

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:<https://orca.cardiff.ac.uk/id/eprint/189816/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Kumar, Raman, Gardner, Alison, Bhattacharjee, Rudrarup, Agarwala, Shreya, van Eyk, Clare L. and Fry, Andrew E. 2026. Variants in TREX complex subunits (THOC2, ALYREF, THOC6, THOC7) define a neurodevelopmental disability spectrum. *Genetics in Medicine* , 102734. 10.1016/j.gim.2026.102734

Publishers page: <https://doi.org/10.1016/j.gim.2026.102734>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



Journal Pre-proof

Variants in TREX complex subunits (*THOC2*, *ALYREF*, *THOC6*, *THOC7*) define a neurodevelopmental disability spectrum

Raman Kumar, Alison Gardner, Rudrarup Bhattacharjee, Shreya Agarwala, Clare L. van Eyk, Mark A. Corbett, Renee Carroll, Thessa Kroes, Tarin Ritchie, Rebekah de Nys, Danielle Mazurkiewicz, Wu Bing, Elizabeth E. Palmer, Michael Field, Jolijn J.A. Verseput, Bert B.A. de Vries, Alexander J.M. Dingemans, Qi Zhang, Fuyi Li, Francois Bernier, Julie Lauzon, Robert Śmigiel, Juan Darío Ortigoza-Escobar, Angeles García-Cazorla, Alejandra Darling, Cornelius F. Boerkoel, Stephanie Huynh, Gregory Costain, Tjakko J. van Ham, Esmee Kasteleijn, Marjon van Slegtenhorst, Tahsin Stefan Barakat, Ype Elgersma, Yvette van Ierland, Danielle Veenma, Dimitar Azmanov, Allison Goetsch Weisman, Carlos E. Prada, Sandhya Parkash, Andrea L. Rideout, W.S. Kerstjens-Frederikse, Jaime Vengoechea, Jan-Christoph Schoene-Bake, Francoise Devillard, Jasper van der Smagt, Noor A.A. Giesbertz, Andrew E. Fry, Jana Jezkova, Antje Kampmeier, Alma Kuechler, Ute Grasshoff, Miriam Bertrand, Scott E. Hickey, Jesse M. Hunter, Mackenzie V. Marr, Juliette Dupont, Stefania Bigoni, Alessandra Ferlini, Rita Selvatici, Nicolas Chatron, Gaetan Lesca, Louis Januel, Massimiliano Rossi, Curtis Rogers, Jessica M. Davis, Fiona McKenzie, Felix Marbach, Laura Geldon, Wolfgang M. Schmidt, Amina Paquay, Reginald E. Bittner, Simone Mahal, Günther Bernert, Ingrid Sinnerbrink, Linda Goodwin, Sandra Cooper, Natalja Bannink, Lachlan A. Jolly, Julie McGaughran, Alexis Bosman, Rebecca Hernan, Wendy K. Chung, Irene Valenzuela, Ivon Cuscó, Agatha Schlüter, Aurora Pujol, Luis Alberto Pérez-Jurado, Jozef Gecz



PII: S1098-3600(26)01052-X

DOI: <https://doi.org/10.1016/j.gim.2026.102734>

Reference: GIM 102734

To appear in: *Genetics in Medicine*

Received Date: 11 July 2025

Revised Date: 24 August 2026

Accepted Date: 18 September 2026

Please cite this article as: Kumar R, Gardner A, Bhattacharjee R, Agarwala S, van Eyk CL, Corbett MA, Carroll R, Kroes T, Ritchie T, de Nys R, Mazurkiewicz D, Bing W, Palmer EE, Field M, Verseput

JJA, de Vries BBA, Dingemans AJM, Zhang Q, Li F, Bernier F, Lauzon J, Śmigiel R, Ortigoza-Escobar JD, García-Cazorla A, Darling A, Boerkoel CF, Huynh S, Costain G, van Ham TJ, Kasteleijn E, van Slegtenhorst M, Barakat TS, Elgersma Y, van Ierland Y, Veenma D, Azmanov D, Weisman AG, Prada CE, Parkash S, Rideout AL, Kerstjens-Frederikse WS, Vengoechea J, Schoene-Bake JC, Devillard F, van der Smagt J, Giesbertz NAA, Fry AE, Jezkova J, Kampmeier A, Kuechler A, Grasshoff U, Bertrand M, Hickey SE, Hunter JM, Marr MV, Dupont J, Bigoni S, Ferlini A, Selvatici R, Chatron N, Lesca G, Januel L, Rossi M, Rogers C, Davis JM, McKenzie F, Marbach F, Gieldon L, Schmidt WM, Paquay A, Bittner RE, Mahal S, Bernert G, Sinnerbrink I, Goodwin L, Cooper S, Bannink N, Jolly LA, McGaughran J, Bosman A, Hernan R, Chung WK, Valenzuela I, Cuscó I, Schlüter A, Pujol A, Pérez-Jurado LA, Gecz J, Variants in TREX complex subunits (*THOC2*, *ALYREF*, *THOC6*, *THOC7*) define a neurodevelopmental disability spectrum, *Genetics in Medicine* (2026), doi: <https://doi.org/10.1016/j.gim.2026.102734>.

This is a PDF of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability. This version will undergo additional copyediting, typesetting and review before it is published in its final form. As such, this version is no longer the Accepted Manuscript, but it is not yet the definitive Version of Record; we are providing this early version to give early visibility of the article. Please note that Elsevier's sharing policy for the Published Journal Article applies to this version, see: <https://www.elsevier.com/about/policies-and-standards/sharing#4-published-journal-article>. Please also note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2026 Published by Elsevier Inc. on behalf of American College of Medical Genetics and Genomics.

Variants in TREX complex subunits (*THOC2*, *ALYREF*, *THOC6*, *THOC7*) define a neurodevelopmental disability spectrum

Raman Kumar,¹ Alison Gardner¹, Rudrarup Bhattacharjee^{1,2}, Shreya Agarwala¹, Clare L van Eyk¹, Mark A. Corbett¹, Renee Carroll¹, Thessa Kroes¹, Tarin Ritchie¹, Rebekah de Nys^{1,3}, Danielle Mazurkiewicz¹, Wu Bing⁴, Elizabeth E Palmer⁵, Michael Field⁶, Jolijn J.A. Verseput⁷, Bert B.A. de Vries⁸, Alexander J.M. Dingemans⁸, Qi Zhang⁹, Fuyi Li⁹, Francois Bernier¹⁰, Julie Lauzon¹⁰, Robert Śmigiel¹¹, Juan Darío Ortigoza-Escobar^{12,13}, Angeles García-Cazorla^{12,13}, Alejandra Darling^{12,13}, Cornelius F. Boerkoel¹⁴, Stephanie Huynh¹⁴, Gregory Costain^{15,16}, Tjakko J. van Ham¹⁷, Esmee Kasteleijn¹⁷, Marjon van Slegtenhorst¹⁷, Tahsin Stefan Barakat¹⁷, Ype Elgersma^{17,18}, Yvette van Ierland^{17,18}, Danielle Veenma¹⁸, Dimitar Azmanov¹⁹, Allison Goetsch Weisman²⁰, Carlos E Prada²⁰, Sandhya Parkash²¹, Andrea L Rideout²², W.S. Kerstjens-Frederikse²³, Jaime Vengoechea²⁴, Jan-Christoph Schoene-Bake²⁵, Françoise Devillard²⁶, Jasper van der Smagt²⁷, Noor A A Giesbertz²⁸, Andrew E Fry^{29,30}, Jana Jezkova²⁹, Antje Kampmeier³¹, Alma Kuechler³¹, Ute Grasshoff³², Miriam Bertrand³², Scott E. Hickey³³, Jesse M. Hunter³⁴, Mackenzie V Marr³³, Juliette Dupont³⁵, Stefania Bigoni³⁶, Alessandra Ferlini³⁶, Rita Selvatici³⁶, Nicolas Chatron^{37,38}, Gaetan Lesca^{37,38}, Louis Januel³⁷, Massimiliano Rossi^{37,39}, Curtis Rogers⁴⁰, Jessica M Davis⁴⁰, Fiona McKenzie⁴¹, Felix Marbach⁴², Laura Gieldon⁴², Wolfgang M. Schmidt⁴³, Amina Paquay⁴³, Reginald E Bittner⁴³, Simone Mahal⁴⁴, Günther Bernert⁴⁴, Ingrid Sinnerbrink⁴⁵, Linda Goodwin⁴⁶, Sandra Cooper⁴⁷, Natalja Bannink⁴⁸, Lachlan A Jolly⁴⁹, Julie McGaughan⁵⁰, Alexis Bosman⁵⁰, Rebecca Hernan⁵¹, Wendy K Chung⁵², Irene Valenzuela⁵³, Ivon Cusco^{13,53}, Agatha Schlüter^{13,54}, Aurora Pujol^{13,54,55}, Luis Alberto Pérez-Jurado^{13,56}, Jozef Gecz^{1,*}

¹School of Medicine and the Robinson Research Institute, College of Health, Adelaide University, Adelaide, SA 5005 Australia; ²Adelaide Centre for Epigenetics, School of Pharmacy and Biomedical Science, South Australian Immunogenomics Cancer Institute (SAiGENCI) and the Robinson Research Institute, Adelaide University, Adelaide, SA 5005 Australia; ³South Australian Health and Medical Research Institute, SA 5000, Australia; ⁴Genetics and Molecular Pathology, SA Pathology, Adelaide, South Australia, Australia; ⁵Rare Diseases NSW, Discipline of Paediatrics and Child Health, Faculty of Medicine and Health, University of New South Wales, Sydney, NSW, Australia; ⁶The Genetics of Learning Disability Service, Waratah, NSW, Australia; ⁷Department of Human Genetics, Radboudumc Nijmegen, the Netherlands; ⁸Department of Human Genetics, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, the Netherlands; ⁹South Australian immunoGENomics Cancer Institute (SAiGENCI), Adelaide University, Adelaide, South Australia, Australia; ¹⁰Department of Medical Genetics, Cummings School of Medicine, University of Calgary, Calgary, Alberta, Canada; ¹¹Department of Pediatrics, Endocrinology, Diabetology and Metabolic Diseases, Wroclaw Medical University, Wroclaw, Poland; ¹²Neurology Department, Institut de Recerca, Hospital Sant Joan de Déu, Barcelona, Spain; ¹³Centre for Biomedical Research on Rare Diseases (CIBER-ER), Instituto de Salud Carlos III, Madrid, Spain; ¹⁴University of British Columbia, Department of Medical Genetics, Children's and Women's Health Centre of BC, Vancouver BC, Canada; ¹⁵SickKids Research Institute, Toronto, Ontario, Canada; ¹⁶Division of Clinical & Metabolic Genetics, The Hospital for Sick Children, Toronto, Ontario, Canada; ¹⁷Department of Clinical genetics, Erasmus MC, University Medical Center, Rotterdam, The Netherlands; ¹⁸Erasmus MC Center of Expertise for Neurodevelopmental Disorders (ENCORE) Rotterdam, The Netherlands; ¹⁹Department of Diagnostic Genomics, PathWest Laboratory Medicine, QEII Medical Centre, Perth, Australia; ²⁰Ann & Robert H. Lurie Children's Hospital of Chicago,

Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA; ²¹Department of Pediatrics, Division of Medical Genetics, Dalhousie University, Halifax, NS, Canada; ²²Maritime Medical Genetics Service, IWK Health Centre, Halifax, NS, Canada; ²³Department Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands; ²⁴Division of Medical Genetics, Department of Human Genetics, Emory University, Atlanta, USA; ²⁵Department. of Pediatric Hematology and Oncology, Hannover Medical School, Hannover, Germany; ²⁶Département de Génétique et Procréation, Hôpital Couple-Enfant, CHU de Grenoble Alpes, Grenoble, France; ²⁷Division of Laboratoriies, Pharmacy en Biomedical Genetics, Section Clinical Genetics, University Medical Center Utrecht, Netherlands; ²⁸Department of Clinical Genetics, Antoni van Leeuwenhoek Netherlands Cancer Institute, Plesmanlaan, Amsterdam, the Netherlands; ²⁹All Wales Medical Genomics Service, Wales Genomic Health Centre, Longwood Drive, Cardiff, United Kingdom; ³⁰Division of Cancer and Genetics, School of Medicine, Cardiff University, Cardiff, United Kingdom; ³¹Institute of Human Genetics, University Hospital Essen, University Duisburg-Essen, 45122 Essen, Germany; ³²Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany; ³³The Ohio State University College of Medicine, Division of Genetic & Genomic Medicine, Nationwide Children's Hospital, Columbus OH USA; ³⁴Nationwide Children's Hospital, Steve and Cindy Rasmussen Institute for Genomic Medicine, Columbus OH USA; ³⁵Serviço de Genética, Hospital de Santa Maria, Lisbon, Portugal; ³⁶Medical Genetics Unit, Department of Medical Sciences and Department of Mother and Child, Ferrara University and Ferrara University Hospital, Ferrara, Italy; ³⁷Service de Genetique, Hospices Civils de Lyon, Bron, France and ERN ITHACA; ³⁸Institute NeuroMyoGène, Pathophysiology and Genetics of Neuron and Muscle, CNRS UMR5261 - INSERM U1315, Lyon, France; ³⁹INSERM U1028, CNRS UMR5292, Centre de Recherche en Neurosciences de Lyon, Lyon, France; ⁴⁰Greenwood

Genetic Center, Greenville, South Carolina, USA; ⁴¹Genetic Health WA, King Edward Memorial Hospital and School of Paediatrics and Child Health, University of Western Australia, Perth, WA, Australia; ⁴²Institute of Human Genetics at the University Hospital Heidelberg, Heidelberg, Germany; ⁴³Neuromuscular Research Department, Center for Anatomy and Cell Biology, Medical University of Vienna, Vienna, Austria; ⁴⁴Department of Pediatrics, Klinik Favoriten, Vienna, Austria; ⁴⁵Speciality of Genomic Medicine, Sydney Medical School, University of Sydney and Department of Clinical Genetics, Nepean Hospital, Sydney, NSW, Australia; ⁴⁶Department of Clinical Genetics, Royal North Shore Hospital, Sydney, NSW, Australia; ⁴⁷School of Medical Science, Faculty of Health and Medicine, University of Sydney, The Children's Hospital at Westmead, The Children's Medical Research Institute, Sydney, NSW, Australia; ⁴⁸Pediatrician-Genetic and developmental pediatrics, Franciscus Gasthuis & Vlietland, Rotterdam & Schiedam, The Netherlands; ⁴⁹School of Pharmacy and Biomedicine, and Robinson Research Institute, Adelaide University, Adelaide, SA, Australia; ⁵⁰The University of Queensland, Genetic Health Queensland, and Royal Brisbane and Women's Hospital, Brisbane, QLD, Australia; ⁵¹Department of Pediatrics, Boston Children's Hospital, Boston, MA USA; ⁵²Department of Pediatrics, Boston Children's Hospital, Harvard Medical School, Boston, MA USA; ⁵³Department of Clinical and Molecular Genetics, Hospital Universitari Vall d'Hebron, Barcelona, Spain; ⁵⁴Neurometabolic Diseases Laboratory, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain; ⁵⁵Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain; ⁵⁶Genetics Service, Hospital del Mar and Hospital del Mar Research Institut (IMIM), and MELIS Department, Universitat Pompeu Fabra, Barcelona, Spain.

