Supporting Information

Evaluation of *SHOX* defects in the era of next-generation sequencing

Mariana F.A. Funari[1], Juliana S. de Barros[2], Lucas S. Santana[3], Antonio M. Lerario[1,3,4], Bruna L. Freire[1], Thais K. Homma[1,3], Gabriela A. Vasques[1,3], Berenice B. Mendonca[1], Mirian Y. Nishi[1], Alexander A.L. Jorge[1,3]

[1]Unidade de Endocrinologia do Desenvolvimento, Laboratório de Hormônios e Genética Molecular/LIM42, Hospital das Clínicas, Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil

[2]Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil

[3]Unidade de Endocrinologia Genética/LIM25, Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil

[4]Department of Internal Medicine, Division of Metabolism, Endocrinology and Diabetes, University of Michigan, Ann Arbor, USA

Table of Content

Appendix S1: Methods: Targeted panel sequencing and data analysis 2

Table S1: Genes and chromosomal regions included in the customized targeted panel 4

Table S2: Depth of coverage of *SHOX* genomic and up- and downstream regulatory regions included in the panel.. 7

Table S3: Depth of coverage of *SHOX* coding regions included in the panel 9

Figure S1: Schematic representation of the pseudoautosomal region 1 region and MLPA probes with the deletion or duplication map of patients with copy number variants in *SHOX* gene and/or regulatory regions. 10

Figure S2: Depth of coverage of *SHOX* gene. 11

Figure S3: CONTRA and Nexus analyses of Cases 5 and 15 (index case and her mother)...12

Appendix S1

Targeted panel sequencing and data analysis

Genomic DNA of the patients was isolated from peripheral blood leukocytes using standard procedures. DNAs were analyzed by a customized panel of targeted sequencing based on the Agilent SureSelect XT (Agilent Technologies, Santa Clara, CA, USA) capture system with genes associated with growth and short stature) that included 97 genes with a target region of approximately 489kb. In this panel, we included the entire genomic region of the *SHOX* gene and some regions with regulatory functions located around it (up- and downstream enhancer regions)[1] (Supporting Information, Table S1). DNA libraries were sequenced in paired-end mode in pools of 96 samples using the Illumina NextSeq 500 platform with NextSeq V2 2x150 kits or pools of 32 samples using the Illumina MiSeq platform with MiSeq V3 2x300 kits (Illumina, Inc., San Diego, CA, USA).

The raw data were aligned to the reference genome (GRCh37/hg19) with BWA tools[2]. The version of the hg19 assembly used is adapted to handle with the pseudoautosomal regions of the sex chromosomes. Accordingly, the corresponding regions on the Y chromosome (Y:10,001-2,649,520 for PAR1 and Y:59,034,050-59,373,566) are “hard-masked” with NNNs. In addition, we analyze XX and XY patients separately. Variant calling for point mutation analysis was performed with Freebayes and annotated with ANNOVAR. The variants were filtered according to frequency (MAF <0.1%) in public (gnomAD, http://gnomad.broadinstitute.org/ and ABraOM http://abraom.ib.usp.br/)[3,4] and in-house databases (739 samples), location (exons and splice site consensus) and consequences to the protein predicted by *in silico* analyses.

CNV analyses were performed using two software packages: COpy Number Targeted Resequencing Analysis (CONTRA)[5] and Nexus Copy Number (BioDiscovery, Inc., El Segundo, CA, USA)[6]. Both software programs are able to call copy number gains and losses for each target region based on the normalized depth of coverage. We considered log ratios of 0.7 or -0.7 and adjusted p values below 0.05 for the detection of heterozygous duplications or deletions, respectively. Regarding Nexus analysis, we applied the SNP-FASST2 algorithm, and a segment was considered duplicated or deleted when the log2 ratio of the test/reference fluorescence intensities of a given region encompassing at least three probes was above 0.3 or

below -0.3, respectively[7]. We also visually inspected *SHOX* coverage using Integrative Genomics Viewer (IGV) software[8].

