The American Journal of Human Genetics, Volume *108*

Supplemental information

Heterozygous *ANKRD17* loss-of-function variants cause a syndrome with intellectual disability, speech delay, and dysmorphism

Maya Chopra, Meriel McEntagart, Jill Clayton-Smith, Konrad Platzer, Anju Shukla, Katta M. Girisha, Anupriya Kaur, Parneet Kaur, Rolph Pfundt, Hermine Veenstra-Knol, Grazia M.S. Mancini, Gerarda Cappuccio, Nicola Brunetti-Pierri, Fanny Kortüm, Maja Hempel, Jonas Denecke, Anna Lehman, CAUSES Study, Tjitske Kleefstra, Kyra E. Stuurman, Martina Wilke, Michelle L. Thompson, E. Martina Bebin, Emilia K. Bijlsma, Mariette J.V. Hoffer, Cacha Peeters-Scholte, Anne Slavotinek, William A. Weiss, Tiffany Yip, Ugur Hodoglugil, Amy Whittle, Janette diMonda, Juanita Neira, Sandra Yang, Amelia Kirby, Hailey Pinz, Rosan Lechner, Frank Sleutels, Ingo Helbig, Sarah McKeown, Katherine Helbig, Rebecca Willaert, Jane Juusola, Jennifer Semotok, Medard Hadonou, John Short, Genomics England Research Consortium, Naomi Yachelevich, Sajel Lala, Alberto Fernández-Jaen, Janvier Porta Pelayo, Chiara Klöckner, Susanne B. Kamphausen, Rami Abou Jamra, Maria Arelin, A. Micheil Innes, Anni Niskakoski, Sam Amin, Maggie Williams, Julie Evans, Sarah Smithson, Damian Smedley, Anna de Burca, Usha Kini, Martin B. Delatycki, Lyndon Gallacher, Alison Yeung, Lynn Pais, Michael Field, Ellenore Martin, Perrine Charles, Thomas Courtin, Boris Keren, Maria Iascone, Anna Cereda, Gemma Poke, Véronique Abadie, Christel Chalouhi, Padmini Parthasarathy, Benjamin J. Halliday, Stephen P. Robertson, Stanislas Lyonnet, Jeanne Amiel, and Christopher T. Gordon