*Correspondence and requests for materials should be addressed to Jozef Gecz at the School of Medicine, College of Health, Adelaide University, Adelaide, SA 5005 Australia.

Telephone number: +61 8 8313 2453 *Email address:* jozef.gecz@adelaide.edu.au

ABSTRACT

Purpose: TRanscription EXport (TREX) is an evolutionarily conserved multifunctional complex with essential roles in regulating transcription, mRNA export, genome stability, cell differentiation and survival. We aimed to determine if variation in the TREX subunits, in addition to *THOC2* and *THOC6* genes, caused Neurodevelopmental Disorders.

Methods: Through international collaboration we aggregated and assessed 40 individuals with high impact variants across four TREX subunit coding genes: *THOC2* and *THOC6* (known disease genes), and *THOC7* and *ALYREF* (*THOC4*) (candidate disease genes).

Results: On the background of considerable heterogeneity of neurodevelopmental and clinical outcomes, we noted that all TREX subunit variants impact cognitive functioning and development, speech and the respiratory systems as well as result in recognisable facial dysmorphism. We identified and studied five recurrent *THOC2* variants, which give overlapping yet variable clinical presentations, including severe perinatal lethal arthrogryposis multiplex congenita (AMC). Using cells from 13 affected individuals, we show that their TREX subunit variants are generally partial loss of function (LoF), which alter protein localisation, protein stability, protein-protein interactions or delete internal or C-terminal protein functional domains. Cells from affected individuals from across all four TREX subunits studied show consistent, yet variable accumulation of DNA damage.

Conclusion: The molecular, cellular and *in silico* data with the clinical presentations define a TREX spectrum disorder due to functional variation across multiple TREX subunits.

Keywords: TREX subunit variants, intellectual disability, neurodevelopmental disorders, partial loss-of-function, splice-site variants

Introduction

Neurodevelopmental disabilities (NDDs) are a heterogeneous group of early onset neurological and neuropsychiatric conditions, including intellectual disability (ID), autism spectrum disorder (ASD), language disorders, and epileptic encephalopathies.^{1,2} Monogenic etiologies are the most common causes.² Although rapid advancements in high-throughput genomic technologies applied to very large patient cohorts and ever-evolving *in silico* pathogenicity prediction tools have accelerated the identification of new NDD causative alleles, only limited progress has been made on molecular and/or functional validation of these variants. The estimated number of genes implicated in NDDs is higher than 1500.^{3,4}

The ATP-dependent multimeric TRanscription-Export (TREX) protein complex couples transcription elongation to mRNA nuclear export and consists of THO subcomplex (THOC1 (HGNC:19070), THOC2 (HGNC:19073), THOC3 (HGNC:19072), THOC5 (HGNC:19074), THOC6 (HGNC:28369) and THOC7 (HGNC:29874)) and several additional proteins (RNA helicase and splicing factor UAP56 (HGNC:13917), mRNA export adaptor ALYREF/THOC4 (HGNC:19071), CIP29/SARNP (HGNC:24432), POLDIP3 (HGNC:23782), ZC3H11A (HGNC:29093), LUZP4 (HGNC:24971), UIF (HGNC:25407), and CHTOP/SRAG (HGNC:24511)).⁵ Although pioneering studies showed that the TREX complex is responsible for exporting mRNA from the cell nucleus to the cytoplasm in all eukaryotic cells^{6,7}, more recent evidence has expanded its functional repertoire to essential roles in transcription, mRNA 3'-end processing, stress response, mitotic progression and prevention of DNA damage.⁸⁻¹⁰ We and others have shown that TREX is involved in

maintaining embryonic stem cell (ESC) self-renewal, pluripotency, and embryogenesis.^{5,11,12}

From among the TREX subunits, only *THOC2* and *THOC6* have been consistently implicated in NDDs, while the other subunits remain unexplored in this regard.¹³⁻¹⁸

THOC2, encoding the highly conserved and largest essential component of TREX, is one of the 172 genes that have been associated with X-linked ID (XLID).^{14,19} We reported missense *THOC2* variants in four families including a large multigenerational family first ascertained in 1971.²⁰ The affected individuals had variable degrees of ID with speech delay, increased body mass index (BMI), short stature, seizure disorders, gait disturbance, and tremors. Subsequently, additional missense, splicing-defective and microdeletion *THOC2* variants were reported, expanding the clinical heterogeneity to include language, growth, movement, and seizure disorders with ID as the core phenotype.^{15,16,20} The *THOC6* gene also encodes for a component of TREX that is highly expressed in the midbrain and eyes. We and others have reported biallelic LoF variants in *THOC6* causing Beaulieu-Boycott-Innes syndrome (BBIS OMIM:613680), an autosomal recessive NDD characterized by moderate to severe ID, dysmorphic facial features, submucous cleft palate and other developmental anomalies affecting the brain (corpus callosum dysgenesis), heart, kidneys and male genitalia.^{13,17,21,22}

Functional data from *THOC2* and *THOC6* NDD genes¹³⁻¹⁷ indicated that other TREX subunits may also be implicated in NDDs.¹⁸ Although the majority of the 33 new variants from 40 individuals assessed herein were in *THOC2* and *THOC6*, we also show that variants in *THOC7* and *ALYREF* (*THOC4*) genes have overlapping cell, molecular and clinical impact. Collectively, the evidence suggests an NDD spectrum disorder due to partial LoF variants across multiple TREX subunits.

Materials and Methods

Recruitment of affected individuals

The cohort of affected individuals with rare genetic variants in the TREX subunit genes was assembled through international collaboration. Some of the patients were ascertained in national programs for undiagnosed rare diseases and through matchmaker exchange.²³ The details on method/s of variant identification, validation, and inheritance pattern are included in Supplemental Tables 1-13. Pedigrees of affected individuals 2, 4, 12, 14A-B, 22, 25A-B, 26 and 29 are presented in Supplemental Figure 1-1.

Variant classification/*in silico* analyses

For predicting pathogenicity, variants were compared using *in silico* tools within dbNSFP4.7a^{24,25} (including CADD, Provean, VEST4, AlphaMissense, BayesDel addAF, and VARITY_R) and MutPred2.²⁶ We graphed scatterplots with a set of predictive tools that used independent algorithms. For AlphaMissense that predicts the effect of missense variants on protein stability, we classified variants as pathogenic and benign if the scores were >0.564 and <0.34, respectively as recommended by the authors.^{27,28} For CADD²⁹, variants were classified based on the PHRED scores using a threshold of >23 for a pathogenic classification. BayesDel (addAF) is a deleteriousness meta-score algorithm, which incorporates population frequency into its model and allows it to identify benign variants more easily with a universal cutoff value of >0.069265 for likely pathogenicity.³⁰ VARITY is an improved pathogenicity prediction algorithm/program for rare human missense variants, which uses a meta-prediction strategy, with scores >0.55 for likely pathogenic. We also used Franklin and GeneBe³¹ on-line tools based on ACMG/AMP⁷⁵ guidelines for classifying the missense variants (Supplemental Table 1-1).

Variant interpretation using facial photographs and Human Phenotype Ontology (HPO)

PhenoScore³² was used for genotype-phenotype analysis. HPO data was collected for 37 individuals. Of these, 36 for which the facial photographs were available and selected for PhenoScore analysis. The 12 of these individuals are part of the current cohort and 24 were previously published.¹⁴⁻¹⁶ The effect of the *THOC2* variant on protein function was known for 20 of these individuals. Their photographs and HPO data were pre-processed for PhenoScore analysis. 8 individuals were then excluded because they were part of the same family (e.g., individual 29 and individual 38), which could cause a bias in the analysis of their facial similarity. The remaining 12 individuals were paired with a sex-, age- and ethnicity-matched same number of controls with an NDD from a Radboudumc in-house database. The photographs and HPO terms of the individuals with a *THOC2* functionally-proven variants were then compared to the controls as described previously.³² A heatmap was generated using local interpretable model-agnostic explanations (LIME), where the prioritized facial features and most important HPO terms were highlighted. Subsequently, the photographs and HPO terms of the remaining 16 individuals with a variant of uncertain significance (VUS) were compared one by one to the trained PhenoScore model of the functionally-proven cohort. This resulted in a match prediction where an AUC score of <0.3 leads to a ‘No Match’ prediction, a score between 0.3-0.7 is seen as a VUS, and a score of >0.7 leads to a ‘Phenotypic Match’ prediction. A schematic overview of which individual was included in which analysis can be seen in Supplemental Table 1-2.

***In silico* mutagenesis and structural modelling**

In silico mutagenesis was carried out using PyMol (Python Molecular) with the published structure of human TREX complex (PDB 7ZNK).³³ The discs indicate van der Waals

overlaps or steric clashes between the mutated amino acids and their vicinity. Only the rotamer with the fewest clashes was used.

RNA extraction, Reverse Transcription-quantitative PCR (RT-qPCR) and RNAseq

RNA was extracted using QIAshredder (Qiagen, 79656) and RNeasy Micro Kit (Qiagen, 74004) with on-column RNase-free DNase treatment (Qiagen, 79254). 2 µg of RNA was reverse transcribed using SuperScript IV reverse transcriptase (Thermo Fisher, 18090010) using manufacturer's protocol. The amount of *THOC6* mRNA was assayed from affected and unaffected sisters, and control LCLs by RT-qPCR using Power SYBR Green PCR Master Mix (Thermo Fisher, 4368577) and the primers listed in Supplemental Table 3-1 and normalised to the housekeeping gene *HPRT1* mRNA. RNAseq was performed as reported previously¹². Sashimi plots were generated using a command-line tool for the visualization of splicing events across multiple samples.³⁴

Generation of expression plasmids

For generating Myc-tagged human THOC6 expression plasmid, THOC6 coding sequence was PCR amplified from fibroblast cDNA with hTHOC6_in fusion_F and hTHOC6_in fusion_R primers (Supplemental Table 3-1) using Platinum Super Fi PCR Master mix (Invitrogen, 12358-010) (95°C/203°F 3min; 98°C/ 208°F 10 sec, 59°C/138°F 10 sec, 72°C/162°F 30sec; 72°C/162°F 10 min/35 cycles) and gel-purified DNA was cloned into the pCMV-Myc vector linearized with *EcoRV* and *NotI* using In-Fusion Snap Assembly cloning reaction (TaKaRa, 638947). p.(Gly46Arg) and p.(Ser185Pro) variants were introduced into the pCMV-Myc-THOC6 WT expression plasmid using PCR products generated by the overlap PCR method and the primers listed in Supplemental Table 3-1. All constructs were validated by Sanger sequencing.

Cell culture and transfections

HEK293T, HeLa and fibroblast cells were maintained in Gibco's Dulbecco's Modified Eagle Medium (DMEM), 1 mM Sodium Pyruvate, 4.5 g/L D-Glucose, 2 mM GlutaMax (Thermo Fisher, 10569010) and 10% Fetal Bovine Serum (FBS). Immortalised lymphoblastoid cell lines (LCLs) were maintained in Gibco's Roswell Park Memorial Institute (RPMI) 1640 Medium, 1 mM Sodium Pyruvate, 2g/L D-Glucose, 2 mM GlutaMax, 25 mM HEPES (Thermo Fisher, 72400047) and 10% FBS.

Protein lysates

LCLs were lysed in 50 mM Tris-HCl pH7.5, 50 mM KCl, 0.1% NP40, 5 mM EDTA, 50 mM NaF, 0.1 mM Na₃VO₄, 1× Protease inhibitor/No EDTA³⁵ and fibroblasts in 50 mM Tris-HCl pH7.5, 250 mM NaCl, 0.1 Triton X-100, 1 mM EDTA, 50 mM NaF, 1× Protease inhibitor/No EDTA, 0.1 mM Na₃VO₄. Cells were sonicated for 15 second 30% amplitude, clarified at 15,000rpm at 4°C/39°F for 15 minutes and assayed by BCA protein assay kit (Pierce, 23225). For preparing the total, and nuclear and cytoplasmic protein fractions of *THOC6* affected p.(Gly46Arg) and unaffected LCLs, 2×10⁶ cells each were suspended in 200 µl 10 mM Tris-HCl pH8.0, 150 mM NaCl and 1× Protease inhibitor/No EDTA. 100 µl of cells were used for preparing the total protein lysates. To the other 100 µl cells, Igepal CA-630 (NP-40) was added to a final concentration of 0.03% and centrifuged at 1000 rpm at 4°C/39°F for 5 minutes. The supernatant was collected as cytoplasmic fraction and the nuclear pellet washed with 200 µl 10 mM Tris-HCl pH8.0, 150 mM NaCl and 1× Protease inhibitor/No EDTA. The pellets of nuclear fraction and LCLs (from 100 µl cells) were resuspended in 100 µl 10 mM Tris-HCl pH8.0, 150 mM NaCl, 1× Protease inhibitor/No EDTA pellet and 0.03% Igepal CA-630, sonicated (1× 10 seconds 30% amplitude on Sonic's

Vibra-Cell VCX) and centrifuged at 15,000 rpm at 4°C/39°F for 15 minutes to collect clarified nuclear and total proteins, respectively. The purity of the cytoplasmic and nuclear fractions was confirmed by western blotting for Hsp90 (cytoplasmic) and histone H3/NONO (nuclear) proteins.

