RECOMMENDATIONS

Application of genome analysis strategies in the clinical testing for pediatric diseases

Yaqiong Jin[1*] | Li Zhang[2*] | Baitang Ning[3] | Huixiao Hong[3] | Wenming Xiao[3] | Weida Tong[3] | Yiran Tao[2] | Xin Ni[1] Tieliu Shi[2] | Yongli Guo[1]

[1]Beijing Key Laboratory for Pediatric Diseases of Otolaryngology, Head and Neck Surgery, MOE Key Laboratory of Major Diseases in Children, Beijing Pediatric Research Institute, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, China
[2]Center for Bioinformatics and Computational Biology, and the Institute of Biomedical Sciences, Shanghai Key Laboratory of Regulatory Biology, the Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai, China
[3]National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR, USA

Correspondence
Xin Ni, Beijing Key Laboratory for Pediatric Diseases of Otolaryngology, Head and Neck Surgery, MOE Key Laboratory of Major Diseases in Children, Beijing Pediatric Research Institute, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, China.
Email: nixin@bch.com.cn
Tieliu Shi, Center for Bioinformatics and Computational Biology, and the Institute of Biomedical Sciences, Shanghai Key Laboratory of Regulatory Biology, the Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai, China.
Email: tieliushi@yahoo.com
Yongli Guo, Beijing Key Laboratory for Pediatric Diseases of Otolaryngology, Head and Neck Surgery, MOE Key Laboratory of Major Diseases in Children, Beijing Pediatric Research Institute, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, China.
Email: guoyongli@bch.com.cn
*These authors contributed equally to the work.

Funding source
National Natural Science Foundation of China (81502144, 81472369, 31671377); National Key Research and Development Program of China (2015AA020108); Beijing Health System Top Level Health Technical Personnel Training Plan (20153079)

Disclaimer
The information in these materials is not a formal dissemination of the U.S. Food and Drug Administration.

Received: 20 April, 2018
Accepted: 22 May, 2018

DOI:10.1002/ped4.12044

ABSTRACT

Next-generation sequencing (NGS) is being used in clinical testing. Government authorities in both China and the United States are overseeing the clinical application of NGS instruments and reagents. In addition, the US Association for Molecular Pathology and the College of American Pathologists have jointly released a guidance to standardize the analysis and interpretation of NGS data involved in clinical testing. At present, the analysis strategies and pipelines for NGS data related to the clinical detection of pediatric disease are similar to those used for adult diseases. However, for rare pediatric diseases without linkage to known genetic variants, it is currently difficult to detect the relevant pathogenic genes using NGS technology. Additionally, it is challenging to identify novel pathogenic genes of familial pediatric tumors. Therefore, characterization of the pathogenic genes associated with above diseases is important for the diagnosis and treatment of rare diseases in children. This article introduces the general pipelines for NGS data analyses of diseases and elucidates data analysis strategies for the pathogenic genes of rare pediatric diseases and familial pediatric tumors.

KEYWORDS

Familial pediatric tumors, Next-generation sequencing, Rare pediatric diseases

INTRODUCTION

The Chinese government supports the development of gene sequencing industry and encourages product innovations to improve its safety and efficiency. In 2014, the China Food and Drug Administration (CFDA) approved the first group of next-generation sequencing (NGS) products, including gene sequencing instruments and fetal chromosome aneuploid detection reagent kits. The CFDA also emphasized that such gene sequencing instruments and reagent kits must be administered by the CFDA and comply with regulatory rules including "Regulations for the Supervision and Administration of Medical Devices (http://www.sfda.gov.cn/WS01/CL0051/97815.html)," "Medical Devices Registration Administration Method (http://www.sfda.gov.cn/WS01/CL0051/103755.html)," "Measures for the Administration of Registration of *In Vitro* Diagnostic Reagents (Interim) (http://www.sfda.gov.cn/WS01/CL0051/169365.html)," and other policies and regulations (http://www.sfda.gov.cn/). In 2017, the CFDA officially released the guiding principles requiring technical review for fetal chromosome aneuploidy (T21, T18 and T13) gene detection reagent kits (high-throughput sequencing) (CFDA Decree No. 52, 2017) (http://www.sfda.gov.cn/WS01/CL0087/171363.html), which provides guidance for the technical review of NGS diagnostic products after their registrations.

The US government has been taking a series of regulatory actions in the gene sequencing industry. The US Food and Drug Administration (US-FDA) approved the first NGS clinical analyzer in 2013.[1] In 2016, it released two guidance documents on *in vitro* diagnostic tests using NGS: "Use of Standards in FDA Regulatory Oversight of Next Generation Sequencing-Based In Vitro Diagnostics Used for Diagnosing Germline Diseases" (https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm509838.pdf, accessed on Jan 2, 2018) and "Use of Public Human Genetic Variant Databases to Support Clinical Validity for Next Generation Sequencing-Based *In Vitro* Diagnostics" (https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM509837.pdf, accessed on Jan 2, 2018). These resources, used to guide human genome NGS and DNA-targeted sequencing, proposed that the pipelines for public gene mutation databases should apply for US-FDA accreditation and periodic reassessments, during the period in which the clinical effectiveness of an NGS test can be evaluated in a pre-marketing application. Thus, the clinical application and standard management of NGS technology are steadily moving forward worldwide.