All identified CNVs were confirmed by MLPA or direct sequencing of the breakpoints. MLPA analysis was carried out using the commercial kit P018-*SHOX*-G1 (MRC Holland, Amsterdam, Netherlands). Sanger sequencing products were bidirectionally sequenced on an ABI PRISM 3130xl automatic sequencer (Applied Biosystems, Foster City, CA, USA).

References

1. Marchini A, Ogata T, Rappold GA. A Track Record on SHOX: From Basic Research to Complex Models and Therapy. *Endocr Rev.* 2016;37(4):417-448.
2. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754-1760.
3. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536(7616):285-291.
4. Naslavsky MS, Yamamoto GL, de Almeida TF, et al. Exomic variants of an elderly cohort of Brazilians in the ABraOM database. *Hum Mutat.* 2017;38(7):751-763.
5. Li J, Lupat R, Amarasinghe KC, et al. CONTRA: copy number analysis for targeted resequencing. *Bioinformatics.* 2012;28(10):1307-1313.
6. Darvishi K. Application of Nexus copy number software for CNV detection and analysis. *Curr Protoc Hum Genet.* 2010;Chapter 4:Unit 4.14.11-28.
7. Villela D, Costa SS, Vianna-Morgante AM, Krepischi ACV, Rosenberg C. Efficient detection of chromosome imbalances and single nucleotide variants using targeted sequencing in the clinical setting. *Eur J Med Genet.* 2017;60(12):667-674.
8. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011;29(1):24-26.

Table S1: Genes and chromosomal regions included in the customized targeted panel

	Gene and chromosomal coordinates	Captured regions
1	*ACAN*	Coding region
2	*ADAMTS10*	Coding region
3	*ADAMTS17*	Coding region
4	*ARNT2*	Coding region
5	*BMP2*	Coding region
6	*BMPR1B*	Coding region
7	*BRAF*	Coding region
8	*CBL*	Coding region
9	*CCDC53*	Coding region
10	*CCDC8*	Coding region
11	*CDH2*	Coding region
12	*CDON*	Coding region
13	*COL2A1*	Coding region
14	*COMP*	Coding region
15	*CREBBP*	Coding region
16	*CUL7*	Coding region
17	*DGCR8*	Coding region
18	*DMXL2*	Coding region
19	*EP300*	Coding region
20	*FBN1*	Exons 42 and 43
21	*FGF8*	Coding region
22	*FGFR1*	Coding region
23	*FGFR3*	Coding region
24	*GDF5*	Coding region
25	*GH1*	Genomic region
26	*GHR*	Coding region and UTRs
27	*GHRH*	Coding region and UTRs
28	*GHRHR*	Coding region and UTRs
29	*GHSR*	Coding region and UTRs
30	*GLI2*	Coding region and UTRs
31	*GNAS*	Coding region
32	*GPR161*	Coding region
33	*HDAC6*	Coding region
34	*HESX1*	Coding region
35	*HHIP*	Coding region
36	*HMGA2*	Coding region
37	*HOXA9*	Coding region

(to be continued)

Table S1: Genes and chromosomal regions included in the customized targeted panel (cont)

	Gene and chromosomal coordinates	Captured regions
38	*HRAS*	Coding region
39	*IGF1*	Coding region and UTRs
40	*IGF1R*	Coding region and UTRs
41	*IGF2BP2*	Coding region
42	*IGFALS*	Coding region
43	*IGSF1*	Coding region
44	*IGSF10*	Coding region
45	*IHH*	Coding region and UTRs
46	*KAL1*	Coding region
47	*KRAS*	Coding region
48	*LHX3*	Coding region
49	*LHX4*	Coding region
50	*LZTR1*	Coding region
51	*MAP2K1*	Coding region
52	*MAP2K2*	Coding region
53	*MEF2C*	Coding region
54	*MLL2*	Coding region
55	*NF1*	Coding region
56	*NPPB*	Coding region
57	*NPPC*	Coding region and UTRs
58	*NPR2*	Coding region and UTRs
59	*NRAS*	Coding region
60	*NSUN2*	Coding region
61	*OBSL1*	Coding region
62	*OTX2*	Coding region
63	*PAPPA2*	Coding region
64	*PITX2*	Coding region
65	*PNPLA6*	Coding region
66	*POU1F1*	Coding region and UTRs
67	*PRKG2*	Coding region
68	*PROKR2*	Coding region
69	*PROP1*	Coding region and UTRs
70	*PTCH1*	Coding region
71	*PTPN11*	Coding region
72	*RAB3IP*	Coding region
73	*RAF1*	Coding region
74	*RASA2*	Coding region

(to be continued)

Table S1: Genes and chromosomal regions included in the customized targeted panel (cont.)

