Review Article

Clinical Application of NGS Tools in the Diagnosis of Collagenopathies

Francesca Cortini[1,2*], Barbara Marinelli[1], Angela Cecilia Pesatori[1], Manuela Seia[2], Agostino Seresini[2,3], Valentina Giannone[2] and Alessandra Bassotti[4]

[1]Department of Clinical Sciences and Community Health, University of Milan, IRCCS Ca' Granda Foundation Via San Barnaba 8, 20122 Milan, Italy; [2]Genetics Laboratory, IRCCS Ca' Granda Foundation, via Francesco Sforza 35, Milan, Italy; [3]Fondazione Grigioni per il Morbo di Parkinson, Via Gianfranco Zuretti 35, 20125 Milan, Italy; [4]Regional Center of Ehlers-Danlos Syndrome, IRCCS Ca' Granda Foundation via San Barnaba 8, Milan, Italy

Abstract

Collagenopathies are heterogeneous diseases that affect collagen proteins, which are ubiquitous in the body and characterized by the distinctive amino acid sequence Gly-X-Y. Next-generation sequencing (NGS) has gained an increasingly essential role in improving our understanding of the molecular bases of heterogeneous diseases like collagenopathies. In the last decades new NGS tools have been developed, such as whole exome sequencing (WES) and custom target sequencing, and they have become efficient and cost-effective methods for clinical diagnosis. In this review, we discuss the relevance of WES and custom target sequencing in the clinical diagnosis of collagenopathies.

Introduction

Collagen proteins are structural proteins in the connective tissue, representing the most abundant proteins in the body. Twenty-nine types of collagen have been described to date, the basic structure of all is composed of three α-chains intertwined in a triple helix. Each chain is composed of Gly-X-Y repeats, with glycine found at third position of the triple helix because it is the only amino acid small enough to fit in the center of the helix. The X is often proline and the Y is often hydroxyproline and they largely contribute to the stabilization of collagen protein structure.

Collagen proteins are divided into three groups based on their structure, as follows: fibrillary collagen, non-fibrillar collagen, and fibril-associated collagens with interrupted triple helixes (FACIT).[1] Collagen proteins are widespread in the body. Consequently, mutations in the genes encoding collagen proteins affect many organ systems (manifesting as collagenopathies). An overview of the collagens, their distribution in the body and associated diseases are shown in Table 1.

Genetic confirmation is important to corroborate a suspected diagnosis. So far, Sanger sequencing has been the gold standard method,[2] but it only allows for analysis of one DNA segment at time and is laborious and time consuming. The gene-by-gene Sanger sequencing approach is neither inexpensive nor efficient for heterogenous diseases such as collagenopathies.[3] Over the past few years, next-generation sequencing (NGS) has experienced a growing role in enabling the analysis of multiple regions of a genome in a single reaction and has been shown to be a cost-reductive and efficient tool in investigating patients with collagenopathies.

In this review, we explore all aspects of NGS tools in the clinical diagnosis of collagenopathies.

NGS: an overview

The NGS process starts with DNA extraction, with sample materials being most commonly obtained from peripheral leukocytes of blood samples or, in rare cases, from other tissues such as saliva or buccal swab. The DNA is broken into short fragments and amplified using PCR or hybridization approaches. The amplified regions could include a particular group of genes (target approach) or all genes in the genome.[4] In the case of sequencing of all genes in the genome, two different approaches are possible: whole exome sequencing (WES), if only the protein-coding regions are amplified; or whole genome sequencing (WGS), if the target is the entire genome.

Keywords: Next-generation sequencing; Whole exome sequencing; Collagen; Bioinformatics analysis; Clinical diagnosis.

Abbreviations: BAM, binary alignment map; EDS, Ehlers-Danlos syndrome; FACIT, fibril-associated collagens with interrupted triple helixes; GATK, genome analysis toolkit; NGS, next-generation sequencing; OI, osteogenesi imperfecta; SAM, sequence alignment map; VCF, variant calling format; VUS, uncertain significance; WES, whole exome sequencing; WGS, whole genome sequencing.