In November 2017, the US Association for Molecular Pathology and the College of American Pathologists jointly released the "Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines".[2] These guidelines put forward 17 recommendations, which covers the design, development, and operation of NGS bioinformatics pipelines and emphasizes the importance of staff training and quality control to ensure the quality of NGS tests. Thus, NGS-based clinical gene testing will be administered in a more standardized manner in the coming years.

At present, using the known gene panels is insufficient to detect the pathogenic genes for some rare pediatric diseases if the association or linkage between disease and genetic variation was not clinically confirmed. Therefore,

characterization of and research on the pathogenic genes for these diseases by using NGS technology combined with genetic research strategies are important for the diagnosis and treatment. In addition, for pediatric tumors, especially for familial pediatric tumors, identification of novel pathogenic genes/mutations by using NGS technology will facilitate the diagnosis of genetic disorder before birth thus promoting children's health. Therefore, the development of basic strategies for analyzing and applying NGS data in pediatric research and clinical practices is urgent.

ANALYSIS PIPELINES AND TOOLS FOR NGS DATA

The NGS data analysis processes for most pediatric diseases, such as common genetic diseases and sporadic diseases are generally similar to those for adult diseases. The basic NGS data analysis processes include base calling, sequence data pre-processing, sequence alignment, variant calling, variant filtering, and variant annotation and prioritization (Figure 1). Table 1 lists some popular tools used in the basic processes of NGS data analysis.

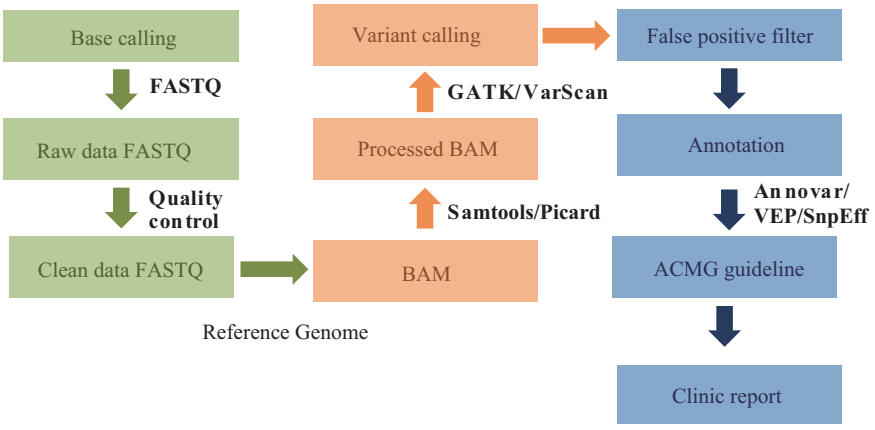

FIGURE 1 NGS-based bioinformatics pipelines. The schematic diagram shows the basic processes and typical tools in an NGS-based bioinformatics analysis. BAM, binary alignment map; GATK, Genome Analysis Toolkit.

TABLE 1 Tools that are used in the basic processes of NGS data analysis

Function	Tool	Website	Reference
Base Calling	naiveBayesCall	http://bayescall.sourceforge.net/	PMID:21385040
	freeIbis	http://bioinf.eva.mpg.de/freeibis/	PMID:23471300
	AYB	http://www.ebi.ac.uk/goldman-srv/AYB/	PMID:22377270
	PyroBayes	http://bioinformatics.bc.edu/marthlab/Software	PMID:18193056
Pre-processing	FASTX-Toolkit	https://packages.qa.debian.org/f/fastx-toolkit.html	unpublished
	FASTQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/	unpublished
	Trimmomatic	http://www.usadellab.org/cms/index.php?page=trimmomatic	PMID:24695404
	NGS-QC Generator	http://www.ngs-qc.org	PMID:27008019
	KMC	http://sun.aei.polsl.pl/kmc	PMID:25609798
Sequence alignment	Bowtie	http://bowtie.cbcb.umd.edu/	PMID:19261174
	BWA	http://bio-bwa.sourceforge.net	PMID:20080505
	SOAP2	http://soap.genomics.org.cn	PMID:19497933
Variant calling	VarScan	http://varscan.sourceforge.net/	PMID:19542151
	GATK	https://gatkforums.broadinstitute.org/gatk/	PMID:21478889
	MuTect	http://www.broadinstitute.org/cancer/cga/mutect	PMID:23396013
Variant annotation	ANNOVAR	http://annovar.openbioinformatics.org/en/latest/	PMID:20601685
	SnpEff	http://snpeff.sourceforge.net/	PMID:22728672
	VEP	http://www.ensembl.org/info/docs/tools/vep/index.html	PMID:27268795

Base calling

The process of base calling may slightly differ in different platforms; however, its common feature is to determine DNA sequences (i.e., four bases: adenine [A], guanine [G], cytosine [C], and thymine [T]) via interpretation of physical signals. Because subsequent analyses are highly dependent on the sequences generated by this process, accurate base calling is crucial for the accuracy of sequencing data analysis. Base calling usually converts original data (e.g., the BCL file) into a FASTQ format that can be used for subsequent analysis.

