Identification of novel genetic causes of monogenic intellectual disability

Francesca Mattioli

▶ **To cite this version:**

Francesca Mattioli. Identification of novel genetic causes of monogenic intellectual disability. Neurobiology. Université de Strasbourg, 2018. English. NNT : 2018STRAJ035 . tel-01963143

HAL Id: tel-01963143
https://theses.hal.science/tel-01963143

Submitted on 21 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

Institut de Génétique et Biologie Moléculaire et Cellulaire

THÈSE

présentée par :

Francesca MATTIOLI

soutenue le : **26 Juin 2018**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Aspects Moléculaires et Cellulaires de la Biologie

Identification of Novel Genetic Causes of Monogenic Intellectual Disability

THÈSE co-dirigée par :

M MANDEL Jean-Louis	Pr, Collège de France, IGBMC, Strasbourg
Mme PITON Amélie	MCU-PH, Université de Strasbourg, IGBMC, Strasbourg

RAPPORTEURS :

Mme RENIERI Alessandra	Pr, Università di Siena
M REYMOND Alexandre	Pr, Center for Integrative Genomics, Lausanne

---

AUTRES MEMBRES DU JURY :

M SERAPHIN Bertrand	DR CNRS, IGBMC, Strasbourg
Mme ZWEIER Christiane	Pr, Institute of Human Genetics, Erlangen

ACKNOWLEDGMENTS

I would like to thank Alessandra Renieri, Alexandre Reymond, Bertrand Seraphin and Christiane Zweier for accepting to be part of the jury, for evaluating my work and for having found time in your busy schedule to come and take part at my PhD defence.

These (almost) four years have been really important to me, for my professional and personal growth. Of these, I would like to thank Jean-Louis and Amélie for accepting me in their équipe and trusting me since the beginning. I'm really honored to be part of your team and you both guided me into becoming closer to the scientist I want to be. Amélie, I cannot thank you enough for your supervision: you patiently taught me everything and you have been always present for any problem. You have been really a model to me of super-woman in science! Jean-Louis, thank for you for your leading by example, your tremendous knowledge that is really inspiring… and for the alsatian gourmet advises!

Thanks to all the people in the small Mandel's lab, for the friendly environment: the current members, particularly Jérémie and Florent for the daily help, but also to the previous members, remarkably Angélique for being always willing to solve any problem (especially with the cells!), and for all the chocolate moments. Thanks to the X-fra group, particularly Eric for his tips on the molecular and cloning techniques, and to the Chelly's team for having welcomed us in the ICS. I would like to particularly thank Nathalie for the help with the libraries (but not only) and to Vicky, especially in the last period, for the microscope expertise.
Thanks to the people of the TMN department for the constant sharing of protocols, reagents and ideas: it has been really of help for my work. Thanks to all the people in the facilities that really simplified the work: the cell-culture, the molecular biology (grazie Paola per i plasmidi!) and especially to the people in the sequencing platform, for being always supportive and willing to help. A big thank to the bioinformaticians that were involved in the different projects. Thanks also to the funding agencies, particularly the Fondation Jérôme Lejeune.

Then, as a PhD is not always about science, thanks to the friends that I found at the IGBMC. Thanks to the Obama's girl: Dami, Lorraine and Samantha. Thanks for the support and all the laugh that we had. Now we are spread over three continents, but I really hope we are going to meet again soon. Thanks also to the koalas Arielle and Tiphaine for sharing our troubles and comfort each other. Thanks to the Italian community and its friends for understanding and sharing our culinary and loud-speak problems.

Ed infine, il mio core parla italiano (*Cit*). Sono stati quattro anni difficili, ma ho sempre avuto persone che mi hanno fatto sentire a casa e ricaricare le pile per affrontare al meglio tutti i problemi. Per questo vorrei ringraziare i colleghi-amici "pavesini", per condividere più o meno le stesse esperienze e rimanere comunque uniti, nonostante le distanze. Ed infine, vorrei ringraziare lo zoccolo duro ferrarese: Annina, Borin, Fede (si, anche tu sei ferrarese), Giammi, Laura, Porrig e Zolia. Nonostante gli anni e le distanze, ritornare al porto sicuro per me vuol dire anche rivedervi.

Non potro mai ringraziare abbastanza Antonio per il totale su(o)pporto in questi anni. Sei stato il miglior compagno di dottorato che potessi avere. Sono contenta di avere condiviso questo percorso difficile con te ma non vedo l'ora di iniziare altre avventure insieme.

Ultimi ma per niente ultimi, un ringraziamento alla mia famiglia, per il costante sostegno e per avermi fatto le spalle larghe attorno a quel famoso tavolo bianco in cucina. Tutto questo non sarebbe stato possibile senza di voi.

SUMMARY

Résumé de la thèse

La déficience intellectuelle (DI) est une trouble du neurodéveloppement caractérisée par une extrême hétérogénéité génétique, avec plus de 700 gènes impliqués dans des formes monogéniques de DI. Cependant un nombre important de gènes restent encore à identifier et les mécanismes physiopathologiques de ces maladies neurodéveloppementales restent encore à comprendre. Mon travail de doctorat a consisté à identifier de nouvelles causes génétiques impliquées dans la DI. En utilisant différentes techniques de séquençage de nouvelle génération, j'ai pu augmenter le taux de diagnostic chez les patients avec DI et identifié plusieurs nouvelles mutations (dans *AUTS2, THOC6*, etc) et nouveaux gènes (*BRPF1, NOVA2,* etc) impliqués dans la DI. Pour les moins caractérisés, j'ai effectué des investigations fonctionnelles pour valider leur pathogénicité, caractériser les mécanismes moléculaires qu'ils affectent et identifier leur rôle dans cette maladie.

Mes travaux de doctorat permettront d'améliorer et d'accélérer la possibilité d'obtenir un diagnostic moléculaire qui donnera accès à un meilleur suivi et à une meilleure prise en charge pour les patients. Cela permettra également de mieux comprendre les mécanismes physiopathologiques impliqués dans ces troubles neurodéveloppementaux. Ces connaissances aideront éventuellement à identifier de nouvelles cibles thérapeutiques.

PROBLEMATIQUE

Les troubles du neurodéveloppement sont les conséquences cliniques d'anomalies survenues à un moment au cours du développement du cerveau. Ces anomalies peuvent être des facteurs environnementaux ou des mutations génétiques dans des gènes fortement exprimés dans le cerveau jouant un rôle dans son processus de développement (prolifération des précurseurs neuronaux, migration des neurones et établissement des connections entre neurones). La déficience intellectuelle (DI), caractérisée par l'apparition de troubles des fonctions cognitives et des capacités adaptatives avant l'âge de trois ans, et les troubles du spectre autistique (TSA), caractérisés par des troubles de la communication, des interactions sociales et la présence d'intérêt et de comportements répétés et restreints, constituent des troubles du neurodéveloppement fréquents dans la population (>2% des enfants) et représentent un enjeu majeur de santé publique. Les connaissances sur les mécanismes génétiques impliqués dans les différentes formes de DI ou de TSA ont considérablement progressé au cours des dernières années, en raison du développement de nouvelles technologies de séquençage permettant d'étudier le génome en partie ou en intégralité. À aujourd'hui, plus de 700 gènes ont été impliqué dans des formes de DI, cependant un nombre important de gènes restent encore à identifier et les mécanismes physiopathologiques de ces maladies neurodéveloppementales restent encore à

comprendre, ce qui complique l'obtention d'un diagnostic moléculaire pour les patients. Cette hétérogénéité au niveau moléculaire rend la recherche exhaustive de mutations souvent très compliquée et une très grande proportion de patients atteints de maladie génétiquement hétérogènes demeurent ainsi non diagnostiqués sur le plan moléculaire. De plus, il peut arriver qu'il ne soit pas possible de conclure en l'état quant à la pathogénicité potentielle de certaines variations identifiées. Il est alors nécessaire de les valider fonctionnellement pour démontrer qu'elles affectent la fonction de la protéine correspondante et qu'elles peuvent affecter des voies moléculaires cellulaires à la base du développement de la maladie. Cette validation fonctionnelle peut prendre beaucoup de temps et retarde d'autant l'obtention du diagnostic. Or l'obtention d'un diagnostic étiologique précis et précoce est primordial pour les familles afin d'être en mesure d'anticiper et de proposer une prise en charge et un traitement adaptés notamment chez le jeune enfant, et à plus long terme de développer des approches thérapeutiques personnalisées (thérapies gène/ mutation spécifique).

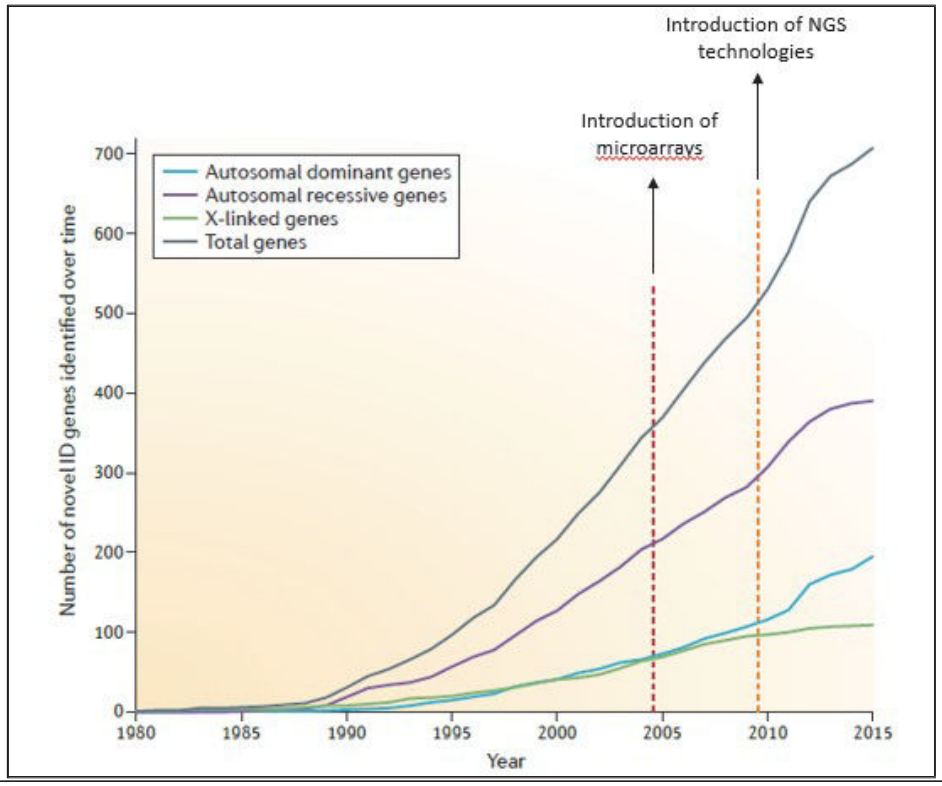
SITUATION DU SUJET DANS LA LITTÉRATURE INTERNATIONALE

En raison du biais de sexe conséquent observé dans la DI (1.3-1.4 hommes atteints pour une femme) et suite à l'identification dans les années 60 de grandes familles avec ségrégation clairement liée à l'X, la recherche des gènes responsable s'est principalement concentrée –et ce jusqu'à récemment– sur les gènes du chromosome X, et donc responsables de DI liée à l'X (XLID). A ce jour, et grâce aux efforts de larges consortiums internationaux, une centaine de gènes de XLID a été identifiée (Gécz et al., 2009; Lubs et al., 2012; Ropers, 2010; Tarpey et al., 2009). Il reste encore des gènes à identifier sur le chromosome X (Hu et al., 2016) qui pourraient être impliqués dans des formes de DI, si l'on en croit les nombreuses familles avec DI liées à l'X mais sans mutation identifiée après séquençage d'exome.

La recherche de gènes autosomiques associés à la DI récessive s'est développé de façon plus récente. La méthode d'identification la plus évidente étant la recherche de régions d'homozygotie dans des familles consanguines couplée à des analyses de liaison, et de suivies du séquençage systématique des gènes candidates. Cette approche a permis l'identification de plus de 100 gènes de DI autosomique récessive (Kaufman et al., 2010; Kuss et al., 2011). Au cours des deniers années, l'utilisation des nouvelles technologies de séquençage ont été utilisés dans des larges cohortes des familles consanguines, soulignent l'efficace de ces techniques pour le diagnostic mais aussi pour l'identification des nouvelle gènes impliquées dans le DI récessive (Najmabadi et al., 2011; Riazuddin et al., 2016).

En raison de l'impact de la DI sur l'aptitude à la reproduction, l'hypothèse d'un mode de transmission autosomal dominant familial avait longuement été délaissée, puisque supposé très rare pour des mutations pleinement pénétrantes. L'identification de syndromes associés à des microdélétions a alors prouvé que des mutations de novo pouvaient être une cause majeure chez les patients sporadiques

(seule personne atteinte de la famille). Ceci a depuis été largement confirmé par l'identification de mutations ponctuelles tronquantes de novo reportées dans des quelques cas simplex de DI, au début dans une petite cohorte de 10 patients (Vissers et al., 2010) dont l'ADN (et celui de leurs parents non-atteints) avait été séquencé par Whole Exome Sequencing (WES). Ensuite, il a été montré sur des cohortes légèrement plus larges (50-100 patients) que le WES permettait d'obtenir un taux de diagnostic important pour la DI (de Ligt et al., 2012; Rauch et al., 2012). Le WES est très efficace également pour identifier de nouveaux gènes impliqués dans la DI, comme cela a été démontré par le consortium « Deciphering Develompental Disorder Study » qui a séquencé par WES plus de 4,000 familles avec troubles du neurodéveloppement et a appliqué une analyse statistique pour identifier des nouveaux gènes associés à ces conditions (Deciphering Developmental Disorders Study, 2015).



Identification des nouveaux gènes impliqués dans la DI en cours des années
(Adapted from (Vissers et al., 2016))

Une étude de séquençage du génome entier (Whole-Genome Sequencing, WGS) sur 50 patients dont l'exome avait été préalablement analysé (de Ligt et al., 2012) a montré que trois quarts des variations de novo avaient été manqué par le WES, principalement dû à des limitations techniques. De même, les auteurs ont aussi détecté des variants structuraux qui n'avaient pas identifié précédemment dans les analyses de microarrays. Dans l'ensemble, avec le WGS, ils ont estimé qu'une grande proportion des cas de DI sévère peut être expliqué par des mutations de novo situées ou affectant une séquence codante (Gilissen et al., 2014).

Dans les dernières années, de nombreux nouveaux gènes impliqués dans la DI ont été identifiés, et de nombreux autres sont encore à identifier. Cette augmentation du nombre de gènes impliqués dans la DI et les TSA aide à mieux comprendre les mécanismes physiopathologiques qui conduisent à ces fréquents troubles du neurodéveloppement. Par exemple, ces dernières années, plusieurs gènes impliqués dans le remodelage de la chromatine ont été récemment identifiés comme impliqués dans la DI, soulignant une voie moléculaire qui avait été jusque-là sous-estimée (*SETD5*, *KDM6A*, *MOZ,* etc).

RÉSULTATS

Mon travail de doctorat consiste à identifier de nouveaux gènes et de nouveaux mécanismes moléculaires impliqués la DI, ainsi qu'à caractériser les signatures cliniques et moléculaires associés à certaines formes génétiques, pour mieux les diagnostiquer, les comprendre, et les prendre en charge.

J'ai utilisé des approches de séquençage ciblé et de séquençage d'exome entier (WES) pour analyser l'ADN de patients avec DI avec ou sans troubles du spectre autistique (TSA) pour identifier de nouvelles mutations et de nouveaux gènes impliqués dans ces deux maladies neurodéveloppementales. J'ai notamment analysé plusieurs centaines de gènes candidats pour 86 patients avec DI et/ou TSA ainsi que l'exome entier pour 36 patients pour lesquels aucune mutation n'avait pu être mise en évidence par séquençage ciblé. J'ai identifié en total 26 variations pathogènes.

Confirmation de l'implication de gènes de DI récemment identifiés et compréhension des mécanismes moléculaires altérés

L'identification de ces mutations m'a permis de confirmer l'implication dans la DI/TSA de gènes récemment publiés comme le gène *ZBTB20*, *THOC6, AUTS2,* et plusieurs d'autres. La description de nouveaux patients porteurs d'une mutation dans ces gènes permet de mieux définir le spectre clinique associé à chacune de ces entités génétiques comme c'est le cas pour le syndrome de Primrose causé par des mutations dans le gène *ZBTB20* pour lequel nous avons mis en évidence une hypothyroïdie associée (***Mattioli et al., 2016)***. Pour certaines mutations faux-sens, c'est-à-dire qui ne changent qu'un ou plusieurs acides aminés de la protéine correspondante, j'ai réalisé des études fonctionnelles pour étudier leurs conséquences sur l'expression et la localisation de la protéine et démontrer leur pathogénicité. En particulier, j'ai étudié trois variations faux-sens identifiées dans le gène *THOC6* et présentes à l'état homozygote chez un garçon avec DI. L'haplotype constitué de ces trois variations est présente chez plusieurs autres patients avec DI et présentant des signes cliniques similaires et représente donc une mutation récurrente pour cette forme de DI. La protéine THOC6 est impliquée dans le transport et la maturation des ARN messagers et normalement est localisée dans le noyau. En

surexprimant la protéine THOC6 avec chacune de ces variations ou l'ensemble des trois dans des cellules HeLa, j'ai montré qu'une de ces variations en particulier et la combinaison des trois conduisait à une mauvaise localisation de la protéine THOC6, restreinte au cytoplasme (***Mattioli et al. Manuscrit en cours d'écriture***).

L'identification de mutations dans des gènes connus permet également d'analyser certains mécanismes physiopathologiques impliques impliqués la DI. Je me suis intéressée en particulier à mieux comprendre le syndrome de DI causée par des mutations du gène *AUTS2,* dans lequel j'ai identifié 4 variations conduisant à l'apparition d'une protéine tronquée. Très peu de mutations ponctuelles ont été identifiées dans ce gène et la plupart des mutations rapportées étant des délétions. J'ai réalisé des études transcriptomiques (séquençage d'ARNm) dans des fibroblastes de 4 patients porteurs de mutations dans *AUTS2* comparé à 4 individus contrôles et j'ai mis en évidence que la régulation du cycle cellulaire et notamment les mécanismes impliqués dans la séparation des chromatides (protéines du centrosome et des kinétochores) étaient altérés chez les patients avec une mutation dans le gène *AUTS2*. De façon intéressante, certains de ces gènes dont l'expression est altérée sont des gènes impliqués dans des formes de DI avec microcéphalie sévère. Je souhaite poursuivre cette étude en analysant s'il existe une éventuelle altération des centrosomes (analyse du fuseau mitotique) dans les fibroblastes de patients. J'aimerais également, en collaboration avec deux équipes de l'institut, analyser chez la souris inactivée pour *Auts2* si l'expression des mêmes gènes est dérégulée (souris *Auts2* +/-, collaboration avec Yann Herault) mais également étudier le rôle d'*AUTS2* et de ces gènes cibles au cours du développement du cerveau (collaboration avec Juliette Godin).

Identification de quatre nouveaux gènes impliqués dans la DI

Les analyses WES ont permis l'identification de mutations candidates dans 4 nouveaux gènes de DI (*BRPF1, NOVA2, UNC13A, NARS*). La réalisation d'études fonctionnelles et l'identification de mutations dans ces gènes chez d'autres patients/familles présentant les mêmes signes cliniques sont nécessaires pour confirmer l'implication de ces gènes dans la DI.

J'ai démontré pour la première fois que le gène *BRPF1* était responsable d'une forme de DI avec retard de croissance, microcéphalie, hypotonie et paupière tombante (ptosis). Une analyse WES a révélé une délétion à l'état hétérozygote dans le gène *BRPF1*, entrainant un décalage du cadre de lecture et l'apparition d'un codon stop prématuré. Cette variation est présente chez tous les individus atteints de la famille (six personnes avec DI légère et d'autres anomalies du développement). *BRPF1* code une protéine impliquée dans la régulation de la transcription, via l'activation d'histones acétyltransférases de la famille MYST, comme les protéines MOZ et MORF qui sont également impliqués dans des formes de DI. J'ai entrepris des validations fonctionnelles pour mieux comprendre les conséquences de la

mutation sur la fonction de *BRPF1*. Dans un premier temps, j'ai montré sur des fibroblastes d'un des individus atteints que l'ARN muté était exprimé et échappait particulièrement au mécanisme de dégradation des ARN non-sens. J'ai observé ainsi que la mutation conduisait à une protéine tronquée. En surexprimant le cDNA muté ou normal de BRPF1 dans des cellules HEK293 avec ses partenaires (MOZ, ING5, MEAC6), j'ai démontré que l'interaction avec MOZ était préservée tandis que l'interaction avec ING5 et MEAC6 est abolie quand BPRF1 est muté. La localisation cellulaire est également affectée. J'ai également montré que la mutation conduisait à une diminution de l'acétylation de l'histone H3 (H3K23) est à une augmentation de l'expression de gènes HOX. Via les bases de données et les nouveaux outils d'échange de données phénotypiques et génotypiques (Decipher et Genematcher), j'ai pu avoir connaissance d'au moins 6 autres patients avec mutations ou délétions du gène BRPF1. J'ai montré qu'ils présentaient tous une DI légère à modérée, avec ptosis uni ou bilatéral et/ou blépharophimosis. L'ensemble de ces résultats, reportant BRPF1 comme un nouveau gène impliqué dans la déficience intellectuelle, a été récemment publié (***Mattioli et al., 2017)***.

Toujours par WES, j'ai identifié, cette fois chez un cas sporadique, une mutation de novo conduisant à l'apparition d'une protéine tronquée dans le gène *NOVA2*. Ce gène code une protéine impliquée dans la régulation de l'épissage, et jouant un rôle important au cours du développement du cerveau. La patiente présente une DI, une microcéphalie et une épilepsie et j'ai pu identifier, en échangeant avec des collègues, quatre autres mutations dans ce gène chez des patients avec signes cliniques similaires. Au niveau fonctionnel, en surexprimant l'ADNc humaine de *NOVA2* muté dans des cellules HeLa ou *NOVA2* n'est pas exprimé, j'ai démontré la présence des protéines tronquées qui ne sont pas capable de réguler l'épissage alternatif de gènes connu pour être ciblé par NOVA2. J'ai démontré aussi que la réduction de l'expression de *NOVA2* altère l'épissage alternatif dans des précurseurs neuronaux humains en culture. En parallèle, en collaboration avec une autre équipe de mon institut de recherche, j'ai entrepris d'étudier les conséquences d'une inactivation de ce gène in vivo en utilisant le modèle zébrafish, qui permettra de mesurer l'effet sur la taille de la tête ou le développement de crises d'épilepsie comme ce qui est observé chez les patients (collaboration avec Christelle Golzio). Les résultats de ces études génétiques et fonctionnelles, rapporteront le gène *NOVA2* comme un nouveau gène de déficience intellectuelle.

Chez une autre patiente, j'ai identifié une mutation tronquante héritée et un faux-sens de novo dans le gène *UNC13A*, qui code une protéine impliqué dans la régulation de la libération des neurotransmetteurs dans les synapses des cellules nerveuses. Une mutation homozygote dans ce gène avaient déjà été identifiées mais cette fois-ci chez un patient avec microcéphalie et une myasthénie fatal. Il sera intéressant d'étudier les conséquences de ces variants sur le cycle des vésicules

synaptiques. Des études fonctionnelles seront effectuées en collaboration avec Dr. Lipstein (Max-Plank Institute for Experimental Medicine).

Dans un autre cas sporadique, j'ai identifié une mutation de novo qui entraine la suppression d'un domaine fonctionnel située à la fin de la protéine NARS, une protéine impliquée dans la formation des tRNAs. En recherchant des autres patients avec une mutation dans le même gène, j'ai découvert 3 jeunes enfants avec la même mutation présente dans mon patient et plusieurs autres cases avec des mutations faux-sens mais homozygote. Nous sommes actuellement en contact avec des autres équipes pour effectuer des études fonctionnelles in vitro –pour vérifier l'activité catalytique- (Hubert Becker, UDS, Strasbourg) et in vivo (Andreea Manole, UCL, London).

CONCLUSIONS ET PERSPECTIVES

Les résultats que j'ai obtenus permettent de mieux comprendre les mécanismes génétiques et moléculaires impliqués dans la déficience intellectuelle et les troubles du spectre autistique, via d'une part l'identification à la fois de nouvelles mutations et de nouveaux gènes, mais également via l'étude des conséquences moléculaires et cellulaires de certaines mutations, in vitro et in vivo.

Les perspectives de ce travail seront d'étudier également d'autres mécanismes génétiques impliqués dans la DI et les TSA. En effet, si les études de séquençage ciblé et d'exome conduisent à l'identification d'une mutation causale dans une portion importante de patients, qui pourrait aller jusqu'à 40% (Gilissen et al., 2014), un nombre non négligeables de patients (et plus encore pour les TSA) reste sans mutation identifiée après ces analyses. Certaines formes de DI sont causées par des mutations localisées dans les introns ou les promoteurs des gènes pouvant avoir des conséquences sur l'expression (épissage, niveau d'expression) des ARNm correspondants. En fait, on a entrepris d'identifier ces mutations en utilisant deux technologies haut-débit, le séquençage d'ARN et le séquençage de génome entier, chez des patients avec DI ou TSA et pour lesquels aucune mutation n'a pu être mise en évidence par WES.

Mes travaux de doctorat permettront d'améliorer et d'accélérer la possibilité d'obtenir un diagnostic moléculaire qui donnera accès à un meilleur suivi et à une meilleure prise en charge pour les patients. Cela permettra également de mieux comprendre les mécanismes physiopathologiques impliqués dans ces troubles neurodéveloppementaux. Ces connaissances aideront à identifier de nouvelles cibles thérapeutiques.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

aCGH: array Comparative Genomic Hybridization
ACMG: American College of Medical Genetic and Genomics
ADHD: Attention Deficit Hyperactivity Disorder
ADID: Autosomal Dominant Intellectual Disability
ARID: Autosomal Recessive Intellectual Disability
ASD: Austism Spectrum Disorder
BBS: Bardet-Biedl Syndrome
bp: base pair
CDG: Congenital Disorder of Glycosylation
cDNA: complementary DNA
CDS: Coding Sequence
ChIP: Chromatin ImmunoPrecipitation
CLIP: Cross-Link ImmunoPrecipitation
CNV: Copy Number Variations
CpG: 5'-C-phosphate-G-3'
ddNTPs: dideoxyribonucleotides
Decipher: DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources
DEG: Differentially Expressed Genes
DNA: Deoxyribonucleic acid
dsDNA: Double Strand Deoxyribonucleic acid
DSM-V: Diagnostic and Statistical Manual of Mental Disorder 5th edition
EJC: Exon-Jucntion Complex
ENCODE: Encyclopedia of DNA Elements
eQTL: Expression Quantitative Loci
ESE: Exonic Splicing Enhancer
ESS: Exonic Splicing Inhibitor
EVS: Exome Variant Server
ExAC: Exome Aggregation Consortium
gDNA: Genomic DNA
GnomAD: Genome Aggregation Database
GoF: Gain-of-funcion
GTEx: Genotype-Tissue Expression
GUS: Gene of Uncertain Significance
HAT: Histone Acetyl Transferase
hNSC: human Neuronal Stem Cell
HPO: Human Phenotype Ontology
ID: Intellectual Disability
IGV: Integrative Genomic Viewer
Indel: Small Insertion or Deletion
iPSC: Induced Pluripotent Stem Cell
IQ: Intelligent Quotient
ISE: Intronic Splicing Enhancer
ISS: Intronic Splicing Silencer
Kb: Kilobase
LoF: Loss-of-function
Mb: Megabase
miRNA: Micro Ribonucleic Acid
MRI: Magnetic Resonance Imaging
mRNA: Messanger Ribonucleic acid
NAHR: Non-Allelic Homologous Recombination
NGS: Next-Generation Sequencing
NHEJ: Non-Homologous End Joining
NMD: Nonsense Mediated Decay
OMIM: Online Mendelian Inheritance in Man
PCR: Polymerase Chain Reaction
PRC1: Polycomb Repressive Complex 1
PSD: Post-Synaptic Density
PTC: Premature Termination Codon
qPCR: Quantitative Polymerase Chain Reaction
RBP: RNA-Binding Protein
RISC: RNA-Induced Silencing Complex
RNA: Ribonucleic Acid
RPKM: Reads per Kilobase Million
rRNA: Ribosomal Ribonucleic Acid
RT: Reverse Transcription
RTT: Rett syndrome
siRNA: Short Interfering Ribonucleic Acid
SNP: Single Nucleotide Polymorphisms
SNV: Single Nucleotide Variant
TAD: Topologically Associated Domain
tRNA: Transfer Ribonucleic Acid
TS: Targeted Sequencing
TSS: Transcription Starting Site
UTR: Untranslated Region
VUS: Variant of Uncertain Significance
WES: Whole-Exome Sequence
WGS: Whole-Genome Sequence
XLID: X-Linked Intellectual Disability

INTRODUCTION

1. DEFINITION OF INTELLECTUAL DISABILITY

Intellectual disability, previously named *mental retardation*, is one of the most common neurodevelopmental disorder that affects about 1% of the worldwide population (Maulik et al., 2011), representing one of the major public health-care and social problem, even though this percentage may vary according to the socioeconomic status and geographical regions. It has been observed a higher prevalence of affected males, with an estimated sex-ratio of 1.3 males/ females (McLaren and Bryson, 1987).