Received: April 30, 2017; Revised: August 14, 2017; Accepted: August 29, 2017

[*]**Correspondence to:** Cortini Francesca, Department of Clinical Sciences and Community Health, University of Milan, Via San Barnaba 8, 20122, Milan, Italy. Tel: +39 02-55032433, Fax: +39 02-55032353, E-mail: francesca.cortini@guest.unimi.it

How to cite this article: Cortini F, Marinelli B, Pesatori AC, Seia M, Seresini A, Giannone V, Bassotti A. Clinical Application of NGS Tools in the Diagnosis of Collagenopathies. *Exploratory Research and Hypothesis in Medicine* 2017;2(3):57–62. doi: 10.14218/ERHM.2017.00010.

Table 1. Collagens genes currently associated with atomic areas of expression and diseases

Collagene gene	Chromosome	Areas of expression	Disease
COL1A1 (collagen type I)	chr17q21.33	Skin, tendon, bone, ligament	Osteogenesis Imperfecta I, II, III, IV, Caffey disease, Ehlers-Danlos type I, VII
COL1A2	chr7q21.3		Osteogenesi Imperfecta II, III, IV, Ehlers-Danlos typeVIIB
COL2A1	chr12q13.11	Cartilagen, vitreous humor of eye, cornea	achondrogenesis, chondrodysplasia, early onset familial osteoarthritis, SED congenita, Langer-Saldino achondrogenesis, Kniest dysplasia, Stickler syndrome type I, spondyloepimetaphyseal dysplasia Strudwick type
COL11A1	chr1p21.1	Cartilage, nucleus pulposus, cornea, inner ear	Stickler syndrome type II, fibrochondrogenesis, Marshal syndrome
COL11A2	chr6p21.32	Cartilage, nucleus pulposus, inner ear	Stickler syndrome type III, fibrochondrogenesis, Deafness dominant and autosomal recessive
COL9A1	chr6q13	Cartilage, vitreous, retina, inner ear	Multiple epiphyseal dysplasia type VI, Stickler syndrome type IV,
COL9A2	chr1p34.2		Multiple epiphyseal dysplasia type II, Stickler syndrome type V,
COL9A3	chr20q13.33		Multiple epiphyseal dysplasia type III, Multiple epiphyseal dysplasia type with myopathy
COL10A1	chr6q22.1	Hypertrophic chondrocytes in calcifying cartilage	Metaphyseal chondrodysplasia, Schmid type
COL3A1	chr2q32.2	Most connective tissue especially vessels, skin and tendons	Ehlers-Danlos type III, type IV
COL5A1	chr9q34.3	Most connective tissue especially skin, cornea, bone, tendon, placenta, fetal membranes	Ehlers-Danlos type I, type II
COL5A2	chr2q32.2		Ehlers-Danlos type I, type II
COL4A1	chr13q34	Basemant membranes	porencephaly, cerebrovascular disease, and renal and muscular defects
COL18A1	chr21q22.3	Basemant membranes	Knobloch syndrome
COL6A1	chr21q22.3	Most connective tissue, tendons, contributes to cell matrix adhesion in skeletal muscle	Bethlem myopathy, Ulrich congenital muscular dystrophy
COL7A1	chr3p21.31	Anchoring fibrils in dermo-epidermal junctions	dystrophic epidermolysis bullosa
COL17A1	chr10q25.1	Component of hemidesmosomes	Junctional epidermolysis bullosa, non-Herlitz type

http://www.proteinatlas.org/

Amplified products can be loaded to various sequencing platforms, such as MiSeq (Illumina), HiSeq (Illumina) and Ion Torrent (ThermoFisher Scientific), to generate millions of short sequence reads (Fig. 1). These products are processed by bioinformatics packages following a previously established workflow. First, reads are aligned to reference genome and compared for similarities and differences at each target position. Then, a list of variants is generated, which is filtered through different software packages to determine significance. Usually, adopted filters are suitable for identifying the presence of rare, unreported or disease causing variants.[5]

NGS in clinical application

Custom target sequencing versus WES

NGS researchers have developed new and very innovative methods and protocols in clinical practice for genetically heterogeneous diseases like collagenopathies. The NGS techniques, such as WES and custom target sequencing, offer several advantages in such applications. Clinicians should choose the right strategy for clinical analysis based on: 1) disease model; 2) region of interest; and 3)