Sequence data pre-processing

In a FASTQ file, there is a Phred quality score for each base. The quality score of base calling for a read typically decreases along the sequence from 5′ to 3′. Thus, reads with low quality scores need to be removed when processing DNA sequencing data. In addition, it is necessary to remove the adapters at both ends of reads before sequence alignment and variant identification. After sequences are filtered, the resulted clean data can be used for sequence alignment. In this section, software parameters, and quality control cutoffs must be reviewed and documented if the options were modified in the newer version software.

Sequence alignment

Sequence reads are aligned onto the relatively complete human reference genome provided by the Human Genome Project,[3] such as GRCh38/hg38 or other versions. It should be noted that decoy sequences should be included in the genome reference, in order to detect variants in patient's genome which are not defined in main chromosomes. The commonly used alignment tools include BWA, bowtie, Novoalign, and MAQ. Specifically, BWA is composed of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM,[4,5] which could be used for aligning 70 bp to 1 Mbp reads. More importantly, BWA supports up to 8 bp indels, and split alignment, which can be used for structural variation calling. In contrast, there is no limitation of mismatches and indel length by using bowtie/bowtie2. Moreover, bowtie/bowtie2 and Novoalign could also trim several bases off at the 3′-end of reads to resolve the problem that sequencing accuracy decreases with the increase of sequencing reaction cycles.[6] In addition, MAQ is specifically designed for mapping very short reads, using scoring methods to derive genotype calls and build-up the consensus sequence of a diploid human genome.[7] For each aligned read, the alignment information includes alignment location, positive or negative strand, and Phred-scale mapping quality score. For germline analysis, the maximum percentage of aligned bases exceeding the minimum Phred score that disagree with reference should be settled to avoid false positives by misaligned bases. The alignment results are often stored in BAM files (binary version of the sequence alignment/map format). The sequence alignment step is important for the quality of aligned reads determines the accuracy of variant calling.

Variant calling

Variant calling is performed to determine and extract single-nucleotide variants (SNVs), copy number variations (CNVs), indels, and large structural alterations including deletions, insertions, inversions, and translocations after accurate sequence alignment. Currently, there are no tools that can detect all types of variants. Typically, different variant types or study designs need different callers. The type of variant calling, applicable study designs, and corresponding variant callers are summarized in Table 1 and Table 2. The variant contents are stored in standard variant call format (VCF) files (https://samtools.github.io/hts-specs/VCFv4.3.pdf). Notably, different range of SNP/indel ratio and transition/transversion ratio should be settled for specific genetic regions to prevent false positives.

Variant filtering

The main purpose of variant filtering is to eliminate false positives from the true positives. The accuracy of gene variation analysis is highly dependent on the base calling quality and read mapping quality. Therefore, the quality scores of the sequences near variants are also stored in the VCF files. Particularly, for Genome Analysis Tool Kit (GATK) best practices, the germline variants filtering is performed by Variant Quality Score Recalibration (VQSR), which uses machine learning to identify real variants. In addition, other parameters such as strand bias, and variant allele frequency (VAF) are also needed to be considered. By contrast, somatic variant filtering is more complicated. The situations including extreme strand bias, within-read position, low mapping quality, flanking homopolymer motifs, too many mismatches to the reference reads, extremely high depth of over-mapping to repeating sequences sites, and the presence of spanning deletions mapped across the site are recommended by Koboldt et al[8] and Saunders et al[9] to use somatic variants filtering. In summary, each caller has its specific filtering parameters, and users should follow the developer's recommendation for variant filtering.