	Gene and chromosomal coordinates	Captured regions
75	*RIT1*	Coding region
76	*RNPC3*	Coding region
77	*ROR2*	Coding region
78	*RUNX2*	Coding region
79	*SHH*	Coding region and UTRs
80	*SHOC2*	Coding region
81	*SHOX*	Genomic region
82	*SHOX2*	Coding region
83	*SIX3*	Coding region
84	*SMO*	Coding region
85	*SOS1*	Coding region
86	*SOS2*	Coding region
87	*SOX2*	Coding region
88	*SOX3*	Coding region
89	*SOX5*	Coding region
90	*SOX6*	Coding region
91	*SOX9*	Coding region
92	*SRCAP*	Coding region
93	*STAT5B*	Coding region and UTRs
94	*TCF7L1*	Coding region
95	*TGIF1*	Coding region
96	*WNT5A*	Coding region
97	*ZIC2*	Coding region
98	chrX:398,100-399,050; chrY:348,100-349,050	SHOX enhancer region
99	chrX:460,100-460,900; chrY:410,100-410,900	SHOX enhancer region
100	chrX:516,400-517,400; chrY:466,400-467,400	SHOX enhancer region
101	chrX:713,900-714,900; chrY:663,900-664,900	SHOX enhancer region
102	chrX:750,700-752,000; chrY:700,700-702,000	SHOX enhancer region
103	chrX:763,900-764,900; chrY:713,900-714,900	SHOX enhancer region
104	chrX:780,400-781,400; chrY:730,400-731,400	SHOX enhancer region
105	chrX:800,700-802,000; chrY:750,700-752,000	SHOX enhancer region
106	chrX:809,000-809,500; chrY:759,000-759,500	SHOX enhancer region
107	chrX:817,500-818,000; chrY:767,500-768,000	SHOX enhancer region
108	chrX:834,500-835,700; chrY:784,500-785,700	SHOX enhancer region
109	chrX:884,500-885,700; chrY:834,500-835,700	SHOX enhancer region

Table S2: Depth of coverage of *SHOX* genomic and up- and downstream regulatory regions included in the panel

ChrX positions	ChrY positions	Region	Enhancer	Size (pb)	Region with >10x depth of coverage (%)*
chrX:398,100-399,050	chrY:348,100-349,050	Upstream	CNE-5	950	97.7
chrX:460,100-460,900	chrY:410,100-410,900	Upstream	CNE-3	800	99.4
chrX:516,400-517,400	chrY:466,400-467,400	Upstream	CNE-2	1000	100.0
chrX:585,079-607,558	chrY:535,079-557,558	*SHOX* genomic**	-	22,480	81.0
chrX:713,900-714,900	chrY:663,900-664,900	Downstream	CNE4	1000	98.4
chrX:750,700-752,000	chrY:700,700-702,000	Downstream	CNE5	1300	100.0
chrX:780,400-781,400	chrY:730,400-731,400	Downstream	CNE7	1000	100.0
chrX:809,000-809,500	chrY:759,000-759,500	Downstream	CNE8	500	100.0
chrX:817,500-818,000	chrY:767,500-768,000	Downstream	CNE8/9	500	100.0
chrX:834,500-835,700	chrY:784,500-785,700	Downstream	CNE9	1200	93.4

The chromosomal coordinates are according to GRCh37/hg19. These coordinates include the major regulatory regions of the *SHOX* already described[1-4].
* This column corresponds to the size of target region (in percentage) with at least 10x depth of coverage.
** Genomic region of *SHOX*'s main transcript NM_000451.
Chr: chromosome; bp: basepair; CNE: conserved non-coding DNA element.

