Applications of next-generation sequencing technique in clinical diagnosis of osteogenesis imperfecta and Wilson's disease

Thesis

Kristóf Árvai

Semmelweis University
Doctoral School of Clinical Medicine

Supervisor: Prof. Dr. Péter Lakatos, DSc., academic lecturer

Offical reviewers: Dr. Péter Reismann, PhD, associate professor
Dr. Veronika Karcagi, PhD, senior research fellow

Head of Final Examination Board:
Prof. Dr. Attila Mócsai, DSc academic lecturer

Members of the Board:
Dr. Klára Werling, PhD, associate professor
Dr. ifj. Csaba Kiss, PhD, head of department

Budapest
2019

1 INTRODUCTION

In the 1970s, Sanger et al. developed a DNA sequencing technique based on fragmentation and chain termination. This marked the beginning of the transformation of molecular biology, with the availability of a tool that made it possible to study first whole genes and then whole genomes. The Human Genom Project required huge financial, time, and human effort, resulting the largest biological-medical collaboration in history and a cost of $ 3 billion. This makes it clear that new technology is needed for DNA sequencing, which is both faster and cheaper. This was the beginning of the development and competition for technologies called generic next generation sequencing (NGS). A common feature of these methods is that due to massive parallelization, hundreds of thousands or even millions of bases are sequenced simultaneously with direct detection without gel electrophoresis. In 2010, Ion Torrent Personal Genome Machine (PGM) sequencer was released. The machine did not contain an expensive camera system and the nucleotides used in the sequencing reaction did not undergo any chemical modification compared to the native form. This made it possible to develop a smaller, cheaper and faster device, the first "desktop" sequencer. The first versions generated readings of 100 base pairs using semiconductor sequencing techniques.

It is now possible to produce sequences up to 600 bp in length, with a maximum capacity of 130 million readings. These discoveries and advances in technology and information technology have led to a revolution in molecular and human genetics that can lead to a truly personalized era of precision medicine. In our study, Ion Torrent sequencing techniques were used to investigate hereditary diseases that are known to be associated with genes which contains a greater number of coding sequences (Wilson's disease) or with multiple large genes (osteogenesis imperfecta), which often results in a very long diagnostic procedure for conventional sequencing with high cost requirements and turnaround times. Furthermore, the phenotypes of these diseases may overlap with the symptoms of other syndromes, so genetic testing can provide important information for differential diagnosis.

2 OBJECTIVE

Osteogenesis imperfecta is a rare inherited disorder that primarily causes bone symptoms. The disease can be manifested in a wide range of signs, such as bluish discolouration of the eye, hearing loss, tooth malformation and frequent fractures of the bone. There are currently nine types. Although some types may have different clinical

manifestations, distinguishing them from other (nearly a dozen) syndromes can often cause difficulties. The importance of accurate genetic diagnosis not only identifies each OI subtype or distinguishes the disease from other overlapping phenotypes with multiple bone fractures or reduced bone mass, but also helps to determine whether recurrent fractures may be due to child abuse. A further diagnostic difficulty is that some clinical manifestations of OI, such as blue sclera, may be normal in infants, or a genetic disease may occur without a family history if a new, de novo mutation occurs.

Wilson's disease is also a rare, recessive disease that causes the loss of the function of a protein that is important for the transport of copper ions. Symptoms typically include progressive liver damage (cirrhosis and fibrosis may occur) and, if untreated, may result in severe CNS damage. In conventional clinical tests, it is often difficult to perform a differential diagnosis of diseases that produce similar symptoms. The gene currently linked to the above syndrome contain dozens of coding sequences without mutation hot spot, so next-generation sequencing may be the most appropriate method to investigate these.

The aim of our study was to develop a fast, precise, cost-effective diagnostic test method for the more common types of osteogenesis imperfecta and genes underlying Wilson's disease using next-generation sequencing techniques. Applying the capability of this new technology the patient and the clinician can reach a clear diagnosis much sooner, which can be of immediate therapeutic importance, especially in Wilson's disease. Genetic diagnostics of patients with osteogenesis imperfecta syndrome may contribute to more accurate prognosis, and prenatal screening in future family members may even avoid the inheritance of the disease or play a role in the early detection of the disease. We intend to implement the evaluation of the genetic differences found by bioinformatics methods, with a multi-step evaluation. If necessary and possible, within-family segregation analysis of mutations is also performed.