Variant annotation

Variant annotation is to interpret the impact of a variant on gene functions, including variant location, cDNA variation, protein sequence alteration, minor allele frequencies in a specific population, and inclusion in the some databases. Thus, a variety of variant annotation databases are needed to thoroughly annotate variants. The commonly used variant annotation tools and databases are listed in Table 2. In general, genes may loss its function if harboring variants resulting in frameshift, premature

TABLE 2 The commonly used variant annotation tools and databases

Database	Website	Reference
HGVS nomenclature	http://varnomen.hgvs.org/	PMID:26931183
dbVar	http://www.ncbi.nlm.nih.gov/dbvar	unpublished
dbSNP	http://www.ncbi.nlm.nih.gov/snp	unpublished
Exome Variant Server	http://evs.gs.washington.edu/EVS	unpublished
1000 Genomes Project	http://browser.1000genomes.org	PMID: 26687719
Catalogue of Somatic Mutations in Cancer (COSMIC)	http://cancer.sanger.ac.uk/cosmic	PMID:27899578
The Cancer Genome Atlas (TCGA)	https://cancergenome.nih.gov/	unpublished
Online Mendelian Inheritance in Man (OMIM)	https://www.ncbi.nlm.nih.gov/omim/	PMID:25428349
ClinVar	http://www.ncbi.nlm.nih.gov/clinvar/	PMID:24234437
ExAC	http://exac.broadinstitute.org/	PMID:27899611
Sorting Tolerant From Intolerant (SIFT)	http://sift.jcvi.org/	PMID:19561590
PolyPhen-2	http://genetics.bwh.harvard.edu/pph2/	PMID:20354512
MutationTaster	http://www.mutationtaster.org/	PMID:24681721
Mendelian Clinically Applicable Pathogenicity (M-CAP)	http://bejerano.stanford.edu/mcap/	PMID:27776117
Combined Annotation Dependent Depletion (CADD)	http://cadd.gs.washington.edu/	PMID:24487276
Genome Wide Annotation of Variants (GWAWA)	http://www.sanger.ac.uk/sanger/StatGen_Gwava	PMID:24487584

stop-gain, or initiation codon loss. For missense variants, their consequences on the function of the encoded need to be analyzed using computational tools, such as SIFT, PolyPhen, MutationTaster, and M-CAP (Table 2) to evaluate the variants' pathogenicity.

Variant prioritization

Variant prioritization is performed primarily to remove non-significant variations that include synonymous variants, intronic variants, and common variants in a population. Only clinically meaningful variants or variants with potential clinical significance will remain after this process. The pathogenicity of these identified variants needs to be further evaluated. A recommended document on this topic is the 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,[10] which proposes to score any potentially pathogenic variants discovered by NGS analysis (http://www.medschool.umaryland.edu/Genetic_Variant_Interpretation_Tool1.html/), thus identifying the true pathogenic variants.

GENOME ANALYSIS STRATEGIES FOR PEDIATRIC DISEASES

Generally, pediatric disorders requiring genetic tests are diseases that may be caused by genetic defects. Thus, whole genome sequencing (WGS) and whole exome sequencing (WES) strategies may be applied in the diagnosis of rare pediatric diseases and familial pediatric diseases using NGS-based data.

Rare diseases

Rare genetic diseases

Rare hereditary diseases in children are sometimes difficult to confirm. Clinical diagnosis is even more difficult if the child's clinical symptoms are atypical. For instance, Chediak-Higashi Syndrome (CHS) often needs to be differentiated from many other hereditary diseases, including oculocutaneous albinism, Griscelli syndrome (GS), and Hermansky-Pudlak syndrome (HPS),[11] In 2014, a girl and her brother visited the Beijing Children's Hospital. Both children had mild albinism symptoms, including thin, soft, and silvery hair. Neither child had a history of serious infection or bleeding tendency at that point. Patients have a healthy sister.[12] No specific diagnosis could be confirmed due to these atypical symptoms; however, genetic test was considered to investigated the possibility of the inherited discover. WGS was then performed on all five members of this family. The results revealed that both pediatric patients carried a compound heterozygous LYST (Lysosomal Trafficking Regulator) mutation, resulting in an LYST functional defect. The findings from subsequent skin biopsy and blood smear examination supported the results of the genetic diagnosis, and a final diagnosis of CHS was made.[12] Therefore, for rare pediatric hereditary diseases with atypical symptoms, WGS/WES strategies, in combination with clinical information, are recommended to search for the associated pathogenic genetic variations. Furthermore, a particularly useful document on this topic is the ACMG interpretation

