Next-generation phenotyping contributing to the identification of a non-coding deletion in KANSL1 causing Koolen-de Vries syndrome

Peter Krawitz[1], Fabian Brand[1], and Claudia Perne[1]

[1]University of Bonn

April 19, 2022

Abstract

Next-generation phenotyping (NGP) is the application of advanced methods of computer vision on medical imaging data such as portrait photos of individuals with rare disorders. NGP on portraits results in gestalt scores that can be used for the selection of appropriate genetic tests, and for the interpretation of the molecular data. Here, we report on an exceptional case of a young girl that was presented at the age of eight and fifteen and enrolled in NGP diagnostics at the latter occasion. The girl had clinical features associated with Koolen-de Vries syndrome and a suggestive facial gestalt. However, chromosomal micro array (CMA), Sanger sequencing, multiplex ligation-dependent probe analysis (MLPA), and trio exome sequencing remained inconclusive. Based on the highly indicative gestalt score for Koolen-de Vries, the decision was made to perform genome sequencing to also evaluate non-coding variants. This analysis revealed a 4.7 kb deletion at the end of intron 6 of the KANSL1 gene, which is the smallest reported structural variant to date for this phenotype. The case illustrates how NGP can be integrated into the iterative diagnostic process of test selection and interpretation of sequencing results.

Introduction

Many genetic syndromes are associated with a distinctive facial gestalt which can be used to expedite the diagnostic process. Although high-throughput sequencing has helped to address the considerable heterogeneity of many syndromes in a single test, the rare expertise of dysmorphologists, which is still required for the data interpretation, is often the bottleneck. In recent years, advances in machine learning have enabled the development of NGP tools, that can be used to analyze facial dysmorphology in patient portrait photos (Ferry et al., 2014; Kuru et al., 2014; Gripp et al., 2016; Wang and Luo, 2016; Dudding-Byth et al., 2017; Hadj-Rabia et al., 2017; Valentine et al., 2017; Liehr et al., 2018; Gurovich et al., 2019; van der Donk et al., 2019; Hsieh et al., 2022). Amongst them is GestaltMatcher, which is a deep convolutional neural network that was trained on thousands of molecularly confirmed cases and achieves high accuracies in the identification of hundreds of syndromes (Hsieh et al., 2022). In this paper, we describe how the results of this artificial intelligence helped to solve a case with a typical phenotype of Koolen-de Vries syndrome but an unusual disease-causing mutation.

Results

We report a female patient who first presented to a syndromic consultation at the age of eight because auf multiple phenotypic abnormalities. The girl had muscular hypotonia since early childhood. During infancy a developmental delay became noticeable and later she scored in the moderate range of intellectual deficiency. Brain MRI showed two heterotopic foci as well as symmetrically clumped hippocampi. Facial dysmorphism, which became more prominent as a teenager, includes a long face, slightly up slanting palpebral fissures, ptosis of the left eye, a prominent, bulbous nasal tip and low-hanging columella (Figure 1). Furthermore, she had pale skin with many moles, thick curly hair, and a missing left upper canine tooth. Her family described her as extremely friendly, but anxious in contact with other children. A chromosome analysis, a chromosomal microarray (CMA) and diagnostics for Fragile X Syndrome, which have been performed after the first consultation at the age of eight years, were unremarkable.

At re-consultation seven years later, the fifteen-year-old female was enrolled into a study protocol that also involved NGP. The computer-assisted assessment of portrait images yielded high gestalt scores for Koolen-de Vries Syndrome (Figure 1). The feature score, which is based on the clinical abnormalities annotated in HPO terminology, was in the lower range reflecting the rather unspecific phenotypic manifestations in the young female (Robinson et al., 2008; Peng et al., 2021). Although some characteristic aspects of the facial gestalt, such as the elongation of the face and the pear-shaped nose, become more prominent with age, the gestalt score for the portrait at the age of eight years was already comparably high (Figure 1).