Several definitions have been proposed for ID. According to the AAIDD (American Association on Intellectual and Developmental Disability), ID is characterized by limitations in intellectual functioning as well as in adaptive behaviour, which includes everyday social, conceptual and practical skills, starting before the age of 18 years (www.aaidd.org). Similarly, the World Health Organization defines ID as "*a significantly reduced ability to understand new or complex information and to learn and apply new skills (impaired intelligence), resulting in a reduced ability to cope independently (impaired social functioning), and begins before adulthood, with a lasting effect on development*" (http://www.euro.who.int). In the DSM-V, ID is renamed as Intellectual Development Disorder and it is defined as a disorder that affects intellectual and adaptive functioning in conceptual, social and practical domains, with onset during the developmental period. The diagnostic criteria of ID have been revised in the DSM-V to accentuate the importance of both the clinical assessment and the use of standardized intelligence tests, which give a numerical output known as Intelligence Quotient (IQ).

ID is described to be two standard deviation or more below the average IQ score of the population, which is considered at 100. Therefore, an IQ score equal to or below 70 is classified as ID (Figure 1).

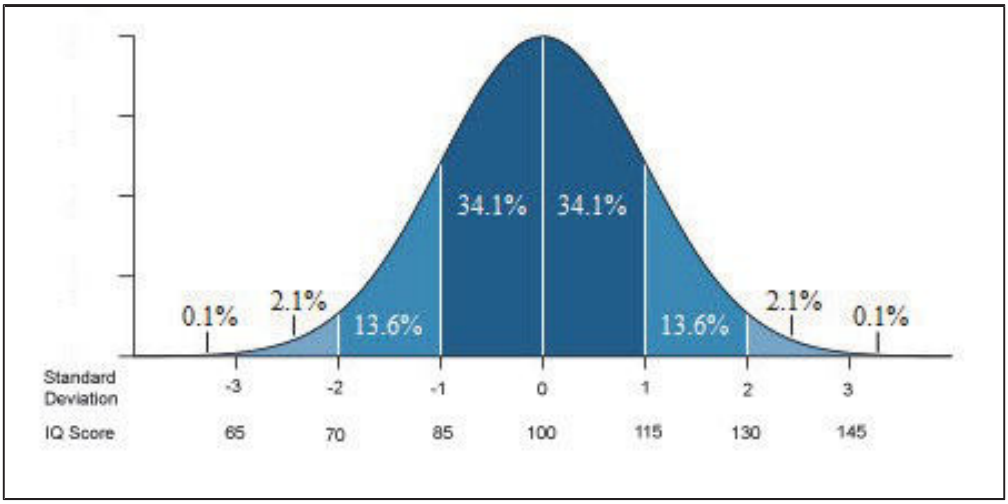


Figure 1: IQ distribution curve in the general population

Several IQ tests exist and they have been standardized in order to compare different individuals of the same age from the same population. One of the most common is the Wechsler test, which has an

adapted version for children ranging from 6 to 16 years - the Wechsler Intelligence Scale for Children (WISC) - and one for adults - the Wechsler Adult Intelligence Scale (WAIS). Both tests return a general IQ score based on four indexes: the verbal comprehension, the perceptual reasoning, the working memory and the processing speed.

According to the IQ score it is possible to classify the degree of severity of ID into 4 main categories: mild, moderate, severe and profound.

- MILD ID (50 < IQ score < 70): this group includes the majority of ID cases (~85%). Individuals are self-sufficient and have developed good communication skills. They can access education although with special needs.
- MODERATE ID (35 < IQ score < 50): it concerns 10% of ID individuals. They are able to communicate, even if the language could be impaired. They may need help for several activities of daily living.
- SEVERE ID (20 < IQ score < 35): it is present in 3-4% of ID patients. Their communication abilities are limited and they may also present a motor delay. They need assistance for everyday routine.
- PRODOUND ID (20 < IQ score): it is the less frequent ID as it affects 1-2% of ID individuals. Patients are not autonomous and they require permanent assistance for daily-life activities. The language is often absent or limited to few words.

In addition to the intellectual functioning, also the adaptive behaviour is tested, by evaluating different domains among which communication, daily living skills, socialization and motor skills. The Vineland Adaptive Behaviour Scales (VABS) is one of the most regular test.

Despite these common guidelines for the evaluation and classification of ID, most of the time these tests are not performed, mainly because they take a long time and require a special consulting, which are not always available. Therefore, the diagnosis and the evaluation of the degree of ID is often done by clinical geneticists or by child neurologists, based on the global skills presented by the patient during the visit.

2. COMORBIDITIES OF ID

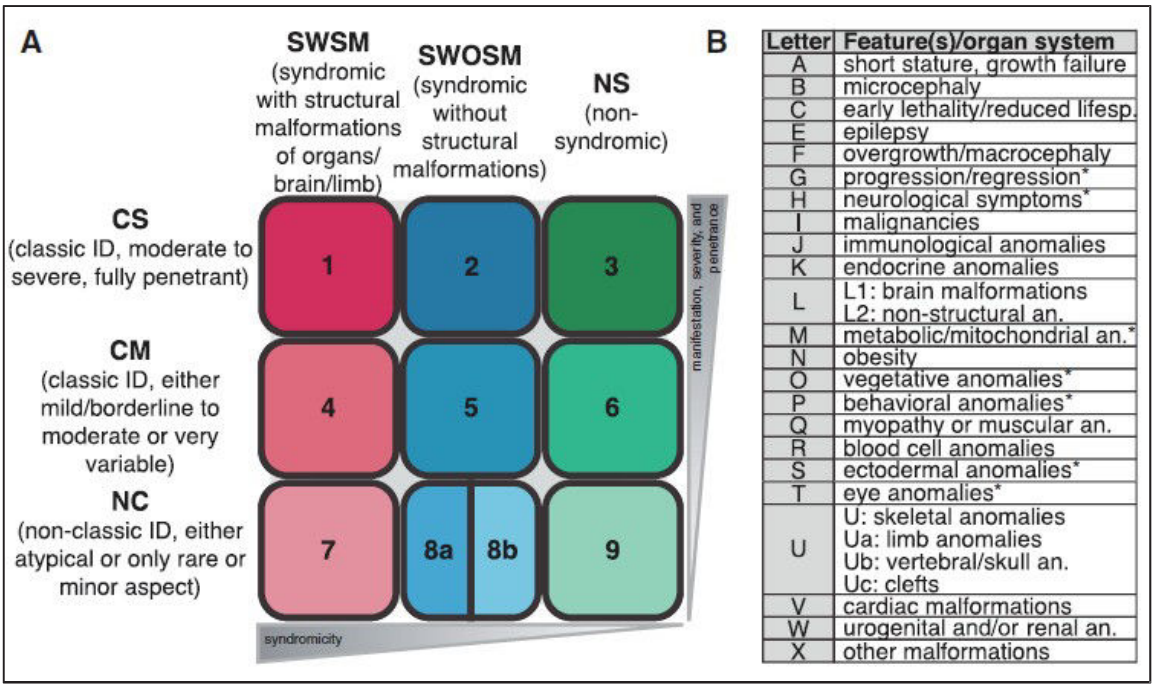
The phenotype of patients affected by ID is extremely heterogeneous, not only because of the severity degree, but also for the occurrence of other symptoms. ID can be present alone - *i.e.* there are no other clinical features - and this is classified as *isolated ID*. Isolated ID is defined by the sole presence of ID, without other clinical features. This condition is extremely rare as it is difficult to rule out the presence of less apparent anomalies, such as neurological or psychiatric ones.

Co-morbid traits		Frequency
Neurological features	Epilepsy	22%
	Balance disorder	20%
Neuropsychiatric features	ASD	24-30%
	Anxiety disorder	17%
	ADHD	10%
Malformations	Cerebellar	20-30%
	Musculo-skeletal	4-8%
	Cardiac	4-6%
	Urogenital	2-3%
	Gastrointestinal	2-4%

Table 1: Frequency of co-morbid traits in ID patients (adapted from (Petterson et al., 2007))

Conversely, ID is most of the time associated with other symptoms, and this is referred to as *syndromic ID*. Patients with syndromic ID have several additional clinical features, most of them affecting the central nervous system, ranging from neurological to neuropsychiatric ones, but other organs might also be affected (Table 1).

The presence of such distinctive symptoms might be helpful for the diagnosis, since an alteration of a specific gene may be associated with an association of different specific clinical manifestations. For this purpose, the use of a standardize vocabulary of phenotypic abnormalities, as proposed by the Human Phenotype Ontology (HPO), is extremely valuable for the better delineation and identification of a specific syndromic ID. Similarly, also the presence of some peculiar facial traits (*facial dysmorphisms*) is particularly important for the diagnosis. As a matter of fact, these peculiar traits represent one of the most specific clinical criteria of the syndrome itself (*e.g.* Down, Kabuki, Noonan syndrome, …).



Letter	Feature(s)/organ system
A	short stature, growth failure
B	microcephaly
C	early lethality/reduced lifesp.
E	epilepsy
F	overgrowth/macrocephaly
G	progression/regression*
H	neurological symptoms*
I	malignancies
J	immunological anomalies
K	endocrine anomalies
L	L1: brain malformations L2: non-structural an.
M	metabolic/mitochondrial an.*
N	obesity
O	vegetative anomalies*
P	behavioral anomalies*
Q	myopathy or muscular an.
R	blood cell anomalies
S	ectodermal anomalies*
T	eye anomalies*
U	U: skeletal anomalies Ua: limb anomalies Ub: vertebral/skull an. Uc: clefts
V	cardiac malformations
W	urogenital and/or renal an.
X	other malformations

Figure 2: Bipartite clinical monogenic ID-classification (Kochinke et al., 2016)

It has been proposed a clinical classification of the monogenic forms of ID according to two main aspect: the severity, clinical manifestation and penetrance on one side and the presence or not of other symptoms (referred to as "*syndromicity*") on the other one (Kochinke et al., 2016). The different ID disorders are grouped into 9 main classes and this gives a practical amount of information (Figure 2). Clinical comorbidities are listed in 27 different categories and they have been linked to the syndromic ID only when the reported frequency was associated to at least 20-30% of the patients (Kochinke et al., 2016). This classification was used to cluster phenotypically similar groups of ID to detect which ID symptoms co-occur the most frequently. This analysis demonstrated also that some specific clinical traits accompanying ID are more significantly associated to alterations in a specific molecular process than others (Kochinke et al., 2016). For example, many ID-genes implicated in a syndromic ID with behavioural anomalies code for proteins with synaptic function.

However, it is not unusual to observe a plethora of different phenotypes in patients with the same disrupted gene or even with the same mutation in one family. The high comorbidity between ID and other neurodevelopmental disorder (*i.e.* Autism Spectrum Disorder) suggests common molecular pathways, which will be discussed in another section (*Genetic Overlap between Neurodevelopmental Disorders,* pg. 32).

3. ETIOLOGY OF ID

The causes of ID are various and variable, and many are still not yet identified, hindering the process of diagnosis. Nevertheless, the causes of ID can be classified into two main groups: the environmental and the genetic factors, which can also overlap.

Both factors might alter one of the neurodevelopmental stages during the pre-, peri- and post-natal period, as depicted in Figure 3. Indeed, brain development is a tightly regulated process that depends on the sequential coordination of proliferation of neuronal precursors, migration and differentiation into neurons, establishment/pruning of synaptic contacts, etc. The myelination is also essential for the correct development and functioning of the brain. The alteration of one of these steps might lead to brain dysfunction, sometimes associated with brain malformations that can be detected with neuroimaging technique, such as the Magnetic Resonance Imaging (MRI). Also environmental factors may affect these crucial mechanisms or even modulate the genetic effects and should not be underestimated.

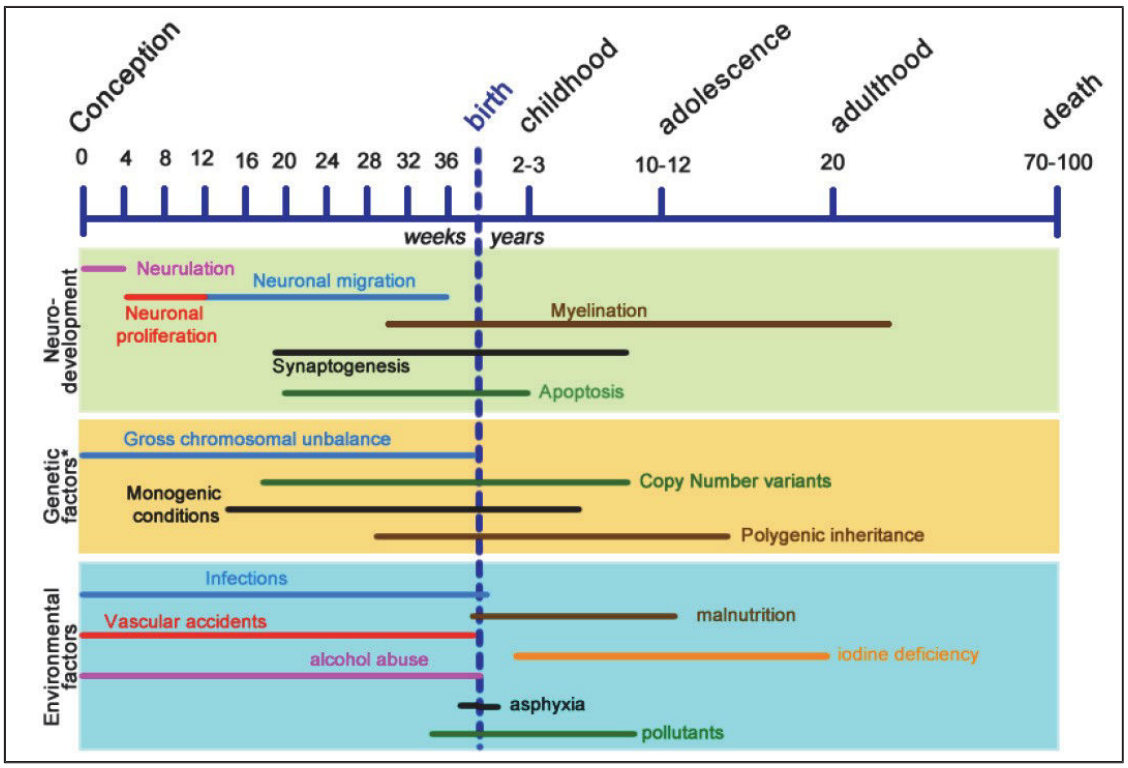


Figure 3: Environmental and genetic factors causing ID over neurodevelopmental stages *(Chiurazzi and Pirozzi, 2016)*

3.1 ENVIROMENTAL FACTORS

It is difficult to calculate the frequency of ID caused by environmental factors since it is not always possible to ascertain the association of some factors, therefore there is no available and reliable diagnostic test. The occurrence of some ID varies according to the socio-cultural background, as some of these causes are linked to the maternal lifestyle as well as to the health-care quality. Moreover, many external risk factors are not yet identified.

Among the known environmental factors of ID, it is possible to distinguish three main groups: toxic, infectious and traumatic causes.

- The toxic causes are related to the exposure to certain chemical compounds during the pre- and post-natal period. One of the most frequent toxic cause is the Foetal Alcohol Syndrome (FAS). Other toxic causes include exposure to drugs (*e.g.* amphetamine, opioids), lead and pollutants. Also maternal metabolic disorders are included in this category.
- The infectious causes are due to exposure to infectious disease during the pre- and post-natal period, such as cytomegalovirus infection during pregnancy and post-natal meningitis.
- The traumatic causes of ID are linked to a physical lesion of the brain (*e.g.* vascular accidents), lack of oxygen (*i.e.* asphyxia) as well as lack of nutrients necessary for the proper brain functioning.

3.2 GENETIC CAUSES

The genetic causes of ID accounts for a large amount of ID cases. In most of the cases a unique genetic event is responsible for the pathology. These genetic anomalies range from whole-chromosome alterations, deletions or duplication of one or several genes, copy number variant, to single nucleotide substitutions in single gene.

3.2.1 CHROMOSOMAL ABNORMALITIES

Chromosomal anomalies are the most frequent cause of ID, therefore karyotype was, during a long time, one of the first diagnostic test to be performed. According to the nature of the anomalies they can be subdivided into different categories: the aneuploidy (abnormal number of chromosome, supernumerary or missing chromosome), large balanced and unbalanced structural variations (translocation, inversion duplication and deletion larger than 5Mb). Chromosomal abnormalities have been implicated in ID since long time, starting with the identification of the trisomy 21 as the cause of the Down syndrome (Lejeune et al., 1959), which is still one of the most frequent cause of ID (Rauch et al., 2006). Initially, due to technique limitations, it was possible to detect only chromosomally abnormal aneuploidies. Beside the trisomy 21, there are few chromosomal aneuploidies since only few of them are able to survive to term and a small number of them are implicated in ID (beside the trisomy 21, trisomy 18 and 13) (Regan and Willatt, 2010). As the molecular cytogenetic technologies advanced - including high resolution karyotyping and fluorescent in situ hybridization (FISH) – it enabled the identification of other cytogenetic rearrangements associated to ID, such as balanced or unbalanced translocation and inversion, deletions and duplications larger than 5Mb. However, they are usually private event, identified in only one patient.

3.2.2 COPY NUMBER VARIATIONS (CNVs)

This category comprises deletions, insertions, duplications and also complex multi-site variants that lead to an imbalanced genetic dosage. They mainly originate from a *non-allelic homologous recombination* (NAHR) or a *non-homologous end joining* (NHEJ) events. The augmented use of array

comparative genomic hybridization (aCGH) improved the discovery and the identification of such submicroscopic rearrangements ranging from kb to Mb.

It has been estimated that CNVs account for about 12% of the human genome in the general population and they encompass many genes, disease loci, functional elements and segmental duplication (Redon et al., 2006), therefore it is complicated to evaluate the CNV contribution to ID. Nevertheless, it has been reported a general enrichment of CNVs in ID cases compared to non-affected individuals, particularly CNVs larger than 400 kb and in patients with a malformation (Cooper et al., 2011). Furthermore, deletions and duplications of dosage-sensitive regions are thought to be responsible for many clinical phenotype observed in genomic disorders (Lupski, 1998).

According to their frequency, CNVs can be subdivided into two groups: the *recurrent* and the *non-recurrent* ones. The non-recurrent CNVs differ in size and they are usually found in a single person. They mainly originate from a NHEJ events, thus they usually have a different breakpoint, which explains the observed differences. Conversely, the recurrent ones are found in multiple individuals, they have the same size and share the same breakpoint of occurrence, since almost all of them arise from a recurrent NAHR rearrangement in the same low-copy repeats genomic regions. For example, in chromosome 7 - involved in the Williams-Beuren syndrome - the region is flanked by highly homologous clusters of genes and pseudogenes, predisposing to a misalignment during meiosis leading to unbalanced recombination. For instance, 98% of patients affected by this microdeletion syndrome have breakpoints occurring in medial and centromeric duplicons, leading to a deletion of approximately 1.5 millions bp (Pober, 2010).

In both cases, patients may have a similar syndromic ID and sharing a minimal region (called *critical region*) encompassing few or only one gene, narrowing down the potential ID candidate genes. It is therefore challenging to dissect these critical regions to identify the gene(s) responsible for the main phenotype as well as the gene(s) that modulate the expressivity of the related clinical features. However, it remains still difficult to predict the pathogenicity of critical regions encompassing only non-coding regions; such variants may have a direct impact on the gene expression regulation (*e.g.* distal promoter and enhancers), the chromatin conformation or in the genome's architecture, resulting in regulatory changes which could lead to alterations of gene expression (Haraksingh and Snyder, 2013; Lupiáñez et al., 2016).

3.2.3 MONOGENIC FORMS OF ID

Monogenic forms of ID are caused by mutations affecting a single gene, which include single nucleotide variants (*SNVs*) and small insertions or deletions (*indels*). They may be classified according to their inheritance mode, mainly: dominant or recessive X-linked, autosomal-dominant or autosomal-recessive.

At the beginning, the research of these genetic causes of ID have been hampered by technical limitations as well as by a high clinical and genetic heterogeneity. Mutations were first identified in recognizable ID syndromes in multiple patients or in large families. However, the investigation of non-syndromic or non-specific ID was more complicated, since they could not be considered as an unique group, precluding the overlap of interfamilial mapping data (van Bokhoven, 2011).

The introduction of the next generation sequencing (NGS) technologies simplified the disease-gene identification. As a matter of fact, there has been a notable increment in the number of identified genes over the years; first with the introduction of genomic microarrays and, then, with the NGS, as illustrated in Figure 4.

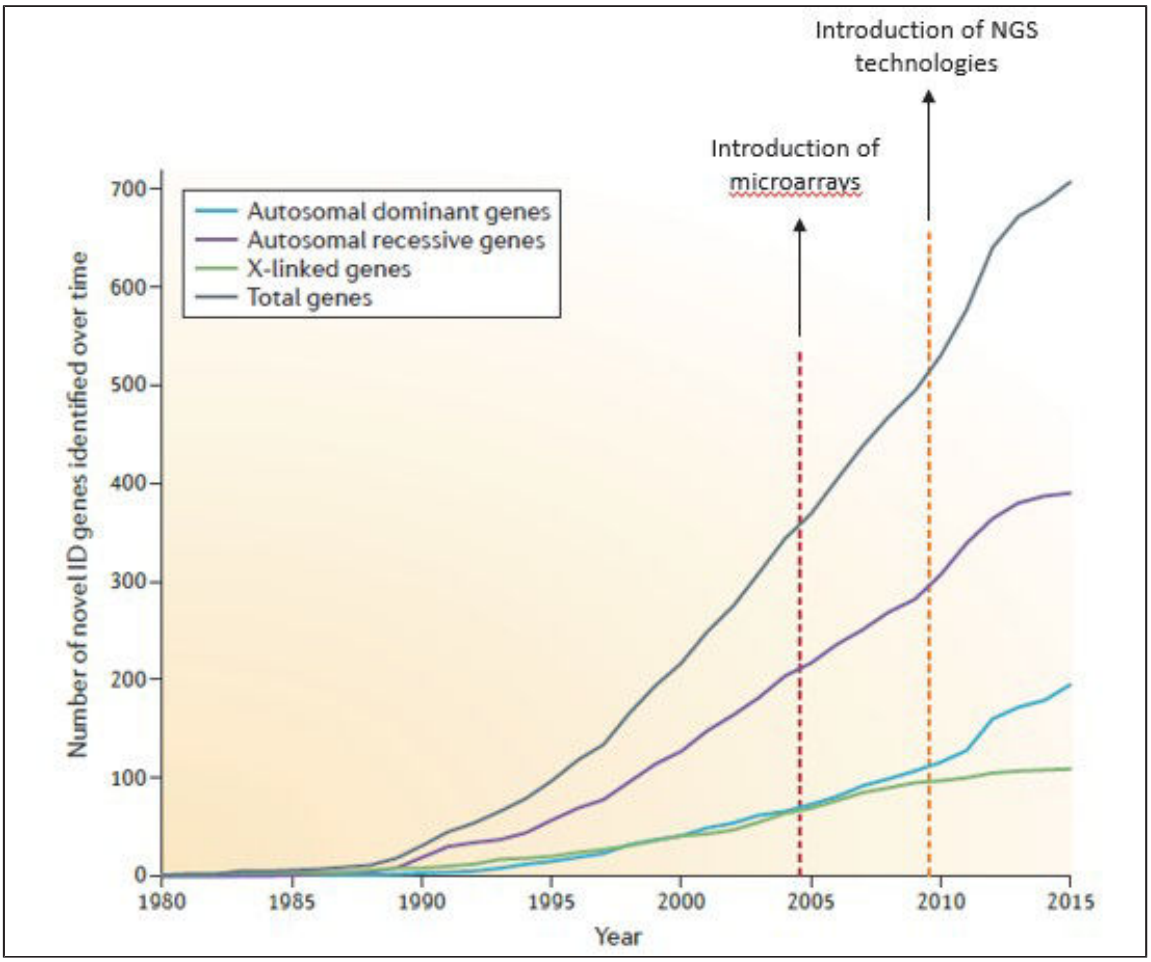


Figure 4: Identification of genes implicated in ID over the years, according to their inheritance mode (Vissers et al., 2016)

The identification of novel ID genes enabled a better understanding of the affected molecular pathways involved in ID, as to pinpoint common molecular pathways or biological processes, it is important to comprehend the role of the different proteins encoded by the ID genes. For example, a large portion of ID genes associated to cortical malformations are related to neurogenesis or neuronal migration.

Up to now, more than 700 genes have been implicated in monogenic forms of ID, across different studies of X-linked, autosomal-dominant and autosomal-recessive ID. Despite this high number, the curve of the number of the novel ID genes did not yet reach a saturation, indicating that there are still genes that remain to be identified.

3.2.3.1 X-LINKED INTELLECTUAL DISABILITY (XLID)

Due to the sex bias observed in ID patients (1.3-1.4 affected males for 1 female) and to the identification of large families with an evocative X-linked segregation, the X-chromosome have been deeply investigated to identify causative ID genes. Over the years, the intensive studies on the X-chromosome of several groups with international consortiums from all over the world have led to the identification of more than 100 genes implicated in XLID (Lubs et al., 2012).

The first implication of the X-chromosome in ID was the observation of a chromatid break in the extremity of the long arm of the X-chromosome in two brothers affected by ID, in 1966 (Lubs, 1969). This cytogenetic marker was recurrently observed in children with a similar syndromic ID, which consisted of a variable degree of ID, speech delay, peculiar facial dysmorphisms (large ears, long face and prominent jaw), macroorchidism and behavioural disorder (among which autism), that has been later named as Fragile X syndrome. *FMR1*, the gene responsible for this syndrome, was identified years later and it encodes the RNA-binding protein FMRP (Oberlé et al., 1991). The most common mutation of this gene is a trinucleotide expansion (> 200 repeats) at the 5'UTR region that leads to the transcriptional silencing of *FMR1*. This results in a drastic reduction of the encoded protein FMRP, affecting the regulation of downstream mRNA targets involved in synaptic structure and function. The fragile X syndrome is still the most frequent monogenic cause of ID, representing ~1% of the total ID cases (Coffee et al., 2009).

After the development of the cytogenetic methods, different strategies have been used for the detection of XLID genes, such as familial linkage analysis or translocation studies, followed by the sequencing of genes located in the linkage region or at the breakpoint. These approaches enabled the identification of many XLID genes, but the advent of NGS drastically increased these investigations; the sequencing of the entire X-chromosome coding regions (X-exome) and the whole-exome sequencing (WES) led to the identification of about a fifth of all XLID genes (Hu et al., 2016; Lubs et al., 2012; Tarpey et al., 2009).

Overall XLID contributes to ~10% of ID in males, but that alone does not explain the male excess in ID. Conversely to previous suppositions, mutations on the X-chromosome play an important role also in female ID patients. First, in several families with a XL recessive ID, certain female carriers may present a mild phenotype. X-skewed inactivation is more frequently observed in family with pathogenic variants (Tzschach et al., 2015), hence X-inactivation analysis in patients' mothers may support a

suspect of an X-linked defect and also help for deciding the further genetic tests. Still random X-inactivation does not exclude *a priori* a XLID.

It exists some typical ID syndromes specific to females and caused by pathogenic variants transmitted by a male carrier (*e.g. PCDH19*). NGS studies in sporadic females affected by ID showed a consistent number of *de novo* mutations in the X-chromosome both in novel ID genes, such as *DDX3X* and *NAA10* (Popp et al., 2015; Snijders Blok et al., 2015), but also in genes previously classified as recessive XL (Alexander-Bloch et al., 2016; Redin et al., 2014). Similarly, also mutations thought to be dominant XL specific to females might finally be viable and responsible for ID in males. For example, mutations in *MECP2* - a gene implicated in the Rett syndrome, which affects mostly females - have been long thought to be lethal in males (Zeev et al., 2002); yet pathogenic variants have been eventually identified in boys, even though they present a distinct phenotype (Couvert et al., 2001; Meloni et al., 2000).

3.2.3.2 AUTOSOMAL RECESSIVE ID (ARID)

Many metabolic diseases can be manifested by alterations in different organs, often including ID. Most of them are transmitted according to an autosomal recessive mode. For example, genetic defects in the synthesis of glycoproteins (*e.g. PMM2* and *ALG8*) lead to congenital disorders of glycosylation (CDG), a genetically heterogeneous group of metabolic disorders causing a severe multisystem disorder in the neonatal period. These diseases are caused by an enzymatic deficit in a metabolic pathway involved in the degradation or in the synthesis of a specific organic compound and consequentially resulting or in an accumulation of a toxic molecule or in the absence of a necessary compound. Metabolic diseases have been assessed to have a prevalence of about 1% in European population, but such disorders may have an even higher incidence in regions of the world with high consanguinity rate.

