encodes 1480 amino acids in 27 exons. The number of genetic variants described in *CFTR* has been increasing since the 1989 cloning of the gene, nowadays there are more than 2000 alterations in the official databases. Although the vast majority of mutations are rare - only 20 reaches 0.1% allele frequency worldwide - 1524 of them are proven to be pathogenic. Among the numerous variations, non-CF causing / asymptomatic and mild types also occur, along with mutations with varying clinical consequences. According to the recently renewed ECFS recommendation, a mutation is considered pathogenic, if accompanied by another known pathogenic mutation *in trans* it leads to clinically confirmed CF. In patients with variable clinical outcome (MVCC), further examination towards recurrent / chronic idiopathic pancreatitis, congenital bilateral vas deferens deficiency (CBAVD) and bronchiectasis is required in specialized CF centers. From the pathophysiologic point of view, CF causing mutations can be sorted into six classes. Among the mutations in class I. are nonsense, splice site and frameshift types that result in an

early stop codon (e.g. c.3484C>T, c.3846G>A, c.1657C>T, c.1679+1.6kbA>G and c.1624G>T). In case of class II. mutations the posttranslational modification of the protein or its transport to the cell membrane is disturbed (e.g. c.1521_1523delCTT and c.3909C>G), while class V pertain to decreased transcription due to promoter or splicing abnormalities (e.g. c .2657+5G>A, c.1364C>A, c.2988+1G>A). The mutation belongs to class VI. if the protein is functional but unstable, therefore its turnover in the plasma membrane is enhanced (e.g. c.-12_10del23, c.859A>T, c.4147_4148insA). To sum up I, II, V and VI. group mutations lead to a severely reduced cell surface CFTR expression. On the contrary, in class III. and IV. genetic defects the amount of protein is adequate but the function is impaired. Class III. mutations decrease the opening probability of the chloride channel or cause its failure (e.g. c.1652G>A, c.1651G>A, c.4046G>A), while in class IV. the structural change of the channel leads to lack of conductance (e.g. c.350G>A, c.1000C>T, c.1040G>C).

Screening and Diagnosis of CF

Setting off in New Zealand in 1979, newborn screening has been introduced in several countries worldwide. This was also promoted by the rapid development of early therapeutic interventional possibilities. The screening has two, more often three-levels and uses hybrid techniques, which typically include a DNA-based method besides the determination of immunoreactive trypsinogen and pancreatitis-associated protein. Depending on the applied combination and selected cut-offs, the positive predictive value of newborn screening ranges from 9 to 19.7%, sensitivity ranges from 87 to 99%, and their specificity usually reaches 99%. Molecular genetic methods for determining the CF mutation status have a well-defined place and a clear role in CF diagnostics. The suspicion of the disease can be raised upon family history or appearance of characteristic symptoms or syndromes (e.g. diffuse bronchiectasia, exocrine pancreatic insufficiency, congenital adrenal hyperplasia, obstructive azoospermia, positive sputum culture, especially *P. aeruginosa*, etc.) According to the current algorithm, a positive sweat chloride test (≥ 60 mmol / L) or the presence of two CF-causing mutations with a genetic test (*in trans*) or demonstration of CFTR dysfunction is required for the diagnosis. In the latter case, nasal potential difference (NPD) and intestinal current measurement (ICM) are the accepted methods, however these can be difficult to carry out and therefore have not spread widely. The latest ECFS Practice Guide also suggests the equipment and capabilities recommended to perform CF genetic testing in a laboratory. Among these criteria we should emphasize the importance of determining the mutational spectrum of the specific population

to be examined, especially in case of molecular genetic screening tests. Since mutations show a very diverse geographical distribution, the identification of pathogenic CF-causing variations occurring more frequently in the given region is unavoidable in order to establish an optimally performing screening panel.

Therapeutic possibilities in CF

Although morbidity / mortality rates have shown significant improvement over the past decades, most CF patients still die because of respiratory failure. Therefore, slowing down of lung disease progression continues to be the primary goal of therapy. Due to the continuously reforming mucus plugs, mucociliary self-purification can not function efficiently, leading to secondary infections. The most common pathogens are *Staphylococcus aureus* and *Pseudomonas aeruginosa*. Chronic inflammation maintained by neutrophil granulocytes is characterized by acute exacerbations and the pulmonary function is unable to return to baseline. Conventional therapy therefore includes antibiotic prophylaxis and prompt, aggressive treatment of exacerbations. This is often supplemented with bronchodilators, DNase and neutrophil elastase inhibitors for the permeability of respiratory tract, and the rate of depletion can be increased by chest physiotherapy. Fortunately new ways of targeted therapeutic options will open up with the detailed knowledge of the background of the disease and the development of molecular diagnostics. In 2012 ivacaftor (VX-770) was the pioneer of genotype-specific therapies, which was able to improve the CFTR function. The significant improvement in the condition of a subpopulation of patients (who had at least one c.1652G> A allele) propagated further research. Nowadays targeted therapeutic strategies follow three main directions: searching for potentiator, corrector and read-through molecules. Potentiator agents (e.g. ivacaftor) improve the efficacy of existing cell surface CFTR, so they are applied for III. and IV. class mutations. Corrector molecules (e.g. lumacaftor and tezacaftor) are used in case of class II. mutations, helping the processing and membrane delivery of mutant CFTR protein. The third group of compounds (e.g. ataluren) enables to skip the early stop codon, resulting in less amount of truncated CFTR protein in patients with Class I. mutations. In general, the class II. mutation c.1521_1523delCTT (p.Phe508del) can be found in two-third of the patients (exact percentage depends on the specific population). In their case, the misprocessed protein is mostly decomposed before reaching the cell membrane, but if it does, then the pathomechanism can actually be considered a type III., "gating" deviation. Perhaps this explains why the efficacy of the combined therapy (corrector + potentiator) in

c.1521_1523delCTT patients is well above the luma- or tezacaftor monotherapy, and became the recommended regimen for patients over 12 years or c.1521_1523delCTT homozygotes. The modulator therapies detailed above allow us to treat nearly half of the CF patients with a considerably favorable outcome than before, so the importance of reliable and more precise genetic analysis is undeniable. The genetic engineering techniques are still in the experimental stage, the dissertation does not cover them.

