Identification of novel deep intronic *PAH* gene variants in patients with phenylketonuria

Xiaohua Jin[1,2], Yousheng Yan[3*], Chuan Zhang[4], Ya Tai [5], Lisha An[1,2], Xinyou Yu[6], Linlin Zhang[7], Shengju Hao[4], Xiaofang Cao[1,2], Chenghong Yin[3], Xu Ma[1,2*]

[1] National Research Institute for Family Planning, Beijing 100081, China;

[2] National Human Genetic Resources Center, Beijing 100081, China;

[3] Prenatal Diagnostic Center, Beijing Obstetrics and Gynecology Hospital, Capital Medical University, Beijing 100026, China;

[4] Gansu Province Medical Genetics Center, Gansu Provincial Maternity and Child-Care Hospital, Lanzhou 730050, China;

[5] Department of Obstetrics and Gynecology, Peking University International Hospital, Beijing, 102206, China;

[6] Department of Prenatal Diagnosis Center, General Hospital of Ningxia Medical University, Yinchuan, 750004, China;

[7] Clinical Lab, The Third Affiliated Hospital of Zhengzhou University, Zhengzhou, 450052, China;

[*]**Corresponding authors:** Yousheng Yan, Prenatal Diagnostic Center, Beijing Obstetrics and Gynecology Hospital, Capital Medical University, Beijing 100026, China, E-mail: yys_521@ccmu.edu.cn; Xu Ma, National Human Genetic Resources Center, National Research Institute for Family Planning, Beijing 100081, China, E-mail: genetic@263.net.cn.

Running title: deep intronic variants of the *PAH* gen

Abstract

Phenylketonuria (PKU) is caused by phenylalanine hydroxylase (PAH) gene variants. Previously, 94.21% of variants were identified using Sanger sequencing and multiplex ligation-dependent probe amplification. To investigate the remaining variants, whole-genome sequencing (WGS) was performed in four patients with PKU with unknown genotype to identify deep intronic or structural variants. Three novel heterozygous variants (c.706+368T>C; c.1065+241C>A; and c.1199+502A>T) were identified in a deep *PAH* gene intron. The c.1199+502A>T variant was detected in 60% (6/10) PKU patients. In silico prediction showed that the three deep variants may impact splice site selection and result in inclusion of a pseudo-exon. The c.1199+502A>T *PAH* minigene and reverse transcription PCR of blood RNA in a patient with PKU and compound heterozygous variants (c.1199+502A>T/c.1199G>A) confirmed that the c.1199+502A>T variant creates a novel branch point and leads to the inclusion of a 25 bp in PAH mRNA (r.1199_2000ins1199+538_1199+562). Furthermore, the c.1199G>A mutation leads to the retention of an additional 17 nt in the *PAH* mRNA transcript (r.1199_2000ins1199+1_1199+17). These results expand the *PAH* genotypic spectrum and highlight that deep intronic analysis of PAH can improve genetic diagnosis in undiagnostic patients.

KEYWORD
Phenylketonuria, *PAH*, whole-genome sequencing, deep intronic variation, RNA splicing

1.Introduction

Phenylketonuria (PKU [MIM: 261600]) is the most frequent inborn error of amino acid metabolism. The incidence of PKU varies among ethnic groups and geographic regions, affecting approximately 1 in 15,000 newborns [Blau et al.,2010]. In mainland China, the average incidence of PKU is 1 in 11,6142, but in the northwest regions of China the incidence of PKU is much higher, and is 1 in 3,420 in Gansu, as measured by newborn screening.

PKU follows an autosomal recessive mode of inheritance and is the result of mutations in the *PAH* gene, which encodes phenylalanine hydroxylase. More than 1,188 PAH variants have been identified and recorded in the PAHvdb (http://www.biopku.org/pah/) locus-specific database. The frequency and distribution of *PAH* mutations differ in different populations. Previous research has identified some PAH hotspot mutations in PKU patients. The most frequent variant types in the PAHvdb were substitutions (80.5%), followed by deletions (12.9%), and duplications (2.1%)[Hillert et al.,2020]. Approximately 95% of all *PAH* mutations are able to be detected in patients with PKU by sequencing all *PAH* exons and exon-intron boundaries [Li et al.,2015], and another 3% pathogenic variants were discovered in patients by multiplex ligation-dependent probe amplification (MLPA)[Yan et al.,2016]. However, in the remaining 2% of patients with PKU, who have classical PKU phenotypes and increased blood Phe concentrations (> 600 mmol/L), only a single deleterious mutation (heterozygous) or no mutations are identified after standard sequencing and MLPA analyses.