Beside the metabolic disorders, the molecular elucidation of ARID has lagged behind. The traditional strategies to map ARID genes include linkage mapping and homozygosity mapping in large size and consanguineous families, which are rare in Western countries where most of these studies are taking place. Furthermore, in outbred population ARID patients are usually sporadic cases and most of them are expected to be compound heterozygotes, thus carrying two different disease-causing alleles (Ten Kate et al., 2010). Conversely, in the so-called *consanguinity belt*, a geographical region that includes North Africa, Middle East and South East Asian countries, large size and consanguineous families are more common; consequently ARID is a relatively frequent cause of ID (Musante and Ropers, 2014).

A large study performed in 2011 comprised a high throughput targeted sequencing of coding exons from homozygosity and linkage regions in 136 consanguineous families (Najmabadi et al., 2011). The authors confirmed over 20 genes previously reported in ID and identified a single homozygous mutation in 50 novel candidate ARID genes, potentially explaining more than 50% of the ID cases.

These results also showed the extreme genetic heterogeneity of ARID. Still, for a large number of families, no gene defects were identified.

Since then, WES (and more recently even whole genome sequencing (WGS)) have been performed on large cohort of consanguineous families, detecting a high number of novel candidate ARID genes (*e.g. FMN2*, *CLIP1* and *SLC6A17*), indicating that the identification of the ARID genes is still at its infancy (Hu et al., 2018; Musante and Ropers, 2014; Riazuddin et al., 2016). However, even if the number of candidate genes is high, many of them have been found in just one family, therefore they have not been replicated and caution is need to interpret their possible implication in ID.

3.2.3.3 AUTOSOMAL DOMINANT ID (ADID)

The identification of ADID genes, compared to the other monogenic forms of ID, is relative novel and it has been developed in recent years. Previously, it was not so much investigated due to the unlikely transmission of a mutation from an affected parent to the offspring, because of the limited fitness associated to ID (especially in the severe forms of ID). Though there are mutations that are transmitted, they are indeed rare and they may present a variable penetrance and expression, ranging from a mild to a severe phenotype (with usually a more severe phenotype in the proband than in his affected parent). Therefore, initial research on ADID was limited.

As the major contribution of *de novo* CNVs to ID became known (de Vries et al., 2005), ADID has started to be more investigated. Indeed, *de novo* mutations could explain why these disorders are still so frequent in the general population despite the reduced fitness of affected individuals. Germline spontaneous mutations lead to an average of 50-100 novel variants in each newborn, with only few altering the protein-coding sequence and not necessarily resulting in a phenotype consequence.

To prove that *de novo* germline mutations are an important cause of ADID, first studies were performed by directly testing candidate genes that were either identified in breakpoints or critical regions of CNVs (*e.g. NSD1, TCF4*) or by directly sequencing genes known to be involved in important synaptic function (*e.g. SYNGAP1*) (Hamdan et al., 2009a, 2009b, 2010). However, this approach precluded an unbiased discovery of unexpected genes. The advent of WES and WGS allowed the identification of novel causative genes in several rare syndromes (Lupski et al., 2010; Ng et al., 2010a; Sobreira et al., 2010) and subsequently the better characterization of the impact and frequencies of *de novo* mutations. Additionally, these NGS technologies enabled a simultaneous comparison of the entire genome of both the parents and the progeny. The first trio-WES study was performed on 10 probands with unexplained ID and their unaffected parents. This family-based analysis pointed out 6 likely pathogenic non-synonymous *de novo* mutations all located in different genes, further demonstrating the high ID genetic heterogeneity. Among them, only two were in genes previously implicated in ID. This study strongly supported the hypothesis that *de novo* mutations are a major cause of sporadic ID (Vissers et al., 2010). Two years later, two independent trio WES studies on larger

cohorts (respectively, 51 and 100 patients) further confirmed the important contribution of *de novo* mutations in sporadic ID, which could explain about 13-35% of the ID cases (de Ligt et al., 2012; Rauch et al., 2012). Moreover, they revealed some recurrently genes that have more *de novo* mutations than others (*e.g. SYNGAP1*, *STXBP1*, *SCN2A*, *TCF4*).

Trio WES approach is now widely used in ID diagnosis, and an impressive number of WES have been performed in ID patients in different centres (*e.g.* GeneDx, Baylor College, Deciphering Developmental Disorders Study, etc.). This has led to the identification of numerous novel ID genes, as demonstrated by the number of publications per year (Figure 4). However, even if *de novo* mutations are indeed a major cause of ADID, a *de novo* variant is not necessarily deleterious, and the interpretation of the pathogenicity of each de novo variant has to be done with caution.

Recent studies on numerous large cohorts of patients affected by severe undiagnosed neurodevelopmental disorders are further disentangling the contribution of *de novo* mutations by also analysing their mechanisms of pathogenicity. Indeed, while some variants lead to a reduction or an absence of the encoded protein (the so-called *Loss of Function* (LoF) mutations, which are mainly caused by truncating variants (*i.e.* frameshifts, nonsense and splice-site), others alter its function (through a *gain-of-function* or a *dominant-negative* effect). It has been estimated that *de novo* mutations causing a severe neurodevelopmental disorders may be roughly split equally between loss of function and altered function (Deciphering Developmental Disorders Study, 2017). However, there is a huge discrepancy between the numbers of genes with truncating mutations identified compared to the genes with protein-altering mutations, probably due to the fact that truncating variants are easier to classify as pathogenic. The identification of genes with a substantial burden and clustering of missense mutations is an efficient strategy to overcome this issue, as proven by the identification of novel genes involved in neurodevelopmental disorders (Deciphering Developmental Disorders Study, 2015; Geisheker et al., 2017).

The identification of ADID with incomplete penetrance is still limited, as it is more difficult to identify and interpret these types of variants. Also somatic mutations could be missed as they are not easily detectable by these types of analysis. For instance, *de novo* mutations in a well-known gene implicated in Cornelia de Lange syndrome (*NIBPL*) were first missed in blood DNA analysis but later identified in DNA from buccal swamp in a relatively large portion of patients (10/44; 23%) (Huisman et al., 2013). We can speculate that somatic mutations in neuronal cells might cause ID. However, identification and studies on these types of variants are hampered by the difficulty of tissue sampling.

A more complicated inheritance form of ID difficult to detect is related to genomic imprinting, an epigenetic event in which only one parental allele is expressed, resulting in a monoallelic expression. Different mechanisms are used to imprint genes, including DNA methylation, antisense transcription

and histone modifications. Imprinted genes are often clustered in a genomic region and the imprinted expression may vary in different tissues. For example, in the 15q11-q13 region there are at least 14 imprinted genes, exclusively expressed by the paternal inherited allele in somatic tissue. However, two of these genes (*UBE3A* and *ATP10A*) have an imprinted expression of the maternal inherited allele in the brain (Chamberlain, 2013). These disorders are then caused by mutations of the parental allele physiologically expressed, either by CNVs, SNVs, epimutation or uniparental disomy. Moreover, deletions of the same locus or uniparental disomy of the same chromosome may lead to different syndromes. That is the case of the Prader-Willi and Angelman syndromes, in which the same genomic region 15q11.2 is disrupted, but while the first one is caused by the loss of the paternal allele, the latter one is caused by the disruption of the maternal one.

3.3 MORE COMPLEX FORMS OF ID

Despite the substantial progress obtained by NGS technologies in the understanding of the genetics of ID, an important portion of affected individuals remain without a molecular diagnosis. This could be explained by mutations in the non-coding region - for which our understanding is still limited – or a genetic scenario more complex than the monogenic one, with an *oligogenic* or *polygenic* model (which may also include environmental factors) that take in account the epistatic interaction among different genes and eventually environment. For instance, it has been showed that patients with 16p12.1 microdeletions have significantly more often a second large CNV compared to controls and some of

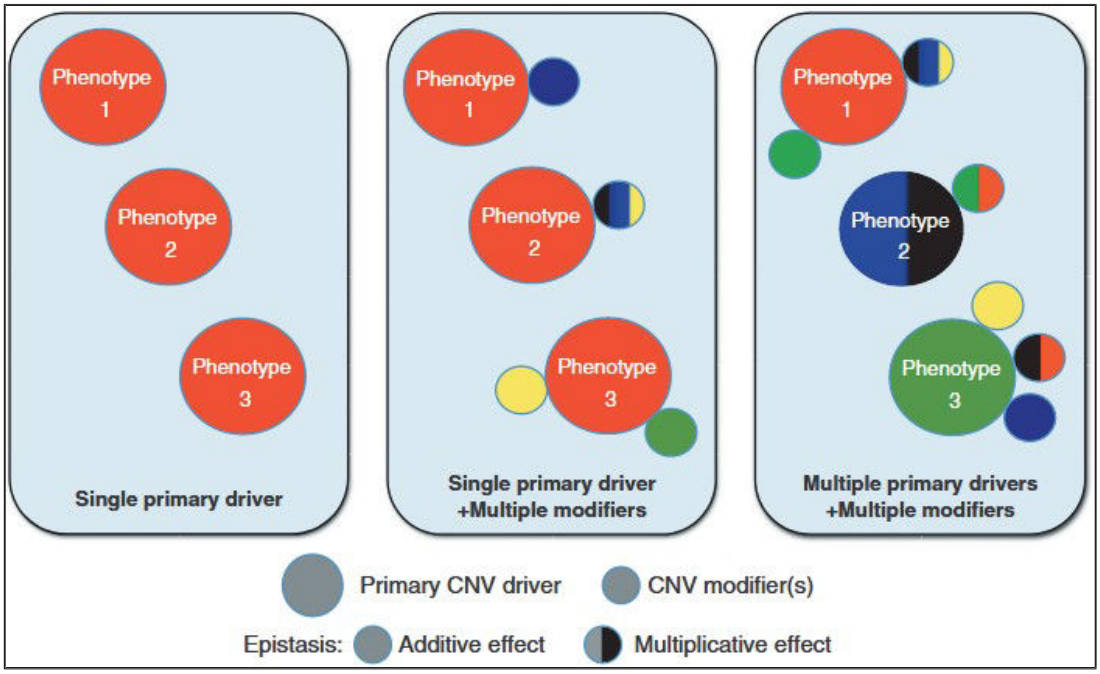


Figure 5: Genetic interaction model (adapted from Golzio and Katsanis, 2013)

them have a more severe phenotype, whereas carrier parents present more neurological or neuropsychiatry anomalies more often than the non-carrier parents (Girirajan et al., 2010).

While mutations with complete penetrance will be the major driver of one specific monogenic ID, there might be other mutations with lower penetrance and variable expressivity, whose effects could be modified by other variants in different genes (*oligogenic* and *polygenic* models), giving rise to a variable phenotype. For the latter case, two models are hypothesized. In the first one the alteration of a single variant is necessary and sufficient to cause ID but the interaction with modifiers loci (that could include also common variants) lead to a variable phenotype. Conversely, in the second model, the phenotype is established by an epistatic interaction of few or more mutations in different genes, which alone could cause some specific clinical traits. Their combined effect may be more severe or qualitatively different and the pattern of inheritance is subsequently more complex (Figure 5).

Studies on CNVs aiming to disentangle the contribution of each encompassed locus to the phenotype further demonstrated the complexity of these epistatic interaction, but little is known about SNVs, as most of the recent studies of large cohort of neurodevelopmental patients were mainly focused on monogenic ID.

4. GENETIC OVERLAP BETWEEN NEURODEVELOPMENTAL DISORDERS

The frequently observed co-morbidities between ID and other neurodevelopmental disorders (*i.e.* ASD, epilepsy and schizophrenia) likewise reflect the common affected molecular pathways and genetic factors among these diseases.

Few years ago, a study on recurrent rearrangements in three different categories of patients (ID, ASD and schizophrenia) reported a large number of recurrent CNVs that are not specific to a disease-category, indicating common affected molecular pathways (Guilmatre et al., 2009). Furthermore, large-scale trio sequencing studies showed an enrichment of *de novo* variants in a restricted number of genes across different neurodevelopmental disorders (Hoischen et al., 2014).

Among the genes involved in ID, some lead to various neurodevelopmental phenotypes. These differences could be explained by the different impact of the mutations (*e.g.* affecting different functional domain or LoF vs GoF mutations), stochastic processes during development and differences among individual genetic backgrounds. Nonetheless, ASD is more present in patients with mutations in some specific ID genes (*e.g. CHD8*, *GRIN2B*). In the same manner, a subset of ID genes seems to be more associated with epilepsy *(e.g. CHD2, SLC2A1*) as illustrated in Figure 6. Therefore, the better identification and characterization of genes more implicated in these specific neurodevelopmental phenotypes may help in the better understanding of the molecular pathways involved in these disorders.

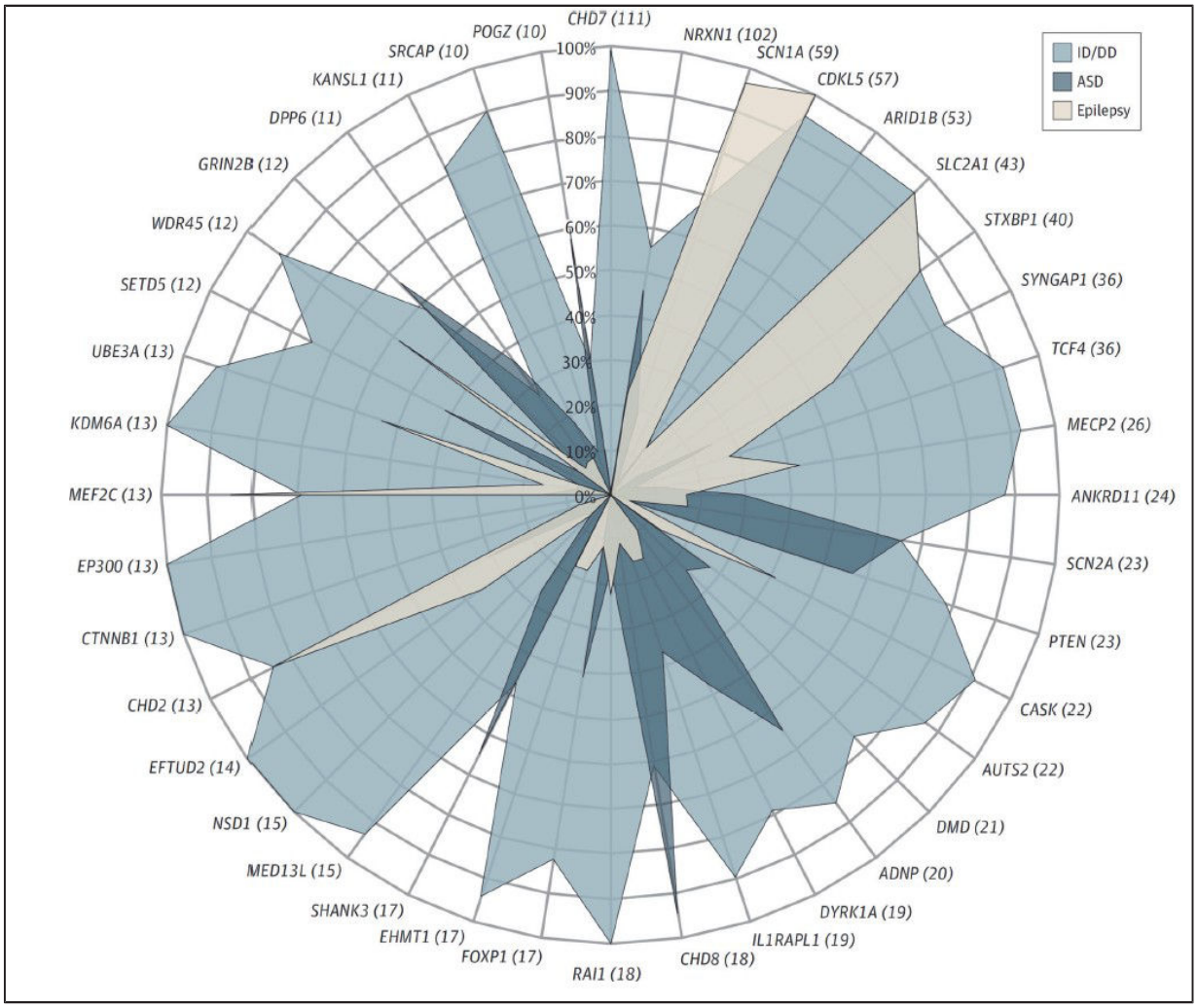


Figure 6: Frequency of ID, ASD and epilepsy associated to LoF mutations (Gonzalez-Mantilla et al., 2016)

ID and ASD have the largest overlap, with 17% of reported ID genes with *de novo* LoF mutations being also found in ASD (Vissers et al., 2016). However, the contribution of these *de novo* events seems to be less important in individuals with ASD with a higher IQ (>90), but it plays a major role in cases with syndromic ASD with ID (Iossifov et al., 2014). To date, all the genes implicated in ASD could also lead to ID with or without ASD.

The genetic model proposed for ASD is more complicated, since it takes into account common and rare variants as well as several other different environmental factors. For example, it has been proposed a model in which rare or *de novo* variants are differentially compensated by each individual genetic background, hence some persons will develop ASD while others will not (Hartman et al., 2001; Rutherford, 2000). As a matter of fact, recent studies reported that common variants are largely implicated in the risk of autism (Gaugler et al., 2014; Klei et al., 2012). In the last study, the authors reported an estimation of about 52.4% of heritability mostly due to common variants while only a 2.6%

of rare variants contribute to individual liability (Gaugler et al., 2014), underscoring the crucial contribution of common variants in ASD risk. However, the identification of such common variants is hindered by their high number and their related low impact risk.

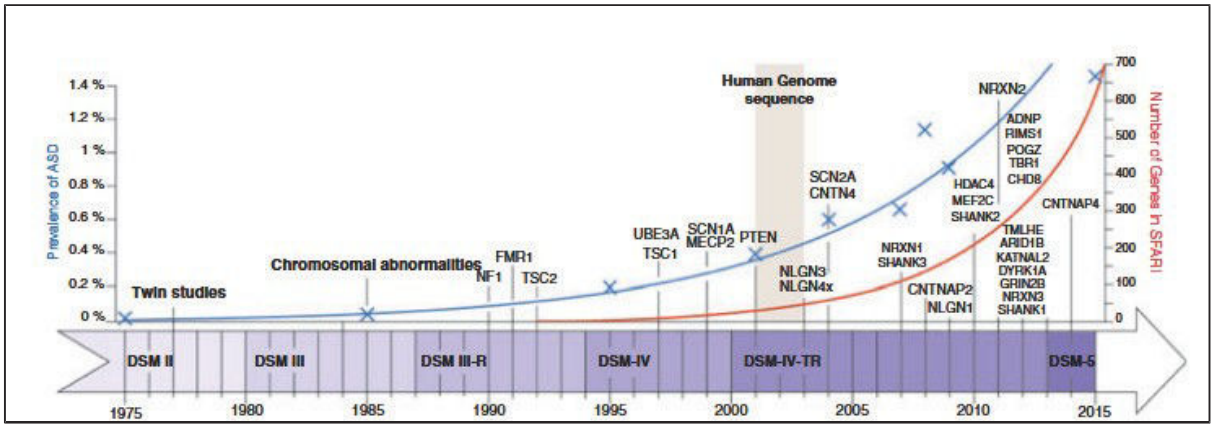


Figure 7: Timeline of ASD Genetics (Adapted from (Huguet et al., 2013))

The ASD heritability has been largely investigated since first evidences from twin and sibling studies indicated a large genetic contribution to ASD-risk. The development of NGS technologies prompted to further investigate on the genetic contribution in ASD at a genome-wide level (Figure 7). Initial studies revealed that individuals with ASD carry a higher number of CNVs compared to controls; furthermore, investigations on family cohorts comparing individuals with ASD to their parents and unaffected siblings showed that *de novo* CNVs are more present in ASD patients, thus increasing the risk of ASD in 5-10% of individuals (Huguet et al., 2013). The research of ASD-risk and candidate genes was further improved by the advent of exome and genome sequencing; more than 4000 families with at least one ASD child have been sequenced, leading to the identification of high-confidence ASD candidate genes. Monogenic forms of ASD have been described (*e.g. NLGN4*, *IL1RAPL1*) and are caused by more than 400 genes (Ronemus et al., 2014). These studies were mainly focused on *de novo* SNVs and they showed that between 3.6 – 8.8% of patients carry a *de novo* causative mutations (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012). Moreover, protein-interaction analyses based on genes implicated in ASD revealed recurrent molecular network comprises synaptogenesis, axon guidance, neuronal motility as well as chromatin remodeling (Gilman et al., 2011; O'Roak et al., 2012). Interestingly, the average mutation rate of individuals with ASD is significantly different from controls only if the analysis is restricted to brain developmental genes (Sanders et al., 2012). A meta-analysis study showed that *de novo* likely-LoF variants are more frequent in patients with ASD than their unaffected siblings (Iossifov et al., 2014).

Many individuals with ASD have been described with multiple mutations in different genes (or even inherited protective alleles), suggesting that even gene-gene interaction could eventually lead to the ASD phenotype (Ziats and Rennert, 2016).

5. MOLECULAR PATHWAYS INVOLVED IN ID

The identification of a large number of genes involved in ID significantly improved the understanding of the affected molecular pathways, crucial for the development of therapeutic targets. ID genes can be clustered into several functional modules according to different parameters, such as a common pathway, a direct physical interaction or co-expression. Enrichment analysis based on gene ontology terms may also be used in the identification of these functional networks.

These analyses revealed the presence of general disrupted molecular and cellular pathways, including neurogenesis, neuronal migration, synapse and gene expression regulation (van Bokhoven, 2011; Chelly et al., 2006; Kleefstra et al., 2014; Kochinke et al., 2016).

5.1 METABOLIC DISORDERS

Among the mutated genes implicated in ID 1-5% of them cause a metabolic disorder, therefore the ID is caused either by an accumulation of a toxic compound or by a lack of a substrate necessary for proper brain development, or by an energy deficit during a critical step of brain development. These disorders are usually diagnosed by proper biochemical testings. A well-known example is the gene *PAH*, which encodes an enzyme that catalyses the hydroxylation of phenylalanine to tyrosine, the rate-limiting step in phenylalanine catabolism. A deficiency of this enzyme causes phenylketonuria, which can be detected by a high ratio between the concentration of phenylalanine and tyrosine in blood.

5.2 SYNAPSE AND CYTOSKELETAL REGULATION AND ORGANIZATION

During brain development, neurons are going thorough different stages that are temporally and spatially tightly regulated. Genes encoding proteins involved in these steps have been reported as mutated in ID with specific associated clinical features, enabling their classification according to the affected neurodevelopmental step.

The neurogenesis starts from the neuroepithelial progenitors that divide to expand the progenitor pool and then giving rise to the intermediate progenitors, which will subsequently divide and give rise to neurons. Defects in progenitor proliferation have been associated to primary microcephaly, which consists in a reduction of the brain size due to a decreased number of neurons. Interestingly, the majority of the genes implicated in primary microcephaly code for centrosomal proteins (*i.e. ASPM*, *CENPJ*), which are important for proper chromosome segregation, showing an important relationship between cell division and neurogenesis (Barbelanne and Tsang, 2014).

For a correct cortical development, neurons migrate from the ventricular zone toward the cortical plate. This process is controlled by various players including cytoskeletal and proteins, for which

several mutations in genes coding for microtubule-associated proteins (*i.e. LIS1* and *DCX*) have been particularly reported in patients affected by cortical malformations (*i.e.* lissencephaly). Moreover, also motor proteins like kinesin, which are required both for cargo transport and for the generation of energy during structural rearrangements, have been associated to brain malformations (*e.g. KIF11, KIF5C*), along with tubulin subunits, highlighting an important role of cytoskeletal dynamics during neuronal migration.

Once neurons reached their proper position, a proper connection among them must be established. Many genes implicated in ID – and often also epilepsy and ASD - codes for synaptic molecules, which are involved in the structure or in the function of neurons, specifically in dendrites and synapses. The majority of the synapses in the nervous system involve chemical signals responding to stimuli (*i.e.* action potential) by releasing neurotransmitters. These neurotransmitters are synthetized by the presynaptic neurons and stored in synaptic vesicles. A critical step is the transfer of these vesicles to the so-called active zone, where vesicles fuse with the presynaptic membrane through a process of exocytosis to release their neurotransmitters in the synaptic cleft. Many genes encoding for pre-synaptic proteins involved in one of these processes have been associated to ID, with or without associated epilepsy. For example, many Rab proteins, which are GTPases regulating the migration and

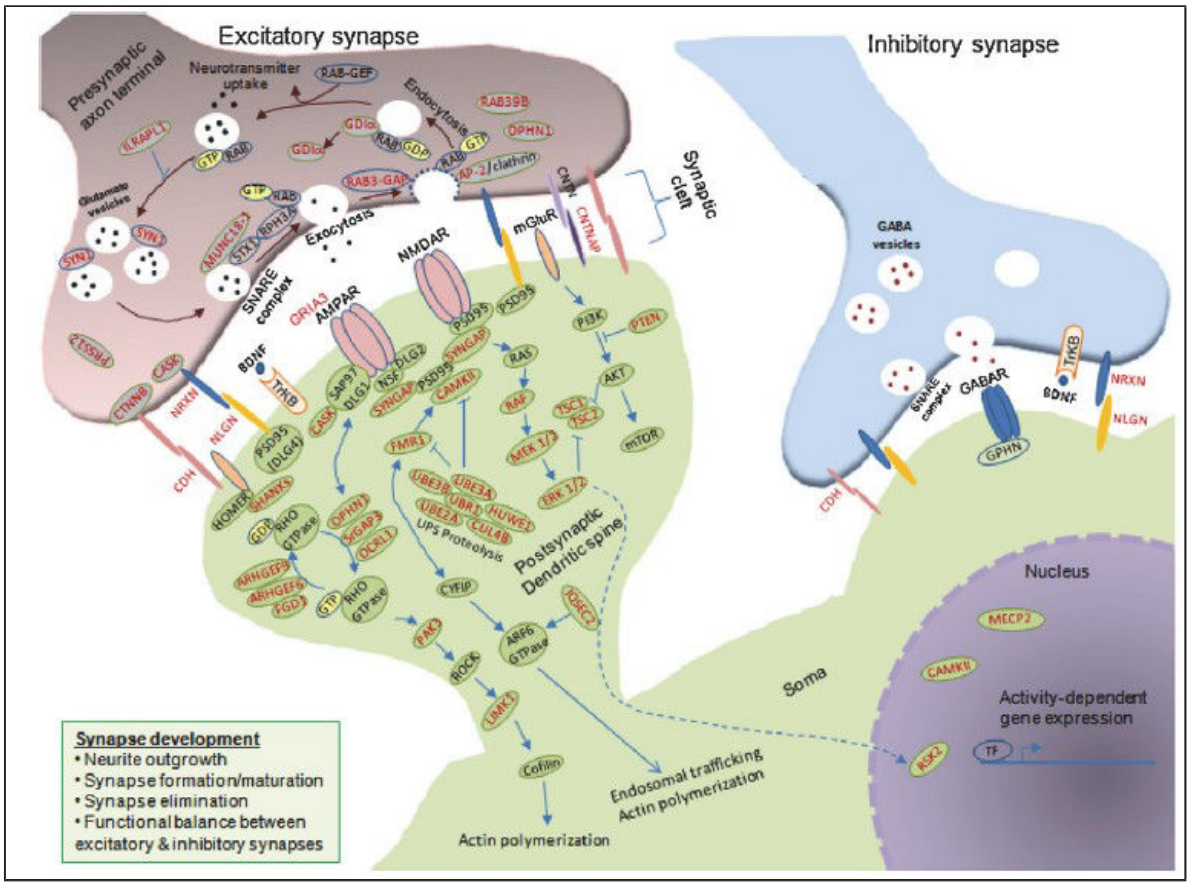


Figure 8: The complexity of the synaptic compartments. Genes implicated in ID or ASD are shown in red (Srivastava and Schwartz, 2014)

circulation of the neurotransmitter vesicles, have been described as mutated in patients with neurodevelopmental disorders (*e.g. RAB3GAB1, RAB39B*).

Cell-adhesion proteins are important for the formation of the contact between the pre- and post-synaptic compartments, which are divided by a gap named *synaptic cleft*. Most of these proteins belongs to the cadherin and integrin family, as well as neurexins and their binding partners, the neuroligins. Mutations in genes coding for these proteins have been identified in patients with ID and ASD (*i.e. NLGN3, NLGN4X*), pointing out the important role of the synaptic cell-adhesion pathways in cognitive and behavioural functions (Srivastava and Schwartz, 2014).

The post-synaptic membrane includes receptors and ionic channels crucial for the conversion of the chemical signal. ID can be the result of the disruption of the signal transduction both for excitatory (glutamatergic) and inhibitory (GABAergic) neurons. Some ID genes encode proteins located in the post-synaptic density (PSD) of glutamatergic synapses, such as glutamate receptors (GRIN2A/B, GRIN1, etc) or scaffolding proteins involved in the PSD architecture (SHANK3, SYNGAP1, DLG3, etc). Finally, the activation of the neurotransmitter receptors activates intracellular post-synaptic signalling, including the Ras-MAPK-ERK and PI3K-AKT-mTOR pathways. The disruption of genes involved in these signalling pathways is at the basis of some syndromic ID (*e.g. PTEN, BRAF*).