SUPPLEMENTAL DATA

Supplemental Figure

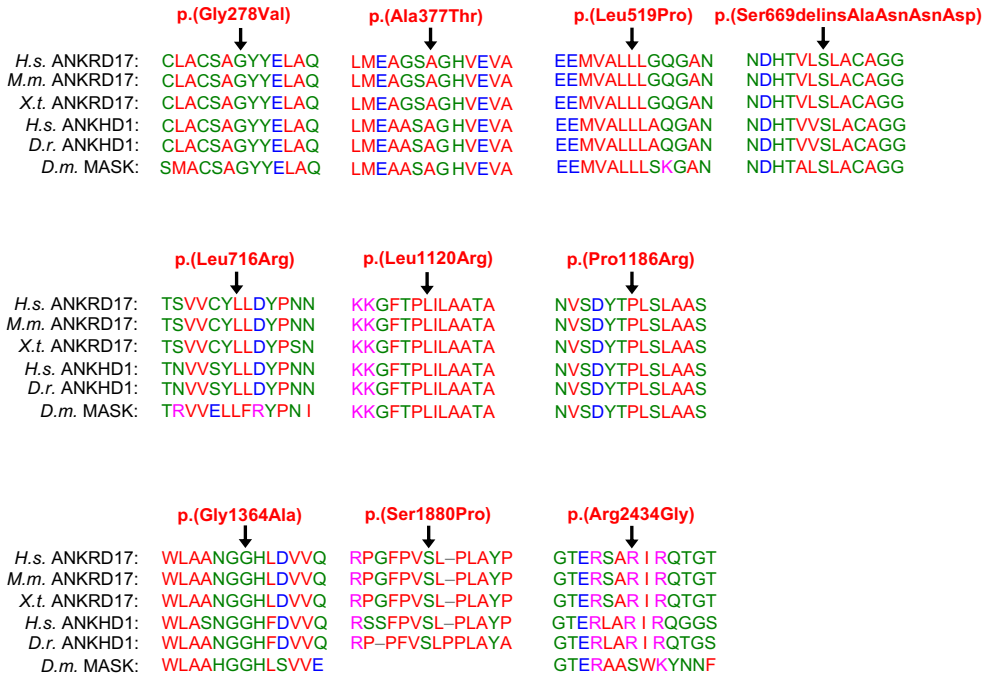

Figure S1 - ANKRD17 non-truncating variants affect highly conserved amino acids. Multiple sequence alignments of proteins homologous to ANKRD17 were performed using Clustal Omega. Arrows indicate amino acids affected by non-truncating variants. Reference sequences used for the alignment: NP_115593, NP_112148.2, XP_002940399.2, NP_060217.1, NP_001186697.1, NP_788733.1. *H.s., Homo sapiens; M.m., Mus musculus; X.t., Xenopus tropicalis; D.r., Danio rerio; D.m., Drosophila melanogaster*.

Supplemental Tables

Table S1 - phenotypic summaries of individuals 1-34. (See separate Excel spreadsheet available on-line).

Table S2 – detailed phenotypic description of individuals 1-34. (See separate Excel spreadsheet available on-line).

Supplemental Methods

Methodology for whole exome sequencing, whole genome sequencing and array-CGH for each centre.

Written informed consent was obtained for either diagnostic or institutional review board-approved research sequencing.

INDIVIDUAL 1: Trio WES was performed according to a previously published protocol[1].

INDIVIDUAL 2: Trio WES was performed according to a previously published protocol[2].

INDIVIDUALS 3 AND 9: Trio WES was performed according to a previously published protocol[3].

INDIVIDUALS 4, 10 and 18: Trio WES was performed according to the following protocol: DNA was enriched using Agilent SureSelect Clinical Research Exome V2 capture and paired-end sequenced on the Illumina platform (outsourced). The average coverage of the exome was ~50x. Duplicate reads were excluded. Data were demultiplexed with bcl2fastq Conversion Software from Illumina. Reads were mapped to the genome using the Burrows-Wheeler Aligner (BWA)-MEM algorithm (reference: http://bio-bwa.sourceforge.net). Variant detection was performed by the Genome Analysis Toolkit (GATK) HaplotypeCaller (reference http://www.broadinstitute.org/gatk/). The detected variants were filtered and annotated with Cartagenia software and classified with Alamut Visual.

INDIVIDUAL 5: Trio WES was performed according to a previously published protocol[4].

INDIVIDUAL 6: Array-CGH (by SurePrint G3 Human CGH Microarray 4x180K Agilent Technologies) was performed on the proband. *De novo* status was established following array-CGH on the parents.

INDIVIDUAL 7: Trio WES was performed according to a previously published protocol[5].

INDIVIDUAL 8: Trio exome sequencing was undertaken at Ambry Genetics under a research protocol on an Illumina sequencing platform in the CAUSES Study (approved by the University of British Columbia ethics review board). Read depth coverages for the proband and parents were 10-fold for >99% of consensus coding sequences. Reads were aligned to UCSC Genome Browser build hg19 with BWA and were called with GATK. Detected variants were annotated, filtered, and prioritized with VarSeq v1.5 (Golden Helix, Bozeman, MT, USA).

INDIVIDUALS 11 AND 12: Trio WGS was undertaken according to the following protocol: DNA was isolated from blood of the patient and parents using QIAsymphony (Qiagen). Sequencing libraries were constructed from whole blood genomic DNA using HudsonAlpha Clinical Sequencing Lab's custom whole genome library preparation protocol. DNA was sequenced on the Illumina HiSeq X sequencer. DNA library fragments were sequenced from both ends (paired) with a read length of 150 base pairs. Genomes were sequenced at an approximate depth of 30X, with at least 80% of base positions reaching 20X coverage. Reads were aligned and variants called according to standard protocols[6,7]. A robust relationship inference algorithm (KING) was used to confirm familial relationships[8]. Sequenced variants were loaded into a custom software analysis application called Codicem for interpretation and variant pathogenicity was determined using American College of Medical Genetics and Genomics (ACMG) criteria. Variants were Sanger confirmed by an external CAP/CLIA laboratory (EGL Genetics, Tucker, GA).