Recently, deleterious mutations have been identified in deep intronic regions in some genes, including *ATP7B*[Woimant et al.,2020], *ABCA4*[Malekkou et al.,2020], and *F8*[Chang et al.,2019]. Moreover, genomic structural variants (SVs) were identified by whole-genome sequencing (WGS) analysis in patients with an inconclusive diagnosis after regular genetic testing [Middelkamp et al.,2019].

Previously, we analyzed *PAH* mutations in 475 patients with PKU in Northwest China by Sanger sequencing and MLPA. A total of 895 alleles with mutation frequency of 94.21% (895/950) were detected in 475 PKU patients. However, the remaining 55 alleles (5.79%) were unknown[Yan et al.,2019]. We hypothesized that these 55 alleles were deleterious deep intronic variants or SVs in *PAH*. In this study, WGS was performed in patients with PKU with unknown genotype to investigate whether they have *PAH* deep intronic variants or SVs.

2. Materials and Methods

2.1 Patients

Form a total of 495 index cases diagnosed with PKU, we selected those patients having only a single mutation or no mutation detected in the *PAH* gene. All selected PKU patients provided from the Medical Genetics Center and Newborn Screening

Center of Gansu Provincial Maternity and Child-care Hospital, and screened for *PAH* mutations by Sanger sequencing and MLPA at diagnosis. Informed consent was obtained for all patients or their families and the study was approved by the Ethics Committee of Gansu Provincial Maternity and Child-care Hospital. All selected PKU patients were identified through Newborn Screening (NBS) program, and diagnosed as classic or moderate PKU according to classification criteria[Association et al.,2020], and excluded from tetrahydrobiopterin deficiency as described in previous research[Yan et al.,2019].

2.2 WGS analysis

Genomic DNA from the proband, family members and control individuals were extracted from peripheral blood. Short read genome sequencing and long read genome sequencing were performed to discovering the deep intronic variants or SVs.

A library for short read genome sequencing was generated from genomic DNA using the Illumina TruSeq DNA PCR-Free Library Prep Kit (Illumina, San Diego, CA) and sequencing was performed on the Illumina HiSeq 2500 System with paired-end 150bp reads to a 36-fold mean depth of coverage. The National Human Genetic Resources Center Service performed the data analysis and variant curation. Single-nucleotide variants and small insertions and deletions were identified using MedGAP v2.0, a pipeline based on GATK best practices for data preprocessing and variant discovery with GATK Haplotype Caller v3.1.1.

Long read genome sequencing with low coverage was performed on the Oxford Nanopore sequencing platform. Genomic DNA is extracted and a library of large-scale insertions is created according to the manufacturer's recommendations (Oxford Nanopore, UK). Libraries were sequenced on R9.4 flowcells by GridION X5. The long reads were aligned to the human reference genome (GRCh37) using NGMLR[Sedlazeck et al.,2018]. Structural variations (SVs) and single nucleotide variants (SNVs) were called by SAMtools[Li et al.,2009] and Sniffles[Sedlazeck et al.,2018], respectively. The alignment results were visualized using Ribbon[Nattestad et al.,2020] and IGV[Thorvaldsdóttir et al.,2013] and possible SV calls were manually validated.

The frequencies of all variants were searched in GnomAD (gnomad.broadinstitute.org) , Exome Sequencing Project (ESP, evs.gs.washington.edu), and dbSNP (www.ncbi.nlm.nih.gov/projects/SNP).

The variants found by WGS were identified by Sanger sequencing, which was performed with Thermo Fisher reagents according to the manufacturer’s protocol on an ABI3730 sequencer (Thermo fisher, Saint Herblain, France) and then, analyzed with Seqscape V4.0 software.