The regulation of the post-synaptic density is strictly associated to the synaptic plasticity, which is the ability to rapidly change structure and morphology in response to a stimulus. This ability is regulated on one side by protein degradation – including proteins involved in ubiquitination (*e.g. UBE3A, CUL4B*) - and on the other one by actin and microtubule polymerization/depolymerisation, which depends on the Rho-GTPase signalling pathways, for which mutations in many genes have been reported to cause ID (*e.g. ARHGEF6, PAK3*).

5.3 GENE EXPRESSION REGULATION: TRANSCRIPTIONAL REGULATION

Brain development requires a tight temporal and spatial gene expression regulation to control neuronal progenitor proliferation, migration, differentiation as well as synapse formation, elimination and plasticity. This regulation should also be dynamic, to rapidly respond and change to extra- and intra-cellular signalling. Gene expression regulation controls the presence and the production of specific gene products and it is an important mechanism for proper cell functioning (among the others). Therefore, most of the factors implicated in these molecular mechanisms are ubiquitously expressed, as they are involved in the regulation of gene expression in the whole body. It is thus common that mutations in genes coding for these factors give rise to a syndromic ID accompanied by other clinical manifestations affecting other systems.

Gene expression regulation could be achieved by a direct control on transcription or by post-translational mechanisms. In this section I will focus on the regulation at the transcription level.

The control of the transcription rate of gene expression programs is accomplished by many players, among which transcription factors and chromatin modifiers. They regulate the transcript production by modulating the recruitment and the activity of the RNA polymerase to specific DNA regions in an extremely coordinated fashion.

Transcription factors bind to specific targeted DNA sequences and facilitate or inhibit the recruitment of the RNA polymerase toward a gene. A large number of them are found mutated in ID (Table 2), such as *FOXP1*, *TCF4*, etc.

The function of transcription factors is also controlled by the distribution of their binding-sites in the genome (Chen et al., 2017), which is regulated by epigenetic modifiers.

Transcription Factors	
ARX, ASCL1, CC2D1A, CTCF, DEAF1, EP300, FOXG1, FOXP1, FOXP2, GATAD2B, GLI2, GLI3, HESX1, HIVEP2, MAF, MEF2C, MYCN, MYT1L, NFIA, NRF21, PAX6, PAX8, RERE, SALL1, SIN3A, SIX3, SOX10, SOX11, SOX2, SOX3, SOX5, TAF1, TBP, TCF4, TBR1, TWIST1, ZBTB16, ZBTB18, ZNF81	
DNA Methylation	
DNMT1, DNMT3B, FTO	
Histone modification	
Writers	*CREBBP, CUL4B, EHMT1, EP300, EZH2, HLCS, HUWE1, KAT6B, KAT6A, KMT2A, KMT2D, KMT2C, NSD1, WHSC1, UBE2A*
Erasers	*HDAC4, HDAC8, KDM5C, KDM6A, PHF8*
Readers	*ASXL, BCOR, CHMP1, CTCF, GATAD2B, HCFC1, KANSL1, MBD5, PHF6, POGZ, SKI, MED12, MED17, MED23, NIPBL, RAD21, SALL1, SMC13A, SMC3*
ATP-dependent chromatin remodeler	*ACTB, ARID1A, ARID1B, ATRX, CHD2, CHD7, CHD4, CHD8, SMARCA2, SMARCA4, SMARCB1, SMARCE1, SRCAP, SS18L1*

Table 2: Main transcriptional regulators involved in ID (Adapted from (Kleefstra et al., 2014))

In recent years, many causative ID mutations have been identified in genes coding for proteins involved in chromatin-mediated control of transcription. A recent study reported a fold enrichment above 2 of chromatin-related genes, similar to the one obtained by synaptic-processes genes group, with around 10% of genes implicated in ID involved in epigenetic transcription regulation (68/650 ID-genes) (Kochinke et al., 2016). In our updated list, containing more than 800 genes implicated in ID (retrieved by different European lists such as SysID, Radboud UMC list and Genome England PanelApp), an even higher percentage of genes implicated in chromatin remodelling is identified (13.3%). Interestingly, patients with mutations in chromatin-related genes showed a significant enrichment of co-morbid traits such as clefts, cardiac problems, limb anomalies and short stature. Moreover, microcephaly and behaviour anomalies were found to co-occur more frequently in chromatin-related genes (Kochinke et al., 2016). Most of the mutations occurring in these genes are heterozygous loss of function, suggesting the importance of gene-dosage in chromatin regulation processes.

Chromatin regulation is important for gene regulation in all the tissues, but is really crucial during neurodevelopment, as it contributes to the dynamic changes required during neuron formation and they are also able to maintain cell fates by providing stable and heritable states of gene expression (Ronan et al., 2013). Different mechanisms have been reported to regulate the chromatin conformation, including DNA methylation, non-coding RNAs, regulation of nucleosome positioning and histone modifications.

5.3.1. DNA METHYLATION

DNA methylation is a well-known regulatory transcription mechanism; it consists in the methylation of DNA at cytosine residue of a CpG. DNA methylation is regulated by three different DNA methyltransferase (DNMT1, DNMT3A and DNMT3B). While two of them are able to methylate cytosine *ex novo*, only one (DNMT1) is responsible for DNA methylation maintenance, which can be inherited through DNA replication and cell division. DNA methylation is implicated in brain plasticity associated with memory and learning abilities. As a matter of fact, a DNA-methylation study revealed epigenomic changes at different development stages in brain, both in mice and human, showing the dynamic of this epigenetic mark (Lister et al., 2013). Autosomal dominant mutations in *DNMT1* can give rise to a cerebellar ataxia with deafness and narcolepsy (Winkelmann et al., 2012), while in *DNMT3A* cause an overgrowth syndrome with ID (Tatton-Brown et al., 2014). Recessive pathogenic variants in *DNMT3B* cause immunodeficiency-centromic instability-facial anomalies syndrome (Xu et al., 1999).

5.3.2 HISTONE MODIFIERS

Post-transcriptional modification of amino acids located in tails of histone proteins is a well-known mechanism that regulates chromatin compaction and therefore transcription. Different type of modifications can occur, such as acetylations, methylations, phosphorylations and ubiquitinations and they control gene expression by influencing the chromatin three-dimensional structure. The effort of several international collaboration studies on genome-wide histone modification profiles combined with transcriptomic analysis revealed a basic *histone code*, in which specific histone modifications are associated with different biological processes, among which repression or expression of specific regions at distinct time. Indeed, some histone modifications are more associated with a thigh nucleosome, hindering gene transcription, whereas other modifications relax the chromatin structure, facilitating the transcription. Post-translational histone modifications are regulated by different actors that could be summarized in four main categories according to their action: writers, erasers, readers and ATP-dependent chromatin remodelers. The coordinated activity of these four groups enables a dynamic regulation of the chromatin structure.

Writers are chromatin modifiers that directly add side groups to histone proteins. Among these, there are histone acetyltransferases (HAT). Lysine acetylation is usually associated to active transcription as it reduces the positive charge of lysine side chains, decreasing its interaction with the negatively charged backbone of the DNA. Mutations in *KAT6A* and *KAT6B* - two lysine acetyltransferases of the same complex - have been associated to two different syndromic IDs, emphasising the role of HAT in neurodevelopment. While HATs acetylate a variety of lysine on histone proteins, histone methyltransferases are more specific, as reflected by their high number. Histone methylation markers are more complex as they are associated to both transcription activation and repression and methylation can be mono or multiple, mainly in lysine and arginine residues. Interestingly, two methyltransferases with different target residues and leading to an opposed effect on transcription are implicated in a similar syndromic ID characterized by overgrowth (*NSD1* and *EZH2*). These apparent discordance can be explained by the different type of identified mutations; while in *NSD1* they give rise to a non-functional protein, mutant *EZH2* protein might have a potential gain of function effect (Tatton-Brown and Rahman, 2013).

Among histone modifications, there is also ubiquitination that results in a much larger modification, as the ubiquitin itself is big. According to its histone target it could lead to both transcription activation (H2B) and silencing (H2A) (Srivastava et al., 2017). For example, the polycomb repressive complex 1 (PRC1) monoubiquitinates H2A leading to transcription repression. It has been shown that when AUTS2 binds to the PRC1, it inhibits its repressive activity by recruiting the caseine kinase 2 (CK2), and activating gene transcription (Gao et al., 2014). Interestingly, disruptions of *AUTS2* have been associated to a syndromic ID (Beunders et al., 2013).

On the other hand, *erasers* chromatin modifiers remove post-transcriptional modifications from histone proteins, reversing writers' action, rendering histone markers dynamic transcriptional regulators. Among the erasers group, histone deacetylates (HDACs) reverse HATs' effect, so they compact the nucleosome and repress transcription. Different types of HDACs are usually associated together and they are commonly present together in multiple specific complexes. Mutations in at least three distinct HDAC (*HDAC4, HDAC6* and *HDAC8*) have been implicated in ID (Table 2). Conversely to HDAC, histone demethylases have been discovered only in 2004, as histone methylation was considered as a stable post-translation histone modification (Bannister and Kouzarides, 2011). Histone demethylases target specific methylated histone residues and they can act both as transcriptional repressor or activator. Some of them modulate their enzymatic activity according to the different protein-complexes they bind to, which confer nucleosomal recognition (*e.g. KDM1A*). Histone demethylase have been linked to neurodevelopmental disorders; for instance, a congenital ID characterized by peculiar facial dysmorphisms (Kabuki syndrome) is caused by mutations in *KDM6A*,

whose activity is strictly coordinated with the one of the histone methylase KMT2D and mutations in this gene give rise to the same syndrome.

Chromatin modifications are recognized by *readers*, which will then exert their function according to the histone marks. They can be chromatin remodelers, core components of transcriptional complexes and proteins that bridges chromatin remodelers with transcription factors. Among the large protein complexes that link transcription factor with chromatin remodelers, two have been implicated in neurodevelopmental disorders. For instance, multiple subunits of the cohesin complex (*e.g.* NIPBL, SMC1A, SMC3, and RAD21) are altered in Cornelia de Lange syndrome, hence giving rise to similar clinical features. The cohesin complex is important during cell division, as it maintains sister chromatids together from S-phase until mitosis or meiosis and it is also involved in the chromatin architecture, as it links distal chromatin segments.

ATP-dependent chromatin remodelers alter nucleosome positioning either by sliding it – hence moving it along the DNA – or by exchanging it on chromatin. These processes are powered by ATP hydrolysis and they clearly play a role in transcriptional regulation. Among the four families of ATP-dependent chromatin remodelling complexes, two of them have been largely implicated in ID and ASD: mutations in members of the SWI/SNF (BAF) complex (*e.g.* ARID1B, SMARCA2, etc) are causing the same syndromic form of ID, the Coffin-Siris syndrome, and mutations in CHD (Chromodomain-helicase-DNA-binding) proteins are involved in different neurodevelopmental syndromes (*e.g.* CHD7, CHD8, etc)

5.4 GENE EXPRESSION REGULATION: POST-TRANSCRIPTIONAL REGULATION

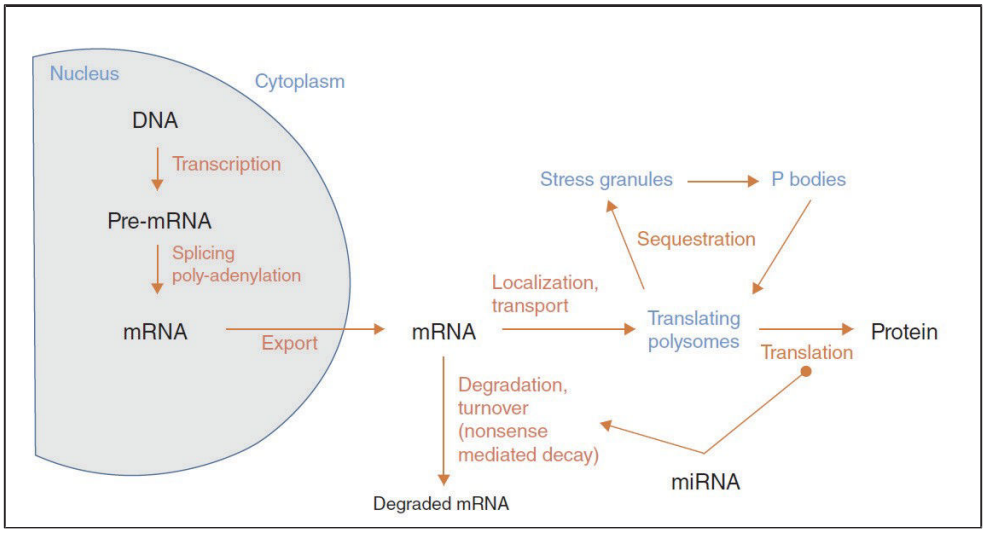


Figure 9: Schematic overview of the post-transcriptional regulation pathways (Adapted from (Cookson, 2017))

Numerous pathways are used for regulating protein synthesis. For instance, it has been showed that protein levels are associated to the mRNA levels; it was estimated that mRNA levels contributes to 56% of the protein variance abundance (with 18% by mRNA degradation and 38% related to transcription

level) while the remaining 44% was explained by translation processes (30% by translation and 14% by protein degradation) (Li et al., 2014). These data underscore the important role that post-transcriptional regulation plays in gene expression regulation. As a matter of fact, around 25% (162/650) of the genes implicated in ID have a role in the RNA metabolism (Kochinke et al., 2016). A similar percentage (~30%; 255/835) have been found also in our updated list of ID-genes.

Post-transcriptional regulation is mainly mediated by RNA-binding proteins (RBPs), which are implicated in different processes through-out all the mRNA life-time (Figure 9). For instance, the most frequent cause of ID – the Fragile X-syndrome – is caused by the absence of the RBP FMRP. About 9% (76/835) of genes implicated in ID code for a RBP. It is therefore clear that RBPs play a central role in ID and mRNA regulation, which is extremely important in highly specialized cells such as neurons for their proper axonal and dendritic growth, spine morphogenesis and synapse formation. In the following sections I will focus my attention on RBPs implicated in post-transcriptional regulation and ID (Tables 3-7).

5.4.1 mRNA MATURATION

In eukaryotes, mRNAs are synthesized while they are processed. mRNA processing consists in additional steps of maturation of the nascent transcript helping in downstream events and also in the ongoing transcription. For instance, a N7-methyl guanosine cap is added to the 5' mRNA at the very beginning of the transcription, speculating that it may be an important signal for mRNA elongation, as it prevents mRNA degradation and enables its nuclear export. On the other end, at the 3' of the newly synthesis transcript between 200 and 300 adenosines are added by the poly(A) polymerase. The same transcript may have different polyadenylation sites, meaning that some mRNAs have the same protein coding sequence but different 3' UTR ends. The existence of alternative polyadenylation sites further increases transcriptome variability and it might be implicated in tissue-specific regulation; for example, transcripts in the brain have generally longer 3' UTR (Licatalosi et al., 2008). Moreover, it could be implicated in several mechanisms of regulation, such as RNA localization and stability, and also miRNA-dependent translational regulation. As transcripts are synthesized, they are spliced and edited in parallel, while the polyadenylation is part of the transcript termination process.

Splicing is essentially a double transesterification reaction, in which introns are removed from the pre-mRNA precursor. This catalytic process is carried out by the *spliceosome*, which is assembled through subsequential steps and interactions among the spliceosomal subunits and numerous other factors. As a matter of fact, spliceosome assembly is also tightly regulated by DHX and DDX proteins (two large families of RNA helicase) that are required for the prespliceosome assembly or to guide the sequential spliceosomal rearrangements by the energy of the ATP hydrolysis (Will and Lührmann, 2011). Mutations in components of the spliceosomal subunits associated to ID have been reported, but they

are linked to a much broader phenotype, due to the essential role of the spliceosome (Table 3). Different auxiliary proteins are involved in the catalysis of the reaction as well as in the correct recognition of the splice sites. Different *cis*-RNA sequencing elements, in combination with protein regulators, help in the site selection of splicing that is performed by the spliceosome. Among the *cis*-regulatory elements there are the exonic splicing enhancers (ESEs) or silencers (ESSs) and the intronic splicing ehancers (ISEs) or silencers (ISSs), which are bound by different proteins that could enhance or inhibit splicing. For example, hnRNPH2 was described in rodents to bind to the pre-mRNA of *Trf2* and to inhibit the splicing of a short isoform of *TRF2* that promotes neuronal differentiation (Zhang et al., 2011). Interestingly, six *de novo* missense mutations have been reported in six unrelated female patients affected by developmental delay, ID, autism, hypotonia and seizures (Bain et al., 2016). *HNRNPH2* is on the X-chromosome and no affected boy has been described, suggesting that these variants may be lethal in males.

Once the pre-mRNA has been spliced, a set of specific proteins is deposited about 20 nucleotides upstream of intron excision of the spliced mRNA, independently of the sequence. This exon-junction complex (EJC) will be stably bound to the formed mRNA until the cytosol, where it will be removed during the first "pioneer" round of translation. The EJC tags the spliced mRNA and it intermediates downstream processes – such as nonsense-mRNA-mediated decay (NMD), translation, mRNA export and transport - by binding to transiently associating factors. The core EJC is composed by four proteins and two of them have been reported as altered in neurodevelopmental disorders (Table 3).

mRNA MATURATION				
Gene	**Function**	**Phenotype**	**Inh.**	**Reference**
ZC3H14	poly(A)	Mental retardation, autosomal recessive 56	AR	(Pak et al., 2011)
RBFOX1	splicing	ID, ASD, ADHD, epilepsy, bipolar disorder and schizophrenia	AD	(Sartor et al., 2015)
PQBP1	splicing	Renpenning syndrome	XLR	(Kalscheuer et al., 2003)
HNRNPH2	splicing	Mental retardation, X-linked, syndromic, Bain type	XL	(Bain et al., 2016)
HNRNPU	splicing	Epileptic encephalopathy, early infantile, 54	AD	(Carvill et al., 2013; Hamdan et al., 2014; de Kovel et al., 2016; Need et al., 2012)
HNRNPK	splicing	Au-Kline syndrome	AD	(Au et al., 2015)
AFF2	splicing	Mental retardation, X-linked, FRAXE type	XLR	(Gecz et al., 1996; Stettner et al., 2011)
RBMX	splicing	Mental retardation, X-linked, syndromic 11, Shashi type	XLR	(Shashi et al., 2015)
EFTUD2	spliceosome	Mandibulofacial dysostosis, Guion-Almeida type	AD	(Lines et al., 2012)
PUF60	spliceosome	Verheij syndrome	AD	(Dauber et al., 2013)
RBM28	spliceosome	Alopecia, neurologic defects, and endocrinopathy syndrome	AR	(Nousbeck et al., 2008)
NONO	spliceosome	Mental retardation, X-linked, syndromic 34	XL	(Mircsof et al., 2015; Scott et al., 2017)
DDX48	EJC component	Robin sequence with cleft mandible and limb anomalies	AR	(Favaro et al., 2014)

Table 3: Main RBPs involved in mRNA maturation implicated in ID

5.4.1.1 ALTERNATIVE SPLICING

Alternative splicing is a well-known mechanism to create proteomic diversity, as it can lead to numerous different transcripts (Figure 10), and many human genes undergo this process. Alternative splicing has a crucial role in gene expression regulation, as it dynamically controls spatial and temporal expression of different isoforms. Indeed, neurodevelopment relies on this mechanism to change gene expression at different developmental stages in a dynamic manner. Moreover, many alternative isoforms are tissue-specific - with the mammalian brain having the highest number of them (Wang et al., 2008) - indicating the important role of the alternative splicing in tissue diversity. The presence of distinct mRNA isoforms among tissues could be the result of different concentration of diverse splicing factors, as each transcript could be regulated by more than one splicing factors. This redundant mechanism increases the RNA-regulation complexity and hinders the characterization of the target gene set of a specific splicing factor. Another explanation of tissue-specific isoforms formation is the restricted expression of certain splicing factors in specific tissues, which are involved in the splicing regulation of target pre-mRNAs whose alternative isoforms have an essential function in that tissue.

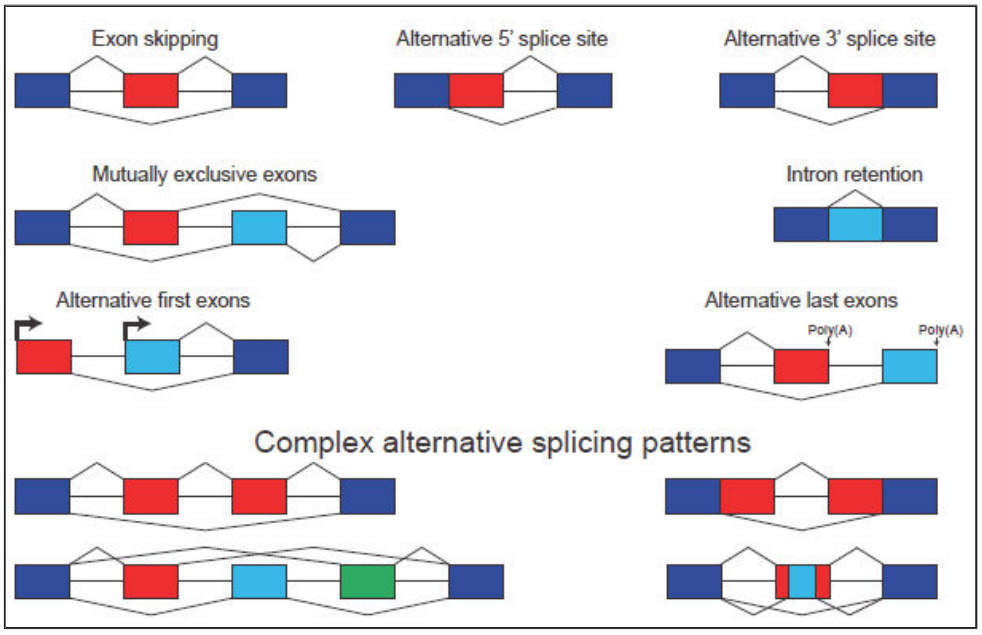


Figure 10: Basic alternative splicing events (adapted from (Park et al., 2018))

Many tissue-specific splicing factors have been identified in brain and neurons, highlighting the importance of the correct expression balance of different isoforms in neurons. For example, it has been showed that in human neuronal stem cells during differentiation, the neuron-specific splicing factor RBFOX1 regulates the alternative splicing of genes involved in neuronal maturation. Moreover, its splicing targets contains transcription factors, other splicing factors and synaptic genes implicated in neurodevelopmental disorder (Fogel et al., 2012). Alteration in RBFOX1 level itself and in its dependent alternative splicing targets were reported in brains of individuals with ASD (Voineagu et al., 2011) and

translocations encompassing *RBFOX1* have been reported in several patients with ASD, epilepsy and ID (Bhalla et al., 2004; Lal et al., 2015; Martin et al., 2007).

5.4.2 mRNA EXPORT

Once the mature mRNA is formed in the nucleus, it must be exported to the cytoplasm to be translated. The mRNA transport from the nucleus to the cytosol occurs through the nuclear pore complex. mRNA export is mediated by export adaptors that recognize and bind to the mRNA during its maturation, and after pass it to the transport factors that will export the mRNA into the cytoplasm. It is interesting to notice that some subunits of the different mRNA export pathways may transport specific classes of mRNAs, suggesting that mRNA export could control gene expression (Wickramasinghe and Laskey, 2015). In humans there are different RNA transport pathways composed by different multisubunit complexes. Mutations in proteins forming these complexes have been associated to ID (Table 4) as for example, *THOC2* and *THOC6*, encoding two subunits of the same complex involved in the transcription and export of mRNA (the TREX complex) (Beaulieu et al., 2013; Kumar et al., 2015).

mRNA EXPORT AND LOCALIZATION				
Gene	**Function**	**Phenotype**	**Inh.**	**Reference**
GANP	export	Charcot-Marie-Totth neuropathy and mild ID	AR	(Schuurs-Hoeijmakers et al., 2013; Ylikallio et al., 2017)
THOC2	export	Mental retardation, X-linked 12/35	XLR	(Kumar et al., 2015)
THOC6	export	Beaulieu-Boycott-Innes syndrome	AR	(Amos et al., 2017; Beaulieu et al., 2013)
XPO1	export	2p15 microdeletion syndrome	AD	(Lévy et al., 2017)
KIF5C	localization	Cortical dysplasia, complex, with other brain malformations 2	AD	(Poirier et al., 2013)
KIF4	localization	Mental retardation, X-linked 100	XLR	(Willemsen et al., 2014)
KIF11	localization	Microcephaly with or without chorioretinopathy, lymphedema, or mental retardation	AD	(Hu et al., 2016; Mirzaa et al., 2014; Ostergaard et al., 2012)

Table 4: Main RBPs involved in mRNA export and localization and implicated in ID

5.4.3 mRNA LOCALIZATION

Not all the exported mRNAs are immediately translated; many of them are maintained in a translationally silent state, waiting for proper subcellular localization or for a timing signal (Moore, 2005). This is particularly important for highly specialized and polarized cells, such as neurons, where the cellular soma is distant from axons and dendrites. Nevertheless, the latters are able to rapidly respond to stimuli and change local protein expression and cytoskeletal structure; this is achieved by axonal mRNA transport and by the presence of a local translational machinery. Localized mRNAs are transported in RNA granules, and several mechanisms are implicated in mRNAs transport among which active transport along the cytoskeleton and in particular microtubules (Kiebler and Bassell, 2006). Studies analysing the proteomic composition of these RNA granules revealed that RNA granules

compositions are highly heterogeneous and contain proteins involved not only in transport but also in translation and degradation of the mRNA. The first study showed that kinesin KIF5 transports RNA granules to dendrites (Kanai et al., 2004); indeed, several kinesins have been implicated in RNA granules transport.

5.4.4 TRANSLATION

Once transcripts reach their proper subcellular localization, they are eventually translated into proteins. Gene expression is regulated also at this step, especially at the initiation of the translation. During translational initiation, the preinitiation complex - made by the 40S ribosome subunit and initiation factors - is assembled and it is recruited at the 5'cap of the mRNA, a step mediated by the cap binding complex eIF4F. This complex is formed by eIF4A, eIF4G and eIF4E. The latter recognizes the 5' cap and replaces in the cytosol the cap-binding proteins CBP20 and CBP80.

TRANSLATION				
Gene	**Function**	**Phenotype**	**Inh.**	**Reference**
EIF4E	intiation	ASD	AD	(Neves-Pereira et al., 2009)
EEF1A2	intiation	Epileptic encephalopathy, early infantile, 33; Mental retardation, autosomal dominant 38	AD	(de Ligt et al., 2012; Nakajima et al., 2015)
FMR1	repression	Fragile X syndrome	XL	(Napoli et al., 2008; Oberlé et al., 1991)
PUM1	repression	Developemental Delay Ataxia and Seizure	AD	(Gennarino et al., 2018)

Table 5: Main RBPs involved in translation and implicated in neurodevelopmental disorder

On the other hand, eIF4G links the 5'cap to the preinitiation complex by bridging it to eIF4E. eIF4G interacts with the poly(A) tail by the binding with PABP (PolyA Binding Protein), circularizing the mRNA and bringing the 5' cap close to the 3' poly(A) tail. To this end, the mRNA translation can be regulated by either inhibiting the binding of eIF4E to the cap or by impeding the interaction between eIF4E and

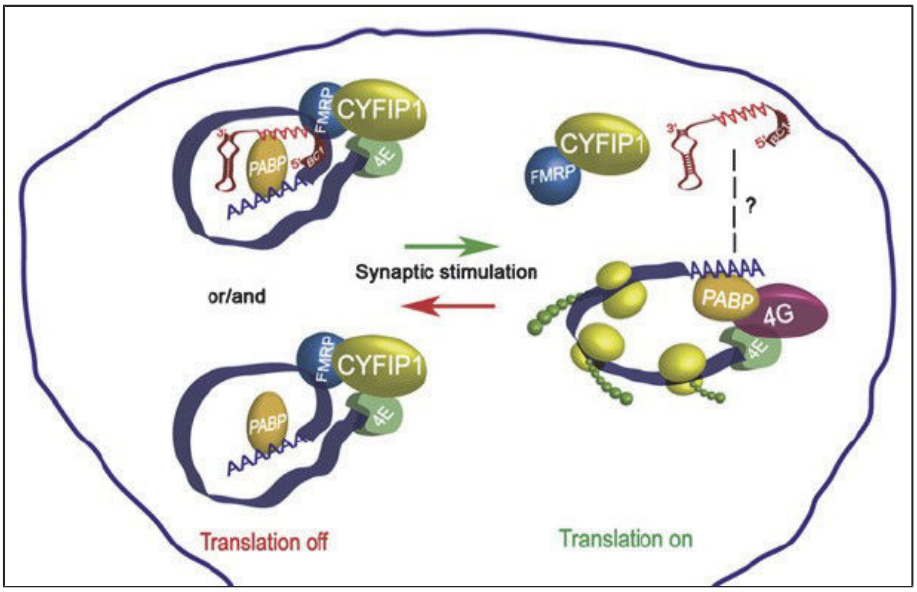


Figure 11: Translation regulation by the RBPs FMRP and CYFIP1 (Adapted from (Napoli et al. 2008))

eIF4G. A RBP implicated in mRNA neuronal translation is FMRP, involved in the fragile-X syndrome. Among the many role of FMRP, it is known to regulate the translation of a subset of mRNAs by inhibiting their translational initiation together with one of its known binding partner CYFIP1. In details, the initiation factor eIF4E is bound by CYFP1, which is recruited by FMRP. In brain, a specifically expressed RNA increases the affinity of FMRP for the CYFP1-eIF4E complex, resulting in a stable complex that repress translation (Figure 11). Upon synaptic stimuli, CYFP1-FMRP dissociates from eIF4E leading to translation activation and subsequent production of proteins encoded by FMRP mRNA targets (Napoli et al., 2008).