Immunoprecipitation (IP)

For immunoprecipitations, HEK293T cells were seeded at a density of 5×10^5 cells/well in 6-well culture plates. The next day, cells were transfected with pCMV-Myc, pCMV-Myc-THOC6-WT, pCMV-Myc-THOC6-Gly46Arg or pCMV-Myc-THOC6-Ser185Pro expression plasmids using Lipofectamine 3000 reagent (Invitrogen, L3000015), harvested 24 hours later, lysed in cold lysis buffer [1× PBS, 0.1% Triton X-100, 5 mM EDTA, 50 mM NaF, 0.1 mM Na_3VO_4 and 1× Protease inhibitor/No EDTA (Roche, 04693132001)], sonicated (1× 10 seconds 30% amplitude on Sonic's Vibra-Cell VCX) and centrifuged at 15,000rpm at 4°C/39°F for 15 minutes. 10% clarified lysate was saved as input, and the remaining lysate was subjected to IP using anti-c-Myc-Magnetic Dynabeads (Pierce, 88843). IP complexes were eluted using 1× SDS protein-loading buffer³⁶ (62.5 mM Tris-HCl, pH 6.8, 2% SDS, 10% glycerol, 5% β-mercaptoethanol) at 95°C/203°F for 5 min and stored at -80°C/-112°F until western blot analyses.

Western blotting

For western blotting, the input and IP samples, and patient cell (LCLs and fibroblasts) lysates were resolved in 4-15% TGX gels (BioRad, 4561086) and transferred onto nitrocellulose membrane. The membrane was blocked in 10% skim milk in 1% Tris-Buffered Saline and Tween 20 (1× TBST; 20 mM Tris-HCl pH7.5, 150 mM NaCl and 0.1% Tween 20) for 40 minutes and then probed with appropriate primary (overnight at 4°C/39.2°F) and secondary

(2 hours at room temperature) antibodies (Supplemental Table 3-2) in 2% skim milk in 1× TBST. The blots were then washed with 1× TBST and target proteins visualized using Clarity Western ECL substrate (Bio-Rad, 1705061) on ChemiDoc XRS+ System (Bio-Rad).

Immunofluorescence staining

For THOC6 protein localization, HeLa cells were plated at 2.5×10^4 /well in a 12-well culture plate containing Poly-L-Lysine coated coverslips (BD, BDAA354085). Next day, cells were transfected with pCMV-Myc-THOC6-WT or pCMV-Myc-THOC6-Ser185Pro expression plasmids using Lipofectamine 3000 reagent (Invitrogen, L3000001). For endogenous THOC2 localization, control or patient fibroblasts were plated at 2×10^4 cells/well on poly-L-lysine coated coverslips (Merck, CLS354085). The following day, cells were washed with 1× PBS for 5 minutes at room temperature, then were fixed with cold 4% paraformaldehyde for 10 minutes, washed 2× with 1× PBS, permeabilized with 0.4% Triton X-100 in 1× PBS for 8 minutes, washed 2× with 1× PBS, and blocked in 10% Horse Serum in 1× PBS for 1.5 hr at room temperature. Cells were then incubated with appropriate primary antibodies (Supplemental Table 3-2; 1 µg/ml) diluted in 10% Horse Serum in 1× PBS overnight at 4°C/39°F. The following day, cells were washed 3× with 1× PBS, incubated with appropriate Alexa Fluor conjugated fluorescent secondary antibodies (Supplemental Table 3-2) for 2 hrs at room temperature, washed 3× times with 1× PBS and then mounted with ProLong glass antifade mounting medium with NucBlue stain (Invitrogen, P36981) followed by overnight curing.

Single cell electrophoresis Comet assay and RNA:DNA immunofluorescence

Comet assay was used to assess DNA damage and RNA:DNA immunofluorescence was performed to visualise R-loops in single cells as reported.¹²

Results

We identified 34 new variants in TREX subunits from 40 individuals affected by NDD, including 25 missense, 3 splicing, and 1 in-frame six nucleotide deletion in *THOC2* (Supplemental Table 4). In some individuals, variants of uncertain significance in other OMIM genes were identified and these are provided in Supplemental Table 4, 13. We present detailed *in silico* and functional assessment of variants of four TREX subunit genes, *THOC2* and *THOC6* (known disease genes) and *ALYREF* (*THOC4*) and *THOC7* (candidate disease genes).

Identification of new *THOC2* variants

Adding to the 19 *THOC2* variants previously reported by us^{14-16,37}, through international collaboration, we collected *THOC2* variants and clinical information on 34 affected individuals with 29 variants (25 missense; 3 splicing defective; and one resulting in a two amino acid in-frame deletion) (Supplemental Figure 1-2). Of these, five were recurrent [NM_001081550.2:c.229C>T p.(Arg77Cys)/individual 1, c.2158A>G p.(Arg720Gly)/individuals 14A-B, c.3223C>T p.(Arg1075Trp)/individuals 19A-B; c.3323C>T p.(Ser1108Leu)/individuals 20A-B and c.2482-1_2484del p.*/individual 29] (Figure 1A, Supplemental Table 1-1, 4-10). Of the previously unreported 29 *THOC2* variants, 3 were *de novo* and the rest maternally inherited (see Supplemental Table 1-1, 4). For the 26 maternally inherited variants, eight were inherited over more than one generation [c.565A>G p.(Ser189Gly)/individual 2, c.587T>A p.(Ile196Asn)/individual 4, c.1844G>A p.(Cys615Tyr)/individual 12, c.2158A>G p.(Arg720Gly)/individuals 14A-B, c.3427G>A p.(Ala1143Thr)/individual 22, c.4442G>A p.(Arg1481Gln)/individuals 25A-B, c.4449+5_4449+6del p.(Leu1438_Arg1483del)/individual 26) and c.2482-1_2484del

p.?

/individual 29] (Supplemental Figure 1-1). We identified one missense variant [c.2757G>A p.(Met919Ile)/individual 16] that resulted in mis-splicing (Figure 4). We also identified two affected females [c.1291G>A p.(Val431Met)/individual 7A, skewed X chromosome inactivation and c.4442G>A p.(Arg1481Gln)/individual 25B, XCI status unknown] who were part of the two brother-sister pairs (Supplemental Table 10). A previously reported^{38,39} recurrent *THOC2* splice-site variant c.2482-1_2484del p.?/individual 29 that causes arthrogryposis multiplex congenita (AMC) was also identified in one individual (Figure 1A).

There were two affected male and female sib-pairs in the cohort: Individuals 7A-B with the c.1291G>A p.(Val431Met) variant and individuals 25A-B with the c.4442G>A p.(Arg1481Gln) variant had shared clinical presentations including developmental delay, ASD, speech and language delay and facial dysmorphism. While individuals 7A-B with the c.1291G>A p.(Val431Met) had severe ID, the individuals 25A-B with c.4442G>A p.(Arg1481Gln) variant did not have formal diagnostic testing (Supplemental Table 10). In addition to the two siblings we published¹⁵, we identified two individuals with the c.3323C>T p.(Ser1108Leu) variant/individual 20A (*de novo*)/20B (maternally-inherited). The four individuals with the c.3323C>T p.(Ser1108Leu) variant showed overlapping phenotypes of moderate to severe ID, developmental delay, speech and language delay, autistic features, gait disturbances and facial dysmorphism along with other features (Supplemental Table 8). The mother of individual 20B has mild cognitive disability and speech impairment; we could not access gDNA to test her XCI status. A female with a heterozygous c.1291G>A p.(Val431Met) variant/individual 7A showed a mildly skewed XCI (80%) in blood gDNA. We can only speculate that skewed XCI may explain why these two females are affected. All recurrent variants were unique, only one c.1291G>A p.(Val431Met) was observed in

gnomAD v4.1 with allele frequency 1.0×10^{-6} originating from a heterozygous female⁴⁰ (Supplemental Table 1-1).

Clinical outcomes of partial loss of *THOC2* function

The detailed clinical data of the new cohort of affected individuals are provided in Supplemental Table 4 and Supplementary information. Overall, the published¹⁴⁻¹⁶ and unpublished cases together showed increased clinical heterogeneity. The core, albeit with some non-specific presentations similar across the combined cohorts, were ID (96.9%), developmental delay (67.1%), speech delay (88.2%), dysmorphic facial features (72.0%), autistic features (61.7%), visual abnormalities (39.2%) and hypotonia (55.5%) (Figure 1B, 1C, Supplemental Table 4 and Supplementary information). Other phenotypes included gait abnormalities (53.1%), microcephaly (27.3%), epilepsy (33.7%), short stature (36.5%), increased body BMI (31.8%), and mild to severe sensorineural hearing loss (32%) (Figure 1B, Supplemental Table 4). Both affected heterozygous females (individual 7A and 25B) displayed developmental, expressive and receptive language delay along with autistic features, facial dysmorphism, and skeletal anomalies (Supplemental Table 9). Interestingly, individual 7A had severe ID, whereas individual 25B had hypotonia.

We also identified an additional affected individual (individual 1) with a maternally inherited c.229C>T p.(Arg77Cys) variant (Supplemental Table 5), which is known to reduce the THOC2 (and THOC1, THOC5) protein stability.^{15,16} The three individuals identified to date with the c.229C>T p.(Arg77Cys) variant have severe to profound ID, developmental delay, speech and language problems, feeding and respiratory difficulties, spasticity (2/3), and epilepsy. Individual 1 has Dandy-Walker syndrome, severe developmental and epileptic

encephalopathy, chronic respiratory failure requiring tracheostomy and persistent mechanical ventilation, macrocephaly, a short and webbed neck (Supplemental Table 5).

Brain MRI data was available for 10 individuals with different *THOC2* variants (Supplemental Table 11). The most prominent were white matter abnormalities (including delayed myelination) and agenesis or thinning of the corpus callosum, consistent with our earlier findings.^{15,16} The MRIs of c.2757G>A p.(Met919Ile)/individual 16, c.3056G>A p.(Arg1019Gln)/individual 18 and c.4449+5_4449+6del p.(Leu1438_Arg1483del) individual 26 showed brain structural abnormalities (see Supplemental Table 11 and Supplemental Figure 1-3). Strikingly, the patients with variants in *ALYREF* (*THOC4*) NM_005782.4:c.374C>A p.(Ser125Ter) and *THOC7* NM_025075.2:c.178T>C p.(Cys60Arg) also showed delayed myelination/cortical maturation, and corpus callosum defects (see Supplemental Table 11 and below). Our aggregate findings suggest that compromised function of TREX subunits leads to diverse, but generally non-specific brain abnormalities.

***In silico* assessment and improved interpretation of *THOC2* variants**

We used the dbNSFP4.7a database of functional predictions and annotations to predict the potential pathogenicity of *THOC2* variants. We compared the outputs of these tools for the *THOC2* missense variants (n=43, including published and unpublished) against the common (n=47; allele frequency of $\geq 10^{-5}$) and rare (n=88; allele frequency of $\leq 10^{-6}$) variants in the gnomADv4.1 database. *THOC2* missense variants identified from individuals with NDDs were scored as significantly more deleterious (padj<0.0005 by ANOVA with Bonferroni correction) than common and rare variants providing supporting evidence of pathogenicity (Supplemental Figure 1-4). The majority of our *THOC2* VUS, and particularly those with functional evidence, clustered in the upper-right quadrants of the scatterplots; for example,

MutPred2/CADD v1.7, VEST4/BayesDel addAF, and AlphaMissense/BayesDel addAF (Supplemental Figure 1-5). This suggested higher confidence in the relevance of *in silico* predictions. Assessing *THOC2* missense variants with seven *in silico* tools predicted 12/25 to be pathogenic (higher prediction cut-offs for ≥ 6 criteria) and 8/25 to be likely pathogenic (higher prediction cut-offs for ≥ 3 to ≤ 5 criteria) (Supplemental Table 1-1). However, using the more stringent American College of Medical Genetics and Genomics (ACMG/AMP) criteria, only two variants were predicted to be pathogenic (c.229C>T p.(Arg77Cys)/individual 1; c.2757G>A p.(Met919Ile)/individual 16) and 11/25 variants were predicted to be likely pathogenic (Supplemental Table 1-1). As an alternative, we assessed the effects of *THOC2* missense variants on the wild-type 3D structure of TREX (<https://www.rcsb.org/structure/7ZNK>). This molecular modelling approach predicted that some of the 12/25 missense variants may affect the protein structure and/or its interaction with the other TREX subunits, potentially destabilising the overall structure and function of the complex (Supplemental Table 1-1 and Supplemental Figure 1-6). All of these 12 variants were also predicted pathogenic by our *in silico* analyses (Supplemental Table 1-1). Overall, we show that multiple *in silico* prediction tools and ACMG/AMP⁷⁵ criteria assess the pathogenicity of *THOC2* missense variants robustly, although functional evidence remains a gold standard for confirming pathogenicity.

We also assessed the phenotype associated with *THOC2* variants using PhenoScore; an Artificial Intelligence-driven phenomics framework that combines facial recognition technology with HPO data to evaluate differences in phenotype.³² When comparing the facial photographs and HPO data of 12 individuals with functionally-validated *THOC2* variants to sex-, age- and ethnicity matched controls, it was shown that individuals with (likely) pathogenic *THOC2* variants have a distinctive phenotype [area under the receiving operator

curve, AUC = 0.86, brier = 0.16, p-value = 3.33E-05) (Supplemental Table 1-2). The prioritized facial features are shown in Supplemental Figure 1-7 using a LIME heatmap. This figure also shows the most important HPO terms for this cohort. After establishing this PhenoScore model for *THOC2*, we examined the likelihood of pathogenicity for the 16 individuals with a VUS by separately matching them to the model. The results of this analysis are shown in Supplemental Table 1-1 and 1-4. For 11 individuals PhenoScore predicted a phenotypic match, 1 individual was predicted not to be a match and for the other 4 individuals the prediction remains a VUS.

Functional assessment of splice altering *THOC2* variants

Five *THOC2* variants, four intronic and one missense, were predicted to alter splicing by SpliceAI (Supplemental Table 1-3). We demonstrated deleterious outcomes on splicing on all 5 variants by RNAseq and assessed the impact of three of these variants (c.4449+5_4449+6del/Individual 26, c.4754+1A>G/Individual 28 and c.4450-2A>G¹⁵) by western blotting on patient-derived cells (see below).

c.4449+5_4449+6del causes aberrant splicing leading to r.4312_4449del p.(Leu1438_Arg1483del)/Individual 26.