2.3 In silico prediction of impact on splice site selection

In silico prediction of the impact of specific variants on splice site choice was performed using Alamut® Visual v.2.11 (Interactive Biosoftware, Rouen, France), which included a splicing module integrating a number of prediction algorithms and splicing prediction data: SpliceSiteFinder-like (SSF), MaxEntScan(MES), NNSplice and GeneSplicer for Splicing signals, ESEfinder and RESCUE-ESE for exonic splicing enhancer (ESE) binding site detection, and Mercer et al. high-confidence branchpoints for identified branchpoints. Individual tools were deemed to predict altered splicing where the change in splice site score was ≥10% (MES and GeneSplicer) or ≥5%(NNSplice and SSF).

2.4 Minigene analysis

For evaluation of in vitro splicing of the novel deep variant *PAH*: c.1199+502A>T, the minigene were constructed by the pcDNA3.1(+) vector. In the minigene vector, a fragment of human *PAH* including full length of exon 10, intron 10, exon11, intron 11 and exon 12, was amplified using primers located in intron 9 and intron 12 from patient's genomic DNA. Gene fragment and flanks region was cloned into the pcDNA3.1(+) vector with HindIII/BamHI. The mutant and wildtype minigene constructs were prepared.

For minigene assays, 293T cells were seeds in $35m^2$ wells at a density of $4◊10^5$ in 2ml 10% MEM and grown overnight. Cells were transfected with a total DNA amount of 4μg per well using the Liposomal Transfection Reagent (Invitrogen, USA). Cells were harvested by trypsinization after 48h. Total RNA was isolated using Trizol Reagent (ThermoFisher) and phenol-chloroform extraction. cDNA synthesis was performed using HiScript II First-Strand cDNA Synthesis Kit (Vazyme, Nanjing, China). Splicing analysis was carried out by PCR amplification with FastStart Taq Ploymerase (Vazyme, Nanjing, China) using specific primers to exclude detection of endogenous *PAH* gene expressinon: 3.1miniRT-F (5'-AACCCACTGCTTACTGGCTTATCG-3') and 3.1miniRT-R (5'-TTAAACGGGCCCTCTAGACTCGA-3') for pcDNA3.1(+) minigene. The PCR products were analyzed by 2% agarose gel electrophoresis, and their identity was confirmed by Sanger sequencing.

2.5 RNA analysis of patients' blood sample

Peripheral blood of the patients was collected with the Tempus™ Blood RNA Tubes (ThermoFisher, Applied Biosystems, US), and whole blood mRNA was extracted using Tempus™ Spin RNA Isolation Kit (ThermoFisher, Applied Biosystems, USA) according to the instructions of the manufacturer. RT-PCR was performed using SuperScript™ IV First-Strand Synthesis System (ThermoFisher, Invitrogen, USA), and the cDNA products obtained from normal control and patients were examined by gel electrophoresis, followed by retrieval and Sanger sequencing of each band of interest. The PCR reaction was performed to amplificated the *PAH* gene cDNA using

TransStart® FastPfu DNA polymerase. The primers for the PCR were according to the previous reported protocol[Okano et al.,1994].

For cloning, the desired amplified PCR fragments were purified from agarose gel and subcloned into pGEM-T-Easy plasmid vector (Promega). Randomly selected individual clones were grown in LB medium at 37°C overnight and the plasmid DNAs were isolated. The amplified PCR products and the cDNA plasmid inserts were bidirectionally sequenced using specific primers. The resultant sequences were aligned and compared to the RefSeq DNA sequence (NC_000012.12) and cDNA sequence (NM_000277.1) of *PAH* gene using the NCBI Blast tool.

3.Results

3.1 Screening deep intronic variants or SVs by WGS

Among the cohort of 494 patients with PKU, we selected 10 patients with one identified mutation in *PAH* and diagnosed as classic PKU (cPKU) and mild PKU (mPKU) (Table 1). The DNA quality of the four patients was sufficient for short and long read genome sequencing, and DNA samples from the patients' parents were also available.