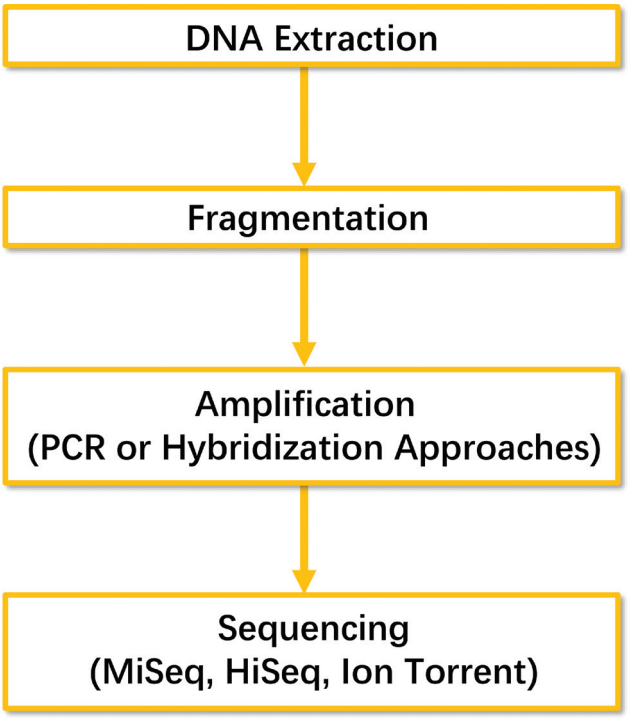

Fig. 1. NGS process: different steps. Abbreviation: NGS, next-generation sequencing.

depth of coverage (the average number of times that a particular nucleotide is present in a data position in a collection of random of sequences).

WES is an appropriate strategy in situations where conventional single-gene sequencing or a genes panel may not be appropriate because a pertinent genetic test has not been developed or because of genetic heterogeneity, atypical clinical presentation or lack of knowledge of the causal gene.[6,7] Moreover, the cost of WES is very attractive compared to that of custom target sequencing. The price of WES is currently around $200–300 per patient to sequence the entire exome, while that of custom target sequencing is around $100–200 to sequence only a few genes (the cost is calculated only for chemical reagents).

Usually, the minimum optimal depth of coverage is 20X spanning at least 80% of targeted bases. CG-rich regions, such as CpG islands, could decrease the depth of coverage because these regions denature, causing difficulties during amplification.[8] It is important for the success of the experiment and for a correct variant analysis to maintain the uniformity of coverage. Nevertheless, custom target sequencing is the best method if the genes clinically related to disease are known. The optimal coverage is 300X, higher than WES can provide, spanning at least the 99% of targeted bases. In this case, analysis will include only genomic regions that are comprised in the custom panel; obviously, complexity zones, such as CpG islands, will be excluded when panels are composed (Fig. 2). Moreover, the main advantage of custom target sequencing is the possibility to personalize the panel (*i.e.* inclusion of certain genes and the possibility to sequence exons, specific intronic regions, promoter regions or the 3′ untranslated region).

Current challenge of NGS tools in the diagnosis of collagenopathies

NGS technologies have revolutionized clinical testing for the rare genetic diseases, such as collagenopathies. As we know, collagenopathies are heterogeneous diseases and clinical phenotypes are often overlapping. NGS tools (WES or custom target panel) give the opportunity to analyze defined regions or the entire exome to identify the genes responsible for the disease.[6,7] Moreover, NGS offers the possibility to sequence multiple genomic regions and a