5.4.5 mRNA DEGRADATION

Transcripts could also be degraded. The mRNA degradation is an important post-transcriptional regulation mechanism, as it alters the transcript level in the cell and enables mRNA turnover regulation. Several mechanisms have been described and most of them require an initial step of deadenylation and decapping.

Deadenylation is one of the first step for degrading mRNA and is hence a crucial step in post-transcriptional regulation. Three different enzymes are involved in deadenylation, all coordinated by RNAse D, an exoribonuclease with 3'-5' activity.

mRNA DEGRADATION				
Gene	**Function**	**Phenotype**	**Inheritance**	**Reference**
CNOT3	Deadenylation	Neurodevelopmental disorder	AD	(Deciphering Developmental Disorders Study, 2017)
TOE1	Deadenylation	Pontocerebellar hypoplasia, type 7	AR	(Lardelli et al., 2017)
RBM8A	NMD	Thrombocytopenia-absent radius syndrome	AR	(Albers et al., 2012)
UPF3B	NMD	Mental retardation, X-linked, syndromic 14	XLR	(Tarpey et al., 2007)
EXOSC3	exosome	Pontocerebellar hypoplasia type 1b	AR	(Halevy et al., 2014; Wan et al., 2012)
EXOSC2	exosome	Short stature, hearing loss, retinitis pigmentosa, and distinctive facies	AR	(Di Donato et al., 2016)
EXOSC8	exosome	Pontocerebellar hypoplasia type 1c	AR	(Boczonadi et al., 2014)
RNASEH2A	Ribonuclease	Aicardi-Goutieres syndrome 4	AR	(Crow et al., 2006; Sanchis et al., 2005)
RNASEH2C	Ribonuclease	Aicardi-Goutieres syndrome 3	AR	(Crow et al., 2006)
RNASEH2B	Ribonuclease	Aicardi-Goutieres syndrome 2	AR	(Crow et al., 2006; Rice et al., 2007)
SAMHD1	Ribonuclease	Aicardi-Goutieres syndrome 5	XLR	(Rice et al., 2009)
DHX30	stress granules	neurodevelopmental disorder with severe motor impairment and absent language	AD	(Lessel et al., 2017)

Table 6: Main RBPs involved in mRNA degradation and implicated in neurodevelopmental disorders

The cytosolic protein PABPC1 influences the first step of polyadenylation as it promotes the activity of the PAN complex, which is composed by PAN2 and PAN3 subunits. The PAN complex is then shortening

the polyA tail and facilitates the binding of the CCR4-NOT complex, which will continue the polyadenylation (Wahle and Winkler, 2013). The CCR4-NOT complex is made of two catalytic subunits: CNOT6 and CNOT6L, which belong to the exonuclease-endonuclease-phosphatase (EEP) protein family, and CNOT7 and CNOT8 that are part of the DEDD class of exonuclease. In addition to them, there is also the NOT modules, whose core subunits are CNOT1, CNOT2 and CNOT3. In the latter, several missense mutations have been reported in a large cohort of patients affected by neurodevelopmental disorders (Deciphering Developmental Disorders Study, 2017) and it is a therefore a candidate gene for ID. Interestingly, homozygous deletion of *Cnto3* in mouse was lethal, suggesting an important role during development while *Cnot3*$^{+/-}$ had cardiomyopathy (Morita et al., 2011). Further studies reported that CNOT3 timely regulates the expression of differentiation genes, by promoting their mRNA deadenylation and subsequent degradation, thus maintaining a pluripotent state (Zheng et al., 2016).

Once the poly(A) tail has been shortened, a holoenzyme composed by DCP1 and DCP2, along with other cofactors among which DDX6, degrades the 5' cap, freeing the mRNA from the translation initiation factors hence resulting in a non-functional mRNA as it cannot be translated.

Non-translating mRNAs can accumulate into two different cytoplasmic mRNPs granules: the P bodies, where there are most of the components of the mRNA decay/degradation machinery, and the stress granules, where there are more translational initiation factors (Decker and Parker, 2012).

5.4.5.1 RIBONUCLEASES

The mRNA can be degraded in two directions. The enzyme XRN1 is a 5'-3' exoribonuclease that binds to the 5' region of the mRNA as the cap and the translational initiation factors are released. On the other hand, mRNA can be degraded in the opposite 3'-5' direction by a well-conserved ribonuclease complex, called the RNA exosome. This complex is formed by a six-subunits ring (going from EXOSC4 to EXOC9), a three subunit cap (EXOSC1, EXOSC2 and EXOSC3) and the catalytically active ribonuclease DIS3. The mRNA to be degraded pass through a central channel from the cap-subunits passing through the ring-subunits and finally reach the catalytic subunit, which will degrade the RNA substrate. The RNA exosome has also different cofactors that help also to direct the exosome to specific target RNAs (*i.e.* the NEXT and the Ski complex). Interestingly, missense mutations in *EXOSC2*, *EXOSC3* and *EXOSC8* have been linked to three different disorders associated to ID (Table 6). Surprisingly, mutations in genes coding for subunits of the same ubiquitous RNA exosome give rise to distinctive tissue-specific phenotypes, caused by an impaired RNA degradation. This could be because mutations might differentially affect the level and the stability of the subunit and of the entire RNA exosome; or missenses might interfere with exosome cofactors binding; or they disturb the interaction with specific RNA substrates in the entry paths (Morton et al., 2018).

5.4.5.2 NONSENSE MEDIATED DECAY

Different surveillance mechanisms are present in eukaryotic cells, preventing the production of aberrant proteins that could have reduced or even damaging functions. In the cytosol, there are different surveillance mechanisms, among which the NMD (Non-Sense Mediated Decay) that degrades transcripts having a premature termination codon (PTC).

The most well-known model of the recognition of the PTC relies on the EJC, which is stably bound to the spliced mRNA from the nucleus together with the interacting splicing factor RNPS1 and the NMD factors UPF2 and UBPF3 paralogous. During the first pioneer round of translation, ribosomes scan the mRNA, remove the EJCs and pause at a stop codon. If a mRNA has an EJC located more than 50-55 nucleotides downstream of a PTC, ribosomes are not able to remove the EJC, hence the translation eukaryotic release factor eRF1 and eRF3 are recruited along with UPF1 – an ATPase helicase - and the kinase SMG1, forming the SURF complex. SMG8 and SMG9 bind to SMG1 to temporarily block its phosphorylation activity of UPF1. At the remaining EJC, UPF3A, UPF3B associate with UPF2, which will be bind by UPF1, resulting in the decay inducing complex (Figure 12). This interaction promotes the phosphorylation of UPF1 by SMG1, leading to the subsequent recruitment of SMG5, SMG6 and SMG7: SMG6 will cleave close to the PTC while SMG5 and SMG7 will promote RNA decapping and deadenylation, leading to its degradation.

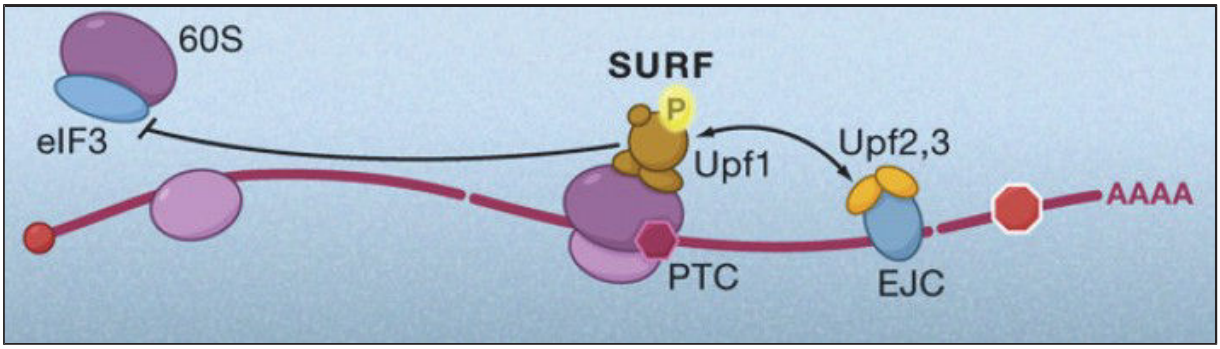


Figure 12: Schematic representation of the NMD mechanism (Adapted from Moore and Proudfoot, 2009)

NMD is also important to control the physiological level of many mRNAs. For instance, it has been described that during neuronal differentiation the NMD activity decreases (Alrahbeni et al., 2015; Lou et al., 2014). For instance, UPF1 promotes the proliferative, undifferentiated cell state by inducing the NMD of transcripts involved in neuronal differentiation but, when this is triggered, it promotes the expression of a neuron-specific miRNA (miR-128) that inhibits UPF1 resulting in a decrease of several NMD factors and activity, enabling the differentiation of neuronal progenitors (Lou et al., 2014). It is therefore not surprising that mutations in NMD factors have been identified in patients with various forms of ID (Table 6). To further understand the implication of NMD in neurodevelopmental disorder, a study to investigate the contribution of CNVs encompassing 18 NMD genes in individuals with ID and/or congenital anomalies was performed (Nguyen et al., 2013). The study reported a significant

enrichment in the patient cohort of copy number losses of *RBM8A* (already implicated in TAR syndrome), *UPF2* and *UPF3A*. Moreover, the authors identified in the cohort a significant enrichment of copy number gain for *UPF2, SMG6, RBM8A, EIF4III* and *RNPS1*, suggesting that CNVs encompassing NMD factors (and even EJC components) might predispose to neurodevelopmental disorders.

5.4.5.3 RNA INTERFERENCE

RNA interference is a molecular mechanism that regulates gene expression by targeting and degrading specific mRNA substrates. It was first identified in 1998 in *Caenorhabditis elegans*, where a double-stranded RNA was showed to interfere with gene expression (Fire et al., 1998). RNA interference is mediated by the hybridization of different type of small RNA molecules to their complementary mRNA target and in this section I will focus on micro RNAs (miRNAs), due to their role in neurodevelopmental disorders.

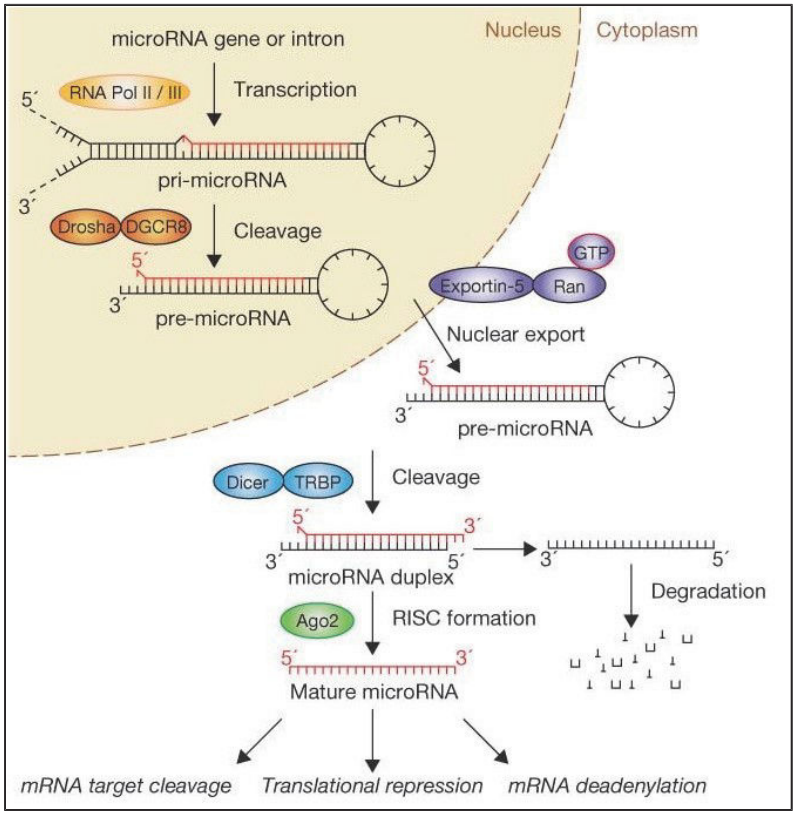


Figure 13: miRNAs biogenesis (Winter et al., 2009)

miRNAs are single-stranded small non-coding RNAs known to regulate the expression of around 60% of protein-coding genes (Esteller, 2011), so they have been implicated in a variety of biological processes. By binding to the 3'UTR of their target mRNAs, miRNAs regulate gene expression as they will promote either mRNA degradation or translational inhibition. miRNA genes are initially transcribed into precursor molecules (pri-miRNAs), which are long double-stranded RNAs forming numerous internal hairpins structures (Figure 13). Pri-miRNAs are then cleaved by the ribonuclease DROSHA guided by DGCR8, which recognizes the dsRNA-ssRNA junction of pri-miRNAs. This cleavage results in single hairpin molecules, referred to as precursor-miRNAs (pre-miRNAs). Pre-miRNA are then exported

to the cytosol through exportin-5 and its cofactor Ran-GTP, where they will be eventually cleaved into mature miRNAs (approximately 21 nucleotides) by the ribonuclease DICER with the RBP TBRP.

The mature miRNAs are then unwound and the mature guide strand is bound by Ago proteins, forming the RNA induced silencing complex (RISC). In humans there are four Ago proteins (AGO1-4) and only AGO2 has a nuclease activity, while all the others are more involved in inducing translational repression and mRNA destabilization (Winter et al., 2009). The mature miRNA will then guide the RISC complex to its target mRNA and, once they hybridize, it silences it by triggering mRNA translational repression, deadenylation and decay.

miRNAs are emerging as important regulators of brain development and function, by modulating the expression of target genes implicated in these processes. Defects in miRNA biogenesis or in miRNA expression itself have been implicated in neurodevelopmental anomalies, such as Rett syndrome and Fragile-X syndrome, whose responsible protein for the latter (FMRP) have been also linked to miRNA pathway (Jin et al., 2004). For instance, several studies using different knockdown animals for genes coding for proteins involved in miRNA biogenesis (*i.e. Ago2* and *Dicer*) showed defects during brain development and subsequent anomalies (Sun and Shi, 2015). On the other hand, numerous brain-specific miRNAs have been described and implicated in different neurodevelopmental stages, as showed by the previous example of miR-128. As a matter of fact, several studies reported a differential miRNA expression profile in individuals with ASD (Abu-Elneel et al., 2008; Talebizadeh et al., 2008) and correlated it with a differential expression of target genes (Sarachana et al., 2010).

5.4.6 tRNAs

tRNAs are ubiquitous non-coding RNAs and are highly abundant in the cell, constituting 4-10% of the total cellular RNA. They have an essential role in the translation process, as they transport the amino acids to be added to the nascent polypeptide to the ribosomes, and allow the translation of a nucleotide sequence into an amino acid. Mutations in genes coding for enzymes involved in tRNA processing have been identified in patients affected by neurodevelopmental disorder, including ID. For instance, a large number of mutations have been identified in the large family of aminoacyl-tRNA synthetase (ARS), the enzymes charging tRNAs with their cognate amino acid (Table 7). Nevertheless, it is not clear the reason of such tissue-specific phenotype nor the pathogenic molecular mechanisms. One explanation could be that some specific cell type are more vulnerable and sensitive to deleterious effect of misfolded proteins, such as postmitotic neurons (Kirchner and Ignatova, 2015). However, mutations in ARS may impair its functional enzymatic activity and cause amino acid misincorporations, but they might also lead to a gain-of-function effect that is currently being explored (Meyer-Schuman and Antonellis, 2017).

RNA and tRNA modification				
Gene	**Function**	**Phenotype**	**Inh.**	**Reference**
ADAR	RNA modification	Aicardi-Goutieres syndrome 6	AR	(Rice et al., 2012)
NSUN2	RNA modification	Mental retardation, autosomal recessive 5	AR	(Abbasi-Moheb et al., 2012)
PUS3	pseudouridine convertion	Mental retardation, autosomal recessive 55	AR	(Shaheen et al., 2016)
AARS	tRNA synthetase	Epileptic encephalopathy, early infantile, 29	AR	(Simons et al., 2015)
HARS	tRNA synthetase	Usher syndrome type 3B	AR	(Puffenberger et al., 2012)
DARS	tRNA synthetase	Hypomyelination with brainstem and spinal cord involvement and leg spasticity	AR	(Taft et al., 2013)
IARS	tRNA synthetase	Growth retardation, intellectual developmental disorder, hypotonia, and hepatopathy	AR	(Kopajtich et al., 2016)
RARS2	tRNA synthetase	Pontocerebellar hypoplasia, type 6	AR	(Edvardson et al., 2007)

Table 7: Main RBPs involved in RNA and tRNA modifications involved in ID

5.4.7 RNA MODIFICATIONS

Different chemical modifications occur on the RNA. For instance, tRNAs and rRNAs are largely modified by pseudouridinylation as well as methylation. Each tRNAs may undergo to 14 modifications; these modifications may alter structural features, hence their function; or they can influence translation efficiency and speed; or tRNA may be cleaved in potential signalling messengers (Nachtergaele and He, 2017). rRNAs are modified in well-conserved position and their modification may influence ribosome biogenesis and protein synthesis. Chemical modifications on the mRNA may change the amino acid sequence, in a process that is called *RNA editing*. RNA editing is majorly present in tissue that require high plasticity, such as the brain, as it can generates several different transcripts. One of the most frequent editing is the deamination of adenosine into inosine, carried by RNA-specific adenosine deaminase (ADAR). Interestingly, several homozygous or compound heterozygotes mutations in *ADAR* have been associated to an inflammatory disorder particularly affecting brain and skin (Aicardi-Goutières syndrome), associated with an upregulation of interferon-stimulated genes, suggesting a role for ADAR in the repression of interferon signalling (Rice et al., 2012). RNA epigenetics is also emerging as an important RNA modification; the discovery of an enzyme (FTO) that can reverse the RNA N6-methyladenoysine modification suggests that also RNA modifications are dynamic, thus they might also play an important role during neurodevelopment (Zheng et al., 2013).

6. NEXT-GENERATION SEQUENCING APPLICATIONS IN ID

6.1 GENERAL PRINCIPLES OF NGS

For almost two decades, Sanger sequencing remained one the most used approach for DNA sequencing, especially after its implementation as automated Sanger sequencing (*e.g.* multicapillary sequencers), which is referred to as the *first generation sequencing* technology. Despite the numerous accomplishments obtained - among which the first human genome sequence and the identification of numerous genes involved in human diseases - Sanger sequencing is a limited technique as it can generate only one sequence at a time, using a polymerisation reaction with ddNTPs. The development of the massive parallel sequencing techniques overcame these main issues by the live neosynthesis of DNA fragments and their immediate detection as they are getting synthesized. This neosynthesis is performed in parallel in millions of independent sequencing reactions, allowing a high-throughput of generated sequences. Different NGS technologies have been developed, and the main differences lie in the template preparation and in the chemistry of the DNA neosynthesis. The most common sequencing methods are the cyclic reversible termination, the sequencing by ligation and the pyrosequencing. The most used one is the cyclic reversible termination, which consists in the cyclic imaging of the incorporation of fluorescently modified nucleotides that block the DNA synthesis. Each incorporated modified nucleotide represents the complement of the template. Then, the terminator fluorescent nucleotide is removed and a next incorporation step is performed. Usually, an amplification step is required before sequencing, in order to pass the imaging detection threshold and the two most common methods are the emulsion PCR or a solid phase amplification. Both amplifications have to stay minimal, otherwise they will produce a biased product.

In parallel to the implementation of the NGS technologies, the bioinformatic field had to improve to face the unprecedented amount of data generated by these technologies that raised some challenges in data management, storage as well as the analysis. The first challenge is the conversion of the image data (*e.g.* the imaging of the fluorescent nucleotide incorporation) into sequence reads, the so-called *base calling*. In parallel to the base calling, it is has been established a score that indicates the quality of the reads (the *phred score*), as it provides important information for the next steps of alignment and assembly, and also for later variant analysis. NGS reads are then aligned and assembled either *de novo* or to a reference sequence, according to the biological investigation. One limitation of the reads alignment is the inability to place regions in repetitive regions or in corresponding regions that may not exist in the reference genome, even if the paired-end sequence is used. Over the years, NGS technologies have constantly improved, increasing the high-through put capacities, speed and accuracy and at the same time significantly decreasing the costs. Companies have developed sequencers able to overcome some previous issues, including the amplification step.

The advent of the NGS technologies revolutionized human genetics as they had a remarkable impact in several research fields due to its wide applicability, ranging from basic research to clinical diagnostic. Consequently, important achievements have been made in the better understanding of the genetic aetiology of ID, as demonstrated by the rapid increase of newly genes associated to ID. NGS technologies also greatly contributed to the increase of the molecular diagnosis of ID patients (Figure 14). The constant evolving of different approaches and new methods have also helped in the delineation of involved molecular mechanisms in ID.

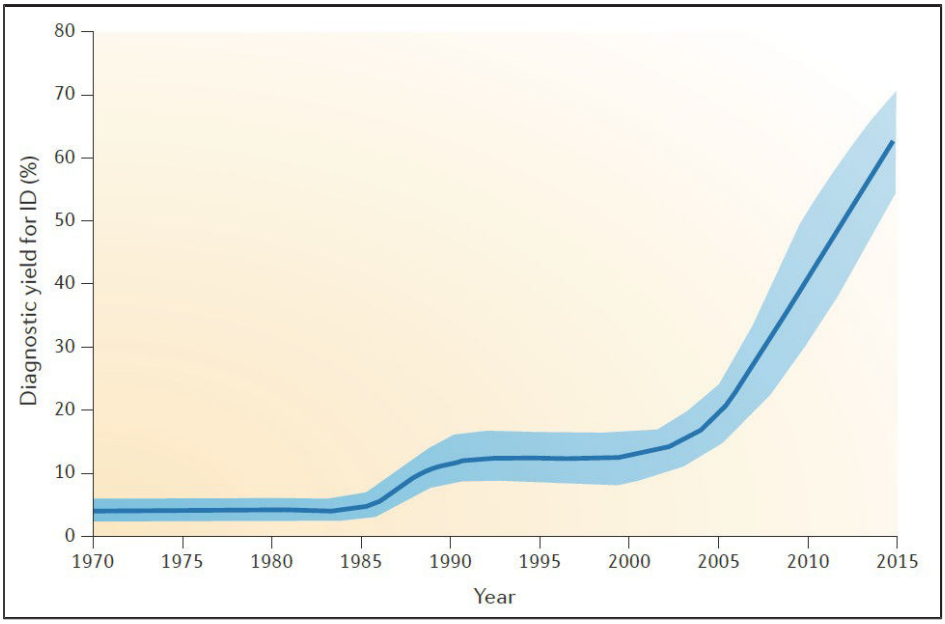


Figure 14: Diagnostic yield over the years (Vissers et al. 2016)

6.2 NGS IN VARIANT AND GENE DISCOVERY IN ID

In the past years, the research of genes implicated in ID was hindered by cost, labour and it was limited to a small number of candidate genes. The main previously used techniques (*i.e.* linkage analysis, homozygosity mapping, analysis of the breakpoints, positional cloning) were laborious and they required large families and numerous cases. The introduction of NGS technologies - and more in particular of WES - significantly ameliorated the identification of novel ID genes. One of the first approach used for gene discovery was to group patients with the same recognizable phenotype, in order to identify a commonly mutated gene. For example, this strategy identified the major genetic cause of the autosomal-dominant Kabuki syndrome (Ng et al., 2010b). However, this strategy is limited by the recognition and characterization of the clinical features, as well as by the variable phenotype and penetrance of a mutation. Moreover, NGS technologies are now routinely used for molecular diagnosis, as their cost, processing time, clinical interpretation and data management became more affordable. Each type of NGS approaches has its advantages and disadvantages and it has to be chosen accordingly to the aim of the study. The most commonly used techniques are the targeted sequencing,

the WES and they will be detailed in the following sections. WGS is for now more used in the research field but might become a diagnostic tool in the next few years.

6.2.1 GENOME ENRICHMENT TECHNIQUES

As WGS is expensive and due to the difficulties to interpret all the genomic variants of one individual, researchers overcame these disadvantages by focusing only on specific genomic regions of interest mainly encoding for proteins, since the majority of the Mendelian disorders are caused by mutations disrupting the protein-coding sequences.

Sequencing library can be enriched by DNA of selected target regions, prior to the amplification step and the subsequent sequencing. In this way, cost and analysis time are drastically reduced with an increase in the coverage of the generated sequences. There are three main enrichment approaches, the enrichment by hybridization (NimbleGen and Agilent Technologies), selective circularization method (molecular inversion probe (MIP)) and PCR-based approaches (RainDance and Fluidigm tehcnologies).

All these methods can be used for the target of a specific region of interest, but they differ in the enrichment specificity (*i.e.* the proportion of sequences in-targets versus sequences off-targets), coverage homogeneity across different samples that importantly influence the reproducibility of the experiment, the coverage homogeneity across all targeted regions, enrichment capacity (*i.e.* the maximum size of total target regions) and the associated time and costs of sample preparation.

6.2.1.1TARGETED SEQUENCING (TS)

This approach consists in the sequencing of a subset of genes or regions of the genome of interest, such as genes known to be implicated in ID. TS is widely used in clinical laboratories because of its robustness and reliability, due to its high coverage. Moreover, its limited cost allows the inclusion of more patients and the data analysis is faster since the attention is focused on a restricted area of interest. Therefore, this targeted approach facilitates the identification of novel and rare variants in ID genes in a large number of patients. On the other hand, TS excludes the unbiased discovery of novel candidate ID genes. The first crucial step of this technique is the selection of the genes to include in the study. This decision has to evaluate different criteria, among which the total size of the targeted regions, which is restricted by manufacturing price thresholds and by the power of the sequencer machine that define the maximal number of patients to multiplex in a single sequencing lane. Once a gene list is made, probes have to be designed according to the portion of the gene that want to be investigated, which could be full-genes, regulatory elements, 5’ and 3’ UTRs or only exons. Usually, coding-exons are sufficient to identify pathogenic variants, since the majority of the identified mutations responsible for Mendelian disorders disrupt protein-coding sequences.

Multigene panels are different and vary on the number of genes (from dozen to several hundreds) and could include genes that are more frequently mutated in ID or even genes that need to be confirmed or better characterized. For example, some gene panels only focus on known X-linked ID genes. Studies using TS as a first diagnostic test in ID patients showed a diagnostic yield of about 20% on average (Figure 15) (Grozeva et al., 2015; Martínez et al., 2017; Redin et al., 2014; Tan et al., 2015).

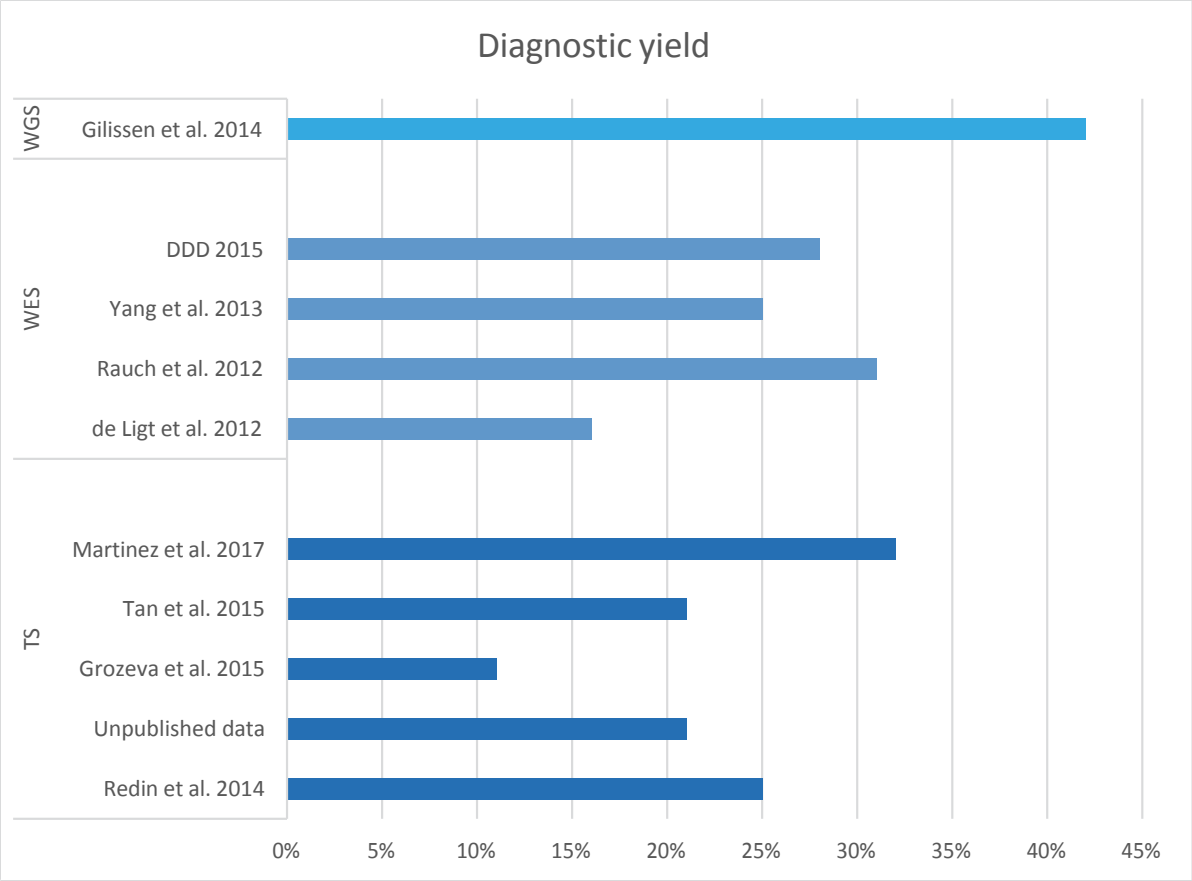


Figure 15: Diagnostic yield using different NGS approaches

(Dark-blue: Targeted sequencing (TS) studies (Grozeva et al., 2015; Martínez et al., 2017; Redin et al., 2014; Tan et al., 2015); Blue: whole-exome sequencing (WES) studies (Deciphering Developmental Disorders Study, 2015; de Ligt et al., 2012; Rauch et al., 2012; Yang et al., 2013); light-blue: whole-genome sequencing (Gilissen et al. 2014).