PAH SVs, and of other genes associated with hyperphenylalaninemia, were excluded by low-coverage genome long read genome sequencing. Short read sequencing confirmed the presence of a heterozygous mutation in a *PAH* exon or exon-intron junction in four patients, which was consistent with the results previously obtained by Sanger sequencing. In four patients, we detected three novel intronic heterozygous variants (c.1199+502A>T; c.1065+241C>A; c.706+368T>C), which are absent in GnomAD, ESP, and dbSNP. These intronic variants were confirmed in patients and their parents and observed to segregate with previously identified mutation in these patients. The c. 1199+502A>T mutation was found in P01 and P02 families, and variants c.1065+241C>A and c.706+368T>C were identified in families P03 and P04, respectively.

We screened for the three intronic variants in the remaining six patients. The c.1199+502A>T variant was identified in four patients and their parents by Sanger sequencing, and none of the three intronic variants were found in the other two patients. So, six of 10 (60%) patients examined carried deep intronic variants.

3.2 In silico analyses show that deep intronic variants cause pseudo-exon inclusion

By in silico prediction, the three deep variants may be impact on splice site selection resulted in abnormal splicing. The c.706+368T>C variant creates a new donor splice site, activates an upstream acceptor splice site, and results in a 36nt intronic sequence inclusion (pseudo-exon) in intron 6(Figure 1a). The variant c.1065+241C>A does not affect the splice sit (Figure 1b), but it creates a novel binding site for SF2/ASF(IgM-BRAC1), thus activating an exonic splicing enhancer (SES) (Figure 1c). This may be led to inclusion of 81nt intronic sequence (pseudo-exon) in intron 10. The variant

c.1199+502A>T creates a high-confidence branchpoint that actives a downstream acceptor splice site and donor splicing site, resulting in inclusion of a 25nt intronic sequence (pseudo-exon) in intron 6(Figure 1d).

3.3 Expression analysis of c.1199+502A>T variant in *PAH*

To investigate the effects of the c.1199+502A>T intronic variants on *PAH* expression, c.1199+502A>T mutant and wildtype minigenes were constructed. These minigenes were then transfected into 293T cells. RT-PCR revealed the presence of a different band produced following c.1199+502A>T mutant minigene expression. The mutant minigene produced a 533-bp band, which was 25 bp longer than that produced by the wildtype minigene (508 bp) (Figure 2a). Sanger sequencing identified that the c.1199+502A>T mutant minigene mRNA included a 25-bp intronic sequence from intron 11 (Figure 2b).

To confirm the deleterious effect of the c.1199+502A>T intronic variants on *PAH* transcript production, peripheral bloods were collected from patient P01, who had a c.1199G>A heterozygous mutation, and their parents for the RT-PCR. Three healthy people were recruited as controls.

RT-PCR and clone sequencing resulted in the discovery of two novel transcripts which markedly differed from those of the controls. Novel transcript A included all 13 exons, plus a 25-nt pseudo-exon between exon11 and exon12. Novel transcript B included all 13 exons, plus a 17-nt retention between exon11 and exon12. Analysis of the patient's parents confirmed the presence of novel transcript A in the father with the deep c.1199+502A>T variant, and the presence of novel transcript B in the mother carrying the heterozygous mutation c.1199G>A.

The c.1199+502A>T intronic variant activated a new branchpoint and a cryptic exon resulting in the inclusion of a 25-nt pseudo-exon from intron 11 into the *PAH* mRNA transcript (r.1199_2000ins1199+538_1199+562). The c.1199G>A mutation disrupted the normal intron 11 splice donor site and activated a downstream cryptic splice donor site, leading to the retention of an additional 17 nt into the *PAH* mRNA transcript (r.1199_2000ins1199+1_1199+17) (Figure 2d).

4. Discussion

Application of next generation sequencing has greatly improved the diagnosis of genetic disease. However, approximately half of all patients with suspected genetic disease lack a precise genetic diagnosis. Routine genetic analysis methods, including target gene panels based on next generation sequencing, whole-exome sequencing, and Sanger sequencing of exons and exon-intron boundaries, were unable to identify deep intron variants that can be located 100 bp away from exons. Pathogenic deep intronic variants have been reported in more than 75 disease-associated genes, commonly resulting in pseudo-exon inclusion due to activation of atypical splice sites or changes in splicing regulatory elements [Vaz-Drago et al.,2017].