With facial dysmorphism typical for Koolen-de Vries Syndrome and some matching phenotypic features such as the friendly personality, this diagnosis was suspected despite the inconclusive CMA results. While ˜95% of the cases with Koolen-de Vries syndrome are due to 500 to 650 kb deletion in 17q21.31, only ˜5% are due to sequence variants in *KANSL1* (Koolen et al., 2006, 2012, 2016; Sharp et al., 2006; Shaw-Smith et al., 2006; Zollino et al., 2012, 2015). Around the microdeletion in 17q21.31 large clusters of low complexity repeats at the breakpoints were described, suggesting an underlying mechanism of non-allelic homologous recombination (NAHR) (Stankiewicz and Lupski, 2002; Dubourg et al., 2011). Up to now, these deletions were found by CMA. So far, only few atypical deletions had been reported for individuals affected by Koolen-de Vries Syndrome, the smallest of these still 68 kb in size (Cooper et al., 2011; Dubourg et al., 2011; Koolen et al., 2012; Zollino et al., 2015). All of these deletions were also detected by CMA.

As the recurrent microdeletion in 17q21.31 was not supported by CMA we initiated Sanger sequencing and multiplex ligation-dependent probe amplification (MLPA) of *KANSL1* . Both analyses did not show any abnormal findings. Next, a trio exome analysis in the patient and her parents was performed. Data for the patient and her parents was generated using the NovaSeq platform (Illumina) and the SureSelect v6 exome capture kit (Agilent). Initial bioinformatics analysis was focused on relevant single nucleotide variants (SNVs) and indels using a local implementation of GATK best practices pipelines optimized for data from the NovaSeq sequencer. Copy number variants (CNVs) were initially generated using cn.MOPS (Klambauer et al., 2012). No variants in*KANSL1* nor any other gene were detected that could explain the phenotype. Following the inconclusive results of the trio exome analysis, a genome sequencing was conducted. The bioinformatics analysis was performed using the NVIDIA Parabricks toolkit. This toolkit enables accelerated genome analysis by utilizing NVIDIA GPU resources. Several algorithms from this toolkit have been used to call SNVs and indels on the genomic data of the patient. In particular, accelerated versions of BWA-mem and the HaplotypeCaller were crucial for fast processing and yielded variant calls of high quality. To determine candidates for structural variants (SVs) and CNVs, we used manta (Chen et al., 2016), delly (Rausch et al., 2012) and lumpy (Layer et al., 2014). Variant calls of all three tools were merged using a vote-based scheme to find candidates supported by all callers. A 4,708 bp deletion affecting the end of intron 6 and only the first 46 bp of exon 7 (NM_015443.4:c.1849-4611_1895del) was detected by all three tools. Furthermore, the deletion was also clearly visible by a drop of coverage and by split reads in the sequence alignment (Figure 2). In a careful re-analysis of the exome data, that was guided by the results from genome sequencing data, the deletion could also be detected using Pindel (Ye et al., 2009) (Figure 2). Changing some alignments preferences in Integrative Genome Viewer enabled the visualization of the deletion in*KANSL1* exome sequencing data (Figure 2). The deletion was also subsequently verified by qPCR.