The principle and limitations of pyrosequencing

New Generation Sequencing (NGS) techniques allow unprecedented throughput, short turnaround times and low cost when determining DNA base sequence. For these reasons their use is extending beyond scientific research to molecular genetic diagnosis. Among the numerous methods, pyrosequencing and ion-semiconductor technologies were the first ones to provide determination of longer sequences, but there are still problems with the quality of base reads. Pyrosequencing is based on bioluminescence detection, where a complementary strand is synthesized enzymatically to a single-stranded DNA template. Within one cycle of synthesis, only one kind of dNTP is added to the reaction zone where the DNA polymerase enzyme catalyzes the incorporation of the bases according to the complementarity rule. During this step nucleotide pyrophosphate is formed and its quantity is measured by a coupled luciferase enzyme reaction. Deoxynucleotides are dispensed in a predefined order and separated from each other by washing steps. Thus, it can be determined how much and which type of dNTP has been incorporated, since the intensity of the detected light will be directly proportional to it. At the end of the analysis, the software generates flowgrams, which are evaluated in order to get the base sequences. It is now common knowledge that in practice, pyrosequencing / ion-semiconductor techniques does not always follow theory. The initially linear relationship between the number of built-in nucleotides and the intensity of the emitted light diminishes as the number of nucleotides increases. As a consequence, when determining repeating bases, the method regularly makes mistakes, which occurs in the form of under- and overcalls. There are other limits to the procedure, e.g. the amount of added dNTP is not sufficient (incomplete extension) or inadequate washing, which can lead to a "carry-forward" phenomenon (different dNTPs incorporate in one cycle) but these can be easily identified and filtered thanks to a large number of parallel measurements.

Homopolymers

In genomics, a homopolymer (HP) is a sequence of consecutive identical bases. Approximately 1.43 million HPs (also known as mononucleotide microsatellites) exist in the human exome, with the size of 4-mer and up. They are believed to play roles in transcriptional regulation and recombination, and the vast majority (96.7%) of them are in the range of 4-mer to 6-mer. HP sequences composed of A:T base pairs are over-represented in the human genome compared to G:C HPs. Although both pairs show structural stability, these loci in the genome are highly mutagenic and have been characterized as hotspots for length change mutations, which has, presumably, contributed to their reduced occurrence in the exome over time.

Bioinformatics

Since HPs are more prone to insertion and deletion mutations (indels), problems are going to aggravate, when utilizing pyrosequencers or ion semiconductor chemistry in diagnostic procedures. It is essential for molecular genetic methods used for diagnostic purposes to be capable of separating true genetic variations from artefacts (i.e. keeping the false positive rate low). No coincidence there are numerous bioinformatic correction tools to overcome this problem. Some of these algorithms are based on clustering the flowgrams; for example Denoiser, which utilizes rank-abundance distributions, or PyroNoise/AmpliconNoise, which calculates a likelihood using empirically derived error distributions. Acacia's main focus is on HP sequences and the algorithm uses a dynamically updated cluster consensus when aligning reads. Coral and ECHO are multiple alignment based techniques, while HECTOR is a homopolymer spectrum based error corrector, with a multistage correction workflow. Another useful software is FlowClus, which provides feedback on the denoising process, allowing the user to apply more suitable analysis parameters for the particular dataset. The most recent tools, such as NoDe (Noise Detector) and DUDE-Seq are believed to produce even lower error rates and are more time-efficient. Even if sophisticated correction tools are used to overcome the difficulties of detection and significantly improve accuracy, it is still very important to estimate the capability of the corresponding NGS system to correctly determine HPs. To avoid uncertainty in the diagnostic testing of patient samples, it is also recommended that the maximum length of stable HP

detection, for reasonable identification of indel mutations in such sequences, be defined before using the NGS instrument in routine clinical practice.

Homopolimers in the *CFTR*

The coding region of the cystic fibrosis transmembrane regulator gene (CFTR) contains 24 homopolymer stretches, involving 17 out of 27 exons. In conjunction with the whole genome, T and A homopolymers vastly outnumber G and C homopolymers; 14 thymine, 8 adenine, 2 guanine and no cytosine HPs are present. In exon 14 there is a seven adenine long homopolymer region (c.2046_2052) and genetic alterations affecting this region could create poly-A tracts of different sizes, e.g. the pathogenic mutation c.2051_2052delAAinsG (2183delAAinsG) results in a five, while c.2052delA (2184delA) results in a six adenine long homopolymer segment. In case of the relatively frequent c.2052_2053insA (p.Gln685Thrfs*4) mutation an eight adenine long homopolymer stretch is formed.

2. OBJECTIVES

The aims of the PhD studies were:

1. To determine the types, frequencies and distribution of mutations occurring in Eastern Hungary on a pre-selected, classical CF patient cohort.
2. To develop a region-specific, rational and cost-effective, multi-stage CF mutation detection panel based on the results.
3. To extend the spectrum survey to other areas of Hungary by enrolling new patients, aiming to obtain data of national validity.
4. Review and test the former CF mutation panel with the combined mutational spectrum.
5. To determine the analytical performance of the pyrosequencer Roche 454 NGS system, with focus on the homopolymer detection capabilities.
6. To develop and validate an NGS-based CF diagnostic method that can be used for routine molecular genetic testing and which is also able to assist in the establishment of a genetic analysis module for future neonatal and/or carrier screening.

3. PATIENTS AND METHODS

The mutational spectrum in Eastern Hungary

Patients

For the regional mutational spectrum analysis, a representative group of 40 patients (mean age ± SD: 14.4 ± 8.7 years) were selected with the classical clinical picture of the disease. Assessing symptoms and recruiting CF suspect patients were done in close clinician collaboration. The following symptoms were the most common in the patients' anamneses: respiratory system signs (tachypnoe, prolonged cough, chest x-ray disorders, recurrent obstructive bronchitis, recurrent pneumonia, bronchiectasis, etc.), gastrointestinal problems (malabsorption, nanosomia, hypoproteinaemia-edema, pancreatic lesions, meconium ileus, etc.), reproductive tract related symptoms (late puberty, early obstruction of vas deferens, azoospermia, etc.) and a positive or grey zone chloride sweat test (≥ 60 mmol / L). To collect sweat samples Macroduct Sweat Collection System (Wescor - ELITechGroup, Logan, UT) was used, to determine sweat chloride concentration Sweat Chek Conductivity Analyzer (Wescor - ELITechGroup, Logan, UT) and / or Sanasol SM-01 sweat analyzer (Sanasol, Zalaegerszeg, Hungary) were utilized. Patients were not related to each other.