3 METHODS

3.1 Designing Gene Panels

Two different techniques were selected for the two assays: the OI gene panel was designed by Agilent HaloPlex, and the ATP7B gene was targeted to be amplified by AmpliSeq using multiplex PCR to construct the DNA library. Target genes or

regions were selected primarily based on information from the NIH Genetic Home Reference and disease specific databases. Given that the alterations in the COL1A1 and COL1A2 genes account for more than 90% of the cases, we designed the HaloPlex probes with the four most commonly mutated genes: collagen type 1 alpha-1 and alpha-2, exons of the P3H1 (LEPRE1) and CRTAP regions were targeted using Agilent SureDesign software.

In addition, the 21 coding exons of the ATP7B gene and a few intronic regions have been considered in the design of the genetic diagnosis of Wilson's disease. A polymorphism of the PRNP gene is known to influence the severity of neurological symptoms in Wilson's disease, and this gene was included in the study.

3.2 Collection of Biological Samples

Six patients (2 females and 4 males) with osteogenesis imperfecta were enrolled in our study with the help of Semmelweis University II. Dept. of Pediatric Clinic. Six patients (1 female and 5 male) with clinical signs of Wilson's disease were referred to Semmelweis University. Peripheral blood samples were collected with the assistance of the I. Department of Internal Medicine.

3.3 DNA Isolation

From the collected peripheral blood samples, genomic DNA was isolated using the Reliaprep Blood gDNA Miniprep System (Promega, Fitchburg, WI, USA) kit.

3.4 Agilent Haloplex Library Construction

The genomic DNA was digested with a mixture of 16 restriction enzymes, and the resulting DNA fragments were hybridized to HaloPlex probes complementing the target region, which also contained sequences of sequencing adapters and barcodes required for sequencing.

3.5 Ampliseq Library Construction

Genomic DNA was amplified by targeted multiplex PCR and the resulting amplicons were partially digested and ligated to the sequencing adapters using the Ion AmpliSeq ™ Library Kit 2.0 (Thermo Fisher, Waltham, MA, USA).

3.6 Next Generation Sequencing on Ion Torrent PGM

DNA libraries cannot be directly sequenced by the Ion Torrent technique, they must attach first to the surface of a sequencing bead and amplified there. A suitable technique is emulsion PCR, whereby PCR reactions inside the oil droplets can be performed in a closed, multi-million-fold parallel fashion. Ion 314 (expected throughput of 500,000 reads) and 316 (expected throughput of 3,000,000 sequences) v2 BC chips were used in

our study. A well on a chip can contain one sequencing bead. Wells have an ion sensor that can detect a change in pH or voltage caused by H + released from the linked nucleotides during synthesis of the complementary strand of DNA. The nucleotides are added to each type, one by one, in a strictly defined flow order. If the four bases were present at the same time, we would have no information about the order of the nucleotides to be incorporated. After the reaction, the device washes the chip surface with wash buffer and removes any remaining nucleotides before the next type of nucleotide arrives.

3.7 Bioinformatics

The raw voltage values collected in DAT files are processed by a platform-specific algorithm and translated to bases. Sequences obtained during base calling are collected by Torrent Suite software in BAM files, sorted by barcode. This is followed by align the readings to the reference genome. In the mapped sequences it is possible to search for alterations from the reference and to extract variants. The final step involves the clinical interpretation of appropriately filtered genetic differences, their comparison with the patient's symptoms, and the processing of information in the literature or databases.