Discussion

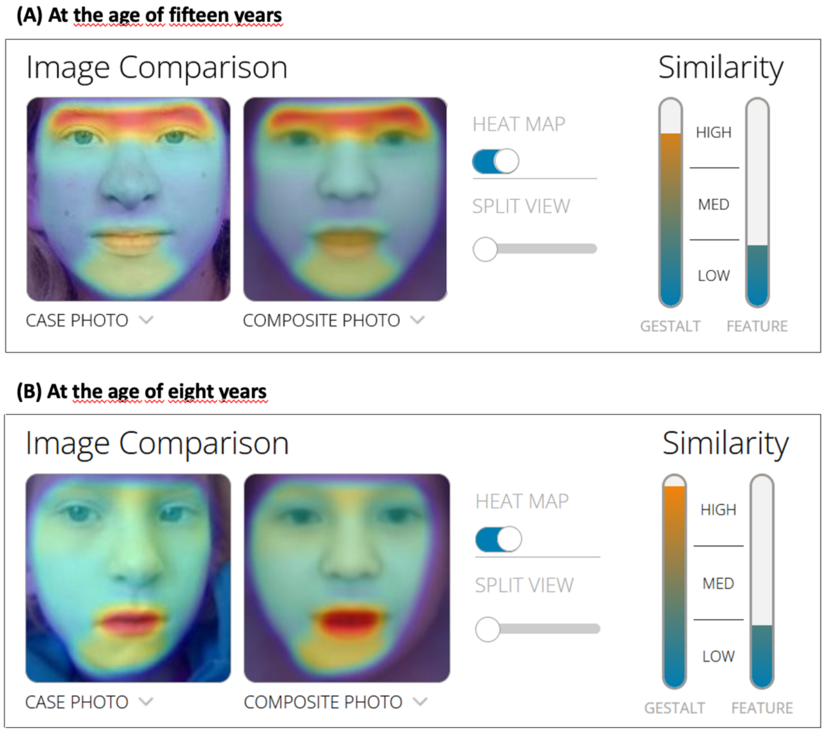
Many SV and CNV tools for exome data rely on depth of coverage signals to identify likely candidates for structural changes in the genome in short read Illumina data. For both, exome and genome data, the effectiveness of this approach is limited by the availability of good normalized control data from other genomic regions in the same individual or other individuals of the same sequencing run. In case of the trio-exome sequencing experiment from our patient, this baseline was formed by other unrelated samples sequenced in parallel. Depth and variability of the coverage in certain genomic regions also has an influence on the ability of those callers to detect structural change to the genome. Other CNV detection methods rely on a mix of other factors to find likely candidates for variation. Pindel incorporates signals from split reads. These are read pairs in which one of the two reads cannot be aligned to the reference genome and is assumed to carry the precise breakpoint information of insertion or deletion events. Similar metrics are used also by other callers that were used for subsequent genome sequencing data analysis (e.g. manta, delly, lumpy).

The initial negative result using other CNV calling methods is due to the suboptimal coverage distribution at some of the *KANSL1* exons and intronic regions and the fact that the deletion reaches only 46 bp into the exon. The variant in question is mainly in the end of intron 7 making coverage-based detection of structural changes based on exome data substantially more difficult than in genome sequencing data. As a result, from sequence analysis, 130 pathogenic or likely pathogenic variants have been reported for *KANSL1* in the database ClinVar (Landrum et al., 2020). In contrast, the 4.7 kb deletion that we identified, is the first entry in ClinVar for a variant length in between 51 bp and 50 kb.

In conclusion, we reported a 4.7 kb deletion in *KANSL1* that is mainly non-coding and was therefore first detected by genome sequencing. However, retrospectively it could also be confirmed in exome sequencing data with fine-tuning of the filter settings. Since high accuracy in CMA analysis is limited to a resolution of 50 kb or higher, and in exome analysis to a resolution of 50 bp or lower, deletions in the order of few kilobases are not detected in the diagnostic tests most often used today. In genome sequencing data, on the other hand SV and CNVs in this size range can be identified more easily, but are usually more difficult to interpret, if they are non-coding.

Therefore, our case exemplifies, how computer-assisted analysis of the portrait can make a significant contribution to the diagnostic process. First, NGP has the potential to speed up data analysis. If our Koolen-de Vries patient would have carried the recurrent microdeletion, a SNV or indel, the high gestalt score would have made the molecular confirmation of the suspected clinical diagnosis straightforward using protocols such as the PEDIA workflow (Hsieh et al., 2019). Second, highly suggestive results of NGP can be used to request genome sequencing if exome or CMA analysis were inconclusive. Third, NGP can help with the classification of the pathogenicity of novel variants found in the genome.