Clinical exome (blood gDNA) sequencing identified a ‘GT’ deletion affecting the intron 34 canonical donor splice site. This variant was inherited from the unaffected heterozygous mother. The maternal grandmother and sister of the affected individual were also germline heterozygous for this *THOC2* variant. Both the heterozygous mother and sister had 95% skewed X-inactivation and no NDD phenotype (Supplemental Figure 1-1). Sashimi plots, RT-PCR (384 bp DNA in control vs 246 bp in the affected individual), and Sanger sequencing showed that the ‘GT’ deletion results in aberrant splicing and an in-frame

deletion of 138 bp (exon 34) (Figure 2A-B, Supplemental Figure 2). The aberrant transcript was predicted to lack the 46 amino acids encoded by exon 34. Consistent with this prediction, the western blot data showed a smaller THOC2 protein in the affected compared to control fibroblasts (Figure 2C).

c.2482-1_2484del recurrent variant causes familial arthrogryposis/Individual 29.

Exome sequencing (ES) and genome sequencing (GS) of gDNA identified a *THOC2* c.2482-1_2484del variant in two terminated male fetuses from non-consanguineous Indian³⁹ and French³⁸ couples, respectively. The authors proposed that the AMC in these fetuses was caused by this variant, however molecular studies confirming the impact of the variant on splicing were not reported. The variant was reported to be maternally-inherited in the first family³⁹ and *de novo* in the second family.³⁸ Using trio ES, we identified a third unrelated Canadian family with multiple similarly affected pregnancies with the *THOC2* c.2482-1_2484del variant, resulting in prenatal onset, progressive AMC and neonatal demise (Supplemental Figure 1-1). The index cases in this family were two male fetuses of a healthy non-consanguineous couple of European descent. Both fetuses presented on detailed mid-trimester ultrasound with evidence of fetal akinesia, multiple large and small joint contractures and progressive hydrops consistent with AMC (Supplemental Figure 3). Both pregnancies resulted in neonatal demises at 32 and 33 weeks, respectively; the boys survived less than an hour. Autopsies were not performed, but both boys had fixed hyperextension of the lower extremities, thin amyotrophic appearing lower limbs with fixed flexion contractures of upper extremities, hands and feet. No additional malformations were present (additional clinical data in Supplemental material). Females from different generations in this Canadian family had neonatal demises of male children (Supplemental Figure 1-1). RNAseq and RT-PCR data from fibroblasts derived from a demised male fetus (III-2) showed two aberrant

splicing events in the *THOC2* transcript. First, 61% of transcripts showed a 6 bp deletion from the 5' of exon 23 (r.2482_2487 p.(Ser828_Lys829del), and second, 39% of transcripts resulted in complete exon 23 exclusion (276 bp, r.2482_2757del p.(Ser828_Met919del) (Figure 3, Supplemental Figure 4). However, we did not observe a change in abundance of THOC2 protein or presence of two bands: a full length ≈ 175 kDa and a smaller ≈ 165 kDa due to deletion of 92 aa) in the fibroblasts of the affected individual compared to controls in western blots probed with three different THOC2 antibodies (Supplemental Figure 7B and *data not shown*).

c.2757G>A p.(Met919Ile) causes aberrant splicing leading to skipping of exon 23 [r.2482_2757del p.(Ser828_Met919del)]/Individual 16.

This maternally inherited variant was identified by trio GS. The affected individual showed severe clinical features (see Supplemental Table 4) and passed away at 9 years of age. The brain MRI showed white matter anomalies (Supplemental Figure 1-3) with prominent involvement of the internal capsules, but no convincing abnormality in the brainstem. Spectroscopy showed large peaks for lactate, and elevated myoinositol. Although consistent with mitochondrial disease, this could be the result of a generalised process of myelinolysis. Whereas *in silico* analysis predicted c.2757G>A p.(Met919Ile) to be tolerable, its location at the 3' edge of exon 23 and splicing predictions (Δ score 0.38 donor loss) suggested that the variant would ablate the adjacent splicing site and thus exon 23 (Figure 4). Consistent with these predictions, the RNAseq data from the affected son and his heterozygous mother's fibroblasts showed partially aberrant splicing resulting in exon 23 skipping in the mother (WT:skipped ratio $\sim 4:1$) and son (ratio $\sim 1:1$). The mother has normal intellectual and behaviour, and her peripheral blood DNA analysis showed severe skewing of X-inactivation (Androgen Receptor/AR allele ratio 89:11 and ZMYM3/ZNF261 93:7). The skipping and in-

frame exclusion of exon 23 (276 bp) is predicted to delete 92 amino acids p.(Ser828_Met919del) from the THOC2 protein (Figure 4).

c.4754+1A>G variant causes aberrant splicing, resulting in r.4755_4761del and replacing p.(His1586_Arg1593) by 13 other amino acids p.(His1586ProfsTer14) at the C-terminus of THOC2 protein/Individual 28.

An analysis of ID gene panel in ES data identified a *de novo* THOC2 c.4754+1A>G variant in a 4.5-year-old boy with developmental delay, recurrent infections, normal height, obesity and hypermetropia. This variant is located in intron 37. Analysis of his blood RNA showed that the variant leads to a frameshift and premature stop codon p.(His1586ProfsTer14) (detailed in⁴¹). The RNA-seq analysis also showed that the level of THOC2 expression in the affected individual was not decreased compared to ~100 control samples, indicating that the aberrant transcript was not degraded via nonsense mediated mRNA decay (NMD). Moreover, the transcript translated a C-terminally truncated THOC2 protein (Figure 5B) that could be detected by the THOC2/HPA antibody that binds the 1097-1200aa (present in the normal and truncated protein) but not by the THOC2/630A antibody that recognised the most C-terminal amino acids (Figure 5B). This result was confirmed by immunofluorescence (IF) staining of the THOC2 protein in the affected (no IF with THOC2/630A antibody) and control (normal IF with all three antibodies) cells (Supplemental Figure 5).

c.4450-2A>G variant predominantly causes exons 35-36 skipping (r. 4450_4677del p.(Asp1484_Lys1559del) and a low level of other aberrant splicing events.¹⁵

We previously reported that the c.4450-2A>G variant abolishes splicing of intron 34, which we argued, would lead to deletion of the C-terminal 58 amino acids.¹⁵ Our RNAseq data now show skipping of exons 35-36 in 87% THOC2 transcripts (Supplemental Figure 6A-B),

resulting in the deletion of 228 bp and thus 76 amino acids [p.(Asp1484_Lys1559del)] coded by these two exons but retaining the 34 C-terminal amino acids of the THOC2 protein. We confirmed this observation by western blotting the affected son and heterozygous mother's fibroblast lysates. The antibodies binding to the most C-terminal part (anti-THOC2/630A) and the 1400-1450 aa (anti-THOC2/629A) region of the THOC2 protein detected a protein of expected size in the fibroblasts of the mother, affected son and control (Supplemental Figure 6C-D). Together, our data showed that the aberrant splicing of exon 35-36 leads to the deletion of 228 bp from the *THOC2* mRNA transcript, which encodes 76 amino acids rather than the deletion of 58 amino acids from the C-terminal end of the THOC2 protein as reported previously¹⁵ (Supplemental Figure 6B-D).

THOC2 variant and TREX subunit protein studies

We analysed levels of THOC2 variant and TREX subunit proteins in fibroblasts of individuals 10, 14A/B, 22, 24, 25A, 27, 29 or LCLs of individual 12 and 13 (Figure 5B, Supplemental Figure 7). We observed reduced p.(Arg522Cys)/individual 10 and accumulation of p.(Ala1143Thr)/individual 22 THOC2 variant (and other TREX subunit) proteins. THOC2 protein appeared to be slightly smaller in c.4425_4430del p.(Glu1475_Lys1476del)/individual 24 compared to the control fibroblasts (Figure 5B). Interestingly, we observed no change in the level of p.(Cys615Tyr) and p.(Gly659Ser) THOC2 proteins although the THOC7 was reduced in these affected LCLs (Supplemental Figure 7). We did not observe mislocalisation of THOC2 variant proteins in fibroblasts from eight affected individuals (Supplemental Figure 5, 7D).

Partial LoF *THOC2* variants cause DNA damage

TREX complex has been shown to prevent replication and transcription-associated genome instability via efficient resolution of triple-stranded DNA-RNA-DNA structures called R-loops.^{42,43} Consistent with these reports, we found that compromised *THOC2* induced R-loop accumulation and DNA damage in both dividing and neuronal cells.¹² Here we used a single cell electrophoresis comet assay to assess the DNA damage in cells derived from the affected individuals. Interestingly, seven missense variants impacting the *THOC2* protein stability and three out of four resulting in C-terminal deletion showed DNA damage (Supplemental Table 12). Overall, we also note that the *THOC2* variants causing deletions near, at, or in the C-terminal of the *THOC2* protein are frequently associated with severe clinical features (Supplemental Figure 8).

Novel *THOC6* variants interfere with THO assembly leading to Beaulieu-Boycott-Innes syndrome (BBIS OMIM:613680), an expanded phenotype of prenatal onset

Many homozygous pathogenic variants in the *THOC6* gene, located on chromosome 16 (Figure 6A), have been associated with heterogeneous neurodevelopmental clinical features with ID, developmental delay and facial dysmorphism as the core phenotypes.

NM_024339.5:c.136G>A p.(Gly46Arg) and c.553T>C p.(Ser185Pro) THOC6 encoded proteins are mislocalised and fail to interact with the other THO subunits.

In 2013, Beaulieu et al implicated *THOC6* c.136G>A p.(Gly46Arg) homozygous variant in four Hutterite females with the autosomal recessive BBIS (OMIM:613680) and showed that exogenously expressed epitope-tagged p.(Gly46Arg) *THOC6* protein was predominantly localised to the cytoplasm instead of to the nucleus in HeLa cells.¹³ Subsequently, LCLs from the two affected and one unaffected sister were generated. Our western blotting on nuclear and cytoplasmic LCL fractions showed presence of p.(Gly46Arg) protein in the cytoplasm in

the two affected LCLs compared to the nucleus in the unaffected sister's LCLs (Figure 6B). This result showed that physiological mislocalisation of p.(Gly46Arg) protein in the affected LCLs is similar to the transiently expressed epitope-tagged p.(Gly46Arg) protein in HeLa cells.¹³ We also observed reduced stability of p.(Gly46Arg) THOC6 and THOC7 but not the other THO subunits in the affected LCLs (Figure 6B). These effects were post-translational as there was no detectable difference in the abundance of THOC6 mRNA between the affected, unaffected sister and unrelated control LCLs (Figure 6C). Finally, co-immunoprecipitation assay showed that the p.(Gly46Arg) protein loses interaction with THOC2, THOC5 and THOC7 (Figure 6D). We also observed significantly higher DNA damage in the affected compared to the unaffected sister's LCLs as assayed by single cell electrophoresis comet assay (Supplemental Figure 9). These data suggested that this variant alters the structure and function of the THOC6 protein.

Using GS (trio), we also identified a new c.553T>C p.(Ser185Pro) homozygous variant in a 7-year-old male born to consanguineous Syrian parents. Clinical features included moderate ID with short stature, severe microcephaly, consistent with BBIS (OMIM:613680). He also had distinctive facial features with prominent broad nasal tip, short columella, short philtrum, triangular face, relative hypertelorism, wide mouth, thick vermilion of upper and lower lip with small teeth, hypotonia, increased reflexes, wide-based stance, and unsteady gait. The p.(Ser185Pro) THOC6 variant was predicted to be damaging by multiple variant analysis tools (Supplemental Table 2). The p.(Ser185Pro) was predicted to abolish the hydrogen bond with Trp195, potentially destabilizing the fourth WD repeat of THOC6 and thereby affecting its interaction with THOC5 (Supplemental Figure 10). The p.(Ser185Pro) protein did not interact with the THO subunits THOC2, THOC5 and THOC7 (Figure 6D), and like p.(Gly46Arg), mislocalised to the cytoplasm (Figure 6E)..

A new biallelic *THOC7* c.178T>C p.(Cys60Arg) variant is associated with NDD

The reanalysis of the genomic sequencing data of a large cohort of individuals recruited to the Catalan Undiagnosed Rare Disease Program (URDCat)⁴⁴ identified a homozygous *THOC7* c.178T>C p.(Cys60Arg) variant in two siblings (a 17year old girl and an 11 year old boy born to consanguineous parents) with a similar phenotype of global developmental delay, moderate ID, sensorineural hearing loss and cerebellar vermis hypoplasia (Supplemental Table 2). Multiple *in silico* variant analysis tools predicted p.(Cys60Arg) to be deleterious (Supplemental Table 2). Although the siblings shared variants in three non OMIM genes [*DYNC1I2* c.393G>T p.(Leu131Phe), *ANKRD30A* c.1357A>G p.(Ile453Val) and *BPIFB2* c.955C>T p.(Arg319Trp)], *THOC7* was considered the most plausible disease-causing gene (see Supplemental Table 13). We found that reduced stability of the p.(Cys60Arg) *THOC7* protein also reduced the stability of *THOC1*, *THOC2*, *THOC5*, and *THOC6* but not *THOC3*, *ALYREF* (*THOC4*) and *UAP56* in the affected compared to control fibroblasts (Figure 7A). As no difference in average *THOC7* transcripts per million (TPM) was observed between affected (106.1 ± 0.6 ; n=2) and control (106.4 ± 4.2 ; n=6) fibroblasts, the reduced *THOC7* variant protein is likely due to proteasome-mediated post-translational regulation. We also observed that this variant caused increased DNA damage as assayed by comet assay and R-loop accumulation (Figure 7B-C, Supplemental Figure 11).

***De novo* LoF *ALYREF* (*THOC4*) variants are associated with multiple developmental malformations**

ALYREF, a highly conserved and ubiquitously expressed gene located on chromosome 17, codes for a multifunctional RNA binding export adaptor protein of TREX complex that is involved in nuclear export of mRNA.⁴⁵ We identified three *de novo* *ALYREF* variants:

ALYREF NM005782.4:c.283C>T p.(Gln95Ter), c.302dup (302-303:agt/agGt) p.(Ser101ArgfsTer35)⁴⁶ and c.374C>A p.(Gln125Ter). The *ALYREF* c.283C>T p.(Gln95Ter) stop gain *de novo* heterozygous variant was identified by GS and validated by Sanger sequencing in a male (Figure 8A-B)⁴⁶. The affected individual has different malformations [congenital diaphragmatic hernia (CDH), ventricular septal defect (VSD)], microcephaly and speech problems but no significant neurodevelopmental issues. The affected fibroblasts were very slow growing and significantly larger than the control fibroblasts derived from unaffected individual (Figure 8C). The RNAseq data showed 22 *ALYREF* TPM in the affected compared to 58±4.4 TPM in the control (n=10) fibroblasts. This along with the observation that, of the *ALYREF* transcripts in the affected fibroblasts only 4% carried the variant suggested that the *ALYREF* c.283C>T transcript is degraded by NMD, leading to haploinsufficiency. The comet assay showed significantly higher DNA damage in affected compared to control fibroblasts (Supplemental Figure 12). The second affected male with a *de novo* c.302dup (302-303:agt/agGt) p.Ser101ArgfsTer35) *ALYREF* variant⁴⁶ died on the first day of life (DoL1) likely exacerbated by prematurity and lung hypoplasia. The third affected individual had a *de novo* c.374C>A p.(Gln125Ter) *ALYREF* variant and presented with several malformations (perinatal oesophageal atresia, supernumerary spleen, colpocephaly, hypoplastic corpus callosum, hypoplastic semicircular canal) and epileptic encephalopathy; their epilepsy may have been related to an additional intragenic *GRM7* homozygous deletion (NC_000003.12:g.(?_7620079)_(7621064_?)del inherited from heterozygous parents causing a more complex blended phenotype.