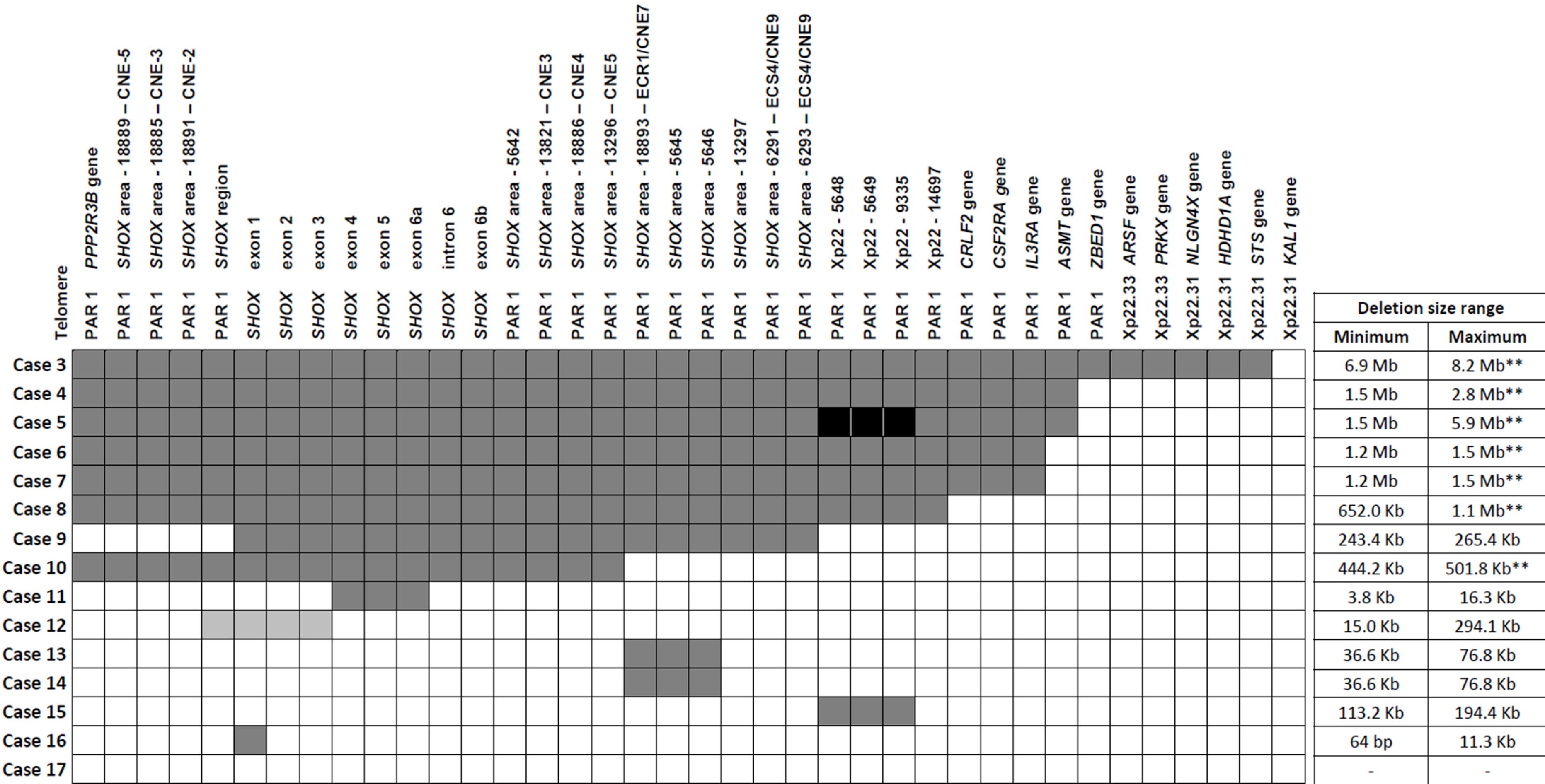
1. Fukami M, Kato F, Tajima T, Yokoya S, Ogata T. Transactivation function of an approximately 800-bp evolutionarily conserved sequence at the SHOX 3' region: implication for the downstream enhancer. *Am J Hum Genet.* 2006;78(1):167-170.
2. Sabherwal N, Bangs F, Röth R, et al. Long-range conserved non-coding SHOX sequences regulate expression in developing chicken limb and are associated with short stature phenotypes in human patients. *Hum Mol Genet.* 2007;16(2):210-222.
3. Durand C, Bangs F, Signolet J, Decker E, Tickle C, Rappold G. Enhancer elements upstream of the SHOX gene are active in the developing limb. *Eur J Hum Genet.* 2010;18(5):527-532.
4. Benito-Sanz S, Royo JL, Barroso E, et al. Identification of the first recurrent PAR1 deletion in Léri-Weill dyschondrosteosis and idiopathic short stature reveals the presence of a novel SHOX enhancer. *J Med Genet.* 2012;49(7):442-450.