According to the guidelines from 2015, a matching phenotype is only considered as supporting evidence for pathogenicity of a sequence variant (PP4) (Richards et al., 2015). However, experienced dysmorphologists may attribute a higher level of evidence to the pathogenicity of a variant in a gene if the associated phenotype is highly specific (Zhang et al., 2020). Most clinicians that are confronted for the first time with such a specific diagnosis will be hesitant to apply these higher weights. Here, computer-assisted analysis could help, since syndromic distinctiveness can be measured and the similarity of a portrait to other molecularly confirmed cases can be quantified (Hsieh et al., 2022). By this means, NGP makes the visual inspection of a patient applicable to a Bayesian classification framework (Tavtigian et al., 2018). Interestingly, the specificity of the facial gestalt of Koolen-de Vries Syndrome ranges only in the upper half of dysmorphic phenotypes and is exceeded for example by the distinctiveness of Baraitser-Winter syndrome or Seckel syndrome. For disorders in this category high gestalt scores should therefore be handled with even greater attention and could justify more comprehensive tests such as genome sequencing if molecular confirmation is still pending.



(A) At the age of fifteen years

(B) At the age of eight years

Figure 1. Graphical heatmaps of the patient´s photos. Photos were taken at A) the age of fifteen and B) at the age of eight. Image comparison shows a high similarity between the descriptor of the patient´s photo and the composite image of Koolen-de Vries syndrome. The heatmaps illustrate the facial elements that informed the DeepGestalt algorithm in this patient. For calculation of the Feature score the following HPO-terms were used: generalized muscle weakness; nevus; intellectual disability, moderate; poor fine motor coordination; tonsillitis; agenesis of canine; abnormal morphology of the hippocampus; decreased serum iron.