Discussion

TREX complex subunits play critical roles in stem cell maintenance, neuronal differentiation, brain neural network activity and synapse development.^{11,12} With the gnomAD database

suggesting that *THOC2* is intolerant to LoF (pLI = 1), this TREX subunit is essential for sustaining life but its partial LoF leads to NDDs with ID, developmental and speech delay as the core phenotype and a range of comorbidities.¹⁴⁻¹⁶ In this study, we present clinical information, pathogenicity predictions and functional data from a larger cohort of individuals affected by NDD that have pathogenic or likely pathogenic *THOC2* variants. We also present functional evidence supporting pathogenicity of a newly-identified c.553T>C p.(Ser185Pro) and a previously published c.136G>A p.(Gly46Arg) *THOC6* variant.¹³ Finally, we report identification of and functional studies on the first *ALYREF* and *THOC7* pathogenic variants, implicating these in new monogenic NDDs. Although the core clinical presentation in the individuals affected by *THOC2* variants remained the same¹⁴⁻¹⁶, we observed some overlap with affected individuals with the *ALYREF*, *THOC6* and *THOC7* variants, which we collectively refer to as TREX spectrum disorder. Our identification of additional *THOC2* variants^{15,16} that result in C-terminal or internal (but closer to the C-terminus) deletions in the *THOC2* protein confirmed that they are frequent (Figure 1, Supplemental Figure 8), often with severe clinical features and occasionally causing death, e.g., c.2757G>A p.(Met919Ile) and c.2482-1_2484del.

Variants in TREX subunits can alter stability¹⁴⁻¹⁶, interaction¹⁷, localization¹⁷ and cause internal or C-terminal deletion of protein. All of the *THOC2*, *THOC6*, *THOC7* and *ALYREF* variants identified to date cause partial LoF. In the current cohort, we observed reduced stability of p.(Arg522Cys) *THOC2*, and p.(Cys60Arg) *THOC7* proteins in cells derived from individuals with these variants (Figure 7A and Supplemental Figure 7). The reduced *THOC2* and *THOC7* protein stability also impacted the stability of other THO subunits in these cells, which was consistent with previous reports showing that siRNA-mediated depletion, or missense variants that reduce *THOC2* protein can lead to reduced levels of other THO

subunits.^{14-16,47} These data suggest that the stability of THO subunit proteins is critical for maintaining the integrity and function of TREX complex. As such, the reduced stability of the TREX complex can be a reliable biomarker for assessing pathogenicity of its subunit variants. Moreover, absence of differences in the levels of transcripts of THO subunits between the affected and control cells suggested that reduction in levels of variant (e.g., THOC2, THOC7) and other TREX subunit proteins likely resulted from their increased susceptibility to proteasome-mediated degradation.⁴⁸ The literature is awash with studies showing that THO subunit depletion can cause mild to severe nuclear mRNA accumulation, for example^{47,49}, causing molecular and cellular impacts in mammalian cells^{10,50}, *Drosophila*⁶, zebrafish⁵¹ and *C. elegans*.⁵⁰ Importantly, *Thoc2*, along with *Thoc5*, regulates mouse embryonic stem cell (mESC) self-renewal by selectively binding and regulating export of mRNAs involved in the maintenance of pluripotency (e.g. *Nanog*, *Sox2*, *Esrrb*, *Klf4*)¹¹. Although mRNA export has been reported to be the most critical function of *THOC2*⁴⁷, our studies on the *Thoc2* exon 37-38 deletion partial LoF mouse model revealed that this essential protein plays a key role in regulating transcription and maintaining genome stability through R-loop homeostasis.¹² Similar to *Thoc2*, *Thoc1* is essential for early embryonic development⁵² and its loss compromises proliferation and lineage-generating capacity of stem cells and genome stability.^{53,54} Furthermore, that one third of the individuals affected by *THOC2* variants have short stature that begins intrauterine, at birth and/or postnatally^{14-16,55} (Figure 1B; Supplemental Table 4) is consistent with the observation that in mice, the *Thoc2* exon 37-38 deletion male embryos and adults have significantly reduced weight and size than the wild-type male littermates.¹²

Although the clinical features such as ID with dysmorphic facial features, microcephaly, teeth anomalies and cardiac malformations associated with *THOC6* ID syndrome - called BBIS

(OMIM:613680) - have remained the same¹³, a number of *THOC6* variants with additional phenotypes have been reported.⁵⁶ We previously showed that *THOC6* variants can lead to mis-localisation and loss of interaction with the TREX complex.^{13,17} Here we confirmed that p.(Gly46Arg) *THOC6* protein is mislocalised in the cells derived from the affected individuals. In addition, we identified a new c.553T>C p.(Ser185Pro) *THOC6* variant for which the protein mislocalised to the cytoplasm and excluded it from interaction with other TREX subunits in the nucleus. The individual carrying this variant had clinical presentations overlapping with BBIS (OMIM:613680) (Supplemental Table 2). We also identified another individual with c.298T>A p.(Trp100Arg), c.700G>C p.(Val234Leu) and c.824G>A p.(Gly275Asp) homozygous variants on the same haplotype (Figure 6A, Supplemental Table 2). We have previously reported this variant and showed that p.(Gly275Asp) and triple variant proteins, like p.(Gly46Arg) and p.(Ser185Pro), lose interaction with the THOC1 and THOC5 subunits.¹⁷ Overall, these data suggest that *THOC6* protein mislocalisation and consequent loss of interaction with other THO subunits impacts TREX function critical for healthy corticogenesis.^{12,22} Shared features across individuals with *THOC2*, *ALYREF*, *THOC6*, and *THOC7* variants, e.g., ID, speech delay, and brain abnormalities support a TREX spectrum disorder.

In the current study, in addition to the three published splicing variants causing C-terminal *THOC2* truncations^{15,16}, we report the identification and molecular events of five new variants causing deletion of a part of the *THOC2* protein due to aberrant splicing, e.g., c.2757G>A (p.Met919Ile)/individual 16 causing exon 23/p.(Ser828_Met919del), c.4449+5_4449+6del p.(Leu1438_Arg1483del)/individual 26, c.4754+1A>G/p.(His1586ProfsTer14)/individual 28, and c.2482-1_2484del/p.(Ser828_Lys829del) or p.(Ser828_Met919del)/individual 29 and, c.4425_4430del

p.(Glu1475_Lys1476del)/individual 24, these now account for 8 of the 50 *THOC2* variants identified so far (Figure 1, Supplemental Figure 8). We showed that *in silico* tools facilitate predicting pathogenicity of *THOC2* missense and splicing variants. The internal or C-terminal protein deletions are highly likely to be pathogenic, again due to altered structure and/or protein-protein interactions impacting protein functions such as protein folding, turnover, trafficking and cellular signaling.⁵⁷ All affected individuals with such splice/deletion variants exhibited severe clinical presentations. Moreover, variants that resulted in loss of exon 23 were particularly severe; individual 16 with the c.2757G>A p.(Met919Ile), p.(Ser828_Met919del) variant died at 16 years of age, and at least three others^{38,39} with the c.2482-1_2484del variant, caused progressive AMC and neonatal demise or termination of pregnancy due to hydrops and fetal immobility, suggesting a critical role of this part of the *THOC2* protein. Overall, the aggregate data from the known and the newly identified (this manuscript) splice and microdeletion *THOC2* variants affecting the C-terminal part of the protein show that these cause severe clinical outcomes^{15,16,37-39} (Supplemental Figure 8). Relative clustering of variants in the C-terminal nucleic acid binding domain (Figure 1) also reinforces the functional importance of this part of *THOC2* protein.⁵⁸

The c.2482-1_2484del is the first recurrent variant that leads to severe fetal AMC and hydrops. Two previous clinical reports implicated this variant in arthrogryposis^{38,39}. Our RNA studies on cells derived from affected individuals showed two major missplicing events resulting in transcripts that would translate into *THOC2* proteins that were 2aa ($\approx$ 175kDa) and 92aa ($\approx$ 165kDa) smaller. We did not detect two *THOC2* proteins or a change in the abundance of *THOC2* protein in the fibroblasts derived from the affected individuals. We speculate, that typical *THOC2* protein levels we detected could result from the stabilisation

of 2aa- and destabilisation/degradation of 92aa-deleted THOC2 proteins. Loss of 2aa or 92aa may also affect the THOC2 protein conformation and thus its function. Finally, the type and level of aberrant splicing in different tissues of the developing embryo can result in conformationally defective forms of THOC2 protein (and thus THO complex), leading to observed clinical features. Together, the four cases of the recurrent c.2482-1_2484del variant^{38,39} and molecular studies suggest that this hypomorphic variant results in a particularly severe fetal presentation of the TREX spectrum disorders with mid-trimester AMC, fetal akinesia with hydrops and neonatal demise. Postnatal examination of the current cases and histological evidence of muscular dystrophy with cytoplasmic bodies³⁸ argues for a primary progressive myopathy in these severe cases. The identified *THOC2* aberrant splicing events likely lead to perinatal deaths similar to the autosomal recessive condition Lethal Congenital Contracture Syndrome 1 (LCCS1) caused by variants in the *GLE1* gene (OMIM: 603371)⁵⁹. *GLE1* encodes an essential RNA export mediator protein conserved from yeast to human and is located on the cytoplasmic face of the nuclear pore complex. That *THOC2* variant causes clinical impacts similar to *GLE1* variants is not surprising as the two proteins play similar cellular roles, for example, depletion of GLE1 or THOC2 causes nuclear mRNA retention^{47,60,61}, cell cycle progression defects^{12,61}, and accumulation of R-loops.^{12,62} This is also consistent with our recent findings from *Thoc2* exon 37-38 deletion mouse model showing that compromised THOC2 function significantly impacted the early embryonic development of *Thoc2* exon 37-38 deletion male mice.¹² Others have also shown an essential role of *Thoc1*, *Thoc2* and *Thoc5* in early embryonic development.^{11,52} Together, these data^{11,12,52} directly suggest the importance of normal TREX functions in early human health and development.

Recently, c.788C>T p.(Pro263Leu)/likely pathogenic, c.3477G>T p.(Arg1159Ser)/likely pathogenic and c.4411G>T p.(Glu1471Ter)/pathogenic *THOC2* variants were reported in male children clinically diagnosed with cerebral palsy (CP)⁶³. CP is a clinical descriptor encompassing a heterogeneous spectrum of nonprogressive disorders of movement and posture which are frequently comorbid with other NDDs such as ID, epilepsy and ASD. Our previous work identified gait disturbances, spasticity and hypotonia as central features of *THOC2* related disorder (OMIM:300957) and this observation is further supported by our findings in the current cohort [e.g., c.623A>G p.(Asn208Ser)/individual 5 (GMFCS level V), and c.1844G>A p.(Cys615Tyr) (GMFCS IV)/individual 12 variants], with >50% of individuals having gait/movement abnormalities and ≈20% with spasticity (Figure 1B). These observations suggest that *THOC2* should be considered in diagnostic testing for CP.

We report two stop-gain c.283C>T p.(Gln95Ter) and c.374C>A p.(Gln125Ter) and one single nucleotide duplication c.302dup p.(Ser101ArgfsTer35) heterozygous *de novo* *ALYREF* variants. The c.283C>T p.(Gln95Ter) variant fibroblasts show slower growth, significantly enlarged cell size (Figure 8), increased DNA damage (Supplemental Figure 12), and NMD-mediated degradation of variant transcripts causing haploinsufficiency. That reduced levels of *ALYREF* (*THOC4*) likely causes slower growth rate of affected fibroblasts is consistent with multiple studies reporting that depletion of *ALYREF* leads to reduced cell growth⁶⁴. That primary human cells increase in size on permanent cell cycle arrest and senesce is supported by the evidence showing that DNA damage induced senescence in human fibroblasts causes significant increase in cell volume⁶⁵. Although excessive size of primary cells impairs cell cycle progression and cell signaling⁶⁵, much less is known about the genetic perturbations that can cause continued increase in cell size. We speculate that *ALYREF* c.283C>T p.(Gln95Ter), and possibly also c.302dup (302-303:agt/agGt) p.(Ser101ArgfsTer35) and

c.374C>A p.(Gln125Ter), induced DNA damage and likely cell cycle arrest causes enlargement of affected fibroblasts. Emerging evidence suggests that atypical activation of senescence due to genetic defects may have a role in the pathogenesis of developmental disorders⁶⁶.

We predict that the c.374C>A p.(Gln125Ter) and c.302dup p.(Ser101ArgfsTer35) transcripts are also degraded as these variants, like c.283C>T, are located within the exon 2. These variants would likely lead to *ALYREF* haploinsufficiency in the affected individuals causing clinical consequences because *ALYREF* is an essential gene and has important roles in human development. *ALYREF* is ubiquitously expressed, conserved from yeast to humans, and harbours a canonical RNA-binding motif through which it directly interacts with RNAs.^{67,68} Compromised *ALYREF* function can lead to impaired nuclear export of both spliced and intronless mRNAs⁶⁹, mRNA export from nucleus⁷⁰, transcriptional regulation⁷¹, genome stability⁴³, RNA 3'-end processing⁷², nuclear export of non-polyadenylated mRNAs⁷² and cell cycle progression and cell proliferation⁷⁰; disruption of one or more of these cellular processes can lead to clinical symptoms.