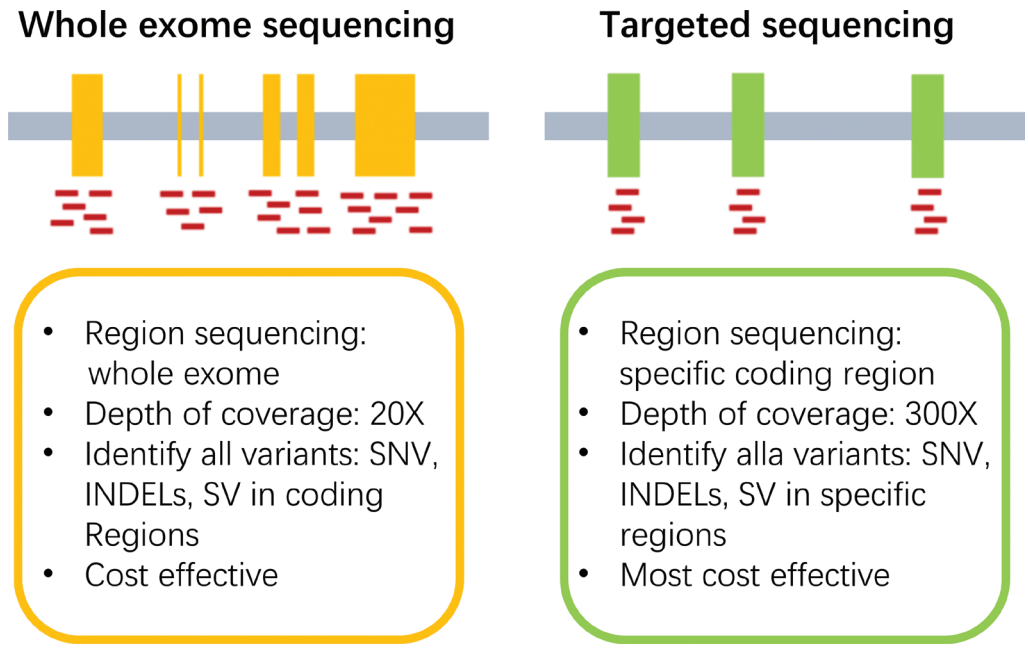

Fig. 2. Different properties of WES and Target Sequencing. The strategy of WES and Target Sequencing is different; WES is for all exons in genome, Target Sequencing is for specific regions. Abbreviation: WES, whole exome sequencing.

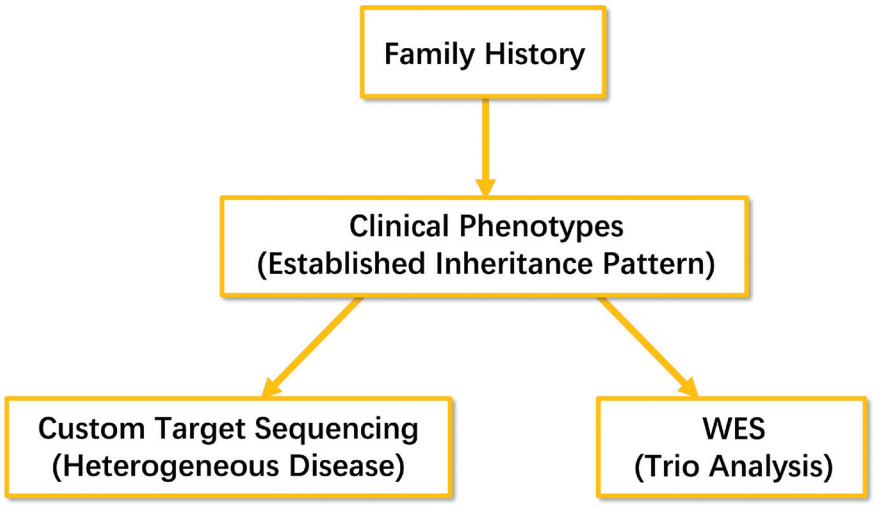

Fig. 3. Workflow for choosing a rigth NGS method in Collagenopathies diseases. Abbreviation: NGS, next-generation sequencing.

multitude of patients in a single reaction. The management of huge amounts of NGS data is the biggest challenge for laboratories. As such, it is important to standardize the workflow to correctly identify and interpret genetic variants.