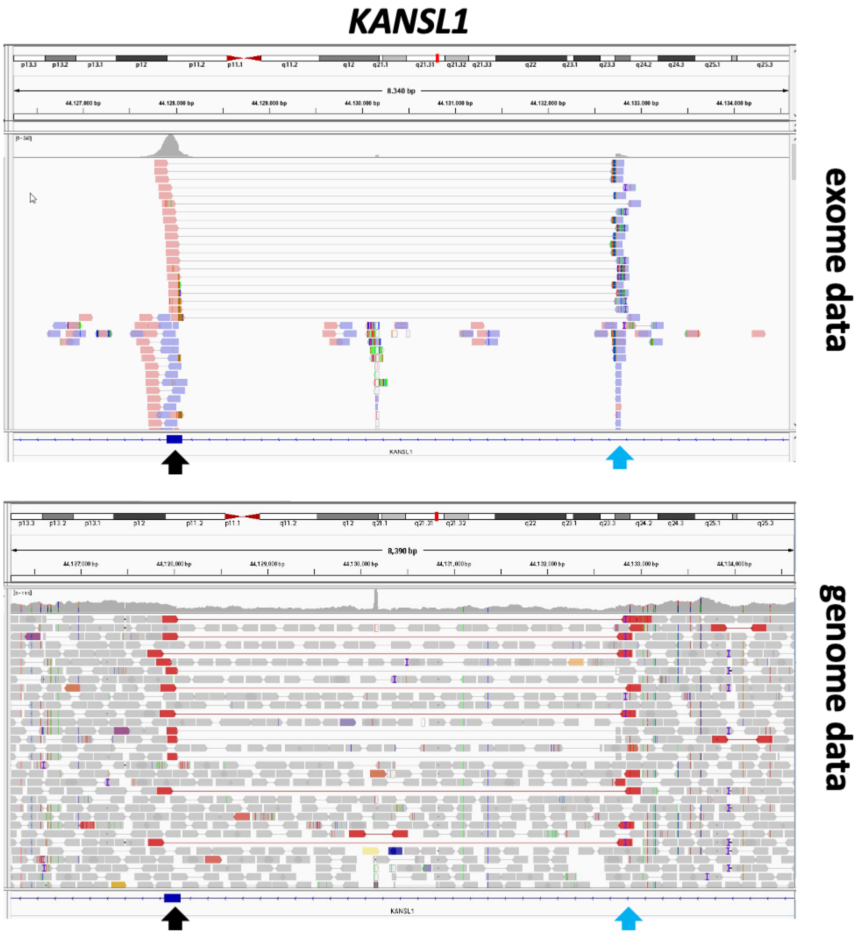


Figure 2. *KANSL1* whole exome and whole genome sequencing data. *KANSL1* sequencing data visualized in IGV. A) A screenshot of the deletion in whole exome sequencing data. Reads are sorted by start location and grouped by read pairs. Soft clipped bases are included, making the exact breakpoint in DNA visible, even against the complete lack of other reads in the region. B) A Screenshot of whole genome sequencing data is shown. The deletion causes a noticeable drop of coverage. Additional support for the detected deletion is provided by split reads, marked in red.

Black arrow = 3‘- end of deletion (Exon 7)

Blue arrow = 5‘-end of deletion (intron 6)

References

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32:1220–1222.

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, Abdel-Hamid H, Bader P, et al. 2011. A copy number variation morbidity map of developmental delay. Nat Genet 43:838–846.

Donk R van der, Jansen S, Schuurs-Hoeijmakers JHM, Koolen DA, Goltstein LCMJ, Hoischen A, Brunner HG, Kemmeren P, Nellåker C, Vissers LELM, Vries BBA de, Hehir-Kwa JY. 2019. Next-generation phenotyping using computer vision algorithms in rare genomic neurodevelopmental disorders. Genet Med 21:1719–1725.

Dubourg C, Sanlaville D, Doco-Fenzy M, Le Caignec C, Missirian C, Jaillard S, Schluth-Bolard C, Landais E, Boute O, Philip N, Toutain A, David A, et al. 2011. Clinical and molecular characterization of 17q21.31 microdeletion syndrome in 14 French patients with mental retardation. Eur J Med Genet 54:144–151.

Dudding-Byth T, Baxter A, Holliday EG, Hackett A, O'Donnell S, White SM, Attia J, Brunner H, Vries B de, Koolen D, Kleefstra T, Ratwatte S, et al. 2017. Computer face-matching technology using two-dimensional photographs accurately matches the facial gestalt of unrelated individuals with the same syndromic form of intellectual disability. BMC Biotechnol 17:90.

Ferry Q, Steinberg J, Webber C, FitzPatrick DR, Ponting CP, Zisserman A, Nellåker C. 2014. Diagnostically relevant facial gestalt information from ordinary photos. Elife 3:e02020.

Gripp KW, Baker L, Telegrafi A, Monaghan KG. 2016. The role of objective facial analysis using FDNA in making diagnoses following whole exome analysis. Report of two patients with mutations in the BAF complex genes. Am J Med Genet A 170:1754–1762.

Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, Basel-Salmon L, Krawitz PM, Kamphausen SB, Zenker M, Bird LM, Gripp KW. 2019. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med 25:60–64.

Hadj-Rabia S, Schneider H, Navarro E, Klein O, Kirby N, Huttner K, Wolf L, Orin M, Wohlfart S, Bodemer C, Grange DK. 2017. Automatic recognition of the XLHED phenotype from facial images. Am J Med Genet A 173:2408–2414.

Hsieh T-C, Bar-Haim A, Moosa S, Ehmke N, Gripp KW, Pantel JT, Danyel M, Mensah MA, Horn D, Rosnev S, Fleischer N, Bonini G, et al. 2022. GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. Nat Genet.

Hsieh T-C, Mensah MA, Pantel JT, Aguilar D, Bar O, Bayat A, Becerra-Solano L, Bentzen HB, Biskup S, Borisov O, Braaten O, Ciaccio C, et al. 2019. PEDIA: prioritization of exome data by image analysis. Genet Med 21:2807–2814.

Klambauer G, Schwarzbauer K, Mayr A, Clevert D-A, Mitterecker A, Bodenhofer U, Hochreiter S. 2012. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res 40:e69.