Molecular genetic methods

The DNA extraction from the leukocytes of the EDTA anticoagulated peripheral blood samples of CF patients was performed with QIAgen Blood Mini Kit (Qiagen, Hilden, Germany). The identification of pathogenic CF-causing mutations was done in a three-step approach, in line with the international recommendations. When using the different analytical methods, we proceeded to the next level until we found both alleles in a given patient. The steps of the analysis followed each other in the below sequence:

1. Elucigene CF29v2 Kit (Tepnel Diagnostics, Manchester, United Kingdom). This commercially available kit covers the 29 most common CF-causing mutations in the Caucasian population and Ashkenazi Jewish diaspora, which include: c.3454G>C (p.Asp1152His), c.1585-1G>A (1717-1G>A), c.1624G>T (p.Gly542*), c.3846G>A (p.Trp1282*), c.3909C>G (p.Asn1303Lys), c.1521_1523delCTT (p.Phe508del), c.3717+12191C>T (3849+10kbC>T), c.262_263delTT (p.Leu88Ilefs*22), c.489+1G>T

(621+1G>T), c.3752G>A (p.Ser1251Asn), c.1652G>A (p.Gly551Asp), c.350G>A (p.Arg117His), c.3484C>T (p.Arg1162*), c.1000C>T (p.Arg334Trp), c.1364C>A (p.Ala455Glu), c.2657+5G>A (2789+5G>A), c.178G>T (p.Glu60*), c.3528delC (p.Lys1177Serfs*15), c.2051_2052delAAinsG (p.Lys684Serfs*38), c.1766+1G>A (1898+1G>A), c.2988+1G>A (3120+1G>A), c.1657C>T (p.Arg553*), c.579+1G>T (711+1G>T), c.948delT (p.Phe316Leufs*12), c.1519_1521delATC (p.Ile507del), c.1040G>C (p.Arg347Pro), c.254G>A (p.Gly85Glu), c.2052delA (p.Lys684Asnfs*38), c.1679G>C (p.Arg560Thr). It is a multiplex allele-specific amplification technique (ARMS), which only generates a PCR product if the mutant allele is present. PCR products are detected in a 3% agarose gel and identified by size compared to the molecule weight marker. One exception is the most frequent c.1521_1523delCTT mutation, where both wild and mutant alleles are amplified, so it is also possible to discriminate between hetero- and homozygous genotypes.

2. Examination of CFTRdele2.3 (21kb), which is also called the "Slavic Deletion". The detection of this common large size mutation (21kb, involving 2 exonic and 3 intronic regions) in Central and Eastern Europe was carried out as described in the literature, with a primer pair designed for the mutational hot spot. To determine the presence of homo- or heterozygous form of this deletion, amplification of exon 3 of *CFTR* was also needed.
3. Sequencing the entire coding region of the *CFTR* was also done according to literary data, which was only altered in one case: exon 7 (formerly 6b) was amplified with a modified pair of primers: 6BF (5'-CTG TAC AGC GTC TGG CAC AT-3') and 6BR (5'-CAA ACA TCA AAT ATG AGG TGG AA-3'). For DNA sequencing, PCR products were purified on ultrafiltration columns (Microcon YM-100, Millipore, Burlington, MA). For sequencing purified products, BigDye Terminator v3.1 Cycle Sequencing Kit was used (Applied Biosystems, Foster City, CA). After gel filtration unreacted nucleotides were removed (DyeEx Kit, Qiagen, Hilden, Germany), and finally capillary gel electrophoresis was performed on ABI Prism 310 Genetic Analyzer (Applied Biosystems, Foster City, CA).
4. At last, intragenic *CFTR* rearrangements (CNVs) were verified by multiplex ligation-dependent probe amplification (MLPA), SALSA MLPA KIT P091-B1 CFTR (MRC-Holland, Amsterdam, The Netherlands). MLPA is a multiplex PCR technique that amplifies probes which were hybridized to template DNA and then ligated. Due to filling sequences, the products only differ in a few nucleotides, so *CFTR* exons can be tested in parallel. The fluorescence intensities (peak heights) of the products are the function of the

initial copy number and are compared to the relative peak heights of the reference (normal control) samples (deletion: <0.7; duplication: >1.3).

The mutational spectrum in Hungary

Additional patients

Similarly to the previous survey, here we also worked with selected patients. To do so, we received help from CF centers in Hungary, mainly from Budapest, Szeged and Debrecen. Our collaborating partners sent samples of patients who demonstrated the classic symptoms of the disease described above and the possibility of CF was solemnly raised. A total of 45 peripheral blood samples were obtained from 22 males and 23 females (mean age ± SD: 10.1 ± 8.1 years). Sample collection took place between 2010 and 2014.

Revising the order of the genetic tests

In the light of our earlier results, the process of genetic testing needed reconsideration. The type of sample and extraction of DNA from leukocytes did not change, but minor modifications were made in the approach protocol. The updated version of the subsequent steps for mutation analysis was as follows:

1. The first line of tests continued to be the Elucigene CF29v2 kit. If only one pathogenic mutation could be detected, Sanger sequencing of that region was performed to find out if one or both alleles are affected, with the exception of c.1521_1523delCTT. If the mutation was present in heterozygous form, the CFTRdele2.3(21kb) assay was carried out as described above.
2. 2. The second level of screening was the sequencing of *CFTR*, but some of the gene sections were given priority. The first exons to be sequenced (e4, e6, e11, e14, e15 and e20) where previous studies have found prevalent mutations in Eastern Hungary; e.g. c.2052_2053insA (p.Gln685Thrfs*4), c. 302T> G (p.Leu101*) and c.3276C> A (p.Tyr1092*). In the case of a negative result, sequencing of the remaining exons took place.
3. Unless Sanger sequencing found both pathological differences, then the MLPA was the next method as already described.

4. In a case where MLPA detected a large rearrangement, additional confirmation was required. The CFTRdele2, c.54-5811_164+2186del273+6780_273+6961inv mutations detected in this patient was confirmed by allele-specific PCR and the bidirectional sequencing of exon-intron boundaries, which was performed according to the literature.

Testing the pyrosequencer and development of a new method

These experiments were carried out using the NGS instrument, Roche 454 (Life Sciences, Penzberg, Germany). Since the measurements can be multiplexed, investigation of the analytical capabilities of the device and the development of a new generation CF mutation detection method were conducted simultaneously. Two test systems were used during the experiment series. One of them was a plasmid system for determining the length of the HP section that the device can still detect with high precision. In the other, we tested the efficacy of self-designed primers on human samples and optimized them. We paid particular attention to the investigation of the ponderous poly-A region in e14 described previously.