In this study, three novel deep variants (c.1199+502A>T; c.1065+241C>A; and c.706+368T>C) were identified in four PKU patients with unknown genotype in trans

with a known pathogenic variant by WGS. The novel c. 1199+502A>T variant was confirmed in four of six selected patients. These results suggest that deep *PAH* gene variants can cause PKU and that WGS technology can be used to identify the genotypes of patients lacking genetic diagnosis. The novel c. 1199+502A>T variant may be a recurrent variant in patients with PKU in Northwest China. According to the PAHvdb records, approximately 15% of the mutations affect conserved 3' and 5' splice sites [Dworniczak et al.,1991], and some studies have revealed that synonymous or missense mutations may cause splicing defects [Chao et al.,2001]. However, deep intron variants of the *PAH* gene, especially at locations more than 100 bp from exons, have not been reported previously.

In silico prediction analyses showed that the three deep variants may impact splice site selection, resulting in pseudo-exon the inclusion in normal manuscript of *PAH*. Pseudo-exon inclusion caused by deep intronic variants has been reported many times. The more common mechanism involves the variant that creates a new donor splice site and activates a pre-existing atypical acceptor splice site [Chang et al.,2019; Tozawa et al.,2019; Malekkou et al.,2020]. Some studies found that deep intronic variants create a new acceptor splice site [Sun et al.,2020] or interfere with splicing regulatory elements resulting a pseudo-exon inclusion [Albert et al.,2018; Fadaie et al.,2019]. In this study, the c.706+368T>C variant creates a novel donor splice site, and the c.1065+241C>A variant activates a splicing enhancer element.

The c.1199+502A>T variant creates a high-confidence branchpoint that actives a downstream acceptor splice site and donor splicing site by in silico prediction. Expression analysis of the c.1199+502A>T variants by minigene and RT-PCR confirmed that results in the inclusion of a 25-nt pseudo-exon inclusion from intron 11 into the *PAH* mRNA transcript. Wang et.al[Wang et al.,2020] found that the F8 c.5999-27A>G variant may disrupt the branch point in intron 18, leading to F8 exon 19 skipping. However, there have been no reports that deep intron variation leads to the establishment of novel branch points. Therefore, our results present a novel mechanism for the pathogenesis of deep intron variation that leads to the inclusion of a pseudo-exon.

In patient P01, RT-PCR identified a 17-nt retention in the mRNA that was caused by the common c.1199G>A mutation. This mutation was regarded as a missense mutation p. (Arg400Lys) in a previous report [Song et al.,2005]. Splicing of *PAH* exon 11 is vulnerable because of a weak 3' splice site [Heintz et al.,2012], and the c.1199+17G>A and c.1199+20G>C variants disrupted U1snRNP70 binding to this intronic region downstream of the natural 5' splice site, causing exon 11 skipping [Martínez-Pizarro et al.,2018]. Our results indicate that the c.1199G>A mutation disrupts the normal intron 11 donor site and activates the downstream cryptic donor splice site, leading to retention of a 17-nt intron into the *PAH* mRNA transcript.

In this study, deep *PAH* gene variants were identified in patients with PKU with unknown genotype using WGS. These results provide new insight and highlight that rare and recurrent variants located deep within *PAH* introns are not uncommon in patients with PKU. *In silico* prediction and *in vivo* expression are powerful approaches to determine the functions and effects of these deep intronic variants, and

deep intronic analysis can improve genetic diagnosis in undiagnostic patients.

Acknowledgements
We would like to thank Professor Shangzhi Huang (from Chinese Academy of Medical Sciences & Peking Union Medical College) for his guidance and valuable suggestions on this study. We would also like to thank the patients and their families for their participation and providing valuable blood samples.

Funding
This study was supported by National Key Research and Development Program of China (Grant No.: 2016YFC1000307) and Natural Science Foundation of Gansu Province (Grant No.: 18JR3RA036).

Conflict of interest
The authors declare that they have no conflict of interest.

Data availability statement
The data that support the findings of this study is available upon reasonable request from corresponding authors. The novel variants have been submitted to the *PAH*vdb (http://www.biopku.org/pah/), and the IDs of *PAH*vdb is PAH1221 (c.706+368T>C), PAH1222 (c.1065+241C>A) and PAH1223 (c.1199+502A>T).