INDIVIDUAL 13: Trio WES was undertaken according to the following protocol: Genomic DNA was fragmented into 200 to 500 bp fragments by means of Adaptive Focused Acoustics (Covaris Inc, Woburn, USA) shearing according to the manufacturer's protocol. Exomes were captured using the

Clinical Research Exome v2 capture library kit (Agilent, Santa Clara, USA) accompanied by Illumina paired end Sequencing on the Hiseq4000 (Illumina, San Diego, USA), generating 2 × 150 bp paired end reads with at least 80x median coverage. An in-house sequence analysis pipeline (Modular GATK-Based Variant Calling Pipeline, MAGPIE) based on read alignment using BWA-MEM and variant calling using the GATK Haplotype Caller and UnifiedGenotyper[9] was used to align reads and call variants on the generated BAM files. Variants were subsequently annotated using the Variant Effect Predictor[10]. Included annotation fields were, amongst others, variant consequence, sift scores, polyphen scores, CADD scores, and allele frequencies in the 1000 Genomes populations. An in-house developed tool additionally annotated variants using dbSNP132, gnomAD allele frequencies, and the Genome of the Netherlands frequencies (GoNL). After annotation, variants with an allele frequency of > 5% in the Genome of the Netherlands or in the 1000 Genomes project were excluded from further interpretation. LOVDplus (Leiden Genome Technology Center, LUMC, Leiden) was used for interpretation of variants.

INDIVIDUAL 14: Trio WES was undertaken according to the following protocol: Exon regions were targeted in extracted genomic DNA from submitted patient and family members using the xGen Whole Exome Panel kit (Integrated DNA Technologies). Targeted regions were sequenced using the Illumina HiSeq 2500 sequencing system (v3 chemistry) with 100bp paired-end reads in rapid run mode. The resulting DNA sequences were mapped to and analyzed in comparison with the published human genome (UCSC hg 19 reference sequence). Sequence variants were filtered and annotated using Ingenuity Variant Software (QIAGEN) and compared to other provided family members to search for causes of Mendelian genetic disease (de novo, homozygous, compound heterozygous and inherited heterozygous disease-causing variants).

INDIVIDUAL 15: Trio WES was undertaken according to the following protocol: Genomic DNA from the submitted specimen was enriched for the complete coding regions and splice site junctions for most genes of the human genome using a proprietary capture system developed by GeneDx for next generation sequencing with CNV calling (NGS-CNV). The enriched targets were simultaneously sequenced with paired end reads on an Illumina platform. Bi-directional sequence reads were assembled and aligned to reference sequences based on NCBI RefSeq transcripts and human genome build GRCh37/USCS hg 19. Using a custom-developed analysis tool (XomeAnalyzer), data were filtered and analysed to identify sequence variants and most deletions and duplications involving three or more exons.

INDIVIDUAL 16: Trio WES, **INDIVIDUAL 20**: Duo WES, I**NDIVIDUAL 21**: Proband only WES. WES in these individuals was undertaken according to the following protocol: Using genomic DNA from the proband or proband and parent(s), the exonic regions and flanking splice junctions of the genome were captured using the IDT xGen Exome Research Panel v1.0. (Integrated DNA Technologies, Coralville IA. Massively parallel (NextGen) sequencing was done on an Illumina system with 100bp or greater paired-end reads. Reads were aligned to human genome build GRCh37/UCSC hg19, and analyzed for sequence variants using a custom-developed analysis tool.

INDIVIDUALS 17, 26, 27 and 28: Trio WGS was performed according to the following protocol: Using the Illumina TruSeq DNA PCR-Free sample preparation kit (Illumina, Inc.) and an Illumina HiSeq 2500 sequencer, a mean depth of 30× was generated. WGS reads were aligned using Isaac Genome Alignment Software and single variant nucleotides and indels called for chromosomes 1 to 22, X, and the mitochondrial genome using the Platypus variant caller[11]. Exomiser was used to identify and prioritise rare, segregating, predicted damaging variants in cases recruited with intellectual disability.

INDIVIDUAL 19: Trio WES was undertaken according to the following protocol: Genomic DNA was extracted from peripheral blood following standard DNA extraction protocols. After extraction of genomic DNA, targeted exons were captured with the Agilent SureSelect XT Clinical Research Exome kit (per manufacturer's protocol) and sequenced on the Illumina HiSeq 2000 or 2500 platform with

100bp paired-end reads. Mapping and analysis were based on the human genome build UCSC hg19 reference sequence. Sequencing data was processed and analyzed using an in-house custom-built bioinformatics pipeline.