The plasmid system

A total of 12 plasmid vectors were constructed, which contained 4-mer, 5-mer and 6-mer HPs of the four nucleotides. The template was pcDNA3.1 (Invitrogen - Life Technologies, Thermo Fisher Scientific, Waltham, MA). For mutagenesis we used the Quikchange II kit (Agilent Technologies, Santa Clara, CA) according to the manufacturer's instructions. Transformation was performed using XL1-Blue supercompetitive cells. In order to avoid the formation of length-change mutations during plasmid replication and amplification of the HP-containing fragments, we employed PicoMaxx DNA polymerase (Agilent Technologies, Santa Clara, CA), which has "proofreading" properties and can correct errors during synthesis. Two colonies per clone were checked by Sanger sequencing whether they contained the corresponding HP segments. Each homopolymer tract in the plasmid system was investigated using three pairs of primers in order to test the hypothesis, that in the beginning of the sequencing reaction a sufficiently high signal-to-noise ratio might enable precise HP length detection. Our primer design therefore was as follows:

- the homopolymer was located in the vicinity of the forward amplification/sequencing primer in 3′ direction (proximal type),

- the analyzed homopolymers were located approximately in the middle of the amplicon (mid type),
- the reverse amplification/sequencing primer's 3′ end was generated to be as close as possible to the HP segment (distal type).

The size of the homopolymer clones varied between 366 and 387 bp, depending on the HP length and the primers used.

Sequencing the *CFTR* gene

In the second set of experiments, a *CFTR* gene mutation detection system was developed and analyzed in detail. 17 clinical samples were used with known *CFTR* mutation status determined by an in vitro diagnostic assay (Elucigene CF29v2, Elucigene Diagnostics, Manchester, UK) and Sanger sequencing. Primer design for *CFTR* mutation analysis was similar to the plasmid system's described above, except for the "HP in the middle" type amplicons, which were left out of this experiment. Exons e3, e14, e15, and e24 have multiple HP sections, most of which could only be covered within the same amplicons; therefore altogether 33 HP-containing amplicons were tested per patient. When designing the primers, all known single-nucleotide polymorphisms (SNPs) were taken into consideration to maximize annealing efficiency and minimize allele drop-out, which was shown to be an issue in a previous test system. We also designed primers for *CFTR* exons that do not contain homopolymer stretches to be able to analyze the complete gene in one run. In this part of the experiment 6-6 patient samples were analysed. Pyro-sequencing was performed using GS Junior Titanium emPCR (Lib-A) and GS Junior Titanium Sequencing Kit (Life Sciences - Roche, Branford, CT) according to the manufacturer's instructions. The most crucial section of the gene (within exon 14) was further tested using 11 human DNA samples; including four wild type and seven c.2052_2053insA heterozygotes. As with the plasmid system, we used three additional primer pairs to generate amplicons for NGS sequencing with "proximal," "mid," and "distal" HP locations. "Proximal" primers were located at a distance of 5 and 11 base pairs from the poly-A tract. In all samples, PicoMaxx enzyme mix was used for the amplification processes.

Mutation nomenclature

Numbering of the *CFTR* exons is based on current recommendations (Ensembl ENSG00000001626). In mutation nomenclature we used the cDNA names suggested by the Human Genome Variation Society (HGVS), but protein names or if they are absent, legacy names are also listed in brackets.

Ethical approval

All human participants gave informed consent for diagnostic genetic analysis. In this study DNA samples were then applied anonymously and procedures were in accordance with the current revision of the Helsinki Declaration. The laboratory is approved by the National Public Health and Medical Officer Service (approval number: 094025024). The candidate carried out the examinations with the help of the staff working at the Division of Molecular Genetics in the Institute of Laboratory Medicine.

Statistical analysis

The individual reads generated by the pyrosequencer were evaluated using the Amplicon Variant Analyzer software (Life Sciences - Roche, Branford, CT). Among the HP section reads, we distinguished "proximal", "mid" and "distal" types, depending on the HP distance from the amplification / sequencing primers. The accuracy of genotyping was characterized by the ratio of valid reads to the total number of reads (acceptable if >75%). Statistical analysis of the results was done with GraphPad Prism (v5.03) (GraphPad Software, La Jolla, CA). Normality was checked by the Shapiro-Wilk test, in case of non-normal distribution, the comparison between the reads was done with the Kruskal-Wallis test with a subsequent Dunn's post-hoc test (significance levels $P<0.05$ and $P<0.01$).

RESULTS

The mutational spectrum in Eastern Hungary

We started the assesment of the mutation spectrum with the East-Hungarian region. In the patients' sweat tests, chloride ion concentrations of 55-173 mmol/L (mean: 108 mmol/L) were measured, with only one patient not reaching the limit of 60 mmol/L proposed by the criterion system. The following mutations were detected on the 80 CF alleles tested using the Elucigene CF29v2 Kit as the first stage of screening:

- 56 alleles (70.0%) carried the c.1521_1523delCTT (p.Phe508del) mutation,
- 4 alleles (5.0%) had the c.3909C>G (p.Asn1303Lys),
- 3 alleles (3.75%) had the c.1624G>T (p.Gly542*),
- 1 allele (1.25%) had the c.1585-1G>A (1717-1G>A) and
- 1 allele (1.25%) had the c.1040G>C (p.Arg347Pro) pathogenic variants.

Subsequently, we searched for the "Slavic" deletion and identified four (5.0%) CFTRdele2.3(21kb) alleles. With the help of the above methods, in 11 patient samples only one pathogenic mutation was found, therefore sequencing of these patients was necessary. The detected mutations by Sanger sequencing of the entire coding region of *CFTR* were:

- 4x (5.0%) c.2052_2053insA (p.Gln685Thrfs*4),
- 2x (2.5%) c.302T>G (p.Leu101*),
- 1x (1.25%) c.658C>T (p.Gln220*),
- 1x (1.25%) c.1397C>A (p.Ser466*),
- 1x (1.25%) c.3276C>G (p.Tyr1092*) and
- 1x (1.25%) c.2491G>T (p.Glu831*).

In 12 out of the 19 compound heterozygote patients testing of the mutation phase was performed in their respective families, with all detected mutations present in *trans*. Only one case remained unclear, having c.1521_1523delCTT (p.Phe508del) together with an unidentified allele, with *CFTR* rearrangement analysis by MLPA being negative. For the analysis of mutations in the cohort under study we used the recommended "cascade

approach". Thus, when using the first line assay 81.25% of CF-causing mutations were identified. Incorporation of the CFTRdele2,3(21 kb) mutation increased the detection rate to 86.25%. The rest of the mutations could only be detected by gene sequencing. If we have started with sequencing of exons 4 and 14, where two frequent mutations are located (c.2052_2053insA and c.302T>G), the hit ratio was already above 90%. Determination of the entire *CFTR* base sequence resulted in a detection rate of 98.75%, since 79 out of 80 alleles were found.