Little is known about the precise role of *THOC7* in the developing human brain. However, that *THOC7* depletion leads to severe polyA⁺ RNA export block⁴⁷ is consistent with the preweaning lethality with complete penetrance observed in *Thoc7* homozygous knockout mice. Therefore, identification of c.178T>C p.(Cys60Arg) that impacts stability of *THOC7* and other TREX subunit proteins has hallmarks similar to variants that reduce *THOC2* and TREX subunit protein stability.¹⁴⁻¹⁶ Our observations of increased R-loop accumulation and DNA damage in *THOC6*, *THOC7*, *ALYREF* and multiple *THOC2* affected cells (Figure 7, Supplemental Figure 9, 11, 12, Supplemental Table 12) are consistent with the reports that

TREX complex prevents genome instability by efficiently resolving the R-loops.⁴² These findings along with our recent data showing that compromised THOC2 function leading to R-loop accumulation and DNA damage may be responsible for the disease pathology¹² although further studies are required to confirm these observations.

There is no dedicated, simple functional assay to confirm the pathogenicity of TREX subunit variants. In our experience, it is a combination of clinical data and molecular studies using affected individuals' cells, *in silico* and protein structure-based predictions, *de novo* or multigeneration inheritance of variants and identification of recurrent variants, which aid the classification of the TREX subunit variants. With identification of additional *THOC2* and *THOC6* variants, the clinical heterogeneity of the associated conditions also expanded. Identification of new disease-causing *THOC7* and *ALYREF* variants has added to the diversity of clinical presentations, which now also includes individuals without ID, but with developmental delay, intrauterine e.g., c.2482-1_2484del *THOC2*, DoL1 e.g., c.302dup p.(Ser101ArgfsTer35) *ALYREF* or childhood e.g., c.2757G>A p.(Met919Ile) *THOC2* demise, and those who succumb to multiple severe complications e.g., c.229C>T p.(Arg77Cys). Common genetic variation might also influence the clinical outcomes seen in individuals with TREX subunit variants.⁷³ As the majority of the TREX subunits are encoded by genes essential for life⁷⁴, mammalian knock-in models are just beginning to emerge¹². Study of clinically accessible tissues of individuals affected by naturally-occurring variants, combined with international collaboration to collect clinical data aided significantly to understand LoF of TREX subunits. This is to be complemented by dedicated mouse¹², and iPSC-derived 2D and 3D neural differentiation models, which will eventually deliver a robust functional diagnostic framework and underpin precision medicine strategies for the individuals affected by TREX spectrum disorders.

Data Availability

Data and materials from the study are available from the authors on application and subject to compliance with our obligations under human research ethics.

Acknowledgements

We acknowledge Dr. Fernanda Fortunato (Medical Genetics Unit, Department of Medical Sciences and Department of Mother and Child, Ferrara University and Ferrara University Hospital, Ferrara, Italy) for help with work on individual 21. The work on individual 21 was supported by the Solve-RD project (European Union's Horizon 2020 grant agreement No 779257) to Alessandra Ferlini. Some of the authors of this publication are members of the European Reference Network on Rare Congenital Malformations and Rare Intellectual Disability ERN-ITHACA [EU4Health Grant Agreement nr. 101085231], on Rare and Complex Epilepsies EpiCARE [EU4Health Grant Agreement nr. 101156820], for Rare Neurological Diseases ERN-RND [EU4Health Grant Agreement nr. 101156814], and for Hereditary Metabolic Disorders MetabERN [EU4Health Grant Agreement nr. 101095426]. Research on individual 16 was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care and data collected as part of their participation in research. Research on several individuals (Supplemental Table 2, 13) was made possible through The Undiagnosed Rare Disease Program of Catalonia (URD-Cat). We thank the Care4Rare Canada Consortium for the cell lines that enabled the THOC6 functional work. The authors also thank all the families for their contribution towards this study.

Funding

This study was financially supported by the National Health and Medical Research Council of Australia (NHMRC) Fellowship APP1155224 to J.G., NHMRC project grant APP1163240 to J.G. and R.K., and Australian Research Council Discovery Project DP170103090 to J.G., R.K. and L.A.J. WKC and RH are supported by NIH grant 1P01HD068250. The Undiagnosed Rare Disease Program of Catalonia URD-Cat was supported by the Department of Health of Catalonia (Grant SLT002/16/00174) to L.A.P-J. The Barakat lab was supported by the Netherlands Organisation for Scientific Research (ZonMw Vidi, grant 09150172110002). Work at the Neurometabolic Diseases Lab (IDIBELL) was supported by the CIBERER (ACCI19-759), the Hesperia Foundation, the Ministry of Business and Knowledge of the Government of Catalonia (2021SGR00899), the Instituto de Salud Carlos III (ISCIII) and 'Fondo Europeo de Desarrollo Regional (FEDER) (FIS PI23/01090) to A.S. The National Genomic Research Library 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. A.E.F. was supported by the Wales Gene Park. Wales Gene Park is a Health and Care Research Wales funded research group. J.J. was supported by a Research Building Capacity Collaboration (RCBC) Wales Postdoctoral Fellowship.

Author Contributions

Conceptualization: R.K., J.G., A.G.; Writing-original draft: R.K., J.G.; Visualisation: R.K., S.A., R.B.; Data curation: all authors; Methodology: R.K., A.G., M.A.C, R.C., T.K., R.B.; Resources: R.K., J.G., RB, S.A., T.R., R.de.N., D.M., W.B.; Formal Analysis: M.A.C., C.L. van E., L.A.J., Q.Z., F.L.; Software: M.A.C., A.G.; Writing-reviewing and editing: all authors

Ethics Declaration

Informed written consent was obtained and archived from all participants as approved by the Institutional Review Board (Women's & Children's Health Network Human Research Ethics Committee, North Adelaide SA 5006, Australia). Clinical and radiological data, photos and/or cells derived from the deidentified affected individuals were obtained after written informed consent by a parent or legal guardian, who granted permission for the publication of photographs, and were provided by clinicians from the participating centres.

Conflict of Interest

The authors declare no conflict of interest.

Additional Information

The online version of this article contains supplementary information, which is available to authorized users.

References

1. Parenti I, Rabaneda LG, Schoen H, Novarino G. Neurodevelopmental Disorders: From Genetics to Functional Pathways. *Trends Neurosci.* 2020;43(8):608-621 <https://doi.org/10.1016/j.tins.2020.05.004>.
2. Huang QQ, Wigdor EM, Malawsky DS, et al. Examining the role of common variants in rare neurodevelopmental conditions. *Nature.* 2024;636(8042):404-411 <https://doi.org/10.1038/s41586-024-08217-y>.
3. Leblond CS, Le TL, Malesys S, et al. Operative list of genes associated with autism and neurodevelopmental disorders based on database review. *Molecular and cellular neurosciences.* 2021;113:103623 <https://doi.org/10.1016/j.mcn.2021.103623>.
4. Jansen S, Vissers L, de Vries BBA. The Genetics of Intellectual Disability. *Brain Sci.* 2023;13(2):231 <https://doi.org/10.3390/brainsci13020231>.
5. Heath CG, Viphakone N, Wilson SA. The role of TREX in gene expression and disease. *The Biochemical journal.* 2016;473(19):2911-2935 <https://doi.org/10.1042/BCJ20160010>.

6. Rehwinkel J, Herold A, Gari K, et al. Genome-wide analysis of mRNAs regulated by the THO complex in *Drosophila melanogaster*. *Nature structural & molecular biology*. 2004;11(6):558-566 <https://doi.org/10.1038/nsmb759>.
7. Reed R, Cheng H. TREX, SR proteins and export of mRNA. *Curr Opin Cell Biol*. 2005;17(3):269-273 [https://doi.org/S0955-0674\(05\)00053-0](https://doi.org/S0955-0674(05)00053-0).
8. Salas-Armenteros I, Perez-Calero C, Bayona-Feliu A, Tumini E, Luna R, Aguilera A. Human THO-Sin3A interaction reveals new mechanisms to prevent R-loops that cause genome instability. *EMBO J*. 2017;36(23):3532-3547 <https://doi.org/10.15252/embj.201797208>.
9. Castellano-Pozo M, Garcia-Muse T, Aguilera A. R-loops cause replication impairment and genome instability during meiosis. *EMBO Rep*. 2012;13(10):923-929 <https://doi.org/10.1038/embor.2012.119>.
10. Yamazaki T, Fujiwara N, Yukinaga H, et al. The closely related RNA helicases, UAP56 and URH49, preferentially form distinct mRNA export machineries and coordinately regulate mitotic progression. *Mol Biol Cell*. 2010;21(16):2953-2965 <https://doi.org/E09-10-0913>.
11. Wang L, Miao YL, Zheng X, et al. The THO complex regulates pluripotency gene mRNA export and controls embryonic stem cell self-renewal and somatic cell reprogramming. *Cell stem cell*. 2013;13(6):676-690 [https://doi.org/S1934-5909\(13\)00454-2](https://doi.org/S1934-5909(13)00454-2).
12. Bhattacharjee R, Jolly LA, Corbett MA, et al. Compromised transcription-mRNA export factor THOC2 causes R-loop accumulation, DNA damage and adverse neurodevelopment. *Nature communications*. 2024;15(1):1210 <https://doi.org/10.1038/s41467-024-45121-5>.
13. Beaulieu CL, Huang L, Innes AM, et al. Intellectual disability associated with a homozygous missense mutation in THOC6. *Orphanet journal of rare diseases*. 2013;8(1):62 <https://doi.org/10.1186/1750-1172-8-62>.
14. Kumar R, Corbett MA, van Bon BW, et al. THOC2 Mutations Implicate mRNA-Export Pathway in X-Linked Intellectual Disability. *American journal of human genetics*. 2015;97(2):302-310 <https://doi.org/10.1016/j.ajhg.2015.05.021>.
15. Kumar R, Gardner A, Homan CC, et al. Severe neurocognitive and growth disorders due to variation in THOC2, an essential component of nuclear mRNA export machinery. *Human mutation*. 2018;39(8):1126-1138 <https://doi.org/10.1002/humu.23557>.
16. Kumar R, Palmer E, Gardner AE, et al. Expanding Clinical Presentations Due to Variations in THOC2 mRNA Nuclear Export Factor. *Front Mol Neurosci*. 2020;13:12 <https://doi.org/10.3389/fnmol.2020.00012>.
17. Mattioli F, Isidor B, Abdul-Rahman O, et al. Clinical and functional characterization of recurrent missense variants implicated in THOC6-related intellectual disability. *Human molecular genetics*. 2019;28(6):952-960 <https://doi.org/10.1093/hmg/ddy391>.
18. Bhattacharjee R, Agarwala S, Mazurkiewicz D, Gecz J, Kumar R. Transcription-Export complex in neurodevelopmental disorders. *Curr Opin Genet Dev*. 2025;94:102380 <https://doi.org/10.1016/j.gde.2025.102380>.
19. Schwartz CE, Louie RJ, Toutain A, Skinner C, Friez MJ, Stevenson RE. X-Linked intellectual disability update 2022. *American journal of medical genetics Part A*. 2023;191(1):144-159 <https://doi.org/10.1002/ajmg.a.63008>.
20. Kerr B, Gedeon A, Mulley J, Turner G. Localization of non-specific X-linked mental retardation genes. *Am J Med Genet*. 1992;43(1-2):392-401 <https://doi.org/10.1002/ajmg.1320430160>.

21. Amos JS, Huang L, Thevenon J, et al. Autosomal recessive mutations in THOC6 cause intellectual disability: syndrome delineation requiring forward and reverse phenotyping. *Clinical genetics*. 2017;91(1):92-99 <https://doi.org/10.1111/cge.12793>.
22. Werren EA, LaForce GR, Srivastava A, et al. TREX tetramer disruption alters RNA processing necessary for corticogenesis in THOC6 Intellectual Disability Syndrome. *Nature communications*. 2024;15(1):1640 <https://doi.org/10.1038/s41467-024-45948-y>.
23. Philippakis AA, Azzariti DR, Beltran S, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. *Human mutation*. 2015;36(10):915-921 <https://doi.org/10.1002/humu.22858>.
24. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Human mutation*. 2011;32(8):894-899 <https://doi.org/10.1002/humu.21517>.
25. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med*. 2020;12(1):103 <https://doi.org/10.1186/s13073-020-00803-9>.
26. Pejaver V, Urresti J, Lugo-Martinez J, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nature communications*. 2020;11(1):5918 <https://doi.org/10.1038/s41467-020-19669-x>.
27. Tordai H, Torres O, Csepi M, Padanyi R, Lukacs GL, Hegedus T. Analysis of AlphaMissense data in different protein groups and structural context. *Sci Data*. 2024;11(1):495 <https://doi.org/10.1038/s41597-024-03327-8>.
28. Cheng J, Novati G, Pan J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*. 2023;381(6664):eadg7492 <https://doi.org/10.1126/science.adg7492>.
29. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*. 2019;47(D1):D886-D894 <https://doi.org/10.1093/nar/gky1016>.
30. Feng BJ. PERCH: A Unified Framework for Disease Gene Prioritization. *Human mutation*. 2017;38(3):243-251 <https://doi.org/10.1002/humu.23158>.
31. Stawinski P, Ploski R. Genebe.net: Implementation and validation of an automatic ACMG variant pathogenicity criteria assignment. *Clinical genetics*. 2024;106(2):119-126 <https://doi.org/10.1111/cge.14516>.
32. Dingemans AJM, Hinne M, Truijen KMG, et al. PhenoScore quantifies phenotypic variation for rare genetic diseases by combining facial analysis with other clinical features using a machine-learning framework. *Nature genetics*. 2023;55(9):1598-1607 <https://doi.org/10.1038/s41588-023-01469-w>.
33. Pacheco-Fiallos B, Vorlander MK, Riabov-Bassat D, et al. mRNA recognition and packaging by the human transcription-export complex. *Nature*. 2023;616(7958):828-835 <https://doi.org/10.1038/s41586-023-05904-0>.
34. Garrido-Martin D, Palumbo E, Guigo R, Breschi A. ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. *PLoS Comput Biol*. 2018;14(8):e1006360 <https://doi.org/10.1371/journal.pcbi.1006360>.
35. Affar el B, Luke MP, Gay F, et al. Targeted ablation of Par-4 reveals a cell type-specific susceptibility to apoptosis-inducing agents. *Cancer research*. 2006;66(7):3456-3462 <https://doi.org/10.1158/0008-5472.CAN-05-0964>.
36. Laemmli UK. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*. 1970;227(5259):680-685 <https://doi.org/10.1038/227680a0>.
37. Aartsma-Rus A, Garanto A, van Roon-Mom W, et al. Consensus Guidelines for the Design and In Vitro Preclinical Efficacy Testing N-of-1 Exon Skipping Antisense