4 RESULTS

4.1 Patient data

Six patients were enrolled for osteogenesis imperfecta study. Four male and two female patients underwent genetic testing. Their age ranged from 1 to 41 years (mean 12 years) and their fractures ranged from 1 to 10 (mean 3). Where available, we also used bone density, ophthalmic and auditory data from patients as well as family history. The following table summarizes patients with OI:

	Gender	Age (yrs)	Number of fractures	BMD (L2-L4) g/cm$_2$	Variants with clinical significance	Family history	Blue sclera	Hearing loss
Patient 1	Male	9	3	0.381	c.391C>T (*COL1A1*)	Father and paternal grandmother has OI	Yes	N/A
Patient 2	Male	41	10	N/A	c.391C>T (*COL1A1*)	Mother has OI	No	Yes
Patient 3	Male	18	3	0.685	c.2072G>A (*COL1A2*)‡	No	No	Yes
Patient 4	Female	2	3	0.390	c.189C>A (*COL1A1*)	Mother has blue sclera, but no fractures	Yes	N/A
Patient 5	Female	1	3	0.170	c.811G>T (*COL1A2*)	Mother had 7 fractures	No	No
Patient 6	Male	1	1	0.167	c.750+1G>A (*COL1A1*)	No	Yes	N/A

History of non-traumatic fractures: Each patient had two fracture events with three broken bones (tibia, fibula and radius). Patient two had about ten fractures. Patient 3 had 3 fractures with three bones (skull, collarbone and femoral neck). Patient 4 also had 3 fractures with three bones (tibia, femur and

humerus) injured. Patient 5 had two bones broken due to three events (tibia, femur). Patient 6 had a fracture of the femur without trauma.

Bone mineral density (BMD) was measured on the lumbar spine by dual energy X-ray absorption metry. (DPX-L, Lunar Corp. Madison, WI, USA). Measured values show a decrease in bone density compared to healthy bone tissue. Patient number one is the son of patient number two.

Six patients (five male and one female) with Wilson's disease symptoms were also enrolled. Their age ranged from 8 to 44 years (mean 18.6 years). Clinical information related to ophthalmic and copper metabolism were also available. Haemolytic anemia did not occur in our patients. Patient data is summarized in the following table:

	Gender	**First signs (year)**	**KFR**	**Neu**	**Urine copper**	**Cerul. (g/L)**	**WD score**	**Phenotype**	**ATP7B genotype**
P1	Female	12	Yes	No	++	0,18	6	T	p.M769fs/ p.H1069Q
P2	Male	17	No	Yes	+	0,05	6	N1	p.A1063V/ p.H1069Q
P3	Male	8	Yes	No	++	0,06	8	H2	p.G1351X/ p.H1069Q
P4	Male	17	Yes	No	+	0,03	5	H2	p.A1135fs/ p.L1305P
P5	Male	44	No	No	++	0,08	4	H1	p.A1270I/ c.1707+2dupT
P6	Male	14	Yes	Yes	NA	0,04	7	N2	p.R969Q/ p.H1069Q

KFR: Kayser-Fleischer ring; Neu: neurological symptoms and/or CT/MRI abnormalities; urin copper: 1-2X ULN: +, 2x ULN or positive D-penicillamine test: ++; Cerul: cöruloplasmine; T: brother of a former patient;

H1: acute liver failure; H2: chronic liver failure; N1: neurological symptoms with liver failure; N2: only neurological symptoms. According to the international score system, a diagnosis of Wilson's disease with a score of 4 or higher is highly likely. Hemolytic anemia was not observed in any patient. Number three also had a liver biopsy, which gave a positive result for Rhodanine staining.

4.2 Quality Control of Sequencing Data

For OI patients investigated with HaloPlex technique, the average number of reads per sample was 436,086, and we succeeded in sequencing 98.65% of the planned regions, reaching an average of 779X coverage. The cut-off value was chosen that at least 90% of the targets had a coverage of at least 20X, which was achieved for each sample. A total of 23 different variants were identified in the patients (4 pathogenic, 1 uncertain, and 18 benign).

During AmpliSeq library construction using multiplex PCR, an average of 134,386 sequences were generated per patient, resulting in an average of 99.46% of the Wilson disease-related regions being sequenced. The average coverage per base was 1883X. The number of identified variants ranged from 8 to 13 per patient, of course, most of which were polymorphisms without clinical effect.