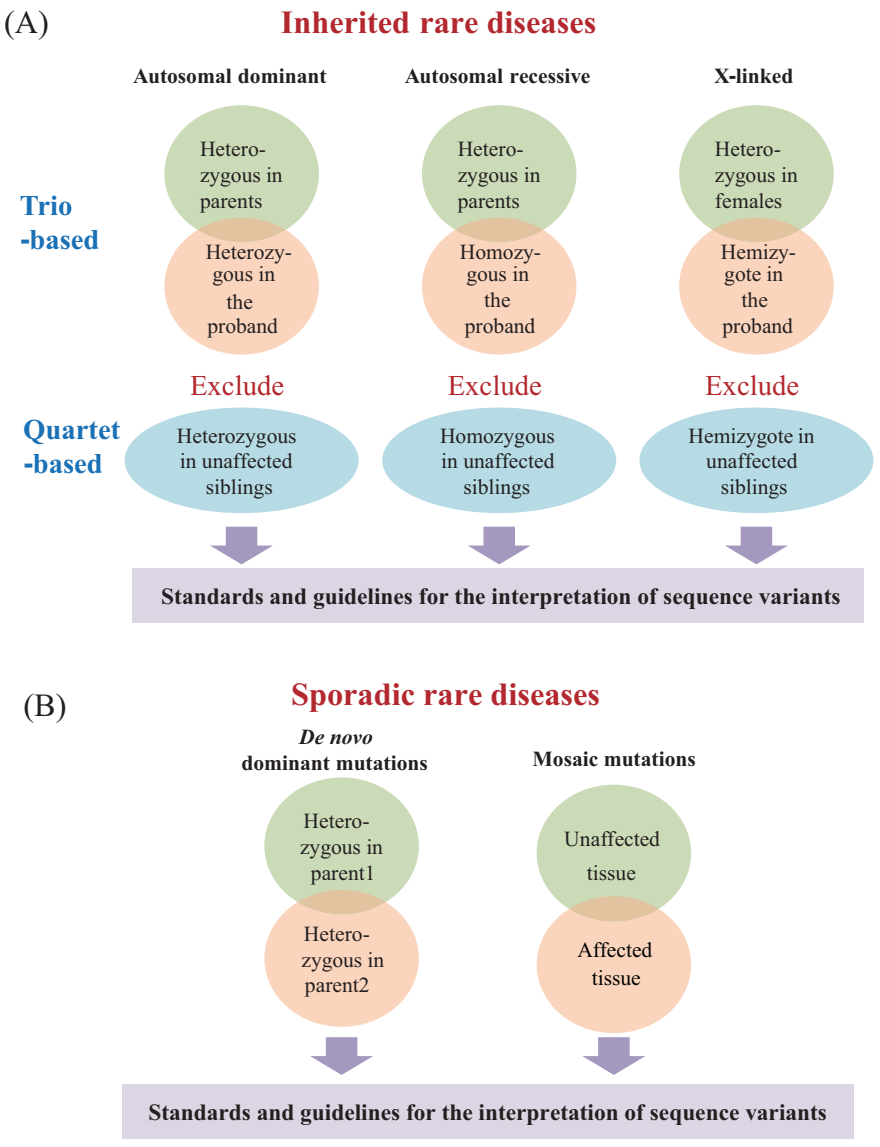

FIGURE 2 Strategies for research on genetic mutations in rare diseases. A, Screening mutations in patients with rare genetic diseases is typically based on three- or four-member families. The analyses are performed based on different genetic models including autosomal dominant, autosomal recessive, and X-linked genetic patterns. Finally, the candidate pathogenic gene mutation is identified in accordance with the ACMG guidelines. B, Screening mutations in patients with rare sporadic diseases usually includes identification of new dominant and mosaic mutations. The relevant genetic mutation is identified by taking steps to reduce the number of false positives and by referring to the relevant guidelines.

guidelines,[10] which suggests to score any potentially pathogenic mutations uncovered by NGS analysis, thus identifying the true pathogenic mutations.

For rare genetic diseases, family-based sequencing analysis is particularly important. In China, most families only have three or four members; thus, trio-based sequencing (sequencing DNA samples for both parents and the patient) and quartet-based sequencing (sequencing DNA samples for both parents and two kids) are the recommended strategies (Figure 2A).

(i) Trio-based sequencing

Trio-based sequencing is applied to families composed of three individuals, typically two parents and the affected child. The pathogenic mutations for autosomal recessive hereditary diseases can be identified by simply screening out the mutants that are homozygous in the proband but heterozygous in the parents. Notably, autosomal recessive inheritance also includes compound heterozygous variants.

In cases of autosomal dominant hereditary disease, both the proband and one of the parents have symptoms. Heterozygous mutations that exist in both the proband and the symptomatic parent, but are absent from the healthy parent, are potentially pathogenic sites.

If symptoms of an X-linked recessive hereditary disease occur mainly in male patients, trio-based sequencing can

be useful to search for a heterozygous mutation on the X-chromosomes of the mother because it would have been transmitted to the affected male offspring. It can also be used to identify homozygous mutants from the affected girl with the X-linked recessive hereditary disease.

(ii) Quartet-based sequencing

Quartet-based sequencing is applied in cases with the possibility that the pediatric patient's sibling is also affected. If the sibling is affected, the mutations are likely to be shared by two patients; if the sibling is not affected, he/she can be used as a normal control, and the mutations shared by two siblings can be ruled out as pathogenic mutations.