6.2.1.2 WHOLE-EXOME SEQUENCING

WES allows the sequencing of all the exons of the human genome and it is predominantly restricted to the coding regions, which consist in the 2% of the human genome. For diagnostic laboratories, it has been developed a version with coding regions of only OMIM genes, or genes already implicated in a disease, speeding up the variant analysis and interpretation. This technique have been used also on a cohort of ID patients, obtaining a diagnostic yield comparable to the one of the TS (Chérot et al., 2018). WES allows the unbiased discovery of novel candidate genes involved in a disorder, as well as the possibility to re-analyse data as the genetic knowledge increases. There are two main approaches for

a WES study: the patient can be sequenced alone or in parallel with his/her parents (*trio*). The first approach is indeed much cheaper and it can include more individuals; however all the identified variants are difficult to interpret, consequently the exome analysis may require much more time and effort. The trio-exome sequencing enables the sorting and filtering of the inherited variants, according to the inheritance scenario: X-linked, autosomal dominant or recessive. The inclusion of the parents in the analysis reduces by 10-fold the number of putative causal variants (Harripaul et al., 2017). It is therefore a powerful tool for the identification of novel pathogenic variants and the discovery of novel ID genes. WES studies in ID patients' cohort obtained on average a molecular yield between 25% and 30% (Figure 15) and they also identified and reported several ADID-genes. Indeed, WES is now currently used in clinical as well as in research laboratories as it allows a rapid variant and gene identification.

6.2.2 WHOLE GENOME SEQUENCING

The most complete NGS approach is undeniably the WGS, as it delivers the whole individual genome. Moreover, the generated uniform coverage allows the detection of any structural variants (balanced and unbalanced) even at a small resolution level and in intronic positions, beside the identification of all personal SNVs. On the other hand, these advantages are not counter parted by its disadvantages, which are essentially the extreme high cost (even if it is now significantly decreasing), storage and manipulation of data. Moreover, it has to be considered that, particularly in diagnostic laboratories, the variant analysis would be restricted in the coding-regions, as the knowledge of non-coding regions are still inadequate for diagnostic purposes, as there are still few and not complete tools to interpret the consequences of variants falling in non-coding regions. Efforts have to be made to develop them, as it is now clear that non-coding regions have crucial role in the regulation of gene expression (promoter, alternative splicing, enhancers, and also for proper chromatin organization (*i.e.* TADs)).

Recently, trio-WGS has been used in a cohort of 50 patients with an unexplained ID - previously tested by genomic microarray, targeted and whole-exome sequencing- and a diagnostic yield of about 40% was obtained. Initially, the authors focused their attention to SNVs and showed a significant enrichment of *de novo* LoF mutations in known ID genes. Moreover, they were able to detect mutations at the mosaic state, which was not possible to identify them with other techniques. They also identified and validated 8 different structural variants that were not detected by previously analysis. However, researchers could not reach a conclusion for variants in non-coding regions, even with the help of resources that provide a rich set of transcription factor-binding sites and chromatin state segments in different tissues and cell types, highlighting the need of follow-up studies for the better comprehension of these variants (Gilissen et al., 2014).

6.2.3 RNA-SEQUENCING

The advent of RNA-sequencing changed the transcriptomic studies, mainly performed by microarrays techniques, enabling, in addition to the quantification of gene expression, the identification of novel mRNA isoforms, and of alternative-splicing events. Beside the transcriptomic analysis and the differential expression studies, it has been recently shown that RNA-sequencing could be a useful technology for variant identification in heterogeneous disorders such as mitochondriopathy and rare muscle disorders (Cummings et al., 2017; Kremer et al., 2017). However, no similar study has been performed for ID.

6.3 NGS FOR UNDERSTANDING MOLECULAR MECHANISMS

Beside the identification of genetic causes of ID, NGS technologies helps also in the understanding and delineation of the implicated molecular mechanisms. To this end, different approaches have been developed in order to study the transcriptome and the epigenome.

6.3.1 TRANSCRIPTOME ANALYSIS

Different methods have been developed according to the type of RNA to be sequenced, such as the polyadenylated mRNA but also small and short interfering RNA. Generally, the RNA is isolated from the sample of interest and then converted into cDNA that will be subsequently sequenced. The generated reads can be either aligned to a genome of reference, compared with known transcripts, or assembled de novo, which can be useful for the identification of new transcripts. Over the years, RNA-sequencing greatly advanced, enabling different applications due to the versatility of this technology as it can be used for studying quantitative and qualitative RNA changes. RNA-sequencing allows the discovery of novel RNA isoforms, the gene expression quantification as well as the identification and estimation of alternative-splicing events. Nevertheless, RNA-sequencing is limited by the generation of short reads. For example, short-reads limit a correct quantification of alternative isoforms. To this end, sequencers have been developed to allow sequencing of long-read mRNAs, enabling a direct resolution of isoform structures, leading to the discovery of novel transcripts and alternative splicing events in different tissues and cell type (Park et al., 2018). As the NGS technologies and bioinformatics tools constantly improved, an increasing number of dataset across tissue and individual were created in order to delineate specific gene expression profiles among different tissues as well as to study the correlation of a genotype to alternative splicing variations. Recently, the genotype-tissue expression (GTEx) dataset has been released (https://www.gtexportal.org/home/), which comprises several human tissues transcriptome from well-genotyped donors, thus providing a powerful resource for the characterization of transcriptional differences among tissues and to assess a genetic correlation between an alternative splicing and to an expression quantitative loci (eQTL).

RNA-sequencing is frequently used to analyse the differential gene expression among patients or among different cells type. For example, a transcriptomic analysis comparison between patients with mutations in two NMD-related genes in *UPF3B* (known to be implicated in ID) and affected individuals with deletion encompassing *UPF2* showed similar transcriptomic consequences, with the identification of neuronal functional proteins among the differentially expressed genes (DEG), suggesting a contribution of deletion of *UPF2* to ID (Nguyen et al., 2013) .The isolation of the RNA from distinct cellular type and also subcellular fractions led to the characterization of specific isoforms present in a specific cells and even of its subcellular localization. The development in recent years of single-cell RNA-sequencing is helping in this characterization of specific cell types within a tissue (*e.g.* neurons, astrocytes, etc.) as well as the better understanding of the complexity inside a single cell. Currently, different technologies are being developed to enable to spatially resolved transcriptomics directly in cells and tissue (*e.g.* In Situ RNA Sequencing), which is a promising complementary tool for studying tissue heterogeneity that could become also useful in diagnosis by checking, for example, for biomarkers (Ke et al., 2016).

6.3.2 EPIGENETIC AND REGULATORY MECHANISMS

NGS technologies also allow the identification of DNA sequences bound by proteins, like transcription factors and regulatory proteins. One of the most common technique is the Chromatin ImmunoPrecipitation (ChIP)-sequencing. In this approach proteins are cross-linked with their genomic binding sites and this complex is then immunoprecipitated with an antibody specific for the protein of interest. DNA is then extracted and purified, prior to be sequenced. In parallel, chromatin markers could be used to delineate the chromatin conformation at a specific stage. This method greatly improved the identification and characterization of binding sites of transcription factors. As previously described, mutations in genes coding for different transcription factors have been associated with ID (as detailed in *Gene Expression Regulation: Transcriptional Regulation,* pg. 37 and Table 2, pg. 38). ChIP-sequencing of some of these factors, revealed the network of genes they regulate. For example, through a ChIP experiment in mouse brain, it has been showed that Tbr1 – for which several truncating mutations in *TBR1* have been identified in ASD patients with a variable phenotype - binds mainly adjacent to ASD genes (Notwell et al., 2016), indicating that alteration of this transcription factor may lead to dysregulation in the expression level of different ASD genes. An analogue approach has been developed for the identification of the RNA-binding proteins targets *(i.e.* CLIP). For instance, this approach has been used for the identification of the RNA targets of FMRP (Tabet et al., 2016), whose absence lead to the fragile X-syndrome.

NGS advanced the epigenetic research, by providing the profile of genome-wide epigenetic marks, such as the methylation. The methyl-sequencing consists in the bisulfite conversion of the genomic DNA and its subsequent sequencing, either of the entire genome or subregions. The bisulfite

conversion transforms the unmethylated cytosine into uracil, while the methylated cytosine will not be converted. By comparing the generated sequencing by methyl-sequencing with the reference genome it is possible to detect the methylated sites. This approach has been used to investigate the consequences of LoF mutations in *NSD1*, a gene coding for a methyltransferase, causing an overgrowth syndromic ID (Sotos syndrome). The methyl-sequencing revealed specific genome-wide DNA methylation alterations in patients, which help to understand the pathophysiological mechanisms involved in this disease and propose a markers for diagnostic (Choufani et al., 2015).

NGS technologies have been used to study chromatin spatial organization at a whole genome level. The combination of the chromosome conformation capture (3C) and NGS revealed general features of genome organization like the presence of hierarchical chromatin structures, compartments, topologically associated domain (TAD), insulated domains and chromatin loops (Schmitt et al., 2016). Generally, cells are first cross-linked to retrieve the three-dimensional spatial proximity of genomic loci. The DNA will then be fragmented, usually by restriction enzymes, and then cross-links will be release to obtain genomic fragments with reshuffled according to their spatial proximity. First approaches (3 and 4C) were limited to analysis of the contact regions of only one genomic loci but the further development of other techniques enabled first, the parallel investigation of contacts between selected sequences (5C) and then eventually enabled the study at a genome-wide scale (HiC). For instance, a study using this technology revealed that CNVs of the 16p11.2 region – frequently associated to a syndromic ASD - lead to an alteration of the three-dimensional positioning of these genes, resulting in gene expression dysregulation involved in the clinical phenotype (Loviglio et al., 2017).

Overall, the integration of the generated data obtained with these NGS technologies (genome, transcriptome and epigenome) will lead to the understanding of functional relationships between chromatin organization, transcription regulation and genome function. Project such as ENCODE are paving the way toward this goal (https://www.encodeproject.org/).

7. IDENTIFICATION AND VALIDATION OF NOVEL VARIANTS OR GENES IN ID

The advancement of the NGS technologies enabled the identification of many variants present in an individual: in the TS around 2,000, WES about 60,000 and in the WGS even 4-5 millions of variants, and many of them are unique. The identification of the pathogenic mutation in monogenic form of ID is hindered by this huge amount of information, causing the classical "*needle in the haystack*" problem. Nevertheless, the improvement of prediction software tools and the collection of large-scale sequencing projects in the general population (detailed in *Variant Annotation*, pg.75) significantly facilitated the process of variant annotation and prioritization. These steps are important for the identification of candidate mutations that will be further analysed.

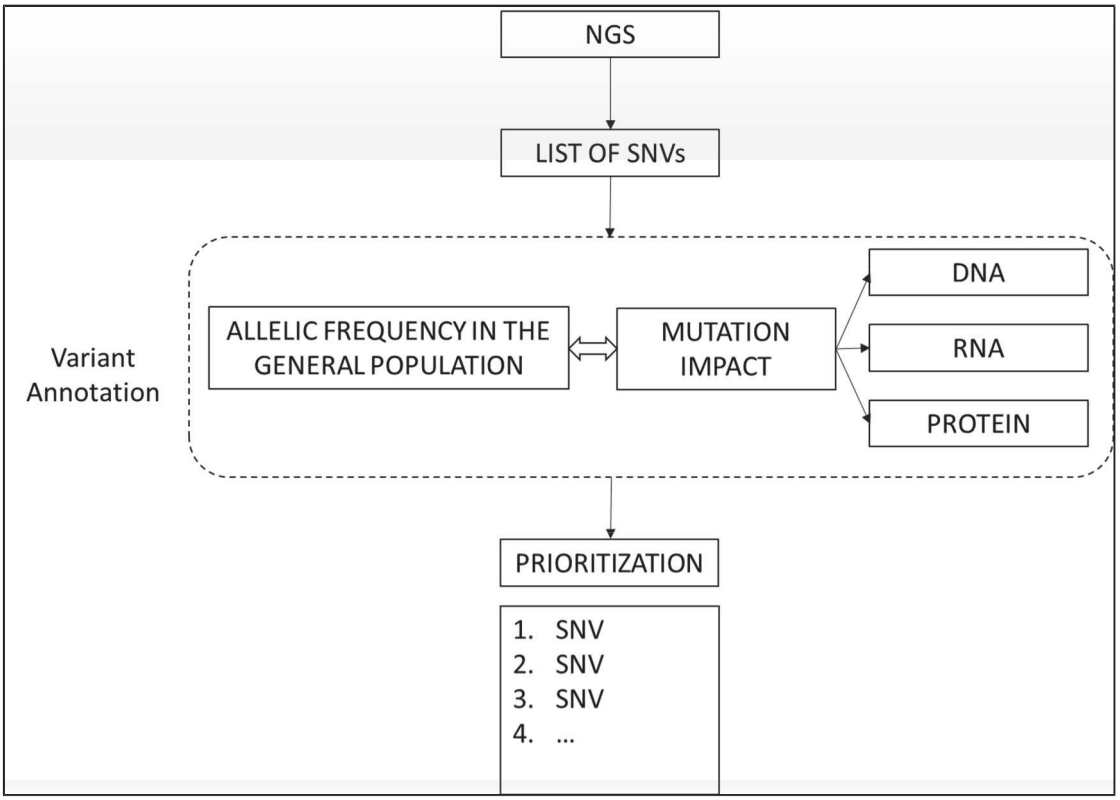


Figure 16. Schematic workflow of variant analysis and identification of candidate variants

As the list of variants is generated, it is crucial to retrieve the maximum number of information possible, including the description of the nature and the potential consequences of each variant (at the DNA, RNA and protein level), the presence of the variants in other affected individuals or unaffected individuals (Figure 16). According to these information, variants are then scored from the most predicted damaging effect (usually, the truncating ones: frameshift, nonsense and splicing) to the less ones (*i.e.* missense, synonymous) (see *Variant Prioritization*, pg. 79).

7.1 VARIANT INTERPRETATION

Variant interpretation comprises the combination of different type of information to conclude about the involvement of the variant in the pathology. Variants might be classified as responsible for the disease (pathogenic; likely-pathogenic according to the strength of this evidence), or not responsible (likely-benign, benign). In certain cases, it is not possible to conclude about variant's implication with the information available, and the variant is classified as variant of unknown significance (VUS). This could be due to several reasons, *e.g.* a missense change never been described before and/or far from the previously identified ones, or a variant in a non-coding region whose consequence is not clear, etc. Using large-scale approaches such as WES or WGS, one can identify a promising variant in a gene never been associated to ID. This gene is considered to be a gene of unknown significance (GUS) for the pathology, and further analysis need to be done to prove that when mutated it causes ID. In the following section I will described common strategies used to go further to interpret a VUS or GUS.

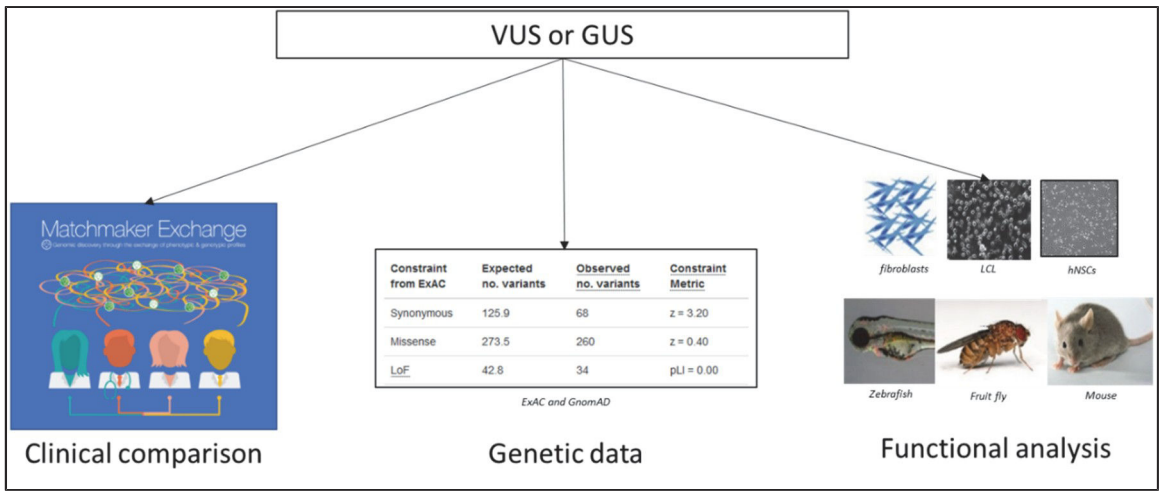


Figure 17: Schematic workflow for VUS or GUS implication in ID

7.1.1 CLINICAL COMPARISON

If the variant is in a gene already associated to ID, a clinical comparison with previously reported patients could further confirm the implication of a VUS, for example, if the phenotypes of the patients overlap. This could also be done for GUS if the gene of interest is included in a CNVs already associated to ID. However, the high genetic heterogeneity of ID has been a limiting factor for gene discovery, especially in sporadic cases, since usually only one mutation in a patient is not sufficient to validate the implication of a gene in a disease. For instance, it is estimated that even the most recurrent genes associated to ID explain less than 1% of the total cases. Thanks to the high-through put capacity of the NGS technologies, a large number of patients have been sequenced, increasing the chances to identify a mutation in the same gene in more than one patient. It is also crucial to exchange information among researchers, genetic counsellors and clinicians from all over the world. To this purpose, Matchmaker

Exchange (matcmakerexchange.org) was created to provide a connection among people that have identified potential mutations in the same genes (Philippakis et al., 2015; Sobreira et al., 2015). *MatchMaker Exchange* includes *Decipher* (https://decipher.sanger.ac.uk/), which is a database that collects and provides public data access to genotype-phenotype data from patients affected by neurodevelopmental disorders (Firth et al., 2009). It contains more CNVs than SNVs, even if the number of the latter category is constantly increasing. The most recent developed platform is *GeneMatcher* (https://www.genematcher.org/) that, as the name suggest, matches clinician and researchers interested in the same gene. The increasing use of this tool enabled a quick clinical comparison of patients with a mutation in a GUS (Table 8, pg.65). The identification of novel genes involved in ID has been facilitated by the emerging of such platforms (*e.g.* Decipher and GeneMatcher) that allows researchers and clinicians to share information on a patient with a potential deleterious variant in a gene never been implicated in ID.

7.1.2 GENETIC DATA

For both VUS and GUS a segregation analysis should be done on the available DNA from parents, siblings and other relatives to check is the variant is present in all affected individuals and absent from all unaffected individuals. As a matter of fact, a *de novo* event in a proband is usually in favour of the contribution of the variant to the phenotype; however, *de novo* variants are not necessarily deleterious. Additionally, a causative mutation for monogenic form of ID should not be frequently observed in the general population; to this purpose, variant databases such as ExAC, GnomAD, EVS and 1,000 Genomes (detailed in *Variant Database*, pg.78) are useful tools for variant interpretation. Furthermore, other variant databases collect variations identified in patients (*e.g.* ClinVar, https://www.ncbi.nlm.nih.gov/clinvar/), thus facilitating the interpretation of some VUS.

Genetic data from the general population could provide further information on a GUS. As a matter of fact, the large amount of data generated by *ExAC* enabled the calculation of a constraint metric for the four different classes of variants (synonymous, missense, loss of function and CNV) in each gene, indicating its intolerance to variation. These values represent the differences observed between the expected and the observed number of variants in a given gene. For synonymous and missense, the Z-score indicates the deviation of the observed counts from the expected ones, according to size of the coding sequence. The Z-score can be either negative, indicating that there are more variants than expected, or positive, revealing fewer variants than expected, hence an increase constraint of the gene. For LoF variants, another value is computed, the pLI, which represents the probability that a given gene is extremely intolerant to LoF variation, based on the number of LoF (splice and nonsense variants observed compared to what is expected. The more pLI is closer to 1, the more the gene is LoF

intolerant. A pLI above or equal to 0.9 reflects a gene that is extremely intolerant to LoF variation (Lek et al., 2016). Thus, these values may indicate if a gene could be potentially implicated in ID.

7.1.3 FUNCTIONAL ANALYSIS

Functional validation should be done on VUS as well as on GUS to confirm their pathogenicity. For a VUS in a known gene, the aim is to show that the variant affect protein function. For a GUS, we should also show that the affected protein function is linked to brain dysfunction. Functional experiments should be designed by considering several factors, among which the nature of the variant (*i.e.* loss- vs gain-of-function) and its subsequent predicted impact (*i.e.* mRNA and protein), and also on available models.

7.1.3.1 CELLULAR MODEL

The cellular model offers a relative easy and fast way to validate that a mutation has an effect on protein function. If patients' cells are available (fibroblasts or immortalized lymphoblastoid cells), experimental studies can be conducted directly on these cells. In this case, protein and/or mRNA analysis could be performed to test if the VUS or GUS affect the expression of the gene or the stability of the corresponding protein. However, some genes involved in ID are specifically expressed in brain, thus it is extremely difficult to obtain cells from the tissue of interest. To overcome this issue, different strategies can be used. A common strategy is to introduce the specific mutation into the cDNA of the transcript of interest (*i.e.* by site-directed mutagenesis) and then transfect human cells with both the wild-type and the mutant transcript, to overexpress the wild-type and mutant proteins and observe any difference in protein level or localization. Functional analysis at the protein or RNA level could then be performed. On the other hand, it is also possible to silence the gene of interest (if expressed in the cell model), for example by siRNA. This strategy can be used to check the consequences of a GUS where truncating mutations have been detected, hence also to prove a haploinsufficiency mechanism. Typically, the cellular model used are neuronal cells as they are a good model to characterize protein functionality and molecular mechanisms involved in ID. For example, neuroblastoma cells such as N2A and SH-SY5Y can be easily differentiated into neurons, and studies on the neurite outgrowth can be performed (*e.g.* numbers of neurites and their length). However, this strategy limits the functional study on a single type of cells (*i.e.* neurons or glias) and the candidate mutation must be introduced. Somatic cells from patients can be reprogrammed to induced pluripotent stem cells (iPSCs) and then differentiate in neurons. This approach is increasingly used due to its broad versatility to model brain disorders and also early brain developmental processes. Currently, three-dimensional modelling of human brain is being developed. Organoids are three-dimensional structures, comprising multiple cell types derived from patients' iPSCs, self-organized to recapitulate brain development. Organoids are a good model for brain development since they well mirror the cytoarchitectural structure as well as the different developmental phases, such as cell proliferation and neuronal migration. The use of these

three-dimensional human brain models is still at its infancy hence some technical problems must be solved, particularly the high variability in quality and brain regions among batches of organoids (Forsberg et al., 2018). Nevertheless, the establishments of new methods to control these variabilities would render organoids as a promising tool to well delineate the pathogenicity of a mutation during neurodevelopment.

7.1.3.2 IN-VIVO ANALYSIS

Animal models are extremely useful for a deeper understanding of the biological role of the variant and gene of interest and its implication in human diseases. Nowadays, the CRISPR/Cas9 technology enables the introduction of specific point mutations in target genes, offering the possibility to study the effect of specific variants. Many animal models have been developed over the years to study genes implicated in neurodevelopmental disorders. The model organism should be able to recapitulate the observed human phenotype, thus it must be chosen according to it. For example, in the zebrafish model it is difficult to perform cognitive analysis while it is possible in the mouse, hence the zebrafish is not a good model for a non-syndromic ID. On the other hand, it takes a long time to obtain a transgenic mouse while it is significantly shorter to genetically modify a fruit-fly or a zebrafish.

The presence in literature or in-house of animal models developed for particular genes helps in the interpretation of the pathogenicity of the variants and to filter out or highlight candidate novel genes involved in ID (Table 8). In this section I will only focus on the mouse, zebrafish and fruit fly models.

Gene	Inh.	Tot. Patient	Matchmaker	Model	Reference
DPF2	AD	8	+	Cellular	(Vasileiou et al., 2018)
RHOBTB2	AD	10	+	Drosophila	(Straub et al., 2018)
RLIM	XL	84	-	Zebrafish	(Frints et al., 2018)
MED13	AD	13	+	Patient cells	(Snijders Blok et al., 2018)
CAMK2A/B	AD	24	+	Mouse	(Küry et al., 2017)
WDR62	AD	15	+	Patient cells	(Skraban et al., 2017)
NAA15	AD	13	-	Drosophila	(Stessman et al., 2017)
CDK10	AR	9	-	Mouse	(Windpassinger et al., 2017)
PPM1D	AD	14	+	Patient cells	(Jansen et al., 2017)
BCLL1A	AD	9	-	Mouse	(Dias et al., 2016)

Table 8: Recent novel ID gene identified by NGS techniques and tool used for validation

7.1.3.2.1 Mus musculus

The establishment in mouse of techniques such as the in-utero electroporation allowed relatively fast studies of genes and human mutations during cortical development, which is one of the processes altered in ID. With this technique it is possible to check different mechanisms known to be involved in the origin of ID, such as the neuronal migration and proliferation, by either silencing the gene of interest or by overexpressing the wild-type and the mutated transcript.

The mouse is also a valuable model, as its genome is highly similar to the human one. As the development of genetic engineering techniques progressed, many mouse models have been developed recapitulating specific genetic human disorders, among which ID. For instance, the presence of a mouse model with a disrupted GUS presenting similar features to human patients it supports the contribution of this gene to the phenotype, as recently demonstrated by the generation of a conditional knock-out mouse for *Cdk10* that displays a similar phenotype observed in patients with mutations in the human paralogue gene affected by a syndromic form of ID with severe growth retardation and spine malformations, further assessing the implication of this gene at the origin of this disorder (Windpassinger et al., 2017). To this issue, the International Mouse Phenotyping Consortium (IMPC) is currently providing a catalogue of gene function by systematically generating and phenotyping each knockout strain for every gene in the mouse genome, obtained by the International Knockout Mouse Consortium (IKMC). Each genotype-phenotype analysed data are then available to the scientific community on the website http://www.mousephenotype.org/. The generation of these mouse models enabled a deeper characterization of the gene function with a focus in neuronal cells. Moreover, thanks to the improvement of the genetic engineering technologies, it is now possible to create knock-in mouse model to study the specific effect of a mutation.

With the mouse model is possible to perform cognitive and behavioural analysis that comprises relative basic functions, such as learning and memory, but also more complex ones, such as social and anxiety behaviours. Therefore, it offers the possibility to perform pharmacological studies and to observe the drug effect at a broader level, from cells to the behaviours. However, the mouse model generation takes a long time and its progeny is not highly numerous, requiring even years to arrive at a statistical significant number of samples.

The large amount of data obtained by the NGS is generating more candidate variants than they can be currently interpreted. It is hence required an animal model that can provide relevant answers to specific mutations in a short time.

7.1.3.2.2 Drosophila melanogaster

The commonly known fruit fly is an animal model extensively used in genetics, since its maintenance as well as the generation of fly mutants is easy, fast and cheap and flies are highly prolific. Despite the evolutionary distance between flies and humans, a strong conservation of genes is observed, with 75% of human disease genes having a related sequence in *Drosophila.* Starting from 2000s, researchers have begun to widely use *Drosophila* to investigate genes involved in ID. Indeed, the *Drosophila* brain is small but enough complex to be an appropriate model to study defects in neuronal morphology and function along with the possibility to assay for cognitive process (van der Voet et al., 2014). *Drosophila* is a useful model to study the role of genes in the brain organization and its nervous system, such as neurotransmitter release, axon growth, synapse formation and physiology. Furthermore, it offers the

possibility to test from relatively simple to more complex behaviours, ranging from fly to cognitive behaviours, such as learning and memory abilities. Over the years, these assays have demonstrated that they are valuable and reliable tool for a deeper investigation and characterization of genes involved in ID. For instance, it has been recently used for the delineation of the pathophysiological mechanisms of several missense variants in a novel ID-gene, *RHOBTB2*, suggesting a role in dendritic formation (Straub et al., 2018).