The mutational spectrum in other regions of the country

Next we tried to asses CF-causing mutations in the whole country by enrolling additional patients. At this step, we used the experience from our previous study and created a modified, more efficient mutation detection algorithm. In this part of the study 27 different mutations were identified. The most common variant, c.1521_1523delCTT (p.Phe508del) was detected in 53.3%. It was followed by four mutations: a c.3846 G>A (p.Trp1282*), c.3909C>G (p.Asn1303Lys), CFTRdele2,3(21kb) and c.2052_2053insA (p.Gln685Thrfs*4) each counting for 4.4% of all alleles. Also worth to mention another four mutations c.1624G>T (p.Gly542*), c.3276C>A (p.Tyr1092*), c.489+1G>T (621+1G>T) and c.2012delT (p.Leu671*), which were found on more than one allele.

Among the detected genetic abnormalities there were two novel ones (according to the Clinical and Functional Translation of CFTR database, cftr2.org and the Human Gene Mutation Database (HGMD)). These newly described sequence alterations are most likely pathogenic. One of them changes the reading frame, generating a premature stop codon 17 amino acids downstream (c.1037_1038insA, p.Leu346Hisfs*17). Pathogenicity of the other detected novel missense mutation c.1394C>T, p.Thr465Ile is supported by the following:

i. the affected residue is located at a phylogenetically highly conserved position according to the orthologs of Bos taurus, Equus caballus, Felis catus,Mus musculus etc.,

ii. the variant may be interpreted as "likely to be pathogenic" (IV) based on the recommendation of the American College of Medical Genetics and Genomics (ACMG) to which the following contribute:

    - the mutation is located in the NBD1 functional domain (PM1),

- control individuals are missing this alteration in the 1000 Genomes Project, the Exome Aggregation Consortium and the Exome Sequencing Project (PM2),
- in case of recessive inheritance, a pathogenic variant in *trans* can be detected. In the given patient it is c.1521_1523delCTT, the child inherited from his mother (PM3). (Result of the study of family members: mother c.1521_1523delCTT carrier, father c.1394C> T carrier),
- the deviation is a new missense mutation that affects the same amino acid residue as a known, already proven pathogenic missense mutation: c.1394C>A vagy p.Thr465Asn (PM5),
- *in silico* studies (SIFT analysis) show results of impaired protein function (PP3),
- for monogenic diseases, the phenotype is characteristic of the particular disease (PP4).

The combined mutational spectrum in Hungary

Finally the results of the two studies were interlaced. Based on the combined data, commercially available assay (Elucigene CF29v2) was able to detect more than three quarters of the mutations. Adding the allele-specific PCR designed for the large deletion CFTRdele2.3(21kb) we experienced a 4.7% increase in our detection rate. Besides c.2052_2053insA, targeted sequencing of e14 of the *CFTR* showed the presence of two other rare mutations (c.2012delT and c.2002C> T) with frequencies of 4.7%, 1.2% and 0.6% respectively. Other variations were mosty detected by sequencing of the remaining *CFTR* exons with the exception of c.54-5811_164+2186del273+6780_273+6961inv (CFTRdele2), which was recognized by the MLPA test. The latter was also confirmed by allele-specific PCR and Sanger sequencing. The pooled database also provided an opportunity to map the geographical distribution of the mutations in the country. Along with already known tendencies (e.g. the north-south gradient for c.1521_1523delCTT) it was observed that the common CFTRdele2.3(21kb) mutation and c.2052_2053insA also show a territorial accumulation. Both mutations occur predominantly in the northern part of the country, which is further limited to the northeastern region in case of the c.2052_2053insA. This finding is not so suprising, given the high frequency of the mutation in Western Ukraine. Of course, this

might be also due to the fact that most of the samples tested came from the northern and eastern parts of Hungary.

The results of NGS sequencing studies

Quality control measures

Since we were trying to determine the analytical limitations of our NGS pyrosequencer, it was indispensable to introduce a quality assurance step. Patients' amplicons and the site-directed mutagenesis experiments were verified using Sanger sequencing, even performed in duplicates in case of the plasmids. No discrepancy was found between the expected and the controlled base sequences, so we could be sure that any inaccuracies in the following measurements could be traced back to the inherent defect of the method and show the real capabilities of the device.

Analysis of HP plasmids

When sequencing homopolymer-containing plasmids, mean coverage was 479 ± 145 and an evident negative correlation was observed between homopolymer length and read accuracy. The average correct genotyping rate of all four nucleotides combined was 95.8, 87.4 and 72.1% in 4-mers, 5-mers, and 6-mers, respectively (with a 79.6–99.3% range in 4-mers, 36.9–98.4% in 5-mers and 14.5–93.8% in 6-mers). While the pyrosequencing-based NGS system was able to detect poly-A 6-mers reliably (with a mean of 76.4%), detection rates fell under 75% for poly-C, poly-G and poly-T 6-mers (means: 71.5, 68.3 and 63.4%, respectively). In general, longer HPs had lower genotyping accuracy, although, the most accurate reads still reached 98.4% for 5-mers and 93.8% for 6-mers, indicating that careful optimization in a given sequence context might help to skip the poor performing primers and find the most functioning ones for the analysis. After testing, primer localization failed to show any association with genotyping accuracy.

Pyrosequencing of human DNA samples

Coding regions and exon-intron boundaries in the *CFTR* gene from 6 cystic fibrosis patients with a known mutation status were sequenced. Mean base coverage was 263 ± 178 using our in-house assay. Altogether 246 amplicons were included in the data analysis (> 40 reads). Although individual read accuracy varied in a wide range (52.2 – 99.1%), average accuracy was generally excellent using the assay (89.3%). The assay was able to detect all 18 small-scale genetic alterations (missense, nonsense, splice site mutations, frameshift/in-frame deletions and insertions) previously identified by Sanger sequencing, providing 100% sensitivity and specificity. Regarding the 24 HP stretches the self-designed primer set yielded good performance with more than 80% genotyping accuracy in all but one HP. The exception was a 7A HP tract (c.2046_2052) with a 52.2% average correctness. The problematic region was further analyzed by using four wild type samples. The amplification of the critical gene section was done using three primer pairs, as described previously (A1, A2, A3) and ended up with very variable results. When evaluating reads from both directions, there were several cases where we did not barely get a 7A signal, while in other amplicons this exceeded 80%. Thanks to our collaborative partner in the Czech Republic (Prof. Dr. Milan Macek Jr., Charles University, Prague), we have been provided seven DNA samples, heterozygous for the c.2052_2053insA (p.Gln685Thrfs*4) mutation. In case of this variant an eight adenine long homopolymer stretch is formed, so the percentage of detected 7A and 8A signals theoretically should have been 50% for each. Among the applied three primer sets for this locus two primer pairs proved to be rather poor performers with 45–50% irrelevant nucleotide calls. On the other hand, sequencing with a third set of primers, correct 7A and proximal 8A calls were detected with satisfactory accuracy, having only 27% misreads on average, but even this set could not identify 8 adenines from an approximate distance of 200 bp.