References

Albert S, Garanto A, Sangermano R, Khan M, Bax NM, Hoyng CB, Zernant J, Lee W, Allikmets R, Collin R, Cremers F. 2018. Identification and Rescue of Splice Defects Caused by Two Neighboring Deep-Intronic ABCA4 Mutations Underlying Stargardt Disease. Am J Hum Genet 102:517-527.

Association WGFPGFDATOGDMGBOCM, Huang S, Song F. 2020. [Clinical practice guidelines for phenylketonuria]. Zhonghua Yi Xue Yi Chuan Xue Za Zhi 37:226-234.

Blau N, van Spronsen FJ, Levy HL. 2010. Phenylketonuria. Lancet 376:1417-1427.

Chang CY, Perng CL, Cheng SN, Hu SH, Wu TY, Lin SY, Chen YC. 2019. Deep intronic variant c.5999-277G>A of F8 gene may be a hot spot mutation for mild hemophilia A patients without mutation in exonic DNA. Eur J Haematol 103:47-55.

Chao HK, Hsiao KJ, Su TS. 2001. A silent mutation induces exon skipping in the phenylalanine hydroxylase gene in phenylketonuria. Hum Genet 108:14-19.

Dworniczak B, Aulehla-Scholz C, Kalaydjieva L, Bartholomé K, Grudda K, Horst J. 1991. Aberrant splicing of phenylalanine hydroxylase mRNA: the major cause for phenylketonuria in parts of southern Europe. Genomics 11:242-246.

Fadaie Z, Khan M, Del Pozo-Valero M, Cornelis SS, Ayuso C, Cremers F, Roosing S, Group TAS. 2019. Identification of splice defects due to noncanonical splice site or deep-intronic variants in ABCA4. Hum Mutat 40:2365-2376.

Heintz C, Dobrowolski SF, Andersen HS, Demirkol M, Blau N, Andresen BS. 2012. Splicing of phenylalanine hydroxylase (PAH) exon 11 is vulnerable: molecular pathology of mutations in PAH exon 11. Mol Genet Metab 106:403-411.

Hillert A, Anikster Y, Belanger-Quintana A, Burlina A, Burton BK, Carducci C, Chiesa AE, Christodoulou J, Đorđević M, Desviat LR, Eliyahu A, Evers R, Fajkusova L, Feillet F, Bonfim-Freitas PE, Giżewska M, Gundorova P, Karall D, Kneller K, Kutsev SI, Leuzzi V, Levy HL, Lichter-Konecki U, Muntau AC, Namour F, Oltarzewski M, Paras A, Perez B, Polak E, Polyakov AV, Porta F, Rohrbach M, Scholl-Bürgi S, Spécola N, Stojiljković M, Shen N, Santana-da Silva LC, Skouma A, van Spronsen F, Stoppioni V, Thöny B, Trefz FK, Vockley J, Yu Y, Zschocke J, Hoffmann GF, Garbade SF, Blau N. 2020. The Genetic Landscape and Epidemiology of Phenylketonuria. Am J Hum Genet 107:234-250.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

Li N, Jia H, Liu Z, Tao J, Chen S, Li X, Deng Y, Jin X, Song J, Zhang L, Liang Y, Wang W, Zhu J. 2015. Molecular characterisation of phenylketonuria in a Chinese mainland population using next-generation sequencing. Sci Rep 5:15769.

Malekkou A, Sevastou I, Mavrikiou G, Georgiou T, Vilageliu L, Moraitou M, Michelakakis H, Prokopiou C, Drousiotou A. 2020. A novel mutation deep within intron 7 of the GBA gene causes Gaucher disease. Mol Genet Genomic Med e1090.

Martínez-Pizarro A, Dembic M, Pérez B, Andresen BS, Desviat LR. 2018. Intronic PAH gene mutations cause a splicing defect by a novel mechanism involving U1snRNP binding downstream of the 5' splice site. PLoS Genet 14:e1007360.