- Oligonucleotides. *Nucleic Acid Ther.* 2023;33(1):17-25
<https://doi.org/10.1089/nat.2022.0060>.
38. Dubucs C, Rendu J, Michel-Calemard L, et al. Muscular phenotype description of abnormal THOC2 splicing. *Neuromuscul Disord.* 2023;33(12):978-982
<https://doi.org/10.1016/j.nmd.2023.09.009>.
 39. Tamhankar V, Tamhankar P, Chaubal R, Chaubal J, Chaubal N. Novel Consensus Splice Site Pathogenic Variation in THOC2 Gene Leads to Recurrent Arthrogryposis Multiplex Congenita Phenotype: A Case Report. *Cureus.* 2021;13(11):e19682
<https://doi.org/10.7759/cureus.19682>.
 40. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581(7809):434-443
<https://doi.org/10.1038/s41586-020-2308-7>.
 41. Drost M, Dekker J, Ferraro F, et al. Routine RNA-based analysis of potential splicing variants facilitates genomic diagnostics and reveals limitations of in silico prediction tools. *HGG Adv.* 2026;7(1):100521 <https://doi.org/10.1016/j.xhgg.2025.100521>.
 42. Luna R, Rondon AG, Perez-Calero C, Salas-Armenteros I, Aguilera A. The THO Complex as a Paradigm for the Prevention of Cotranscriptional R-Loops. *Cold Spring Harb Symp Quant Biol.* 2019;84:105-114 <https://doi.org/10.1101/sqb.2019.84.039594>.
 43. Dominguez-Sanchez MS, Barroso S, Gomez-Gonzalez B, Luna R, Aguilera A. Genome instability and transcription elongation impairment in human cells depleted of THO/TREX. *PLoS genetics.* 2011;7(12):e1002386
<https://doi.org/10.1371/journal.pgen.1002386>.
 44. Bullich G, Matalonga L, Pujadas M, et al. Systematic Collaborative Reanalysis of Genomic Data Improves Diagnostic Yield in Neurologic Rare Diseases. *J Mol Diagn.* 2022;24(5):529-542 <https://doi.org/10.1016/j.jmoldx.2022.02.003>.
 45. Zhao Y, Xing C, Peng H. ALYREF (Aly/REF export factor): A potential biomarker for predicting cancer occurrence and therapeutic efficacy. *Life Sci.* 2024;338:122372
<https://doi.org/10.1016/j.lfs.2023.122372>.
 46. Qiao L, Xu L, Yu L, et al. Rare and de novo variants in 827 congenital diaphragmatic hernia probands implicate LONP1 as candidate risk gene. *American journal of human genetics.* 2021;108(10):1964-1980 <https://doi.org/10.1016/j.ajhg.2021.08.011>.
 47. Chi B, Wang Q, Wu G, et al. Aly and THO are required for assembly of the human TREX complex and association of TREX components with the spliced mRNA. *Nucleic acids research.* 2013;41(2):1294-1306 <https://doi.org/10.1093/nar/gks1188>.
 48. Lopitz-Otsoa F, Rodriguez-Suarez E, Aillet F, et al. Integrative analysis of the ubiquitin proteome isolated using Tandem Ubiquitin Binding Entities (TUBEs). *J Proteomics.* 2012;75(10):2998-3014 <https://doi.org/10.1016/j.jprot.2011.12.001>.
 49. Katahira J, Inoue H, Hurt E, Yoneda Y. Adaptor Aly and co-adaptor Thoc5 function in the Tap-p15-mediated nuclear export of HSP70 mRNA. *EMBO J.* 2009;28(5):556-567
<https://doi.org/10.1038/emboj.2009.5>.
 50. Di Gregorio E, Bianchi FT, Schiavi A, et al. A de novo X;8 translocation creates a PTK2-THOC2 gene fusion with THOC2 expression knockdown in a patient with psychomotor retardation and congenital cerebellar hypoplasia. *Journal of medical genetics.* 2013;50(8):543-551 <https://doi.org/10.1136/jmedgenet-2013-101542>.
 51. Amsterdam A, Nissen RM, Sun Z, Swindell EC, Farrington S, Hopkins N. Identification of 315 genes essential for early zebrafish development. *Proceedings of the National Academy of Sciences of the United States of America.* 2004;101(35):12792-12797
<https://doi.org/10.1073/pnas.0403929101>.

52. Wang X, Chang Y, Li Y, Zhang X, Goodrich DW. Thoc1/Hpr1/p84 is essential for early embryonic development in the mouse. *Molecular and cellular biology*. 2006;26(11):4362-4367 <https://doi.org/10.1128/MCB.02163-05>.
53. Li Y, Lin AW, Zhang X, Wang Y, Wang X, Goodrich DW. Cancer cells and normal cells differ in their requirements for Thoc1. *Cancer research*. 2007;67(14):6657-6664 <https://doi.org/67/14/6657>.
54. Pitzonka L, Wang X, Ullas S, Wolff DW, Wang Y, Goodrich DW. The THO ribonucleoprotein complex is required for stem cell homeostasis in the adult mouse small intestine. *Molecular and cellular biology*. 2013;33(17):3505-3514 <https://doi.org/10.1128/MCB.00751-13>.
55. Bai X, Ni J, Beretov J, et al. THOC2 and THOC5 Regulate Stemness and Radioresistance in Triple-Negative Breast Cancer. *Adv Sci (Weinh)*. 2021;8(24):e2102658 <https://doi.org/10.1002/advs.202102658>.
56. Gupta N, Yadav S, Gurramkonda VB, VI R, Sg T, Kabra M. First report of THOC6 related intellectual disability (Beaulieu Boycott Innes syndrome) in two siblings from India. *Eur J Med Genet*. 2020;63(3):103742 <https://doi.org/10.1016/j.ejmg.2019.103742>.
57. Sharma S, Schiller MR. The carboxy-terminus, a key regulator of protein function. *Crit Rev Biochem Mol Biol*. 2019;54(2):85-102 <https://doi.org/10.1080/10409238.2019.1586828>.
58. Quinodoz M, Peter VG, Cisarova K, et al. Analysis of missense variants in the human genome reveals widespread gene-specific clustering and improves prediction of pathogenicity. *American journal of human genetics*. 2022;109(3):457-470 <https://doi.org/10.1016/j.ajhg.2022.01.006>.
59. Nousiainen HO, Kestila M, Pakkasjarvi N, et al. Mutations in mRNA export mediator GLE1 result in a fetal motoneuron disease. *Nature genetics*. 2008;40(2):155-157 <https://doi.org/10.1038/ng.2007.65>.
60. Folkmann AW, Collier SE, Zhan X, Aditi, Ohi MD, Wentz SR. Gle1 functions during mRNA export in an oligomeric complex that is altered in human disease. *Cell*. 2013;155(3):582-593 <https://doi.org/10.1016/j.cell.2013.09.023>.
61. Okamura M, Yamanaka Y, Shigemoto M, et al. Depletion of mRNA export regulator DBP5/DDX19, GLE1 or IPPK that is a key enzyme for the production of IP6, resulting in differentially altered cytoplasmic mRNA expression and specific cell defect. *PloS one*. 2018;13(5):e0197165 <https://doi.org/10.1371/journal.pone.0197165>.
62. Sharma M, Wentz SR. Nucleocytoplasmic shuttling of Gle1 impacts DDX1 at transcription termination sites. *Mol Biol Cell*. 2020;31(21):2398-2408 <https://doi.org/10.1091/mbc.E20-03-0215>.
63. Wang Y, Xu Y, Zhou C, et al. Exome sequencing reveals genetic heterogeneity and clinically actionable findings in children with cerebral palsy. *Nat Med*. 2024;30(5):1395-1405 <https://doi.org/10.1038/s41591-024-02912-z>.
64. Seigel GM, Onwumere O, Sauane M, et al. ALYREF/THOC4 expression and cell growth modulation in retinoblastoma. *Pathol Res Pract*. 2024;260:155392 <https://doi.org/10.1016/j.prp.2024.155392>.
65. Neurohr GE, Terry RL, Lengefeld J, et al. Excessive Cell Growth Causes Cytoplasm Dilution And Contributes to Senescence. *Cell*. 2019;176(5):1083-1097 e1018 <https://doi.org/10.1016/j.cell.2019.01.018>.
66. Klein A, Rhinn M, Keyes WM. Cellular senescence and developmental defects. *FEBS J*. 2023;290(5):1303-1313 <https://doi.org/10.1111/febs.16731>.
67. Gromadzka AM, Steckelberg AL, Singh KK, Hofmann K, Gehring NH. A short conserved motif in ALYREF directs cap- and EJC-dependent assembly of export

- complexes on spliced mRNAs. *Nucleic acids research*. 2016;44(5):2348-2361
<https://doi.org/10.1093/nar/gkw009>.
68. Shi M, Zhang H, Wu X, et al. ALYREF mainly binds to the 5' and the 3' regions of the mRNA in vivo. *Nucleic acids research*. 2017;45(16):9640-9653
<https://doi.org/10.1093/nar/gkx597>.
69. Rodrigues JP, Rode M, Gatfield D, Blencowe BJ, Carmo-Fonseca M, Izaurralde E. REF proteins mediate the export of spliced and unspliced mRNAs from the nucleus. *Proceedings of the National Academy of Sciences of the United States of America*. 2001;98(3):1030-1035 <https://doi.org/10.1073/pnas.98.3.1030>.
70. Okada M, Jang SW, Ye K. Akt phosphorylation and nuclear phosphoinositide association mediate mRNA export and cell proliferation activities by ALY. *Proceedings of the National Academy of Sciences of the United States of America*. 2008;105(25):8649-8654 <https://doi.org/10.1073/pnas.0802533105>.
71. Stubbs SH, Conrad NK. Depletion of REF/Aly alters gene expression and reduces RNA polymerase II occupancy. *Nucleic acids research*. 2015;43(1):504-519
<https://doi.org/10.1093/nar/gku1278>.
72. Fan J, Wang K, Du X, et al. ALYREF links 3'-end processing to nuclear export of non-polyadenylated mRNAs. *EMBO J*. 2019;38(9)
<https://doi.org/10.15252/embj.201899910>.
73. Niemi MEK, Martin HC, Rice DL, et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature*. 2018;562(7726):268-271
<https://doi.org/10.1038/s41586-018-0566-4>.
74. Wang T, Birsoy K, Hughes NW, et al. Identification and characterization of essential genes in the human genome. *Science*. 2015;350(6264):1096-1101
<https://doi.org/10.1126/science.aac7041>.
75. 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. *Genetics in medicine : official journal of the American College of Medical Genetics*. 2015;17(5):405-424. <https://doi.org/10.1038/gim.2015.30>

Figure legends

Figure 1 Genetic and clinical features of *THOC2* syndrome. A. Schematic illustration of the human *THOC2* protein showing nucleic acid binding domain (NBD) and the new (red) and published^{14,15,16} (blue) *THOC2* variants (NM_001081550.2). Missense and splicing, frameshift or deletion variants are shown in the upper and lower panel, respectively. The # after each variant indicates study participant's number. B. Clinical presentations in *THOC2* affected individuals. C. Photographs of the current cohort of affected individuals carrying the *THOC2* variants.

Figure 2 *THOC2* c.4449+5_4449+6del variant causes aberrant splicing leading to p.(Leu1438_Arg1483del). A. Schematic of *THOC2* gene showing the location of GT deletion in intron 34 that causes aberrant splicing and skipping of 138 bp exon 34 (deleting 46 amino acids). B. Sashimi plots showing *THOC2* mRNA splicing events in the control and affected fibroblasts. Splicing junctions are displayed as arcs connecting the exons. The number of reads for each junction is indicated within the arcs. Y-axis show read depth. C. Western blot showing smaller *THOC2* protein in the affected compared to control fibroblasts. Lanes 1 and 2 showing *THOC2* proteins in fibroblasts from two different passages.

Figure 3 *THOC2* c.2482-1_2484del causing aberrant splicing is associated with arthrogryposis multiplex congenita 7, X-linked (OMIM:301127). A. Schematic of *THOC2* gene showing the location of CAGT deletion at the intron 22-exon 23 boundary. Percentage of misplicing events is shown on the right. Sanger sequencing chromatograms showing genomic (PCRed with P582/P585 primers) and cDNA sequences (PCRed with P394/P397 primers) of affected and control fibroblasts are shown. B. Sashimi plots showing *THOC2* mRNA splicing events in the control and affected fibroblasts. Splicing junctions are displayed as arcs connecting the exons. The number of reads for each junction is indicated within the arcs. Y-axis show read depth.

Figure 4 *THOC2* c.2757G>A p.(Met919Ile) causes aberrant splicing leading to skipping of exon 23. A. Schematic of *THOC2* gene showing the location of the G>A variant in exon 23. B. Sashimi plots showing *THOC2* mRNA splicing events in the affected son, heterozygous mother and control fibroblasts. Splicing junctions are displayed as arcs connecting the exons. There is skipping of exon 23 in mother and son compared to controls.

The ratio of reads (wild-type:skipped) is ~4:1 in heterozygous mother and ~1:1 in the affected son.