NGS and collagenopathies

Collagenopathies are a heterogeneous group of hereditary disorders of connective tissue. Genetic defects of collagen formation affect almost every organ system and tissue in the body. They can be grouped based on clinical phenotypic characteristics. Moreover, collagenopathy phenotypes are often overlapping, and it is very difficult to distinguish them. NGS is the best and most efficient methodology to analyze patients affected by collagenopathies (Table 1). By means of NGS, it is possible to sequence the whole exome or a particular list of genes (the custom target sequence) to detect genetic variants, in situations where Sanger sequencing (the gene-by-gene approach) would be too costly and time consuming.[2]

The usefulness of NGS is helping clinical counseling and identifying genetic variants in patients with unclear phenotypes. Moreover, NGS tools are relevant to understanding the genotype-phenotype relationship in heterogeneous diseases, such as Ehlers-Danlos syndrome (EDS).[9] Weerakkody *et al*.[9] analyzed 177 EDS patients with two different custom panels composed of five collagen genes and aortopathy genes (aortopathy represents the vascular component of the EDS phenotype) respectively. The researchers identified 28 pathogenetic variants in *COL5A1/2, COL3A1, FBN1* and *COL1A1* and 4 likely pathogenetic variants in *COL1A1, TGFBR1/2* and *SMAD3* through their NGS assays. Twenty-two variants of uncertain significance were detected, seven of which were in aortropathy genes. Thus, NGS panels could represent a new method for molecular diagnosis beyond the expected EDS genotype-phenotype relationship and reveal new clinical variants in aortopathy genes.

Osteogenesi imperfecta (OI) is a disorder related to the collagenopathies. It is a heterogeneous bone disorder characterized by frequent fractures and seems to be inherited both in dominant and recessive manners. Mutations in the *COL1A1* and *COL1A2* genes are the demonstrated causes of different forms of OI and show autosomal dominant inheritance.[10,11] To date, a plethora of genes, responsible for both conditions of OI, have been identified as dominant and recessive in this disease.

WES is the best method to analyze such genes in a single experiment. Indeed, Caparros-Martin *et al*.[12] analyzed 42 OI probands, all offspring parents, to determine the spectrum of mutated genes and variants detected for these cases. This work confirmed that *COL1A1* mutations are responsible for the OI dominant form. It is necessary to investigate *COL1A1* in parents if the proband carries *COL1A1* mutations. Moreover, WES gives useful information on the positive role of genes such as *SCN9A* and *NTRK1* for the proband with no familiar history and for differential diagnoses of OI.

Clinical information: counseling, familiar history, clinical phenotype and best genetic approach

It is important that patients and healthcare familiars are counseled by clinicians or genetic counselors about the most appropriate NGS method to use. Moreover, it is imperative that clinicians or genetic counselors declare the specificity and limitations of the NGS method. It also has to be emphasized that positive results may not change the treatment or the prognosis.

The process of NGS organization starts with a detailed family history, to identify if there are individuals with the same phenotype and to individuate the possible inheritance pattern. In fact, it often happens that there is no relationship between genotype-phenotype within the same family, and these data confirm a different penetrance.[13]

The next step is describing the detailed phenotype of affected individuals. This could include evaluations by other specialists and application of other clinical exams or radiological tests. For example, for EDS, patients are checked by different specialists, such as a radiologist, neurologist and cardiologist.[14] Given the collected data from familiar history and phenotypes, the specialist decides the right NGS method for genetic analysis. In the case of genetic heterogeneity, custom target sequencing may be preferred. On the other hand, WES will be preferred for trio analysis in a case of a family group with different phenotypes (Fig. 3).

Interpretation of NGS results

To date, there are different sequencing chemistries available to ob-

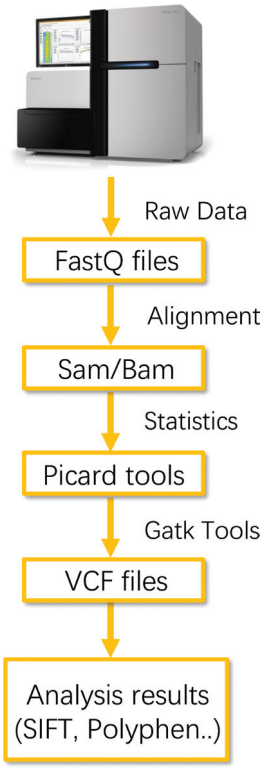

Fig. 4. NGS pipeline: Fastq files are mapping versus genomic reference to obtain SAM/BAM files. Statistics interpretation by Picard tools and manipulation by GATK tools generate VCF files. They include all variants that are analyze by in silico software as SIFT, Polyphen. Abbreviations: BAM, binary alignment map; GATK, genome analysis toolkit; NGS, next-generation sequencing; SAM, sequence alignment map; VCF, variant calling format.