Its relatively easy gene manipulation allows high throughput studies, and it offers the possibility to check the pathogenicity of specific variants by over- and re-expressing either the human mutated gene or - if the affected residue or gene is conserved – the mutated fly gene. The fly can also be used to identify genetic modifiers as well as functionally related genes (Cukier et al., 2008; Schenck et al., 2003), further contributing to the better delineation of the molecular pathways and networks involved in ID.

7.1.3.2.3 Danio rerio

The zebrafish model has been widely used in embryogenesis and organ development studies, since its embryo transparency facilitates its observation and manipulation, allowing *in-vivo* visualization of cell and organ processes. The zebrafish genes have about 70% of orthologues in human and, due to its easy and cheap maintenance as well as its short generation time, it is now becoming a common tool also for genetic studies, particularly in neurodevelopmental disorders. However, the modelling has mainly focused on embryonic development and associated disorder comorbidities, such as the head size (Kozol et al., 2016). The genetic manipulations in zebrafish not only validate the pathogenicity of a candidate variant but also provide information on its effect (*i.e.* loss- or gain- of function), by combining reduction of gene expression (using Morpholino or CRISPR/Cas9) and injection of wild-type or mutant human transcript. Overall, the zebrafish model is an efficient tool to delineate the genetic mechanisms at the origin of a disease and it is also a useful model to better understand the molecular and cellular mechanisms that underlie behavioural phenotypes in developmental disorders and cause at the same time non-cognitive comorbidities traits, such as epilepsy and gastrointestinal distress. For example, *chd8* morphants have a reduced numbers of enteric neurons resulting in an impaired gut motility, which can be at the base of the gastrointestinal discomfort often reported in patients with a mutation in *CHD8* (Bernier et al., 2014). Due to its small size and large population, the zebrafish is increasingly used for high throughput drug screens, thanks also to the improvement and standardization of the high throughput techniques for behavioural screens, such as the swimming trackers software.

AIMS OF THE PROJECT

ID is a neurodevelopmental disorder that affects around 1% of the general population, representing a major public health and social problem. ID is an extremely genetic heterogeneous disorder, with single genetic events accounting for a large number of cases, ranging from chromosomal abnormalities to CNVs (which may affect several genes) to SNVs in single genes. The increasing use of NGS technologies in the research and in the clinical practice significantly helped the identification of the genetic causes of this disease and up to now more than several hundred of genes have been reported to be implicated in monogenic forms of ID. Nevertheless, there are still some genes that are not identified yet. Genetic investigations allow the identification of a large number of variants, not always easy to interpret, especially those falling in a gene never linked to any human disease. Moreover, for most of the monogenic forms of ID, little is known about the physiological mechanisms that can lead to brain dysfunction.

Therefore, the main goals of my PhD project are:

- To perform genetic investigation in patients with ID and/or ASD, in order to identify novel mutations and novel genes involved in monogenic forms of ID/ASD using NGS techniques (targeted, whole-exome and RNA sequencing).
- To prove the pathogenicity of the mutations identified and to confirm their involvement in ID, but also to study their consequences to understand the pathophysiological mechanisms involved.

PART 1: GENETIC INVESTIGATIONS IN PATIENTS WITH ID/ASD

MATERIALS AND METHODS
PART 1

1. PATIENTS RECRUITMENT

Patients with ID and ASD were recruited through the molecular genetic laboratory of Strasbourg. They were previously tested with routine genetic analysis, including X-fragile test, aCGHs. Clinical forms were obtained from the clinicians following the patients along with consent from parents. Ethical approval was obtained from the local ethics committees. Samples from patients, including blood or saliva, were collected at Strasbourg hospital along with the parental ones and other relatives if available. For some patients, fibroblasts and blood cells were also collected for subsequent analysis.

A cohort of 38 patients affected by ASD was included in this study, among which 9 females and 29 males. All patients were diagnosed with ASD by an autism diagnostic interview- revised (ADI-R). The majority of the probands are sporadic cases (~76%; 29/38), while 9 individuals belongs to multiplex families, with more than one relative affected with a similar phenotype (~24%).

Patients with ID who did not receive a molecular diagnosis after the targeted sequencing analysis were passed to the whole-exome sequencing. Patients with a highly syndromic form of ID or belonging to a family with more than one individuals affected by the same phenotype were directly passed to WES, as these characteristics suggest a genetic origin of ID. Overall, 29 males and 10 females were included in the cohort, for a total number of 39 patients sequenced by whole-exome sequencing. Individuals can be grouped according to different criteria: sex (*Males vs Females*); if they were previously passed through the targeted-sequencing or not (*Negative TS vs Direct WES*); and if they were sequenced alone or in parallel with their parents (*Solo vs Trio*) (Table 9).

	Males	**Females**
Solo	8	1
Trio	21	9
Negative TS	23	8
Direct WES	6	2
total	*29*	*10*

Table 9: Classification of the patients' cohort passed to the WES

Patients that did not receive a diagnosis after WES were passed either to whole-genome sequencing (not detailed in this manuscript) or to RNA-sequencing, if RNA was accessible. RNA from patients affected by ID or BBS was collected from 15 patients, 9 from fibroblast and 6 from whole blood. RNA from these patients was sequenced in two batches: in one batch samples were in duplicates while in the second one only one RNA sample per individual was sequenced.

2. DNA-SEQUENCING

2.1 LIBRARY PREPARATION AND SEQUENCING

DNA were extracted at the diagnostic laboratory (Strasbourg University Hospital) either from peripheral blood, using QIASymphony from Qiagen®, or from saliva extracts, using Oragene kits from DNAgenotek®. Upon their arrival, DNA integrity was verified on a 1% agarose gel by electrophoresis. Quantification and further quality analysis were performed using the Nanodrop®. Samples should have a 260/280 ratio above 1.8 and a 260/230 ratio above 1.7.

2.1.1 TARGETED-SEQUENCING

The DNA library preparation was performed using the SeqCap EZ Library from NimbleGen (Roche®). Briefly, 1 µg of genomic DNA from samples was fragmented using the Covaris E220 AFA sonicator, to obtain an average fragment size between 180 and 200bp. After the sonication, the ends of the gDNA fragments were repaired to produce blunt-ended fragments and subsequently adenylated at the 3'-end. Pre-captured dsDNA indexing adapters were ligated to the A-tailed library fragments and the indexed library was then amplified with primers complementary to the sequencing adaptors, so to amplify only the fragments carrying the appropriate adapter sequencing at both ends. The quality and quantity of the amplified DNA library was assessed on a DNA Chip 1000 on a Bioanalyzer (Agilent Technologies). Samples were mixed together at an equal quantity to obtain a multiple DNA samples library pool. The multiplexed DNA sample library pool was hybridized to the biotinylated long oligonucleotide designed probes provided by Roche NimbleGen. The hybridized DNA was then captured using streptavidin beads and further amplified by PCR using the primers complementary to sequencing adapters. The quality and quantity of the amplified, captured multiplex DNA sample was finally checked on a DNA Chip 1000 on a Bioanalyzer (Agilent Technologies).

The amplified, captured multiplex DNA library were then sequenced 0,25 per lane of the flowcell. Sequencing runs of 100bp paired-end were performed on a HiSeq4000 (Illumina).

The bioinformatics pipeline, variant annotation, prioritization, visualization, interpretation and validation were carried out as the ones used in the WES.

2.1.2 WHOLE-EXOME SEQUENCING

The DNA library preparation was performed at the GenomEast platform at the IGBMC starting from 1µg of genomic DNA using the SureSelect XT Human all exon V5 (Agilent Technologies). Samples were then sequenced 100bp paired-end, 4 per lane on a HiSeq2500 (until April 2016) or on a HiSeq4000 (Illumina).

2.2 BIOINFORMATIC PIPELINE

This bioinformatics pipeline was developed in-house by bioinformaticians of the IGBMC sequencing platform, mainly Stephanie Le Gras.

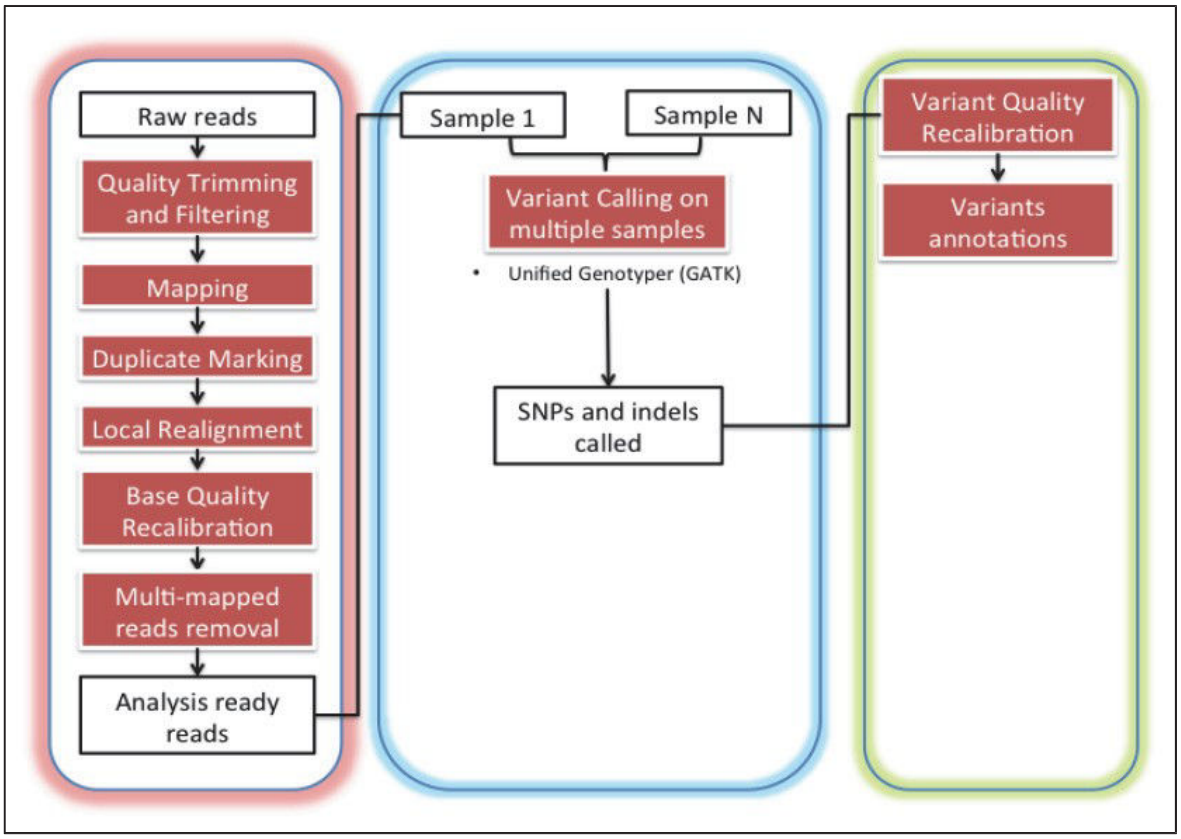


Figure 18: Bioinformatic pipeline used to detect SNVs and small indels developed by Stephanie Le Gras

Imaging analysis and base calling were performed using CASAVA v.1.8.2 (Illumina). Barcode and part of reads having a quality below 10 were discarded, so to avoid false positive variations. Reads were then mapped onto the reference genome hg19/CRCh37 using BWA v0.7.5a (Li and Durbin, 2009). Reads that shared the same genomic coordinates and sequence (called *duplicate reads*) were marked, so that they were not considered for further analysis, as they were most probably a PCR artefact, as the probability of having twice the same DNA fragment is extremely low. Reads were then locally re-aligned to improve the alignment quality, especially in the proximity of indels. A base quality score recalibration was performed to avoid systemic biases in quality score assignment from the sequencers. Finally, reads that mapped to several position in the genome (*multi-mapped reads*) were removed from the analysis using Samtools v0.1.19 (Li et al., 2009), as their right position is uncertain.

Once reads were prepared, the variant calling was done by using GATK 3.2-2 UnifiedGenotyper (DePristo et al., 2011), which allows to perform it on multiple samples, helping in the detection of low covered variants in a given sample but are observed in other samples of the same project. A last step of variant quality score recalibration was done to assign a well-calibrated probability to each variant call.

2.3 VARIANT ANNOTATION

Variant annotation was done via VaRank, a tool developed *in-house* (Geoffroy et al., 2015). VaRank provides an accurate nomenclature of the variations, as it uses the HGVS mutation nomenclature (http://www.hgvs.org/mutnomen/) and it calculates the frequency of each variant in the cohort of patients analysed, indicating its allelic frequency in individuals with the same disease. Moreover, VaRank computes also a family barcode, which enables for each variant a quick overview of its presence/absence and its zygosity status (represented as "0" for homozygous wild-type; "1" for heterozygous; "2" for homozygous for the variant allele), facilitating the analysis, especially in trio-family exomes. VaRank provides a score of the variants based on the variation type and the predicted impact, ranking the variants from the most likely to the less likely pathogenic. VaRank uses Alamut-HT (Alamut-High Throughput, Interactive Biosoftware) to integrate several information for each variant, such as: the putative effect of the variant at the protein and RNA level, conservation cues and the frequency in the main variant databases.

2.3.1 BIOINFORMATIC TOOLS AND DATABASES

To understand the pathogenicity of the identified variant, different bioinformatics prediction tools have been developed. These tools have different domains of predictions and all together provide an overview of the possible consequence of a variant. The software Alamut Visual gather all the different information on the analysed variant, including many prediction tools hence it is possible to have a general overview of the variant all at once. In the following sections I will detailed the bioinformatics tool and database used for the variant annotation, grouped on the predicted effect of the variant.

2.3.1.1 IMPACT AT THE DNA LEVEL

Two main tools are used for the prediction of the level of conservation of a specific nucleotide and they are both freely provided for the entire human genome by the UCSC browser.

PhyloP is based on the multiple alignments of 46 species and it take in consideration not only the single nucleotide conservation but also the one of the neighbouring bases. Then it calculates a probability score that indicates the confidence that the nucleotide belongs to a conserved region (Pollard et al., 2010). On the other hand, PhastCons considered only the single nucleotide and do not take in consideration the adjacent residues. The output value ranges from -14.1 to 6.4. The positive values indicate a slower evolution of the site than expected under neutral drift (*i.e.* conservation), while negative scores reflects a faster evolution than the expected one, thus a fast evolving site (Siepel et al., 2005).

2.3.1.2 IMPACT AT THE mRNA LEVEL

Alamut displays the result of five splice prediction software: MaxEntScan, NNSplice, HumanSplicingFinder, GeneSplicer and SpliceSiteFinder-like. However, in VaRank only three of them are computed, as two of them (namely GeneSplicer and SpliceSiteFinder) have been found to be

redundant with the results of the other prediction tools, mainly due to the use of a similar computational method.

MaxEntScan is based on the maximum entropy principle, whose parameters were estimated from known signal sequences (Yeo and Burge, 2004). *NNSplice* analyses the structure of donor and acceptor sites using a separate neural network recognizer for each site. Only genes with constraint consensus splice site were considered in the training set, specifically GT for donor splice sites and AG for acceptor splice sites (Reese et al., 1997). *HumanSpliceFinder* is based on position weight matrices with some position-dependent logic. It asses the strength of 5' and 3' splice sites and branch points (Desmet et al., 2009). *GeneSplicer* considers only a small regions around the splice junction (Pertea et al., 2001). It uses a combination of the maximal dependent decomposition, to cluster a group of aligned signal sequences into subgroups containing significant motifs (Lee et al., 2011), and a stochastic model that capture additional dependencies among neighbouring bases. *SpliceSiteFinder-like* is based on position weight matrices computed from a set of human constitutive exon/intron junctions for donor and acceptor sites.

All the software calculate a score that indicate the strength of the acceptor/donor splice site. Splice sites are considered to be affected when a decrease of more than 10% is observed in at least two different prediction software. These prediction tools are only used in the exon/intron junction defined by the cis-elements the 5'- and 3' splice site and branch sites. There are also cis-elements located in the coding-regions that contribute to the correct alternative splicing, by either stimulating (ESEs) or by repressing (ESSs) splicing. ESEs are generally recognized by a class of protein (the SR proteins) that promote the recruitment of the spliceosome to the correct position, hence stimulating the splicing. Alamut includes also some prediction tools for these elements, in particular ESEFinder and RESCUE-ESE. The *ESEFinder* method computes putative binding sites for ESEs, by using weight matrices corresponding to the motifs for four different human SR proteins (Cartegni et al., 2003). Similarly, in the computational/ experimental method *RESCUE-ESE* specific hexanucleotide sequences are identified as candidate ESEs (Fairbrother et al., 2004).

2.3.1.3 IMPACT AT THE PROTEIN LEVEL

To help in the interpretation of the pathogenicity of the missense variants, several prediction tools have been developed. In Alamut four of them are displayed: Sift, Align GVDV and PolyPhen2.

Sift prediction tool consists in the protein alignments of highly similar sequences and it is based on the principle that the more the amino acid position is conserved, the more important is its function in the protein (Kumar et al., 2009; Ng and Henikoff, 2003). The calculated scores range from 0 to 1 and the pathogenicity threshold is below 0.05. *Sift* returns also a median value that indicates the diversity of the sequences used in the alignment. It goes from 0 to 4.32, where the higher value designates that

the full set of used sequences is highly similar, thus caution must be taken in the pathogenicity score as all positions are considered as highly conserved.

Align-GVGD uses the protein multiple alignments, combined with the biophysical characteristic of amino acids. It calculates two types of conservation scores: the Grantham Variation (GV) and the Grantham Deviation (GD). The GV measures the degree of biochemical variation among the amino acids found at a given position, while the GD indicates the biochemical distance of the mutated amino acid from the wild-type one at a particular position (Mathe et al., 2006). Variants are then classified in seven classes from C0 to C65, according to their risk of pathogenicity. The higher the class the higher risk that the missense variant is deleterious.

PolyPhen-2 predicts the impact of a missense variant based on the structure and function of the protein. It utilizes the multiple alignment of homologous sequences combined with functional annotation and structural information if available (Adzhubei et al., 2010). To check if the protein function may be altered, *PolyPhen-2* uses the UniProtKB/Swiss-Prot dabaseses to verify if a specific protein feature is altered and - if a 3D structure is available- it also checks whether the variant is spatially close to one of these critical domains. For each variant, it calculates a Bayesan probability that the mutation is damaging and returns an estimation of the false positive rate, which is the chance that the mutation is classified as damaging when it is not. Variants are then classified in three different qualitative categories: benign, possibly damaging and probably damaging. Two different training datasets are used: HumDiv and HumVar. HumDiv is assembled with all damaging alleles with known effects on the molecular function causing Mendelian diseases from the UniProtKB databases, together with differences between human proteins and their closely related homologous, apparently not deleterious. The HumDiv model uses the 5% /10 % false positive rate as a threshold to determine the probably or possibly damaging impact of a variant. On the other hand, HumVar uses all human mutations causing a disease from the UniProtKB database in combination with common non-synonymous SNPs that are considered as non-pathogenic. The cut-off for the HumVar model is 10%/20%, a little bit higher than the one used in HumDiv.

Conversely to the other software, *MutationTaster* is an integrated tool that returns predictions for any DNA alterations, by collecting information from different biomedical databases and uses established analysis tools. Therefore, it provides several information at once, ranging from evolutionary conservation, splice site prediction, the NMD probability as well as loss of protein features and changes (Schwarz et al., 2010). In Alamut Visual, *MutationTaster* is considered as a prediction tool for missense variants. To predict the pathogenicity of a variant, *MutationTaster* uses a training set of known disease mutations from Human Gene Mutation Database (HGMD) and SNPs and indel polymorphisms from the 1,000 Genome Project. Then, the variant is classified in three Bayesian models: silent alteration, alterations affecting a single amino acid or alterations causing complex changes in the amino acid

sequence. The disease potential is scored as a probability value representing the confidence level of the prediction that is divided in two main categories: polymorphism and disease causing.

2.3.1.4 VARIANT DATABASES

The introduction of NGS technologies has led to a huge amount of human genetic information that are collected in different variant dabases.

The oldest polymorphism database is *dbSNP*, freely provided by the NCBI (https://www.ncbi.nlm.nih.gov/SNP/) and it collects SNVs as well as small indels. dbSNP is directly linked to ClinVar, a database that collects data from genomic variation and its relationship to human diseases, helping in discriminating SNPs from mutations. However, cautions should be taken since ClinVar is not always properly updated, hence some mutations may be listed as pathogenic when they are not and viceversa. Similarly, *1,000 Genomes* provides a large set of common human genetic variation from multiple populations obtained by whole-genome sequencing. The 1,000 Genomes project reported a total of 84.7 million of SNPs, 3.6 million of indels and also 60000 structural variants in more than 2000 of individuals from 26 different populations (1000 Genomes Project Consortium et al., 2015). This project was a pioneer in genome-sequencing of a large number of individuals and its initial goal was to identify the most genetic variants with frequencies of at least 1% in the population. These two variant databases are based on a general population which is not affected by a specific disorder. Over the years, as the NGS technologies started to be widely used in clinical laboratories, many genetic data have been accumulated, hence new databases were created collecting data from individuals affected by a specific disease. Since many of these disorders are not related to cognitive dysfunction, they were used as a control reference.

Exome Variant Server (*EVS*) collects data from more than 6,000 exomes from individuals with European American or African American origins. This server is provided by the National Health Lung and Blood Institute, and it collects data from individuals affected by cardiac, lung and metabolic phenotypes that can be considered as *bona fide* controls for patients affected by ID (http://evs.gs.washington.edu/EVS/). The EVS displays not only the allele frequencies but also the genotype of the individuals, which is extremely helpful when analysing X-linked genes.

In the late 2014, the *Exome Aggregation Consortium* (*ExAC*) database was released (http://exac.broadinstitute.org/). The aim of this consortium is to gather the largest amount and to harmonize all the exome-sequencing data obtained from several whole-exome sequencing projects, making it available to the biomedical scientific community. All the collected raw data from the different projects were reanalysed using the same pipeline and jointly variant-called to increase the consistency across them. In the ExAC browser, the average coverage sequencing of a specific gene is showed, followed by a list of all identified variants. It is also possible to visualize on a genome browser the generated reads of a specific variant in order to discriminate a false variant calling. Moreover, a

summary statistic with the allele frequency, the allele count and the number of homozygotes and hemizygotes is reported, and they are reported as a total as well as divided in the different population groups. Currently, there are data from over 60,000 unrelated individuals that are part of a specific disease cohort, except for severe paediatric disorders.

The natural evolution of ExAC led to the release in the early 2017 of the *Genome Aggregation Database* (*gnomAD*) that contains both exome-sequencing data from more than 120,000 individuals as well as whole-genome sequencing from 15,496 unrelated individuals (http://gnomad.broadinstitute.org/). As in ExAC, all the raw data were re-analysed with the same pipeline and the same variant-calling protocol. Even in this dataset, patients with a severe paediatric disorder and their first-degree relatives were not considered. The gnomAD browser is like the ExAC one, and it combines information both from the genome and from the exome callset. When looking to a gene, the coverage obtained both from the exome and from the genome are showed, with a list of all the reported variants and the corresponding NGS technology that identified it. When a variant is present in both dataset, it has a combined summary statistic (that includes the allele count and frequency, the number of homozygous and hemizygotes), but it is possible to select which data to display. The variant reads on the genome browser are still displayed.

2.4 VARIANT PRIORITAZION

The tool used for the variant annotation – *VaRank* - ranks the variants from the most likely to the less likely pathogenic, according to the variation type and the coding effect. Known mutation are the first ones in the list, followed by truncating variants (*i.e.* nonsense, frameshift), essential splice site, start and stop loss, intron-exon boundary (donor site is -3 to +6; acceptor site -12 to +2), missense, in-frame, deep intronic changes and synonymous. Each variant score is then adjusted according to the additional information coming from different prediction tools: a +5 is added if the conservation at the genomic level is high and a +10 for each deleterious prediction to missense variants (Geoffroy et al., 2015).

Once variants are ranked, they are then filtered according to different criteria.

One of the first criteria was based on the read quality of the variant, so to avoid false positive results. The corresponding base should be covered at least by 10 sequencing reads, in which the variant allele should be seen in a minimum of 15% of all reads. Moreover, the allele read variant must be seen in at least two reads to be considered. However, this filtering criteria may lead to negative false variants, especially in poorly covered regions.

Variants were filtered according to the inheritance scenario and hence to a compatible frequency in the general population. For an X-linked variant, the family barcode was set so that the male proband resulted as homozygous, his mother heterozygous while absent in the father. The variant frequency in

Variant Category	VaRank Score	Definitions
Known mutation	110	Known mutation as annotated by HGMD and/or dbSNP (rsClinicalSignificance="pathogenic/probable-pathogenic")
Nonsense	100, 105	A single-base substitution in DNA resulting in a STOP codon (TGA, TAA or TAG).
Frameshift	100	Exonic insertion/deletion of a non-multiple of 3bp resulting often in a premature stop in the reading frame of the gene.
Essential splice site	90, 95	Mutation in one of the canonical splice sites resulting in a significant effect on splicing (at least 2 out of the 3 programs indicate a relative variation in their score compared to the wild type sequence).
Start loss	80, 85	Mutation leading to the loss of the initiation codon (Met).
Stop loss	80, 85	Mutation leading to the loss of the STOP codon.
Intron-exon boundary	70, 75	Mutation outside of the canonical splice sites (donor site is -3 to +6', acceptor site -12 to +2) resulting in a significant effect on splicing (at least 2 out of the 3 programs indicate a relative variation in their score compared to the wild type sequence).
Missense	50, 55, 60, 65, 70, 75	A single-base substitution in DNA not resulting in a change in the amino acid.
Indel in-frame	40	Exonic insertion/deletion of a multiple of 3bp.
Deep intron-exon boundary	25, 30	Intronic mutation resulting in a significant effect on splicing (at least 2 out of the 3 programs indicate a relative variation in their score compared to the wild type sequence).
Synonymous coding	10, 15	A single-base substitution in DNA not resulting in a change in the amino acid.

Table 10: VaRank scoring criteria (Adapted from the VaRank Manual)

the sequenced cohort was set as no more than 1 homozygous, as we did not expect to see the same variant more than once. The frequency allele in the general population should not exceed 2 males and the EVS number of total cases below 10 was used for filtering. The number of hemizygous individuals in the gnomAD or ExAC database was then checked manually on the few obtained variants.

In the autosomal dominant *de novo* scenario, only the unique allele variants present in the patient were considered, thus both the family barcode and the cohort allele count were set accordingly. Only rare variants were considered, hence variants present in the ExAC general population more than 1% were filtered out. Usually, the remaining variants were still too many to be analysed, thus I restricted the analysis to variants not reported in ExAC.

For the autosomal recessive variants, the ExAC allele frequencies was set below 0.45%, so with a frequency of homozygotes less than 0.002%. Then, for the homozygous variants, the family barcode and the frequencies in the sequenced cohort were set in order to not have more than the variants present in the family. For the variants at the compound heterozygote state, the ones that were present more than 10 times in the sequenced cohort population and the ones in the same parent were filtered out. On the other hand, the analysis of the compound heterozygote variant was not possible for the proband sequenced alone.

The best candidate variants were then further analysed in the next steps.

2.5 DATA VISUALIZATION

Even if a quality filter is applied, the quality of the candidate variants is manually checked in a genome browser, like IGV (Integrated Genome Viewer, Broad Institute). The variant is visualized in parallel with other unrelated samples and, if available, the parents. This step gives a general overview of the

sequencing quality of the region and whether this variant is also observed in another samples (though it is not annotated) or not.

2.6 VARIANT INTERPRETATION

Once the filtering step is done, usually a limited number of variants is obtained. At this step, it was assessed if they could be implicated in monogenic ID or not. Variant interpretation comprises the collection of all the evidences supporting the pathogenicity (or not) of a variant in a specific gene and their connection to a specific phenotype. To do that, different and several data must be combined.

First, it was checked if the gene had been already reported in ID or in a related disorder. Alamut provides a direct link to OMIM (https://omim.org/), which is an online catalogue of all human genes implicated in monogenic disorders. OMIM provides several information on a gene, among which its implication in specific disease. Moreover, it describes all the clinical features associated to a disorder, enabling the users to check for a specific symptom, in our case ID.

The expression and the function of the gene was also checked in several databases, such as *UniProt* (http://www.uniprot.org/) and *Human Protein Atlas* (https://www.proteinatlas.org/). Variants in genes that appear to be not related to ID - *e.g.* genes not expressed in brain - were noted as likely benign.

For each prioritized variant, an extensive bibliography was done to look for possible reported patients (*i.e.* patients with a CNVs that encompass the gene of interest), previous studies on animal models, and earlier investigation to verify if the mutated gene may be implicated in ID.

When a variant was identified in a gene that never been implicated in ID (GUS) or its effect is not so well understood, the variant is reported as Variant of Unknown Significance (VUS), until other evidences are gathered. GUS were uploaded on the *MatchMaker platform* (including *Decipher* and *GeneMatcher*). If a match was not find at the moment of the submission, the genes will continue to be queried by new entries (Sobreira et al., 2015).

2.7 VALIDATION AND DIAGNOSIS

Once a candidate variant(s) is highlighted, it is validated by Sanger sequencing. In parallel, a co-segregation analysis was performed.