Table S3: Depth of coverage of *SHOX* coding regions included in the panel

ChrX positions	ChrY positions	Region	Mean coverage (x)	Maximum coverage (x)	Minimum Coverage (x)
chrX:591,633-591,909	chrY:541,633-541,909	Exon 2	480	913	241
chrX:595,353-595,561	chrY:545,353-545,561	Exon 3	394	805	160
chrX:601,556-601,613	chrY:551,556-551,613	Exon 4	407	843	217
chrX:601,734-601,822	chrY:551,734-551,822	Exon 5	465	905	234
chrX:605,126-605,368	chrY:555,126-555,368	Exon 6	218	686	44

The coding regions correspond to the *SHOX*'s main transcript NM_000451. All *SHOX* coding region of our cohort of patients was sequenced at least 44x. The chromosomal coordinates are according to GRCh37/hg19.
Chr: chromosome; x: number of times that the region was sequenced.

Figure S1: Schematic representation of the pseudoautosomal region 1 (PAR1) and multiplex ligation-dependent probe amplification (MLPA) probes (kit P018) with the deletion or duplication map of patients with copy number variants in *SHOX* gene and/or regulatory regions. The numbers indicated in the upper part of the figure correspond to the identification of MLPA probes. The dark gray squares indicate regions deleted in the heterozygous state; the black squares indicate regions deleted in the homozygous state; the light gray squares indicate a duplication; and the white squares indicate retained regions. Minimum and maximum approximated deletion interval, determined by MLPA data, is indicated adjacent to each deletion. Cases 3 to 15 correspond to individuals with previously known *SHOX* defects; Cases 16 and 17 correspond to individuals with deletions initially detected by the NGS panel in the prospective evaluation. The deletion identified in Case 17 is not indicated in the figure, because it was not detected by MLPA. The Case 5 has two deletions: the smallest deletion (in black) is similar to the deletion detected in Case 15 who is her mother. CNE: conserved noncoding element; ECR: evolutionarily conserved region; ECS: evolutionarily conserved sequence.



Deletion size range	
Minimum	Maximum
6.9 Mb	8.2 Mb**
1.5 Mb	2.8 Mb**
1.5 Mb	5.9 Mb**
1.2 Mb	1.5 Mb**
1.2 Mb	1.5 Mb**
652.0 Kb	1.1 Mb**
243.4 Kb	265.4 Kb
444.2 Kb	501.8 Kb**
3.8 Kb	16.3 Kb
15.0 Kb	294.1 Kb
36.6 Kb	76.8 Kb
36.6 Kb	76.8 Kb
113.2 Kb	194.4 Kb
64 bp	11.3 Kb
-	-

Figure S2: Depth of coverage of *SHOX* gene. This IGV image shows the real coverage of the *SHOX* sequencing in our panel. The chart at the top of the figure indicates the depth of coverage. On the bottom, inside the boxes, we can see the exons of the main transcript of *SHOX* (1 to 6). The gray bars, located above the exons, correspond to the sequencing reads. The regions without reads are the regions with no coverage in the sequencing. From this image we can see a good coverage of the coding region of the gene. The uncovered regions are all intronic regions.