4.3 Mutations in the patients diagnosed with osteogenesis imperfecta

Three new genetic variants were identified. Two of these (NM_000088.3 (COL1A1): c.189C>A and NM_000089.3

(COL1A2): c.811G>T) were evaluated as pathogenic. The heterozygous c.189C>A variant is located in the second exon of the COL1A1 gene and creates a premature stop codon. The heterozygous c.811G>T alteration causes the 271 glycine amino acid to be replaced by cysteine in the product of the COL1A2 gene. This genomic position was previously described in exon 17 as a site of deleterious mutation, but in that case a G>C base exchange and thus a glycine>arginine amino acid change took place. So far, G>T substitution has not been reported in the literature. All of our protein function prediction software (SIFT, PolyPhen-2) have classified this variant as deleterious. For both amino acid substitutions, the neutral charged, nonpolar glycine is replaced by a single charged, polar side chain amino acid. A targeted analysis of the c.811G>T mutation in the affected family was performed and the mutation was found in a heterozygous form only in members with OI phenotype.

In exon 31 of the COL1A2 gene, a heterozygous (NM_000089.3 (COL1A2): c.2072G>A) variant was identified that causes the replacement of glycine 691 with asparagine. All of the prediction software used indicate that the mutation will alter the protein structure and does not occur in healthy

population. However, we did not have segregation data, so we classified this finding as an uncertain variant.

Two heterozygous pathogenic mutations previously described in the literature were found in two of our patients: NM_000088.3 (COL1A1): c.391C>T and NM_000088.3 (COL1A1): c.750 + 1G>A. The c.391C>T substitution induces an early stop codon (p.Arg131Ter), whereas the c.750+1G>A mutation in intron 10 causes a splicing site change that will lead to abnormal exon-intron splicing.

4.4 Mutations in patients diagnosed with Wilson's disease

A total of nine disease-causing mutations were found. Most frequently, c.3207C>A (p.H1069Q) was found in exon 14 of the ATP7B (NM_000053.3) gene in a heterozygous form in four patients. The base exchange results in an amino acid change in the protein's ATP loop structure, with glutamine being incorporated into the protein chain instead of histidine.

We identified three missense heterozygous mutations previously reported in the literature. In exon 13, arginine 969 was exchanged for glutamine (p.Arg969Gln, TM6 domain), in exon 14, valine (p.Ala1063Val, ATP loop) was substituted in position 1063, and in exon 19 at position 1305, leucine mutated to proline (p.Leu1305Pro, ATP hinge / TM7 domain). We also found a new mutation in our patient cohort: in exon 18, the

1270th alanine changed to isoleucine (p.Ala1270Ile, ATP hinge domain).

We found three heterozygous frameshift mutations: in exon 4 a c.1707+2dupT abnormality disrupt the exon-intron splicing process, in exon 8 and in exon 15 occurred a single nucleotide deletion (c.2304delC and c.3402delC).

In exon 20, a heterozygous stop codon mutation causing early protein chain termination was identified, in the 1351th glutamine (p.Gln1351Ter). All mutations were successfully confirmed by bidirectional Sanger sequencing, and there was no false positive result in the next generation sequence datasets.

5 CONCLUSIONS

As a result of the recent advances in laboratory and information technology, genetic tests that previously consumed enormous resources can be carried out at a fraction of the initial cost, in much less time.

Our aim was to investigate the applicability and feasibility of next-generation sequencing in the clinical diagnosis of rare bone and metabolic diseases.

According to our results, DNA library preparation based on both sequence capture and multiplex PCR is a robust, highly efficient method. Its cost requirements are comparable, but obtaining the larger amount of primers required for multiplex

PCR is a greater investment at the start of the assays. However, later this will not increase the cost of the tests. With HaloPlex technology, the cost of testing is evenly distributed, with the amount you need to buy for each kit.