Rare sporadic diseases

Among rare sporadic pediatric diseases, de novo mutations in patients with autosomal dominant diseases can be easily identified but hard to annotation. To search for a novel pathogenic mutation related to this type of disease, cases may be collected in at least two affected probands, and WES/WGS can be performed to find the shared novel pathogenic genes/mutations. For instance, in Weaver syndrome cases, the causative mutant in the *EZH2* gene was found from two three-member families.[13] Furthermore, targeted gene sequencing for sporadic cases with known pathogenic genes can identify novel pathogenic mutations (Figure 2B). For example, by sequencing the *CARD15* gene in 30 children with Blau syndrome, Li et al identified a total of 10 mutations, of which 5 were unreported.[14] In addition, Sturge-Weber syndrome is a chimerism syndrome usually caused by *the GNAQ* gene mutations in some cells during development.[15] For diseases caused by this type of chimerism mutation, the *de novo* mosaic disease-causing mutations may be found by comparing the DNA sequences in diseased tissues with that in normal tissues (Figure 2B).

(i) *De novo* mutations

De novo mutations are typically identified by trio-based sequencing, which can find out mutations that are carried by patients but not by their parents. Probands in multiple families can further confirm the shared *de novo* mutation when parents–child trios are not possible to obtain. Notably, the mutation frequency of *de novo* mutation in each cell passage is extremely low (only 10^{-9}),[16] and NGS may yield a large amount of false positive mistakes in base determination due to sequencing errors. Therefore, false positives should be removed partly by adjusting the threshold of mutation allele frequency.

(ii) Mosaic mutations

The diseased tissues and normal tissues of pediatric patients often need to be stratified to identify mosaic mutations. With the normal tissues as the controls, mutations that occur only in the diseased tissues can be screened out. The identification of mosaic mutations is related to the purity of diseased tissues; therefore, the sensitivity of the identification can sometimes be increased by increasing the sequencing depth in the diseased tissues. While sequencing errors may introduce some false positives, and can be reduced by using the same technique described above in de novo mutations, that is, by adjusting the threshold of the mutant allele frequency.

Familial pediatric tumors

Couples who have one or two children with a malignant tumor simultaneously or successively are generally eager to know the risk of such disease in the third child, and typically they want to know the risk during pregnancy. For instance, in 2016, a boy was diagnosed as ataxia-telangiectasia (AT) with lymphoma in Beijing Children's Hospital. Further genetic test found compound heterozygous ATM gene mutations (unpublished case). After the boy passed, his mother got pregnant and would like to know if the fetal has both mutations. Fortunately, genetic counselling result shows that her second child only carries one mutation, thus avoid AT and lymphoma. Although some familial pediatric tumors, such as retinoblastoma (RB) and neuroblastoma (NB), have familial pathogenetic mutations,[17] these pathogenic mutations cannot fully explain the incidences of these diseases. The germline DNA of some familial NB children may carry the *PHOX2B* gene mutations, which are also associated with Hirschsprung disease and congenital central hypoventilation syndrome (CCHS). However, the NB children with germline *PHOX2B* mutation are not necessarily accompanied with CCHS and congenital myopathy. In contrast, some ganglioneuroblastoma (GNB) or ganglioneuroma (GN) (both belong to NB) children carrying a *PHOX2B* mutation could have ROHHAD (rapid-onset obesity, hypothalamic dysfunction, hypoventilation, and autonomic dysfunction) syndromes; similarly, ROHHAD patients could carry a *PHOX2B* mutation.[17,18] Therefore, other mutations in different genes may contribute the occurrence of familial NB/GNB/GN. In such cases, WGS/WES should be performed to examine germline DNA samples from family members, and various genetic models should be applied to analyze any mutation shared by two or more children within a family. To perform such an investigation, the document "Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer" proposed standardized procedures.[19]

Cancer predisposition genes

Characterization of cancer predisposition genes is mainly used for tumors with clear gene annotations. Generally, familial pediatric tumors are associated with cancer predisposition syndromes, including hereditary paraganglioma/pheochromocytoma syndrome (HPPS),