tain genomic libraries and new NGS platforms to load samples, including MiSeq (Illumina), Ion Torrent (ThermoFisher Scientific) and NextSeq (Illumina). NGS platforms generate millions of reads that are processed bioinformatically. It is relevant to define a good pipeline to analyze these data. Fastq files (file storing biological sequence and its quality score), output data generated at the end of an NGS run, are processed by their quality scores, aligned to a reference genome and reads are filtered based on statistics and other kinds of information to obtain variant calling format (VCF) files for storing genetic variation data (Fig. 4).[15] The VCF files contain a list of variants that are classified as 1) pathogenetic, 2) likely pathogenetic, 3) uncertain significance (VUS), 4) likely benign and 5) benign (Table 2).[16]

Pathogenetic variants alter the protein functions and may have been previously reported in other affected individuals. In EDS types IV and VII and OI, pathogenetic variants often have different penetrance within the same family.[17] These could be missense, frameshift, small insertions or deletions. Benign variants are found in many individuals, including healthy subjects; in addition, they are often found in subjects tested by NGS. These are missense, intronic, synonymous or intergenic variants.[16] VUS are variants that could possibly affect protein function based on results from *in silico* software prediction tools (*i.e.* SIFT,[18,19] Polyphen,[20] *etc*.) and that are not described to affect other individuals. On the other hand, VUS are described in literature but the *in silico* analysis has revealed controversial results.

It is important that all genomic variants are compared to specific databases and literature data to understand their specific significance. Databases such as dbSNP and the Exome Aggregation Consortium are important to evaluate if these variants are polymorphisms (allelic frequency is >0.5%) or mutations. Disease databases such as the OI and EDS variant databases (http://www.le.ac.uk/ge/collagen/) or ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) are useful to interpret the properties of variants if there are relationships between variants and human health.

Future research perspective

NGS, also known as massive parallel sequencing, is being incorporated rapidly in the clinical laboratory testing routine. Recently, academic companies and institutions have continued technological research to improve NGS applications, such as WES and custom target sequencing. WES and custom target sequencing are useful for clinicians to understand the heterogeneous diseases like collagenopathies. We expect that NGS tools will become a routine clinical diagnostic test. The laboratory's challenge is now to standardize bioinformatic analysis, so as to make the NGS data interpretation more easily accessible. NGS produces a substantial amount of data, and the important issue is to have a unique and simple

Table 2. Variants classification[16]

Classification	Description
Pathogenetic	Contribute to the development of disease (some pathogenetic variant may not be fully penetrant) In the case of recessive or X-linked, a single pathogenetic variant may not be sufficient to cause disease. Additional data is not expected to alter the classification of this variant.
Likely pathogenetic	Very likely to contribute to the development of disease, however the scientific data is insufficient to confirm the pathogenicity Additional data is necessary to confirm this assertion of pathogenicity, but we cannot fully exclude the possibility that new evidence may demonstrate whether this variant has clinical significance or not
Uncertain significance (VUS)	Not enough information to support a definitive classification of this variant
Likely benign	Not expected to have a major effect on disease At this time, the scientific evidence is currently insufficient to prove its pathogenicity Additional evidence is expected to confirm this assertion, but we cannot fully exclude the possibility that new data may demonstrate that this variant can contribute to disease
Benign	Does not cause the disease

analysis workflow for genetic testing to allow rapid and correct identification of mutations and variants.

Acknowledgments

The authors wish to express their gratitude to Dr. Bice Strumbo for excellent technical assistance.

Conflict of interest

The authors have no conflict of interests related to this publication.

Author contributions

Having the idea of the review and adding lots of data about genetic of collagenopathies and different methods of NGS (FC), manuscript writing (BM, AS), supervisor of the NGS project (ACP), coordinator of analysis of NGS data (MS), organizing figures and tables (VG), collecting clinical data of collagenopathies and coordinator of clinicians (AB).