A significant difference between approaches is the necessary sequencing data of individual DNA libraries. Because the DNA fragments captured by the overlapping HaloPlex probes significantly extend beyond the targeted coding regions, the required minimum number of reads may be significantly higher. Due to overlapping probes, a variant found may result from multiple DNA fragments, reducing the likelihood of false positives. Sequence capture is a slightly more complex laboratory procedure that requires more practice.

In the multiplex PCR approach, only the targeted regions are amplified and thus examined, making it easier to design a sequencing run. The disadvantage of this method is that as the number of amplicons increases, so does the initial primary demand and, on the other hand, the likelihood of the amplicons unbalance increases. This may result in unnecessarily high number of reads of some fragments, while insufficient sequence data on other amplicons.

In the study of osteogensis imperfecta disease, we did not target all the genes known to be associated with the disease on

purpose. We aimed to detect mutations in the genes responsible for more than 90% of the disease, which was successful in all six cases. HaloPlex sequence capture has been highly effective in enriching selected genomic regions and has proven to be reliable in variant information. Based on the found mutations, it can be concluded, in accordance with literature, that osteogenesis imperfecta is primarily caused by mutation of one of the COL1A1 and COL1A2 genes encoding the α-chains of collagen type I. Mutations typically affect the structure of the protein, most often causing amino acid changes. There may be nonsense mutations that cause early termination of protein synthesis or damage to the exon-intron splicing process.

The genes tested consist of a large number of coding exons (126 in total), which analysis would be a very time consuming and costly process to perform by conventional Sanger sequencing, and there is always the possibility of false negativity using the pre-screening method due to their lower analytical values.

For Wilson's disease, a multiplex PCR-based AmpliSeq technique was used. Creating a DNA library is slightly simpler than HaloPlex, but it has proven to be equally reliable. All of the patients in the study were able to diagnose the compound heterozygous genotype causing Wilson's disease. Acute liver failure, which is a symptom of the disease, is also accepted by

international organizations as an indication of emergency liver transplantation, but requires rapid diagnosis, which can be ensured by next-generation sequencing. The mutations found cause deleterious changes in the various domains of Atp7b, and mutation hotspot has not been confirmed in this disease, in accordance with literature data.

The Ion Torrent semiconductor-based next-generation sequencing technique has also proven to be capable of identifying diverse molecular structure substitutions and insertion/deletions. The device requires little space on the bench and does not require any special infrastructure to operate. Preparation of the sequencing run takes approximately two and a half hours, followed by sequencing itself. The graphical interface integrated into the platform-specific bioinformatics algorithm collection is easy to use and easy to customize. The variant calling is fully parameterizable but requires in-depth knowledge of the platform. By pairing the latest sequencing chemistry and software, homopolymer errors can be almost completely eliminated

Overall, next-generation sequencing allowed for previously unavailable clinical genetic testing, while its reliability remained comparable to conventional Sanger sequencing.

6 LIST OF PUBLICATIONS

Publications related to the thesis:

1. **Árvai K**, Horváth P, Balla B, Tobiás B, Kató Karina, Kirschner Gy, Klujber V, Lakatos P, Kósa JP: Next-generation sequencing of common osteogenesis imperfecta-related genes in clinical practice. Sci Rep. 2016 Jun 23;6:28417.
   IF: 4,259
2. Németh D, **Árvai K**, Horváth P, Kósa JP, Tobiás B, Balla B, Folhoffer A, Krolopp A, Lakatos P, Szalay Ferenc: Clinical Use of Next-Generation Sequencing in the Diagnosis of Wilson's Disease, Gastroenterology Research and Practice, vol. 2016, Article ID 4548039, 6 pages, 2016.
   IF: 1,863
3. **Árvai K**, Dr. Kósa J, Dr. Horváth P, Dr. Balla B, Dr. Tobiás B, Dr. Takács I, Dr. Nagy Zs, Dr. Lakatos P: Osteogenesis Imperfecta rutin genetikai diagnosztikája új generációs szekvenálási (NGS) technológiával. Magy Belorv Arch 2013; 66:
   IF: --

Publications not related to the thesis:

1. H Barti-Juhász, A Pazsitka, B Jóri, **K Árvai**, D Erös, Gy Kéri, I Peták, R Mihalik: Altered Expression of a Broad Range of Protease Genes in SH-SY5Y Neuroblastoma Cells after Bortezomib Treatment Roche Cancer Research Application Note No. 2 2009.
   IF: --
2. **Árvai K**, Nagy K, Barti-Juhász H Peták I, Krenács T, Micsik T, Végső Gy, Perner F, Szende B: Molecular profiling of parathyroid hyperplasia, adenoma and carcinoma. Pathol Oncol Res. 2012 Jul;18(3):607-14.
   IF: 1,555
3. Balla B, **Árvai K**, Horváth P, Tobiás B, Takács I, Nagy Zs, Dank M, Fekete Gy, Kósa J P, Lakatos P: Fast and Robust Next-Generation Sequencing Technique Using Ion Torrent Personal Genome Machine for the Screening of Neurofibromatosis Type 1 (NF1) Gene. J Mol Neurosci. 2014 Jun;53(2):204-10.
   IF: 2,343
4. **Árvai K**, Horváth P, Balla B, Tőkés A, Tobiás B, Takács I, Nagy Zs, Lakatos P, Kósa J P: Rapid and cost effective screening of breast and ovarian cancer genes using novel sequence capture method in clinical samples. Fam Cancer. 2014 May 23.
   IF: 1,977
5. Balla B, Tobiás B, Kósa J P, Podani J, Horváth P, Nagy Zs, Horányi J, Járay B, Székely E, Krenács L, **Árvai K**, Dank M, Putz Zs, Szabó B, Szili B, Valkusz Zs, Vasas B, Győri G, Lakatos P, Takács I: Vitamin D-neutralizing CYP24A1 expression, oncogenic mutation states and histological findings of human papillary thyroid cancer. J Endocrinol Invest. 2014 Sep 9.
   IF: 1,994
6. Donáth J, Speer G, Kósa J P, **Árvai K**, Balla B, Juhász P, Lakatos P, Poór Gy: Polymorphisms of CSF1 and TM7SF4 genes in a case of mild juvenile Paget's disease found using next-generation sequencing. Croat Med J. 2015 Apr 20;56(2):145-51.

IF: 1,483

7. Tobiás B, Halászlaki Cs, Balla B, Kósa J P, **Árvai K**, Horváth P, Takács I, Nagy Zs, Horváth E, Horányi J, Járay B, Székely E, Székely T, Győri G, Putz Zs, Dank M, Valkusz Zs, Vasas B, Iványi B, Lakatos P: Genetic Alterations in Hungarian Patients with Papillary Thyroid Cancer. Pathol Oncol Res. 2016 Jan;22(1):27-33.
   IF: 1,736
8. Kirschner Gy, Balla B, Horváth P, Kövesdi A, Lakatos G, Takács I, Nagy Zs, Tóbiás B, **Árvai K**, Kósa J P, Lakatos P: Effects of imatinib and nilotinib on the whole transcriptome of cultured murine osteoblasts. Mol Med Rep. 2016 Sep;14(3):2025-37.
   IF: 1,692
9. Balla B, Sárvári M, Kósa J P, Kocsis-Deák B, Tobiás B, **Árvai K**, Takács I, Podani J, Liposits Zs, Lakatos P: Long-term selective estrogen receptor-beta agonist treatment modulates gene expression in bone and bone marrow of ovariectomized rats. J Steroid Biochem Mol Biol. 2019 Apr;188:185-194.
   IF: 3.785
10. Kocsis-Deák B, dr. Balla B, **Árvai K**, dr. Tobiás B, dr. Győri G, dr. Járay B, dr. Székely E, Podani J, dr. Kósa J, dr. Lakatos P: A pajzsmirigygöbök genetikai vizsgálata újgenerációs szekvenáláson alapuló platformon kifejlesztett génpanel segítségével. Orvosi Hetilap, 2019 Sep;160(36):1417-1425.
    IF: 0.564