DISCUSSION

Cystic fibrosis is a prevalent autosomal recessive hereditary disorder, whose diagnosis still holds challenges, despite the fact, that since its first description in 1938, the information available on both phenotype and genetic background has multiplied severalfold. Data from newborn screening programs, patient registers, clinical databases and functional research contributed greatly to a more complete picture of the *CFTR* gene and the nature of the disease, but complicated molecular genetic laboratory diagnostics as well. In the first part of this work, we focused on the epidemiology of cystic fibrosis and the frequency of pathogenic mutations in Hungary. In the light of the results, we attempted to develop a multi-stage, "cascade" type test system that is capable of identifying genetic variants in a cost-effective way. Finally, with the help of a new generation sequencing instrument we tried to simplify the molecular genetic testing, also hoping that it might serve as a basis of a genetic module in a future newborn screening.

CF Mutations in Eastern Hungary

This comprehensive study involved a preselected cohort of CF patients (and samples) diagnosed at the University of Debrecen, partly in a retrospective manner. We also tried to fill a gap here, since the most recent publications on the Hungarian mutational spectrum were born in the mid 1990s. According to previous genetic studies describing European and North American populations, we assume that the found allele frequencies can be applied to other parts of the country either. At present, there are approximately 10 million citizens living in Hungary. Population admixture increased when subsequently Romanian shepherds, Flemmish and Slovakian settlers colonized this region. According to a local survey from 1910 the population reported to be of 54.5% Hungarian, 16.1% Romanian, 10.7% Slovakian, and 10.2% German origin, including several other minorities. We would like to use the above survey for illustration, since self-reporting of ethnicity substantially changed during the last century. For example in 2001, over 94% of the inhabitants were declared to be of Hungarian origin, making this data less relevant to us, which does not provide tangible aid in interpreting the geographical distribution of mutations. For the analysis of mutations in the cohort under study we used the recommended “cascade approach”: searching for common mutations first, followed by sequencing and rearrangement analysis as specified in the Methods section. Thus,

when using the Elucigene CF29v2 assay 81.25% of CF-causing mutations were identified. Incorporation of the CFTRdele2,3(21 kb) mutation increased the detection rate to 86.25%. Sanger sequencing was necessary for detecting the rest of the mutations and reaching 98,75% sensitivity. In case of one patient we were unable to identify one of the aberrancies, the most likely explanation is that an intronic or promoter mutation which was not analyzed by our methods might be producing the null allele. In general, the observed degree of mutation heterogeneity is between the reported Northern and Southern European mutation spectra, whereby all mutations were previously detected in South German, Ashkenazi Jewish and other Balkans populations and filed in the Cystic Fibrosis Mutation Database. Altogether six mutations reached a higher prevalence than 1.25%: c.1521_1523delCTT (p.Phe508del), c.3909C>G (p.Asn1303Lys), c.54-5940_273+10250del21kb [CFTRdele2,3(21kb)], c.2052_2053insA (p.Gln685Thrfs*4), c.1624G>T (p.Gly542*) and c.302T>G (p.Leu101*), in decreasing order of their frequencies. As the population under study is not primarily of Slavic origin, it is interesting, that the Slavic mutation CFTRdele2,3(21 kb) was found on 5.00% of CF alleles, which is the third highest prevalence after Czech Republic (6.37%) and Russia (5.69%). In this respect Eastern Hungary was formerly inhabited by Slavic tribes who later gradually assimilated with Hungarians (from 895 AD), which likely explains the high frequency of this allele. Another interesting finding that the c.2052_2053insA frameshift mutation was found at a particularly high frequency (5.00%). In this regard a paper of Makukh et al. (2010) reported that this allele is the second most common mutation in Western Ukraine, comprising 7.20% of all mutated CF alleles. Since Western Ukraine is bordering the area from which our cohort was drawn, this result shows population relatedness of both regions given their close longterm historical ties. Therefore, our data confirm the “Galician origin” of this mutation given its decreasing gradient towards the region from which our patients were drawn. It would also be worthy to study similar cohorts of CF patients in neighboring Eastern Slovakia, Southeastern Poland, Belarus and Northwestern Romania in order to further substantiate this most likely regional founder effect. It would be very interesting to assess the occurrence of these two mutations in all Hungarian CF patients, which will be possible with the current genetic revision of the patient registry.

The mutation spectrum in Hungary

In the second part of our studies we extended the mutation analysis nationwide. Based on the results of the previous study in Eastern Hungary, we have applied a modified,

"cascade" approach, detailed in the Methods chapter. Altogether, 90 aberrant alleles were identified containing 27 different mutations. Among them two new, most likely pathogenic variants were found: c.1394C>T (p.Thr465Ile) and c.1037_1038insA (p.Leu346Hisfs*17), which were not previously described in the databases and the literature. The pathogenicity of the mutations are supported by the following: c.1037_1038insA changes the reading frame, generating a premature stop codon 17 amino acids downstream (p.Leu346Hisfs*17); while in case of c.1394C>T the affected residue is located at a phylogenetically highly conserved position, another pathogenic mutation (c.1394C>A vagy p.Thr465Asn) affecting the same amino acid residue has already been described, SIFT analysis predicts a damaging effect on the protein function and c.1394C>T may be interpreted as a "likely to be pathogenic" (IV) variant based on the recommendation of the American College of Medical Genetics and Genomics (ACMG) as detailed before. We updated our existing CF database by merging recently acquired nationwide results with our previous data from Eastern Hungary (n=85). Altogether 31 different mutations were identified, among which eleven reached a frequency higher than 1%. In accordance with the literature, the decreasing North-to-South gradient stands for the distribution of c.1521_1523delCTT (p.Phe508del) found on 104/170 CF alleles, compared to the Czech Republic (61,2% vs. 67,4%), but not to Poland (61,2% vs. 54,5%). c.1521_1523delCTT was followed by the “Mediterranean mutation” c.3909C>G (p.Asn1303Lys), the “Slavic” mutation c.54-5940_273+10250del21kb [CFTRdele2,3(21kb)] and the “Galican” mutation c.2052_2053insA (p.Gln685Thrfs*4), each responsible for 4.7% of all CF alleles, which meets our previous observations. c.1624G>T (p.Gly542*) was detected in 2.9%, the “Israeli” mutation c.3846G>A (p.Trp1282*) in 2.4%. Other relatively frequent mutations were c.3276C>A (p.Tyr1092*) and c.302T>G (p.Leu101*) found in 1.8%, while c.489+1G>T (621+1G>T), c.1397C>G (p.Ser466*) and c.2012delT (p.Leu671*) found in 1.2% of the patients. One gross rearrangement (0.6%) in the *CFTR* gene c.54-5811_164+2186del273+6780_273+6961inv (CFTRdele2) was detected by MLPA analysis and confirmed by allele specific PCR and bidirectional sequencing of the junction regions, as described in the literature. According to the mapped mutation frequencies in patients originating from different regions of Hungary using the commercially available assay 75.9% of the mutations can be identified. The allele specific PCR for the detection of the common “Slavic” deletion, CFTRdele2,3(21kb) adds 4.7%, while the sequencing of exon 14 adds 6.5% to the proportion of detected mutations. MLPA analysis revealed one rearrangement (0.6%), while direct sequencing of the entire coding region of *CFTR* gene identified 20 CF alleles (11.8%). Overall, the cascade screening achieved 99.5% sensitivity on the preselected patient