INDIVIDUAL 22: Trio WES was undertaken according to the following protocol: Exome sequencing was performed using genomic DNA isolated from whole blood (MagnaPure24, Roche Diagnostics) from both proband and parents. Enriched libraries were prepared using KAPA HyperPlus Kit (Roche Diagnostics) and SeqCap EZ MedExome Kit (Roche Diagnostics). Sequencing was performed using massive parallel sequencing on a NextSeq550 (Illumina). The analysis of the obtained sequences was performed with the GenoSystem Variant Analysis software. This software was developed by Genologica and includes an optimized algorithm with the following features: a) Initial sequence quality control, b) Sequence filtering by elimination of indeterminacies, adapters and low quality areas, c) Second quality control of the sequences, d) Mapping on the reference genome Hg19, e) Obtaining variants and CNVs, f) Study of coverage of the mapping and g) Annotation of variants. Variant filtering and prioritization were performed with the same software. Candidate variants were evaluated based on stringent assessments at both the gene and variant levels, considering both the patient's phenotype and the inheritance pattern. Variants were classified following the guidelines of the ACMG. Identified variants were confirmed by Sanger sequencing.

INDIVIDUALS 23 and 24: Trio WES was performed using a BGI Exom kit (59M) for capture followed by sequencing on a BGISEQ-500, generating paired-end 100bp reads. Mapping was performed using BWA and calling with the GATK Haplotype caller. Analysis of the raw data was performed using the software Varfeed (Limbus, Rostock, Germany) and the variants were annotated and prioritized using the software Varvis (Limbus, Rostock, Germany).

INDIVIDUAL 25: Trio WES was performed using the IDT xGEN Exome Research Panel v1.0 with added custom clinical content and the Illumina NovaSeq 6000 platform. Samples were sequenced using 2x150 bp paired-end reads. Sequencing-derived raw image files were processed using a base-calling software (Illumina) and the sequence data was transformed into FASTQ format. Clean sequence reads of each sample were mapped to the human reference genome (GRCh37/hg19). BWA-MEM software was used for read alignment. Duplicate read marking, local realignment around indels, base quality score recalibration and variant calling were performed using GATK algorithms (Sentieon). Variant calling covered all coding regions and flanking introns with 20 bp upstream and downstream of each exon. The median exome-wide sequencing depth for the proband sample was 180x and 99.5% of target nucleotides were covered with >20x sequencing depth.

INDIVIDUAL 29 Trio WES was performed by the Genomics Platform at the Broad Institute of MIT and Harvard. Using Twist exome capture (~38 Mb target), samples were sequenced (150 bp paired reads) to cover >80% of targets at 20x, with a mean target coverage of >100x. Exome sequencing data was processed through a pipeline based on Picard and mapping to the human genome build 38 was performed using the BWA aligner. Variants were called using GATK HaplotypeCaller package version 3.5.

INDIVIDUALS 30, 31 and 32 Trio WES was undertaken according to the following protocol: Genomic DNA was extracted from blood samples of affected individuals and their parents. Whole-genome samples were prepared using a KAPA HyperPrep PCR-free Kit (Roche) and sequenced using a HiSeq X (Illumina) by the Kinghorn Centre for Clinical Genomics (Sydney, Australia). Sequence data was obtained in FASTQ format, with paired-ended 150 basepair reads. Alignment of reads and variant calling was processed in accordance with the GATK best practice guidelines.[7,9] Reads were aligned to the human reference sequence (GRCh37 assembly) using the BWA-MEM algorithm (v0.7.17). Duplicate reads were removed using Picard MarkDuplicates (v2.18.11) and outputted in BAM format using GATK (v3.8) BaseRecalibrator. GATK HaplotypeCaller was used to generate single-sample GVCFs, containing single nucleotide variants (SNVs) and short insertions/deletions (indels). Multi-

sample VCFs were produced using GATK GenotypeGVCFs and subsequently annotated with gene context information using SnpEff (v4.3S)[12]. Population allele frequencies from the gnomAD project (v2.1.1) were added using BCFtools annotate (v1.9)[13]. Variants were filtered assuming full penetrance of the causal variant. Variant sites where the variant allele fraction (VAF) in the affected individual(s) was ≥ 0.20, and with an overall read depth ≥ 8 for all individuals in the family were included. Rare variants (gnomAD population allele frequency < 0.001) were selected. Variants that impacted the coding sequence or a canonical splice site of a gene, based on the annotations provided by SnpEff, were included. Variants observed in homozygosity in either parent were excluded.