Koolen DA, Kramer JM, Neveling K, Nillesen WM, Moore-Barton HL, Elmslie FV, Toutain A, Amiel J, Malan V, Tsai AC-H, Cheung SW, Gilissen C, et al. 2012. Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. Nat Genet 44:639–641.

Koolen DA, Pfundt R, Linda K, Beunders G, Veenstra-Knol HE, Conta JH, Fortuna AM, Gillessen-Kaesbach G, Dugan S, Halbach S, Abdul-Rahman OA, Winesett HM, et al. 2016. The Koolen-de Vries syndrome: a phenotypic comparison of patients with a 17q21.31 microdeletion versus a KANSL1 sequence variant. Eur J Hum Genet 24:652–659.

Koolen DA, Vissers LELM, Pfundt R, Leeuw N de, Knight SJL, Regan R, Kooy RF, Reyniers E, Romano C, Fichera M, Schinzel A, Baumer A, et al. 2006. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. Nat Genet 38:999–1001.

Kuru K, Niranjan M, Tunca Y, Osvank E, Azim T. 2014. Biomedical visual data analysis to build an intelligent diagnostic decision support system in medical genetics. Artif Intell Med 62:105–118.

Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, Hoffman D, Jang W, Kaur K, Liu C, Lyoshin V, Maddipatla Z, et al. 2020. ClinVar: improvements to accessing data. Nucleic Acids Res 48:D835–D844.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 15:R84.

Liehr T, Acquarola N, Pyle K, St-Pierre S, Rinholm M, Bar O, Wilhelm K, Schreyer I. 2018. Next generation phenotyping in Emanuel and Pallister-Killian syndrome using computer-aided facial dysmorphology analysis of 2D photos. Clin Genet 93:378–381.

Peng C, Dieck S, Schmid A, Ahmad A, Knaus A, Wenzel M, Mehnert L, Zirn B, Haack T, Ossowski S, Wagner M, Brunet T, et al. 2021. CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. NAR Genom Bioinform 3:.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:i333–i339.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, et al. 2015. 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 17:405–424.

Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. 2008. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 83:610–615.

Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, Fitzpatrick CA, Segraves R, et al. 2006. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat Genet 38:1038–1042.

Shaw-Smith C, Pittman AM, Willatt L, Martin H, Rickman L, Gribble S, Curley R, Cumming S, Dunn C, Kalaitzopoulos D, Porter K, Prigmore E, et al. 2006. Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. Nat Genet 38:1032–1037.

Stankiewicz P, Lupski JR. 2002. Genome architecture, rearrangements and genomic disorders. Trends Genet 18:74–82.

Tavtigian SV, Greenblatt MS, Harrison SM, Nussbaum RL, Prabhu SA, Boucher KM, Biesecker LG. 2018. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet Med.

Valentine M, Bihm DCJ, Wolf L, Hoyme HE, May PA, Buckley D, Kalberg W, Abdul-Rahman OA. 2017. Computer-Aided Recognition of Facial Attributes for Fetal Alcohol Spectrum Disorders. Pediatrics 140:.

Wang K, Luo J. 2016. Detecting Visually Observable Disease Symptoms from Faces. EURASIP J Bioinform Syst Biol 2016:13.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25:2865–2871.

Zhang J, Yao Y, He H, Shen J. 2020. Clinical Interpretation of Sequence Variants. Curr Protoc Hum Genet 106:e98.

Zollino M, Marangi G, Ponzi E, Orteschi D, Ricciardi S, Lattante S, Murdolo M, Battaglia D, Contaldo I, Mercuri E, Stefanini MC, Caumes R, et al. 2015. Intragenic KANSL1 mutations and chromosome 17q21.31 deletions: broadening the clinical spectrum and genotype-phenotype correlations in a large cohort of patients. J Med Genet 52:804–814.

Zollino M, Orteschi D, Murdolo M, Lattante S, Battaglia D, Stefanini C, Mercuri E, Chiurazzi P, Neri G, Marangi G. 2012. Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype. Nat Genet 44:636–638.