cohort. At the same time geographical tendencies can be recognized in the distribution of both CFTRdele2,3(21kb) and c.2052_2053insA mutations. All but one CFTRdele2,3(21kb) and c.2052_2053insA positive samples originate from the northern regions of the country and c.2052_2053insA even seems to be restricted to the northeastern territories of Hungary, which is not suprising if we consider the high frequency of this mutation in Western Ukraine. In case of CFTRdele2,3(21kb), the first millennium-end migration of Slavic tribes may have played a role, while the fact that Transcarpathia was part of Hungary for about a thousand years might be responsible for the localization of the c.2052_2053insA mutation. On the basis of the combined results, we believe that the utilized multi-step approach can be highly effective in our country. Moreover, we believe it also could be applied for the Hungarian diaspora, as more than 2 million Hungarians live beyond our borders in Europe and North America. Finally, we would like to point out that our genetic laboratory fulfills all the diagnostic criteria set by the ECFS in 2018. It is also important to note that the acquaintance of the mutation spectrum defined by this study is a basic requirement for the development of a two-, or three-level newborn screening program or a carrier-screening program. In our opinion (which is supported by recent literature data) a combination of Elucigene CF29v2, c.2052_2053insA, and CFTRdele2.3(21kb) would obtain a sensitivity of 85-90%, if used during neonatal screening.

The modern laboratory diagnosis of CF

For cystic fibrosis, a screening program has been introduced in a number of countries. Among monogenic diseases, CF is the most common indication of prenatal and preimplantation genetic tests. The development of diagnostic techniques has undergone dramatic changes over the last twenty years, which is still taking place today. From traditional Sanger sequencing we have arrived to high performance NGS technologies that allow rapid and effective investigation of the whole *CFTR* locus and other factors affecting the clinical appearance of the disease. When appropriate bioinformatics background is available to interpret the provided results and to understand the biological significance of each described variant, serious advances in patient care and therapeutic decision making are expected. In the third phase of our work, we aimed to introduce a new generation sequencing device into our routine laboratory diagnostics. This NGS instrument is based on pyrosequencing. The inherent drawback of the method is that it identifies the length of the HP sections larger than 4-5 base pairs with incorrect accuracy. Although the commercially available CF Mutation

Detection CE IVD Kit (CFTR MASTR DX, Multiplicom, Agilent Technologies) was effective on the Illumina NGS platform, our preliminary studies showed that it did not yield acceptable results with respect to the large number of HP sections in the *CFTR*, which are often the predilection sites for some common mutations. To test the analytical performance of the Roche benchtop NGS pyrosequencer a series of plasmid vectors were generated. Altogether 12 clones were produced (4-mer, 5-mer, and 6-mer homopolymers of all four nucleotides) using site directed mutagenesis. After transformation and replication of plasmids each homopolymer tract was amplified using three pairs of primers, then the amplicons were confirmed by Sanger sequencing before loading them in the NGS instrument. As expected, NGS sequencing was the most accurate in determining 4-mers, while the reliability of the method gradually decreased when analysing 5-mers and 6-mers. We observed that there is also a difference between the base types, the system was most successful in detecting the adenine bases, least precise with thymine nucleotides. Based on previous experiments we hypothesized, that in the beginning of the sequencing reaction a sufficiently high signal-to-noise ratio might enable more accurate HP length detection. That is why we used three amplifier primer sets, to have homopolymer reads at the initial, middle or late stages of the sequencing reaction. The results of our plasmid system did not confirm the hypothesis, there was no significant difference between the "proximal", "mid" and "distal" types of reads. However, it should be mentioned, that despite the highly variable accuracy, the detection efficiency has reached the required level (> 75% correct reads) for some amplicons, which calls attention to the importance of prudent primer design and optimization during method development. In the second part of the NGS experiments, we concentrated on the development, optimization and testing of a molecular genetic method to be used in the laboratory diagnostics of cystic fibrosis. Our goal was to set up a method which is fast, cost-effective and can detect mutations in our country with high sensitivity, so that it could be used for neonatal and/or heterozygote screening. For this purpose, pyrosequencing of custom designed *CFTR* amplicons was performed. To maximize the reliability of the system, HP regions were typically amplified with 2-3 pairs of primer pairs. This method has detected all small scale mutations previously identified by Sanger sequencing (synonym, missense, nonsense, splice site mutations, frameshift / in-frame deletions and insertions). Although read accuracy varied in a wide range, the average reading rate of 89.3% is considered acceptable. Regarding the HP sections in the gene, the self-designed method provided satisfactory performance, as the pyrosequenced reads were more than 80% compliant with Sanger sequences. The only exception was a unique 7A tract (c.2046_2052) with a 52.2% average

correctness. However a 7-mer in *CFTR* e1 could be identified with 83% accuracy. Therefore, it is likely, that homopolymer detection depends not only on HP length or primer distance, but also on several other factors, such as the nucleotide microenvironment in the DNA sequence or the spatial location of beads on PicoTiter plates. It is extremely important to identify even small scale insertions or deletions in homopolymer sections with high reliability. In the above mentioned 7A (c.2046_2052) region there are several mutations, of which the detection of c.2052delA (2184delA) and c.2051_2052delAAinsG (2183AA>G) did not encounter any problems with pyrosequencing. This is what we expected, because in these cases homopolymers of 5A and 6A are formed. The more critical the situation is when the patient carries the c.2052_2053insA (p.Gln685Thrfs*4) variant, because the adenine HP to be detected is eight nucleotides in length. Unfortunately, as our epidemiologic data demonstrates, the frequency of c.2052_2053insA allele in Hungary is high, with a prevalence of 4.7%, which is the second most common mutation after c.1521_1523delCTT. This particular issue, whether c.2052_2053insA can be detected with pyrosequencing with great certainty, was investigated on additional patient samples. The results obtained with the sequencing of three different amplicons show that the identification of patients carrying this specific mutation can not always be reproduced by pyrosequencing. In light of all this, we have to state that although the method we developed provides reliable sequences for the vast majority of cases, investigation of the c.2052_2053insA mutation clearly shows that Sanger sequencing is still required in specific situations and thus, cannot yet be eliminated from the molecular diagnostic workflow.