INDIVIDUAL 33 Trio WES was performed according to the following protocol: Sequencing was performed using the NextSeq 500 Sequencing System (Illumina, San Diego, CA), with a 2 × 150 bp high output sequencing kit after a 12-plex enrichment with SeqCap EZ MedExome kit (Roche, Basel, Switzerland), according to the manufacturer's specifications. Sequence quality was assessed with FastQC 0.11.5, then the reads were mapped using BWA-MEM (version 0.7.13), sorted and indexed in a bam file (samtools 1.4.1), duplicates were flagged (sambamba 0.6.6) and coverage was calculated (picard- tools 2.10.10). Variant calling was done with GATK 3.7 Haplotype Caller. Coverage for these samples was >93% at a 20× depth threshold. Variants were annotated with SnpEff 4.3, dbNSFP 2.9.3, gnomAD, ClinVar, HGMD, OMIM, and an internal database. Filtering was performed with criteria based on the consequence on the protein and frequency in gnomAD.

INDIVIDUAL 34: Trio WES was undertaken according to a previously published protocol[14].

Supplemental acknowledgements

DDD statement: The DDD study presents independent research commissioned by the Health Innovation Challenge Fund [grant number HICF-1009-003], a parallel funding partnership between Wellcome and the Department of Health, and the Wellcome Sanger Institute [grant number WT098051]. The views expressed in this publication are those of the author(s) and not necessarily those of Wellcome or the Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by the Republic of Ireland REC). The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network. This study makes use of DECIPHER (https://decipher.sanger.ac.uk), which is funded by Wellcome.
100,000 Genomes Project statement: This research was made possible through access to data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by individuals and collected by the National Health Service as part of their care and support.

Supplemental References

1. Gordon, C.T., Petit, F., Kroisel, P.M., Jakobsen, L., Zechi-Ceide, R.M., Oufadem, M., Bole-Feysot, C., Pruvost, S., Masson, C., Tores, F., et al. (2013). Mutations in Endothelin 1 Cause Recessive Auriculocondylar Syndrome and Dominant Isolated Question-Mark Ears. The American Journal of Human Genetics *93*, 1118–1125.

2. Girisha, K.M., Shukla, A., Trujillano, D., Bhavani, G.S., Hebbar, M., Kadavigere, R., and Rolfs, A. (2016). A homozygous nonsense variant in IFT52 is associated with a human skeletal ciliopathy. Clin. Genet. *90*, 536–539.

3. de Ligt, J., Willemsen, M.H., van Bon, B.W.M., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N. Engl. J. Med. *367*, 1921–1929.

4. Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature *519*, 223–228.

5. Hempel, M., Cremer, K., Ockeloen, C.W., Lichtenbelt, K.D., Herkert, J.C., Denecke, J., Haack, T.B., Zink, A.M., Becker, J., Wohlleber, E., et al. (2015). De Novo Mutations in CHAMP1 Cause Intellectual Disability with Severe Speech Impairment. Am. J. Hum. Genet. *97*, 493–500.

6. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

7. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation

discovery and genotyping using next-generation DNA sequencing data. Nat Genet *43*, 491–498.

8. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.

9. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297–1303.

10. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol *17*, 122.

11. Wei, W., Tuna, S., Keogh, M.J., Smith, K.R., Aitman, T.J., Beales, P.L., Bennett, D.L., Gale, D.P., Bitner-Glindzicz, M.A.K., Black, G.C., et al. (2019). Germline selection shapes human mitochondrial DNA diversity. Science *364*, 1–12.

12. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.

13. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

14. Pezzani, L., Marchetti, D., Cereda, A., Caffi, L.G., Manara, O., Mamoli, D., Pezzoli, L., Lincesso, A.R., Perego, L., Pellicioli, I., et al. (2018). Atypical presentation of pediatric BRAF RASopathy with acute encephalopathy. Am J Med Genet A *176*, 2867–2871.