References

[1] Caretr EM, Raggio CL. Genetic and orthopedic aspects of collagen disorders. Curr Opin Pediatr 2009;21:46–54.
[2] Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 1975;94(3):441–448. doi:10.1016/0022-2836(75)90213-2.
[3] Neveling K, Feenstra I, Gilissen C, Hoefsloot LH, Kamsteeg Ej, Mensenkamp AR, *et al*. A post-hoc comparison of the utility of Sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. Hum Mutat 2013;34(12):1721–1726. doi:10.1002/humu.22450.
[4] Ballester LY, Luthra R, Kanagal-Shamanna R, Singh RR. Advances in clinical next-generation sequencing: target enrichment and sequencing. Expert Rev Mol Diagn 2016;16(3):357–372. doi:10.1586/14737159.2016.1133298.
[5] Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing. N Engl J Med 2014;370(25):2418–2425. doi:10.1056/NEJMra1312543.
[6] Alazami AM, Patel N, Shamseldin HE, Anazi S, Al-Dosari MS, Alzahrani F, *et al*. Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families. Cell Rep 2015;10(2):148–161. doi:10.1016/j.celrep.2014.12.015.
[7] Yu TW, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, Ataman B, *et al*. Using whole-exome sequencing to identify inherited causes of autism. Neuron 2013;77(2):259–273. doi:10.1016/j.neuron.2012.11.002.
[8] Veal CD, Freeman PJ, Jacobs K, Lancaster O, Jamain S, Leboyer M, *et al*. A mechanistic basis for amplification differences between samples and between genome regions. BMC Genomics 2012;13:455. doi:10.1186/1471-2164-13-455.
[9] Weerakkody RA, Vandrovcova J, Kanonidou C, Mueller M, Gampawar P, Ibrahim Y, *et al*. Target next-generation sequencing makes new molecular diagnoses and expands genotype-phenotype relationship in Ehlers-Danlos syndrome. Genet Med 2016;18(11):1119–1127. doi:10.1038/gim.2016.14.
[10] Chu ML, Williams CJ, Pepe G, Hirsch JL, Prockop DJ, Ramirez F. Internal deletion in a collagen gene in a perinatal lethal form of osteogenesis imperfecta. Nature 1983;304(5921):78–80. doi:10.1038/304078a0.
[11] Forlino A, Marini JC. Osteogenesis imperfecta. Lancet 2016;387(10028):1657–1671. doi:10.1016/S0140-6736(15)00728-X.
[12] Caparros-Martin JA, Aglan MS, Temtamy S, Otaify GA, Valencia M, Nevado J, *et al*. Molecular spectrum and differential diagnosis in patients referred with sporadic or autosomal recessive osteogenesis imperfecta. Mol Genet Genomic Med 2016;5(1):28–39. doi:10.1002/mgg3.257.
[13] Cortini F, Marinelli B, Romi S, Seresini A, Pesatori AC, Seia M, *et al*. A new *COL3A1* mutation in Ehlers-Danlos syndrome Vascular type with different phenotypes in the same family. Vasc Endovascular Surg 2017;51(3):141–145. doi:10.1177/1538574417692114.
[14] De Paepe A, MAlfait F. The Ehlers-Danlos syndrome, a disorder with many faces. Clin Genet 2012;82(1):1–11. doi:10.1111/j.1399-0004.2012.01858.x.
[15] Danecek P, Anton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al*. The variant call format and VCF tools. Bioinformatics 2011;27(15):2156–2158. doi:10.1093/bioinformatics/btr330.
[16] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, *et al*. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015;17(5):405–424. doi:10.1038/gim.2015.30.
[17] Kuivaniemi H, Tromp G, Prockop DJ. Mutations in fibrillary collagens (types I, II, III and XI), fibril-associated collagen (type IX), and network forming collagen (type X) cause a spectrum of disease bone, cartilage, and blood vessels. Hum Mutat 1997;9(4):300–315. doi:10.1002/(SICI)1098-1004(1997)9:4<300::AID-HUMU2>3.0.CO;2-9.
[18] Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31(13):3812–3814. doi:10.1093/nar/gkg509.
[19] Flanagan SE, Patch AM, Ellard S. Using SIFT and Polyphen to predict loss of function and gain of function mutations. Genet Test Mol Biomarkers 2010;14(4):533–537. doi:10.1089/gtmb.2010.0036.
[20] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using Polyphen-2. Curr Protoc Hum Genet 2013;Chapter 7:Unit7.20. doi:10.1002/0471142905.hg0720s76.