Figure 5 *THOC2* c.4754+1A>G variant causes aberrant splicing resulting in replacement of p.(His1586_Arg1593) (r.4755_4761del) by 13 new amino acids

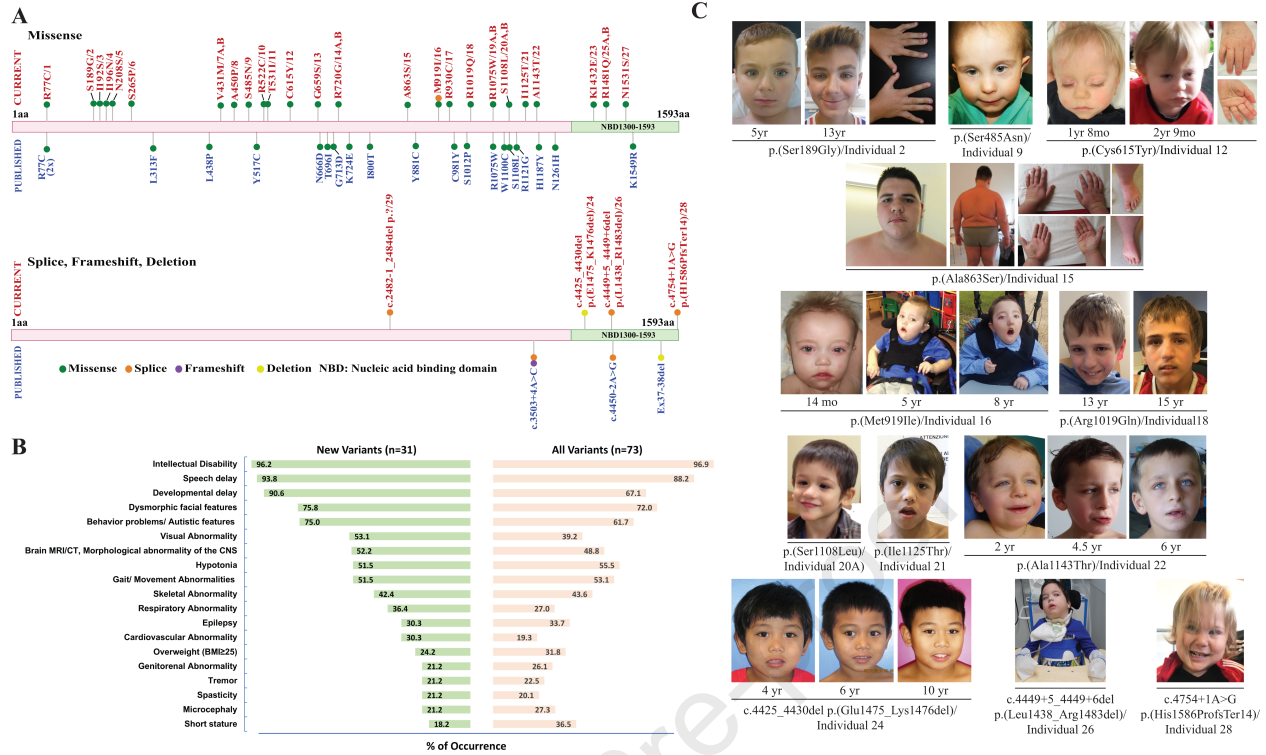
p.(His1586ProfsTer14) at the C-terminus. A. Schematic of *THOC2* gene showing the location of A>G splice variant in intron 37. B. Western blot confirming the deletion of C-terminal eight amino acids of the *THOC2* protein. A schematic showing location of the two antibody binding sites within the *THOC2* protein (bottom). NBD, nucleic acid binding domain.

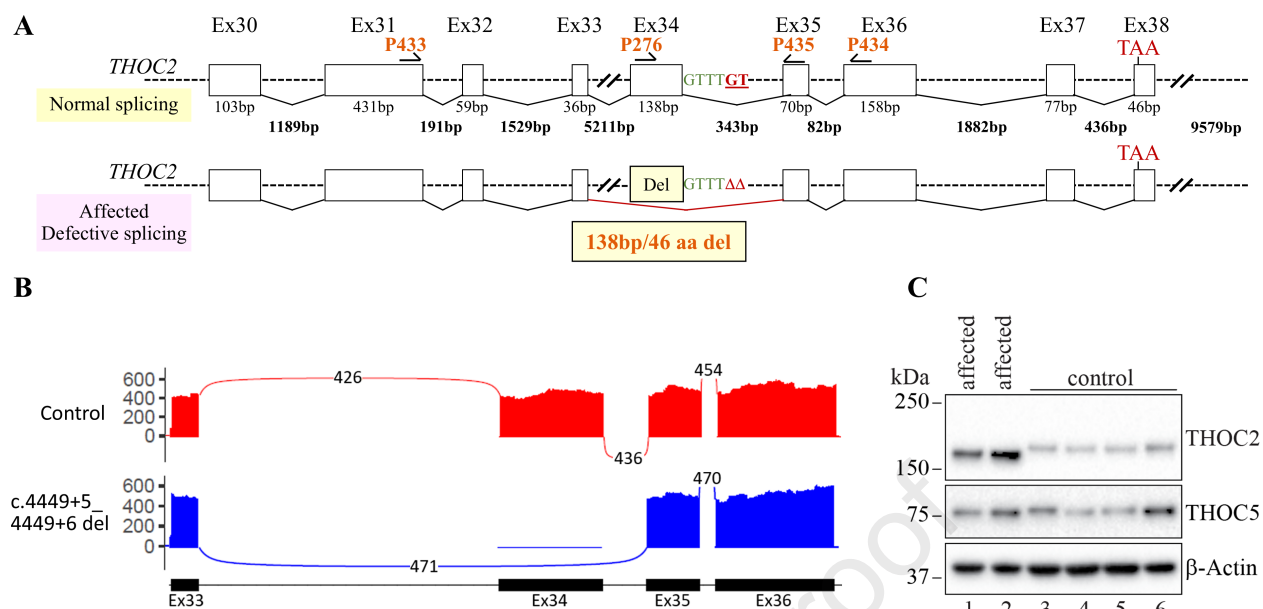
Figure 6 *THOC6* c.553T>C p.(Ser185Pro) and c.136G>A p.(Gly46Arg) variant proteins

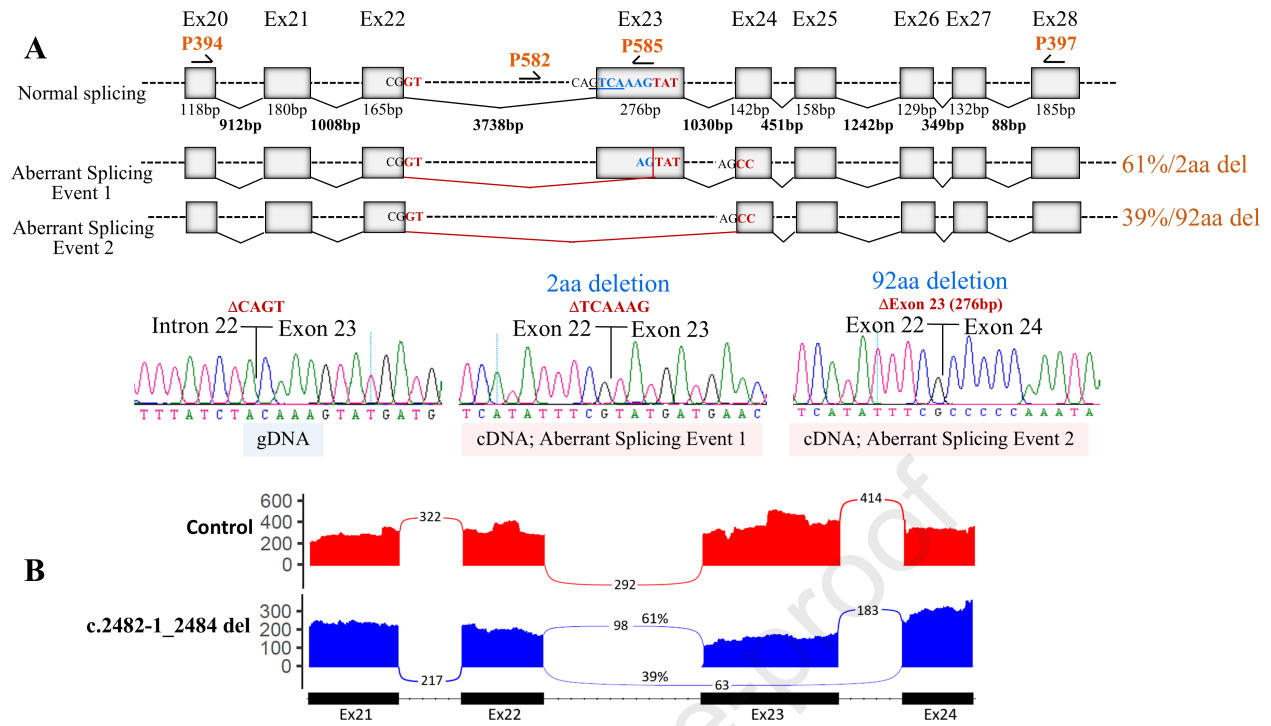
are mislocalised to cytoplasm and lose interaction with the THO subunits. A. Schematic showing locations of published (black text) and new (red text) *THOC6* variants. c.664T>C p.(Trp222Arg) and c.945+1 G>A (• not shown) are compound heterozygous variants. B. Endogenous p.(Gly46Arg) *THOC6* protein mislocalises to cytoplasm and has reduced stability in the affected LCLs. C. Reduced p.(Gly46Arg) *THOC6* protein is not due to altered levels of mRNA in the affected LCLs. Data is represented as the mean±SEM. D. p.(Ser185Pro) and p.(Gly46Arg) *THOC6* proteins loose interaction with the *THOC2*, *THOC5* and *THOC7* subunits. E. Immunofluorescence showing mislocalisation of the p.(Ser185Pro) *THOC6* protein in HeLa cells transfected with plasmids expressing Myc-tagged WT or p.(Ser185Pro) *THOC6* protein. Cells were stained for *THOC6* (green, mouse anti-c-Myc antibody), *THOC2* (red, rabbit anti-*THOC2* antibody, A303-630A) and nuclei (blue, DAPI).

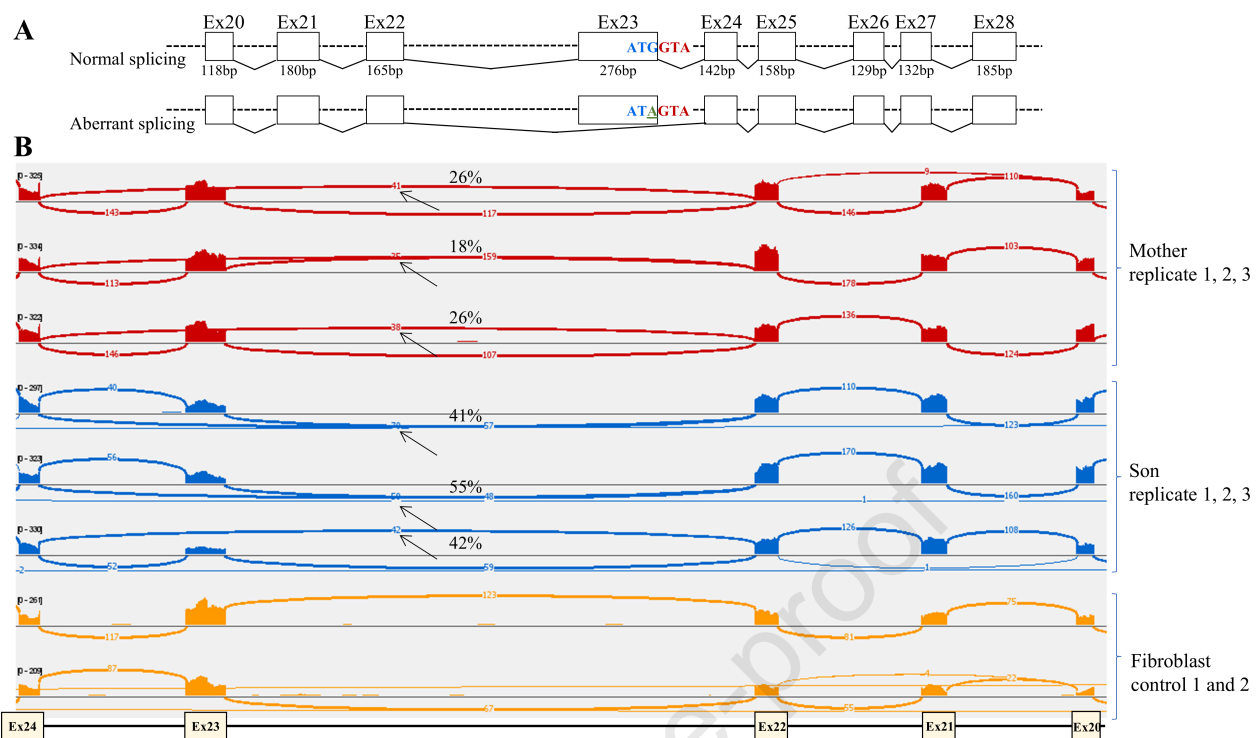
Figure 7 *THOC7* c.178T>C p.(Cys60Arg) variant reduces protein stability, increases DNA damage and R-loop accumulation. A. Western blot showing reduced *THOC7* variant and *THOC1*, *THOC2*, *THOC5*, *THOC6* but not *THOC3*, *ALYREF* and *UAP56* protein stability in the two affected compared to the normal control fibroblasts. B. % Comet cells are significantly higher in affected than control fibroblasts (n=2). One-way ANOVA. *p<0.05, **p<0.01. C. R-loop accumulation is significantly increased in affected compared to control fibroblasts (n=3; >600 cells each). S9.6 antibody-stained immunofluorescence images are shown in Supplemental Figure 11. Each data point indicates the number of S9.6 foci/nucleus that were calculated by dividing the total number of S9.6 foci by the number of cells in one microscopic field. Mann-Whitney U-test. ***p<0.001, ****p<0.0001

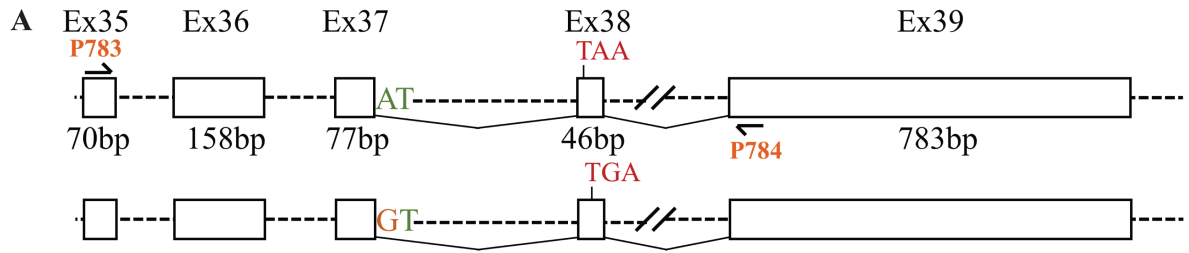
Figure 8 *de novo* stop gain *ALYREF* exon 2: c. 283C>T p.(Gln95Ter) heterozygous variant causes microcephaly and speech problems. A. Schematics showing location of the two variants. UBM, *UAP56*-binding motif; RBD, RNA-binding domain; RRM, RNA-recognition motif. B. Sanger sequencing chromatograms of genomic DNA sequence showing heterozygous c.283C>T variant. C. Immunofluorescence detection of *ALYREF* protein in control and affected fibroblasts. The affected fibroblasts are significantly larger than the control fibroblasts (n=23). Mann-Whitney U-test. ****p<0.0001











B