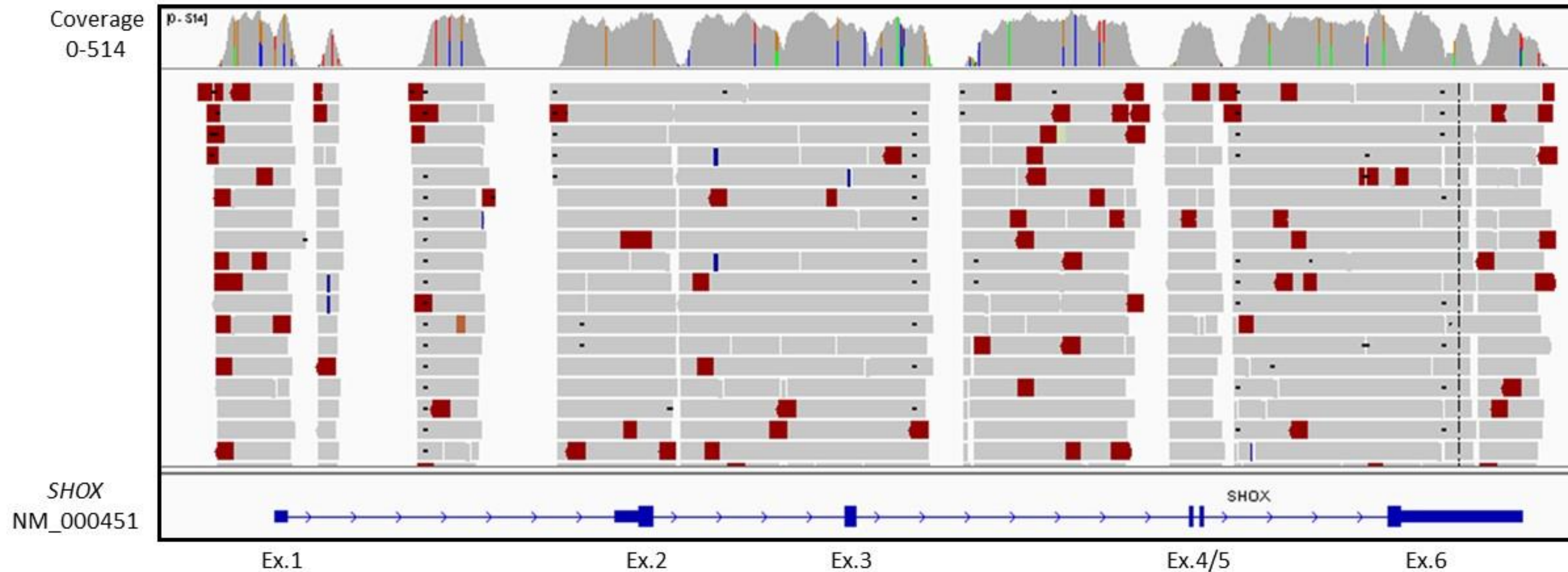


Figure S3: CONTRA and Nexus analyses of Cases 5 and 15 (index case and her mother). CONTRA (A) and Nexus (B) plots of Case 5, who has two deletions: a large deletion in the paternal allele and a second deletion located downstream of *SHOX* in the maternal allele. Thus, she has a homozygous deletion exactly in this downstream region. In the CONTRA plot of the index case (A) the heterozygous deletion is indicated by the upper arrow. The log ratios of those dots in the *SHOX* region were around -1.0. The homozygous deletion is indicated by the bottom arrow. The X (in the lower part of the plot) indicates a homozygous deletion between ChrX: 884,500-885,700 (GRCh37/hg19). In the CONTRA plot of her mother (C) we can see dots in this same region with log ratios near -1.0, indicating a heterozygous deletion. In the Nexus plot of the index case (B) the large heterozygous deletion is indicated by dots with log ratios near -1.0 from the beginning until ~835,000, including the *SHOX* gene. In the Nexus plot of her mother (D) the dots in this same region have log ratios near zero, indicating normal copy number. Nexus did not detect the downstream deletion, probably located between 835,000 and 885,000 (GRCh37/hg19) (indicated by the rectangle), in homozygous state in Case 5, and in heterozygous in Case 15. We can see one single dot with log ratios below -2.0, in Case 5 (B), and other single dot with log ratios below -0.5, in Case 15 (D). Probably those dots are located in the deleted region, but they are not sufficient for the software to call a deletion, so we can say that the Nexus was not able identify this downstream deletion.