SUMMARY

The dissertation focuses on two topics: the prevalence of cystic fibrosis causing mutations in Hungary and the development of a modern molecular genetic diagnostic approach to the disease. The Hungarian mutation spectrum was determined in multiple steps. This part of the study included 85 patients altogether. We carried out the investigations in a cascade manner, which was continously refined during the process. The final form of our CF mutation analysis was the following: Elucigene CF29v2 assay (targeted sequencing in case of any positive result except for c.1521_1523delCTT) combined with the examination of the CFTRdele2,3(21kb) mutation, sequencing of e4, e6, e11, e14, e15 and e20, followed by sequencing of the remaining *CFTR* exons and finally MLPA analysis of the gene. During the investigation 169/170 CF alleles were identified, with 31 different pathogenic variants. Among these were included two newly found mutations (c.1394C>T and c.1037_1038insA), which are likely to be CF-causing as well. When depicting the geographical distribution of the mutations, we observed distinctive localisation patterns, which have been described in the literature previously. In the second part of the study we investigated the analytical performance of a pyrosequencing-based NGS instrument. At first, the homopolymer-detecting capabilities were tested using a plasmid system. The results showed decreasing reliability when detecting 5- and 6-mers. Our hypothesis, that the initiation of the sequencing reaction provides more accurate HP detection was not supported by the results. Thereafter 17 human DNA samples originating from CF patients were examined. We performed the NGS analysis of the *CFTR* gene using different amplification settings. In light of the obtained data, it can be stated that the instrument is applicable in the molecular genetic testing of CF patients, but a few cumbersome alterations must be verified by Sanger sequencing. At last we think, that our findings may contribute to the laboratory diagnostics of cystic fibrosis and the introduction of the newborn/carrier screening in Hungary in merits.

GRANT SUPPORT

Gazdaságfejlesztési és Innovációs Operatív Program (GINOP-2.3.2-15-2016-00039)

LIST OF PUBLICATIONS

UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRAY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/136/2018.PL
Subject: PhD Publikációs Lista

Candidate: Gergely Ivády
Neptun ID: TWPPFH
Doctoral School: Doctoral School of Molecular Cellular and Immune Biology
MTMT ID: 10037475

List of publications related to the dissertation

1. **Ivády, G.**, Madar, L., Dzsudzsák, E., Koczok, K., Kappelmayer, J., Krulisova, V., Macek, J. M., Horváth, A., Balogh, I.: Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system.
   BMC Genomics. 19, 1-8, 2018.
   DOI: http://dx.doi.org/10.1186/s12864-018-4544-x
   IF: 3.729 (2016)
2. **Ivády, G.**, Koczok, K., Madar, L., Gombos, É., Tóth, I., Győri, K., Balogh, I.: Molecular Analysis of Cystic Fibrosis Patients in Hungary - an Update to the Mutational Spectrum.
   J. Med. Biochem. 34, 1-6, 2015.
   IF: 0.742
3. **Ivády, G.**, Madar, L., Nagy, B., Gönczi, F., Ajzner, É., Dzsudzsák, E., Dvorakova, L., Gombos, É., Kappelmayer, J., Macek, J. M., Balogh, I.: Distribution of CFTR mutations in Eastern Hungarians: relevance to genetic testing and to the introduction of newborn screening for cystic fibrosis?
   J. Cyst. Fibros. 10 (3), 217-220, 2011.
   DOI: http://dx.doi.org/10.1016/j.jcf.2010.12.009
   IF: 3.19

UNIVERSITY AND NATIONAL LIBRAY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

List of other publications

4. Szánthó, E., Kárai, B., **Ivády, G.**, Bedekovics, J., Szegedi, I., Petrás, M., Ujj, G., Ujfalusi, A., Kiss, C., Kappelmayer, J., Hevessy, Z.: Comparative Analysis of Multicolor Flow Cytometry and Immunohistochemistry for the Detection of Disseminated Tumor Cells.
Appl. Immunohistochem. [Epub ahead of print], 2017.
DOI: http://dx.doi.org/10.1097/PAI.0000000000000519
IF: 1.634 (2016)

5. Oláh, A., Asztalos, L., **Ivády, G.**, Varga, É., Kovács, M. Á., Kappelmayer, J., Varga, J.: Monitoring of mycophenolic acid and kidney function during combined immunosuppressive therapy.
Clin. Chem. Lab. Med. 49 (11), 1849-1853, 2011.
DOI: http://dx.doi.org/10.1515/cclm.2011.678
IF: 2.15

6. **Ivády**, G., Bekéné Debreceni, I., Kissné, S. V., Hevessy, Z., Kappelmayer, J.: A timidin kináz aktivitás összehasonlító elemzése egyéb prognosztikai markerekkel krónikus lymphocytás leukémiában.
Klin. Kísérl. Lab. Med. 33 (2), 7-11, 2008.

7. Oláh, A., **Ivády, G.**, Kappelmayer, J.: Kényes paraméterek a kardiovaszkuláris betegségek diagnosztikájában.
Metabolizmus. 6 (Suppl.), 91-94, 2008.

8. Antal-Szalmás, P., **Ivády, G.**, Molnár, A., Hevessy, Z., Kissné, S. V., Oláh, A., Lenkey, Á., Kappelmayer, J.: "Turnaround time": a laboratóriumi eredménykiadás hatékonyságának új paramétere.
Orv. Hetil. 148 (28), 1317-1327, 2007.
DOI: http://dx.doi.org/10.1556/OH.2007.28087

Total IF of journals (all publications): 11,445
Total IF of journals (publications related to the dissertation): 7,661

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of Web of Science, Scopus and Journal Citation Report (Impact Factor) databases.

07 May, 2018