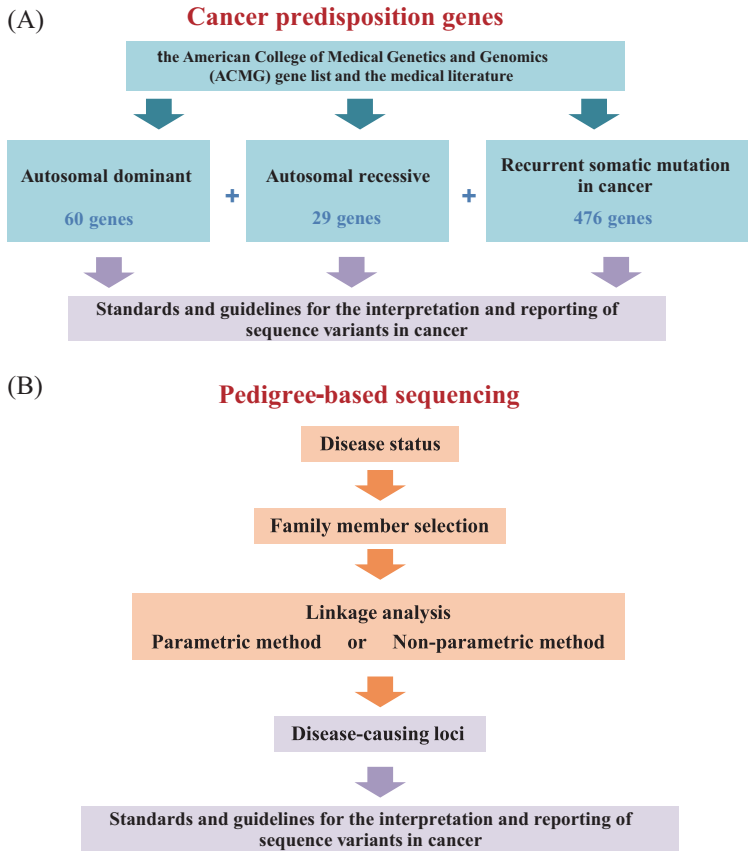

FIGURE 3 Strategies for research on the mutation loci of familial clustered pediatric tumors. A, One strategy for identifying the mutation loci of familial clustered pediatric tumors is searching for potential gene mutation loci via the gene mutation list summarized by ACMG [PMID: 26580448]. B, Another strategy is conducting pedigree analysis on patient's family members.

retinoblastoma, rhabdoid tumor predisposition, DICER1-related pleuropulmonary blastoma familial tumor predisposition syndrome, and Li-Fraumeni syndrome.[20,21] Cancer predisposition syndromes are genetic disorders that can be either dominant or recessive. Zhang et al listed 60 dominant, 29 recessive genes and 476 recurrent somatic mutations (565 genes in total) associated with pediatric tumors.[22] Gene panels often include somatically mutated genes with high-frequency mutations to maximize the chances of identifying cancer predisposition syndrome associated genes (Figure 3A). Thus, incorporating the 565 genes in the panels will decrease the cost of gene test and find as many mutations potentially associated with tumor predisposition syndromes as possible.

Pedigree-based WGS/WES

In contrast with the cancer predisposition syndromes that have annotated pathogenic genes, the pathogenic genes in many familial pediatric tumors remain unknown. In these cases, the application of pedigree-based WGS/WES may effectively detect new pathogenic genes. Pedigree-based sequencing is designed to screen the potential pathogenic loci in the genome via linkage analysis. There are two types of linkage analyses: parametric (requiring an inheritance model for the trait locus) and non-parametric (or parameter-free model, allele-sharing analysis).[23] A large number of statistical methods based on the parametric and non-parametric analyses have been developed and applied to identify potential pathogenic loci (Figure 3B).[23] Unlike in rare pediatric diseases, there is the possibility of incomplete dominance or delayed dominance for familial tumors. Therefore, the following two factors need to be considered during pedigree-based sequencing:

(i) Pedigree members to be tested must have a clear disease status

Identifying the disease status is helpful to screen potential pathogenic genes. The disease status is mainly determined by doctors based on the results of clinical characterization. In addition, family members with precancerous lesions in some specific organs should also be considered.

(ii) Appropriate family members should be selected for sequencing

The selection of family members for sequencing is directly related to the ability to find out tumor pathogenic genes. Typically, at least two affected family members need to be sequenced. Specifically, for autosomal dominant inheritance, two distantly related family members should be selected, if possible. Because the selection of family members for sequencing is a complex process, computer simulation is often applied. Common tools for this process include GIGI-Pick and PRIMUS.[24,25] Additionally, as more sequencing projects are being conducted to identify disease causative variants, many pedigree analysis tools, including Merlin, pVAAST, GIGI-Check, and SEQLinkage,[26-29] have been developed, which has dramatically increased the chance to identify pathogenic genes associated with familial tumors.

SUMMARY AND PROSPECTS

The China Food and Drug Administration (CFDA) started to review and approve NGS products in 2014, and these products include gene sequencing instruments and fetal chromosome aneuploid detection reagent kits (http://www.sfda.gov.cn/). The US-FDA has been oversighting a variety of NGS-based detection kits and related technologies. Two guidelines on NGS-based *in vitro* diagnosis were issued to address standard procedures for using NGS in human genome sequencing and DNA-targeted sequencing for the individualized diagnosis of germline diseases. Both the CFDA and US-FDA have made efforts in promoting and regulating NGS-based clinical gene testing.