Middelkamp S, Vlaar JM, Giltay J, Korzelius J, Besselink N, Boymans S, Janssen R, de la Fonteijne L, van

Binsbergen E, van Roosmalen MJ, Hochstenbach R, Giachino D, Talkowski ME, Kloosterman WP, Cuppen E. 2019. Prioritization of genes driving congenital phenotypes of patients with de novo genomic structural variants. Genome Med 11:79.

Nattestad M, Aboukhalil R, Chin CS, Schatz MC. 2020. Ribbon: Intuitive visualization for complex genomic variation. Bioinformatics .

Okano Y, Hase Y, Shintaku H, Araki K, Furuyama J, Oura T, Isshiki G. 1994. Molecular characterization of phenylketonuric mutations in Japanese by analysis of phenylalanine hydroxylase mRNA from lymphoblasts. Hum Mol Genet 3:659.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. 2018. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 15:461-468.

Song F, Qu YJ, Zhang T, Jin YW, Wang H, Zheng XY. 2005. Phenylketonuria mutations in Northern China. Mol Genet Metab 86 Suppl 1:S107-118.

Sun W, Xiao X, Li S, Jia X, Zhang Q. 2020. A novel deep intronic COL2A1 mutation in a family with early-onset high myopia/ocular-only Stickler syndrome. Ophthalmic Physiol Opt 40:281-288.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178-192.

Tozawa Y, Abdrabou S, Nogawa-Chida N, Nishiuchi R, Ishida T, Suzuki Y, Sano H, Kobayashi R, Kishimoto K, Ohara O, Imai K, Naruto T, Kobayashi K, Ariga T, Yamada M. 2019. A deep intronic mutation of c.1166-285 T > G in SLC46A1 is shared by four unrelated Japanese patients with hereditary folate malabsorption (HFM). Clin Immunol 208:108256.

Vaz-Drago R, Custódio N, Carmo-Fonseca M. 2017. Deep intronic mutations and human disease. Hum Genet 136:1093-1111.

Wang X, Hu Q, Tang N, Lu Y, Deng J. 2020. Deep intronic F8 c.5999-27A>G variant causes exon 19 skipping and leads to moderate hemophilia A. Blood Coagul Fibrinolysis .

Woimant F, Poujois A, Bloch A, Jordi T, Laplanche JL, Morel H, Collet C. 2020. A novel deep intronic variant in ATP7B in five unrelated families affected by Wilson disease. Mol Genet Genomic Med e1428.

Yan Y, Zhang C, Jin X, Zhang Q, Zheng L, Feng X, Hao S, Gao H, Ma X. 2019. Mutation spectrum of PAH gene in phenylketonuria patients in Northwest China: identification of twenty novel variants. Metab Brain Dis 34:733-745.

Yan YS, Yao FX, Hao SJ, Zhang C, Chen X, Feng X, Yang T, Huang SZ. 2016. [Analysis of large deletion of phenylalanine hydroxylase gene in Chinese patients with phenylketonuria]. Zhonghua Yi Xue Za Zhi 96:1097-1102.

Figure Legends

Figure 1. In silico prediction of deep intronic variants (a) The variant c.706+368T>C creates a novel donor splice site and actives an upstream acceptor splice site lead to inclusion of a 36nt intronic sequence (pseudo-exon) in intron 6. (b) The variant c.1065+241C>A does not affect the splice sit. (c) it creates a novel binding site for SF2/ASF(IgM-BRAC1), thus activating a splicing enhancer element. (d) The variant c.1199+502A>T creates a high-confidence branchpoint that activates a downstream acceptor splice site and donor splicing site, resulting in inclusion of a 25nt intronic sequence (pseudo-exon) in intron 6.

Figure2. **The splice effect of the variant c.1199+502A>T and c.1199G>A by Minigene or RT-PCR analysis** (a) Abnormal transcript were detected in the c.1199+502A>T mutant pcDNA3.1 minigenes that transfected in 293T cells. The mutant minigene detected a 533bp production, 25bp more than wildtype minigene 508bp. (b) Sanger sequence identified the 25bp pseudo-exon inclusion in transcript of mutant minigene. (c) Normal transcript of wildtype minigene. (d) RT-PCR of patients' blood show that the common mutation c.1199G>A result in 17nt retention into *PAH* mRNA transcript.