In the context of the further application and standardized administration of NGS-based clinical testing and analysis, this review has summarized the current analysis strategies for identifying pathogenic mutations in pediatric diseases, particularly in rare diseases using NGS technologies (including familial clustered pediatric tumors). We hope this information will help clinicians and researchers to create a decision tree to make a correct diagnosis, and optimize the construction of pediatric disease pathogenic gene databases (the Pediatric Disease Annotations & Medicines [http://www.unimd.org/pedam/] and Encyclopedia of Rare Disease Annotations for Precision Medicine [http://www.unimd.org/eram/]),[30,31] so as to promote the application of NGS-based gene testing for rare diseases in clinical settings. The identification of pathogenic genes for rare pediatric diseases is particularly important for the future diagnosis of these diseases. Although children with rare diseases may be the first beneficiaries of NGS technology, these experiences are also valuable for the application of precision medicine in other disciplines. In addition to identifying the associated pathogenic genes, a better understanding of the relationships between these mutations and the disease phenotypes is critical for the diagnosis and treatment of rare diseases.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

REFERENCES

1. Collins FS, Hamburg MA. First FDA authorization for next-generation sequencer. *N Engl J MED.* 2013;369:2369-2371.
2. Roy S, Coldren C, Karunamurthy A, et al. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J MOL Diag.* 2018;20:4-27.
3. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature. 2*001;409:860-921.
4. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25:1754-1760.
5. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics. 2*010;26:589-595.
6. Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. *Bioinformatics.* 2012;28:3169-3177.
7. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 2008;18:1851-1858.
8. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012;22:568-576.
9. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics.* 2012;28:1811-1817.
10. Richards S, Aziz N, Bale S, 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;17:405-424.
11. Introne WJ, Westbroek W, Golas GA, Adams D. Chediak-Higashi Syndrome. In: Adam MP, Ardinger HH, Pagon RA, et al., eds. GeneReviews®. Seattle (WA);1993-2018.
12. Jin Y, Zhang L, Wang S, et al. Whole Genome Sequencing Identifies Novel Compound Heterozygous Lysosomal Trafficking Regulator Gene Mutations Associated with Autosomal Recessive Chediak-Higashi Syndrome. *Sci Rep.* 2017;7:41308.
13. Gibson WT, Hood RL, Zhan SH, et al. Mutations in EZH2 cause Weaver syndrome. *Am J hum Genet.* 2012;90:110-118.
14. Li C, Zhang J, Li S, et al. Gene mutations and clinical phenotypes in Chinese children with Blau syndrome. *Sci China Life Sci*. 2017;60:758-762.
15. Martins L, Giovani PA, Reboucas PD, et al. Computational analysis for GNAQ mutations: New insights on the molecular etiology of Sturge-Weber syndrome. *J Mol Graph Model.* 2017;76:429-440.
16. Veltman JA, Brunner HG. *De novo* mutations in human genetic disease. *Nat Rev Genet.* 2012;13:565-575.
17. Kamihara J, Bourdeaut F, Foulkes WD, et al. Retinoblastoma and Neuroblastoma Predisposition and Surveillance. *Clin Cancer Res.* 2017;23:e98-e106.
18. Bourdeaut F, Trochet D, Janoueix-Lerosey I, et al. Germline

mutations of the paired-like *homeobox 2B (PHOX2B)* gene in neuroblastoma. *Cancer Lett.* 2005;228:51-58.
19. Li MM, Datto M, Duncavage EJ, et al. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J Mol Diagn.* 2017;19:4-23.
20. Schiffman JD, Geller JI, Mundt E, Means A, Means L, Means V. Update on pediatric cancer predisposition syndromes. *Pediatr Blood Cancer.* 2013;60:1247-1252.
21. Saletta F, Dalla Pozza L, Byrne JA. Genetic causes of cancer predisposition in children and adolescents. *Transl Pediatr.* 2015;4:67-75.
22. Zhang J, Walsh MF, Wu G, et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. *N Engl J Med.* 2015;373:2336-2346.
23. Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet.* 2015;16:275-284.
24. Cheung CY, Marchani Blue E, Wijsman EM. A statistical framework to guide sequencing choices in pedigrees. *Am J Hum Genet.* 2014;94:257-267.
25. Staples J, Qiao D, Cho MH, et al. PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *Am J Hum Genet.* 2014;95:553-564.
26. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet.* 2002;30:97-101.
27. Hu H, Roach JC, Coon H, et al. A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. *Nat Biotechnol.* 2014;32:663-669.
28. Cheung CY, Thompson EA, Wijsman EM. Detection of Mendelian consistent genotyping errors in pedigrees. *Genet Epidemiol.* 2014;38:291-299.
29. Wang GT, Zhang D, Li B, Dai H, Leal SM. Collapsed haplotype pattern method for linkage analysis of next-generation sequence data. *Eur J Hum Genet.* 2015;23:1739-1743.
30. Jia J, An Z, Ming Y, et al. PedAM: a database for Pediatric Disease Annotation and Medicine. *Nucleic Acids Res.* 2018;46:D977-D983.
31. Jia J, An Z, Ming Y, et al. eRAM: encyclopedia of rare disease annotations for precision medicine. *Nucleic Acids Res.* 2018;46:D937-D943.

How to cite this article: Jin Y, Zhang L, Ning B, et al. Application of genome analysis strategies in the clinical testing for pediatric diseases. *Pediatr Invest.* 2018;2:72-81. https://doi.org/10.1002/ped4.12044