If the variant co-segregates with the disease-status in the family and either it was a missense previously described or a truncating variant in a known ID gene it was classified as certainly pathogenic mutation. All certainly pathogenic mutations were then transmitted to the diagnostic laboratory of Strasbourg University Hospital, to ascertain official diagnosis reports to the clinicians that they relayed the information to the patients and their families.

All the other variants with potentially deleterious effect were considered as likely-pathogenic, waiting to be further confirmed or excluded, by performing additional co-segregation analysis or further

studies. In particular for variants listed as likely-pathogenic in genes never been implicate in ID further functional and characterization analysis were carried out.

3. RNA-SEQUENCING

3.1 RNA LIBRARY PREPARATION AND SEQUENCING

RNA samples were extracted either from blood by PAXgene RNA kit (Qiagen®) from blood at the diagnostic laboratory (Strasbourg University Hospital) or from fibroblast using TRI reagent® (Molecular Research Center) or using the RNeasy Mini Kit (Qiagen®). Both protocols included an additional step of DNAse I recombinant treatment (Sigma-Aldrich®). The integrity of the RNA was visualized on a 1% bleach agarose gel by electrophoresis (Aranda et al., 2012). Quantification and further quality analysis were performed using the Nanodrop®. Samples should have a 260/280 ratio around 2 and a 260/230 ratio above 1.7. The integrity and quality of the RNA were also evaluated by running samples on a RNA 6000 Nano Chip on the Bioanalyzer (Agilent Technologies) and samples should have a RNA integrity number (RIN) equal or above 8.

Library preparation was performed at the GenomEast platform at the IGBMC, using the TruSeq® RNA sample preparation v2 protocol (Illumina) starting from 1µg of extracted total RNA.

Libraries were then 2x100bp paired-end sequenced, 2 samples per lane on an Illumina Hiseq4000 sequencer. For the transcriptomic analysis on patients with the same mutated gene, libraries were sequenced 2x100bp paired-end sequenced, 4 samples per lane on an Illumina Hiseq4000 sequencer.

3.2 BIOINFORMATIC PIPELINE

This bioinformatic pipeline was developed and currently curated by bioinformaticians at the IGBMC sequencing platform, mainly Céline Keime and Damien Plassard, who ran it.

Image analysis and base calling were performed using CASAVA v1.8.2 (Illumina). Sequence reads were mapped onto the reference genome hg19/CRCh37 using Tophat 2.0.14 (Kim et al., 2013) and bowtie version 2-2.1.0 (Langmead et al., 2009). Only uniquely mapped reads were retained for further analysis, to avoid PCR artefacts and only non-ambiguously assigned reads were considered for further analysis. Three different type of analysis were carried out on the generated data and the bioinformatics pipeline will be detailed in the proper section.

3.2.1 DIFFERENTIAL EXPRESSION

Comparisons of expression level were performed using the statistical method proposed by Anders and Huber implemented in the DESeq v.1.6.1 Bioconductor package (Anders et al., 2015). Once the normalization was validated, the generated data were explored and visualized to assess and check data quality and to eventually remove bad quality data. Variance was then stabilized using the regularized log transformation method (Love et al., 2014). The Wald test was used to estimate the p-values that were subsequently adjusted for multiple testing with the Benjamini and Hochberg method (Benjamini and Hochberg, 1995). Results of the statistical analysis were plotted on scattered and volcano plots.

Genes were filtered based on their log2 fold-change, which should be above 0.6 or below -0.6, to detect respectively up- and down-regulated genes. As using only the fold-change does not control the

false positive rate, another filter was added based on the adjusted p-value that should be statistical significant, thus below 0.05.

Up- and down-regulated genes were analysed on the functional annotation tool of the Database for Annotation, Visualization and Integrated Discovery v6.8 (DAVID) and on the Ingenuity Pathways Analysis (IPA®, Qiagen®). The common aim of these software is to convert large gene list into biologically meaningful modules to understand how genes are connected to each other and to the functional annotation. Their common strategy is to link the provided gene list to the associated biological annotation and then statistically highlight the most overrepresented ones (Huang et al., 2009).

DAVID is a bioinformatic resource available online since 2003 (https://david.ncifcrf.gov/home.jsp) that consists of an integrated biological knowledgebase and analytical tool. The user can upload a gene list and select the genome of reference, which is then considered as the background. The genes in the list are mapped to the relevant biological annotation in the DAVID database, which integrates in a non-redundant way more than 40 publicly available annotation categories. With the gene functional clustering function, genes with a related biological or cellular role are grouped together, enabling the users to explore the larger biological networks. For each cluster, the enrichment score is calculated and each enriched term functional annotations are showed with its gene counting and its proportion relative to the background, their fold enrichment and their enrichment adjusted p-values.

IPA is a software developed by Qiagen Bioinformatics that is based on the manually curated Ingenuity Knowledge Base and, similarly to DAVID, it identifies the most significant pathways present in the gene list. The user uploads the list of genes with their associated fold-change and IPA returns the altered pathways. For each pathway, a table reports the genes implicated and their expression values and predictions. Moreover, IPA draws potential affected signal cascades and protein networks.

3.2.2 SPLICING ANALYSIS

To detect alteration in the alternative splicing among patients, three different tools were used: JunctionSeq, rMATS and LeafCutter. According to the approach, different filtering criteria were used.

JunctionSeq is part of the Bioconductor package and it tests for the differential exon and splice junction usage, which consists in the differential expression of a particular sub-unit of a gene relative to the whole gene expression (Hartley and Mullikin, 2016). JunctionSeq detects only skipped exon events. However, it is designed to consider replicates and it enables the detection of novel splice site, since an additional isoform assembly step is not required. For each sample, the number of reads mapping to each exon are counted and they are compared to the number of reads mapping to any other exons of the gene. The ratio of these two counts indicates the relative exon usage and its changes are inferred from the differences across the conditions. JunctionSeq also provides an automated visualization of the expression profiles that helps in their interpretation. Once data were generated, exons and splice-

junctions with a mean dispersion above 0.05 were filtered out to avoid false positive events. The remained data were then ranked based on the adjusted p-value that was set to be at least less than 0.01. The absolute value of the log2 fold-change –which indicates the differential expression of an exon- should be higher than 3, at a first analysis, or 0.6, during a second investigation.

rMATS, which stands for replicate Multivariate Analysis Alternative Splicing, also identifies novel differential alternative splicing from replicate RNA-sequencing data. In addition to the other tools, it detects several alternative splicing: skipped exon, 5' and 3' alternative splice sites, mutually exclusive exons and retained introns events. For a skipped exon event, rMATS estimates the exon inclusion level by counting the reads specific to the exon inclusion isoform –which are the reads from the upstream splice junction, the alternative exon itself and the downstream splice junction- and the count of reads specific to the exon skipping isoforms – which are the reads that directly connects the upstream exons to the downstream exons. The other types of alternative splicing events can be similarly modelled with this framework (Shen et al., 2014). The filter criteria applied to the obtained data were very stringent, to retrieve a reasonable amount of data that could be manually checked. At the beginning, only data with a p-value corrected for multiple testing below the order of E-10 were considered. The differences of the average ratio of the exon inclusion transcripts among exon inclusion transcript between the two conditions were considered if the absolute value was more than 0.05. The threshold of the likely-hood ratio estimated by rMATS as the p-value of these differences was set at an absolute value of 5%.

LeafCutter is a relatively new tool for detecting alternative splicing events. While the other approaches focus on the isoform ratio or exon inclusion levels, LeafCutter focuses on intron excisions to identify and quantify known and novel alternative splicing events, such as exon skipping, 5' and 3' alternative splice-site and additional complex events, apart from the intronic retention ones, since no split-reads are present in this scenario. LeafCutter does not require read assembly or inference of isoforms. It uses split reads to detect alternatively excised introns by connecting overlapping introns demarcated by split reads into clusters, which represent alternative intron excision events. Rarely used introns are then removed based on the ratio of reads supporting a given intron compared with other introns in the same cluster. To infer the differential splicing among conditions, LeafCutter compares the counts from the clustering step between the defined conditions (Li et al., 2018). Variants were filtered so that the p-value adjusted with the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) are below 0.01. The retained absolute value of the difference in the usage proportion among the two conditions was above 0.2 and the absolute value of the log effect size was considered if above 0.6.

3.2.3 MONOALLELIC EXPRESSION

Variant calling has been performed as recommended using the GATK workflow for SNP and indel calling on RNA-sequencing data, followed by several filtering steps to avoid as many false positives as possible. In the next future, variants obtained by RNA-sequencing will be compared with the heterozygous

variants identified by TS or WES in genes with a sufficient expression level in the RNA-sequencing. Variants will be annotated and ranked from the most likely- to the less likely- pathogenic by VaRank. Thanks to the family-barcode, only variants reported as heterozygous in the WES but homozygous either for the alternative or the reference allele in the RNA-sequencing will be analysed. The loss of heterozygosity of a genomic variant at the mRNA level could also be the result of a NMD mechanism that could correlate to a decreased gene expression level, which could be subsequentially checked.

3.3 VARIANT VALIDATION

If a differential gene expression level is identified, this can be confirmed by an RT-qPCR. Database (*i.e.* GTEx) and literature could also be reviewed to check if an alteration of this gene expression is common or it has been already related to a disease. If the gene is a good candidate, the genomic regions involved in gene expression regulation (*i.e.* promoter, 5' and 3' UTR) could be amplified and Sanger sequenced to detect any potential pathogenic variant that could affect RNA expression or stability.

In case of the identification of an effect on the alternative splicing, all the different isoforms reported for this gene are listed, to check its potential pathogenicity. The genomic DNA could then be amplified to retrieve the causative variant. As for any other genomic DNA variant, if this variation occurs in a gene already implicated in ID, clinical data could be compared to the already described phenotype to check if they overlap. Otherwise, if the candidate mutation is in a gene never been implicated in ID, we looked for other variants in this gene to replicate the results via MatchMaker Exchange.

To further verify the mono-allelic expression, an RT followed by Sanger sequencing could be performed.

RESULTS
PART 1

I performed genetic investigations on patients affected by ID/ASD combining different NGS technology. Patients with an unexplained form of ID or ASD were sequenced by TS. If no candidate mutation were identified in 220-275 known ID genes, some of these patients have been sequenced by WES. This strategy increases the chance to identify new genetic causes of ID. Some additional patients with a highly syndromic form or coming from multiplex families were passed directly to WES. If still no candidate mutation was identified, we passed some of these patients to whole-genome sequencing (not detailed in this manuscript). Another strategy was to sequence the transcriptome of patients either with a pathogenic variants affected splicing or mRNA level (positive control), variant of unknown significance (VUS) or with no potential candidate variant. The goal of this analysis was to evaluate the efficiency of the RNA-sequencing as a complementary tool for the diagnosis of genetically heterogeneous disorder, such as ID.

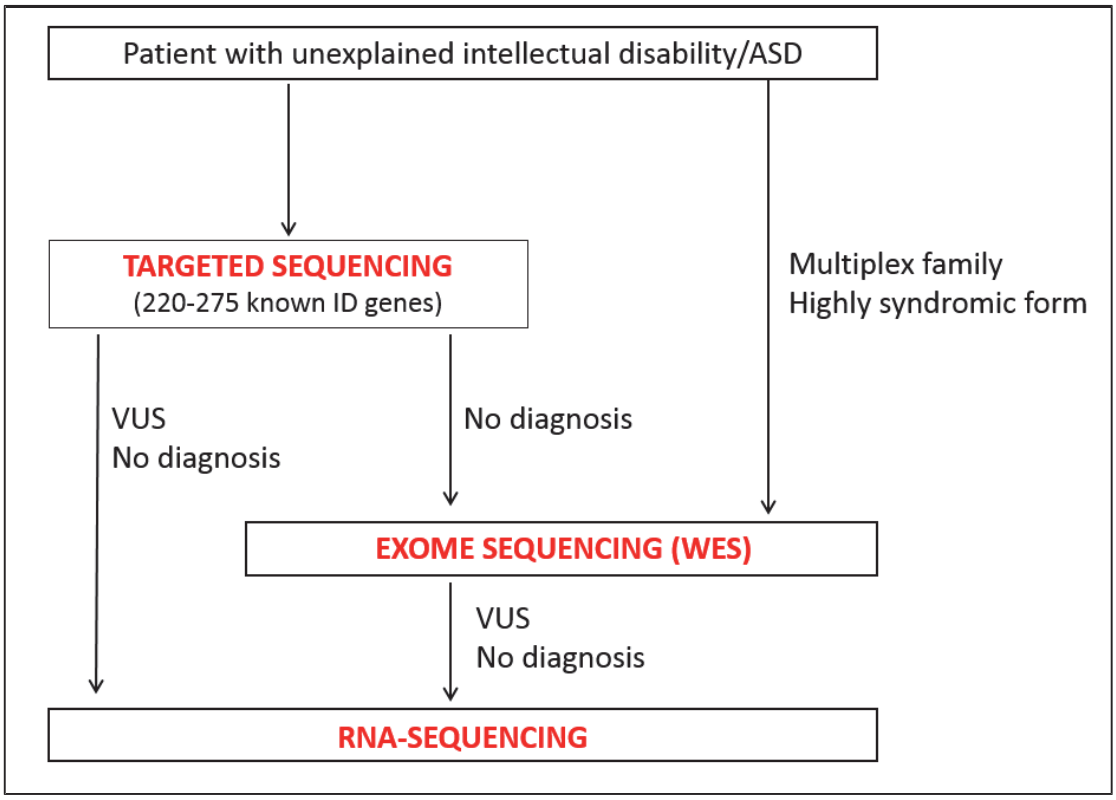


Figure 19: NGS strategy used for genetic investigations in patients with ID/ASD

1. TARGETED SEQUENCING IN PATIENTS WITH ID/ASD

When I joined the lab, it has been developed a targeted sequencing on 220-275 genes (depending on the version of the gene panel) that, on an initial cohort of 248 patients previously negative for routine genetic analysis, gave a diagnosis yield of about 25% (Redin et al., 2014), comparable to other TS studies with a similar number of genes in the gene panel (Figure 15, pg.56). The great advantages of the TS are the time of processing and analysing data, and the extremely low-cost compared to other NGS techniques that enables the inclusion of a major number of patients. To further confirm the efficiency of the developed targeted sequencing workflow, a larger cohort of patients has been recruited and tested at the diagnostic laboratory of the Strasbourg University Hospital, for a total of about 1,500 individuals. I participated to these genetic investigations on some of the ID patients at the beginning of my PhD project. During the genetic analysis, I identified several novel variants in known ID-gene. The identification of these variants enlarged and better delineated the phenotype caused by alterations in specific genes, such as *TCF4* or *ZBTB20* (Mary et al., 2018; Mattioli et al., 2016) (*Appendix 1 and 2*, respectively pg.166 and pg.167).

Even if it has been showed a genetic overlap between ID and ASD and monogenic forms of ASD have been described, their contribution to ASD seems to be lower compared to the ID ones, and the genomic architecture of ASD is consider more complex, due to a large interplay between common and rare variants as well as environmental factors (previously described in the section *Genetic Overlap Between Neurodevelopmental Disorders*, pg. 32).

ASD patients can be classified in several groups: patients with ID vs patients without ID, with syndromic vs non-syndromic ASD, and girls and boys affected. We sequenced patients mainly affected by ASD using the same TS strategy and gene panel, in order to compare the two diagnostic rates and to understand if one group of ASD-patients presents a higher rate of causative mutation.

1.1 RESULTS

The obtained sequencing data were of good quality, with 95% of the selected exonic regions covered more than 30x in all patients. On average, 2,000 SNVs were identified per individual that were annotated using Varank. After the filtering step, usually less than 10 variants were highlighted per patient, even if this number was highly variable among individuals.

Overall, five certainly-pathogenic variants were identified in sporadic cases. All of them are *de novo* and are in genes implicated in autosomal dominant form of ID often associated to ASD (Table 11). Three of them are truncating mutations, while one affects splicing and the remaining one is a missense change.

Gender	Cohort (n=38)	Positive diagnosis			
		XLID	ADID	ARID	Tot. Mutations
Male	29	(1)	3		3 (1)
Female	9		2		2
tot	*38*	*0*	*5*	*0*	*5*

Table 11: Classification of the identified mutation in TS

In a boy affected by ASD but no ID, a *de novo* frameshift mutation was identified in *ZNF292* (Figure 20B), a gene whose implication in ID and ASD is still not confirmed. In the cohort of ID patients analyzed at the diagnostic laboratory and through GeneMatcher, we have been in contact with several clinicians who mainly identified *de novo* truncating variants in patients sharing a similar phenotype. In total, 25 cases have been identified and clinical features are currently being collected by MD M. Mirzaa at Seattle Children's Hospital, USA.

In a boy affected by ASD and mild ID, we identified a frameshift in *SETD5* (Figure 20A), a well-known gene causing ID by loss-of-function mutations. Interestingly, in the first paper that linked this gene to an autosomal dominant form of ID, 5 out of the total 7 patients present also ritualized behaviour and/or autism in addition to ID (Grozeva et al., 2014). Another frameshift mutation was identified in *FOXP1* in a boy affected by ASD and ID, with an evocative fragile X-syndrome phenotype (Figure 20C). *FOXP1* is a well described gene implicated in ID with language impairment, with or without ASD. Moreover, a variable degree of ID has been also reported.

A mutation predicted to affect a donor site was identified in *SOX5* in a girl affected by ASD and mild ID (Figure 20E). This mutation is predicted to lead to an exclusion of the entire exon or to the usage of an alternative donor site. An investigation on the RNA of the proband is required to delineate the exact effect of the mutation. Several patients have been reported with deletions encompassing *SOX5* but only recently a *de novo* nonsense mutation in *SOX5* was identified, further confirming the implication of this gene in the previously described haploinsufficiency syndrome, which includes ID, language and motor impairment and facial dysmorphisms (Nesbitt et al., 2015) but no ASD was described. Nevertheless, we have been in contact with Dr. C. Depienne who identified and collected several patients with ID and/or ASD.

In a girl affected by ASD and ID, we detected a missense mutation in *SYNGAP1* (Figure 20D), a well-characterized gene known to cause ID with ASD. The missense is in a highly conserved amino acid and moderately conserved nucleotide (PhyloP: 2.79; PhastCons: 1). The mutation is predicted to be pathogenic by two different software (SIFT: Deleterious; PolyPhen-2: Probably Damaging) and it has a

large physiochemical difference between the wild-type and the substituted amino acid (Grantham Score: 155). Furthermore, the missense variant is in an important functional domain of the protein.

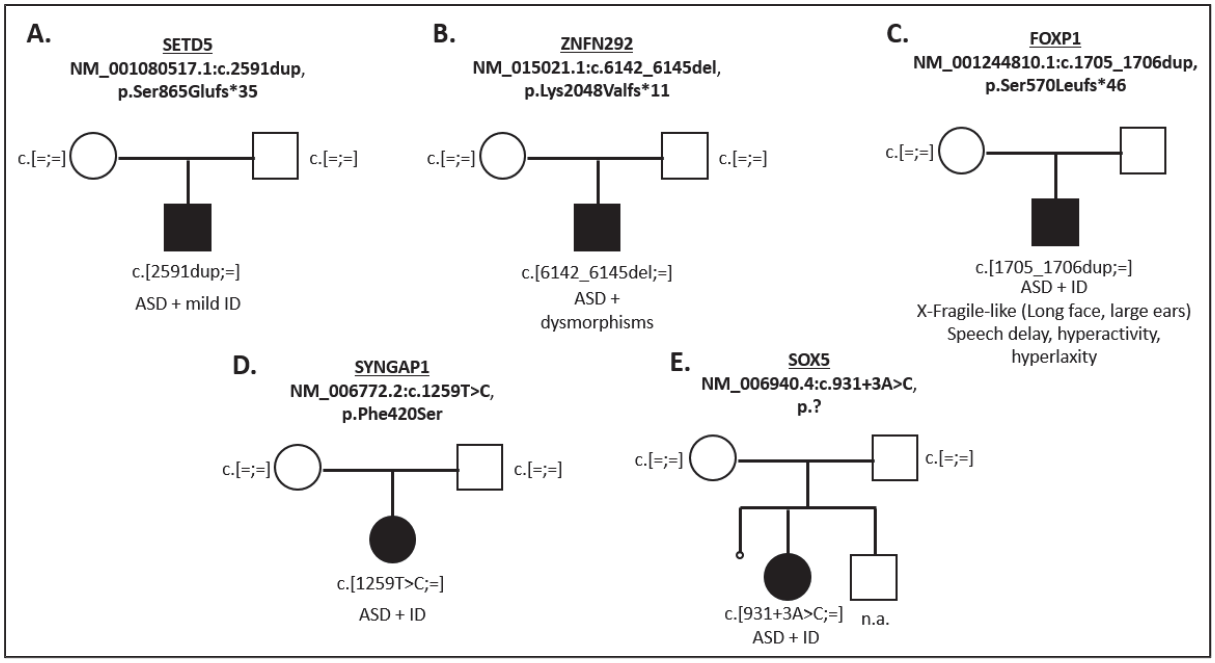


Figure 20: Mutations identified in the ASD cohort

Moreover, a missense variant was identified in *ARHGEF9*. This mutation affects a highly conserved amino acid that have been already reported as mutated in a patient affected by ID and seizure (Lemke et al., 2012). Mutations in *ARHGEF9* have been reported in several patients affected by epileptic encephalopathy, which is coherent with its function as a brain-specific guanine nucleotide exchange factor. The missense mutation identified in our patient is transmitted from the heterozygote mother (*ARHGEF9* is on the chromosome X) and a deeper segregation analysis is needed to eventually conclude on the variant pathogenicity. Interestingly, two other missense variants have been identified in other males from the cohort. This gene does not well tolerate missense variants (ExAC z-score: -2.97; 47 observed variants/ 110.9 expected variants). On the other hand, no putative deleterious variants in this gene were identified in previous WES on ASD patients (De Rubeis et al., 2014; Iossifov et al., 2014; O'Roak et al., 2014; Sanders et al., 2012). Further investigations are required to understand if missense variants in this gene could be predisposing factors for ASD.

	NM_001173479.1	NM_015185.2	NM_015185.2
cDNA	**c.16G>A**	**c.868C>T**	**c.1453G>A**
Protein	**p.Gly6Ser**	**p.Arg290Cys**	**p.Gly484Ser**
PhyloP	2.79	1.98	5.05
SIFT	Tolerated (score: 0.73)	Deleterious (score: 0)	Tolerated (score: 0.57)
Mutation Taster	n.d.	Disease causing (p-value: 1)	Disease causing (p-value: 1)
PolyPhen-2	Benign (HumVar: 0.004)	Probably damaging (HumVar: 1)	Benign (HumVar: 0.005)
Grantham Score	56	180	56

Table 12: Prediction scores for the detected missense variants in ARHGEF9

1.2 DISCUSSION AND CONCLUSION

Overall, we obtained a diagnostic yield of 13 (16% if we count the p.Arg290Cys in ARHGEF9) that is twice lower compared to the one obtained in cohorts of patients mainly affected by ID (usually around 25%) (Redin et al., 2014), but still significant. The majority of the patients for which a certainly-pathogenic variant was identified were affected by a mild or borderline ID and they were all sporadic cases and surprisingly no mutations has been detected in the familial cases. The obtained diagnostic yield is higher than the one obtained by another similar study (13% vs 3.7%) (Chérot et al., 2018). This difference could be explained by a different patient's recruitment: for instance, Chérot et al. considered in the ASD group only children meeting the criteria of autism diagnosis interview (ADI) without any early delay development, while in our study we considered patients diagnosed with stringent ASD criteria but we did not exclude patients for the presence of other comorbidity. Indeed, a stratification of ASD patients based on their clinical morphological categorization in a trio-WES previously showed that ASD individuals with minor physical anomalies have a low yield of diagnosis (3.1%; 2/64) compared to the ones with more (respectively, 28.6% in the intermediate group and 16.7% in patients with more complex morphological phenotypes) (Tammimies et al., 2015). Moreover, most of the patients for whom a certainly-pathogenic variant was identified were affected by a mild or borderline ID (4/5); similarly, one out of the 2 diagnosed patients in the study of Chérot et al. (2018) is also affected by mild ID. However, both studies have small cohorts (less than 55 individuals) hence it is really difficult to conclude that ASD with ID or other features are majorly caused by a single genetic cause compared to ASD with no other comorbid traits.

Nevertheless, despite the large genetic overlap between these two neurodevelopmental disorders, both studies showed a higher diagnostic yield in ID patients than in ASD ones, further supporting the genetic complexity of ASD. All the identified pathogenic variants were in genes previously associated with ASD, suggesting that only a subgroup of ID genes is more involved in ASD than others. The absence of pathogenic mutation in the majority of patients could be explained by various reasons: a pathogenic mutation located in a gene not included in our gene-panel or in a non-coding region; or structural

variants difficult to identify. On the other hand, it could also be explained by the implication of several risk factors – including the environmental ones - with a moderate or variable risk rather than the full penetrance of a single genetic cause. However, our TS strategy is not optimized to investigate these non-monogenic genetic forms and other approaches should be developed (*i.e.* analysis of variants in a network of genes rather than gene by gene). Our diagnostic yield should be confirmed on a much larger cohort of patients. Nevertheless, these results highlighted the importance and effectiveness to perform genetic analysis in patients mainly affected by ASD even in those not presenting a severe ID.

2. WES ON ID-PATIENTS WITH NO MUTATION IDENTIFIED IN TS

Despite the relatively high diagnostic yield obtained with the TS, the majority (~75%) of the patients remained without a molecular diagnosis. This could be explained by several reasons other than the non-genetic origin of ID; for instance, the causal mutation may lie in a gene not included in the gene panel – *e.g.* it is newly associated to ID – or it is in a gene never been implicated in ID. 31 patients who did not receive a diagnosis after the TS were sequenced by trio-WES. This strategy increases the chance to find potential pathogenic mutation never been implicated in ID, a major goal of this study. Additionally, some patients with a highly syndromic form of ID or belonging to a multiplex family were directly passed to WES without going through TS.

2.1 RESULTS

The mean coverage obtained for all the sequenced patients and parents was around 94x and the 93% of the targeted regions were covered by more than 20x. On average, more than 60.000 SNVs were identified per individual. After the filtering criteria, the potential candidate variants in trio were less than 5 for XLID, and below 10 for ADID and ARID, even if these numbers were extremely variable among individuals, and much higher when the proband was sequenced alone.

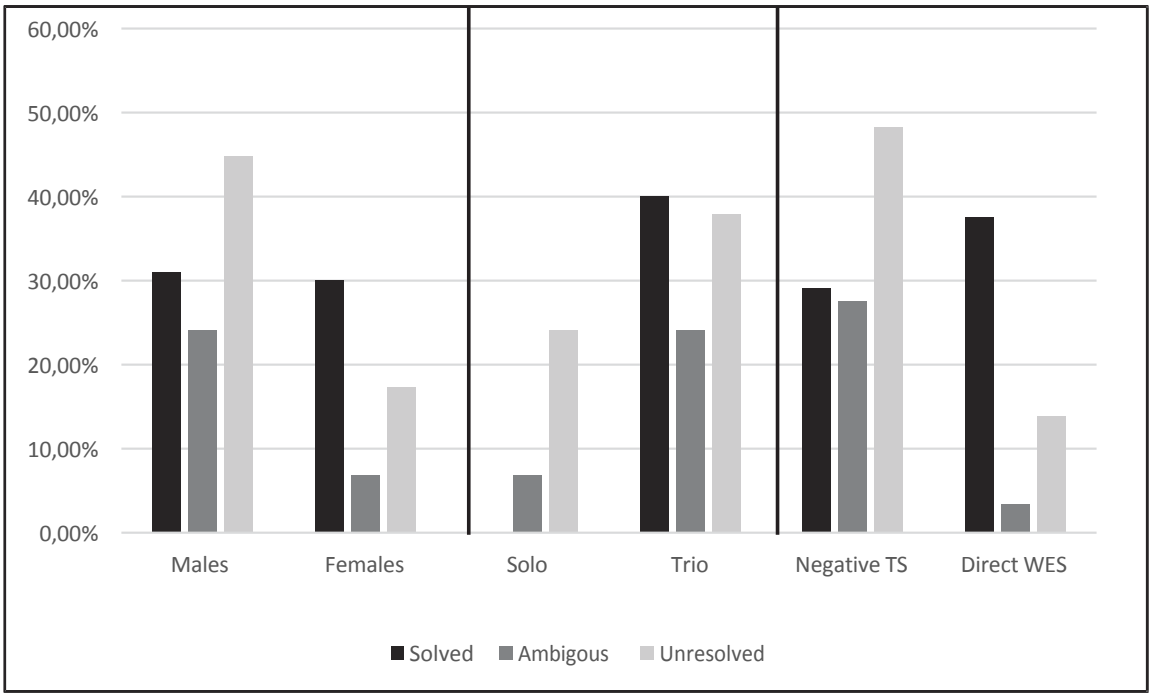


Figure 21: Percentage partioning of patients classification in WES

During the variant analysis, we classified patients into three main groups: *solved*, which includes patients for which we identified a pathogenic mutation either in a known or novel ID gene; *ambiguous*, which comprises individuals for which we detected a variant of unknown significance; and *unresolved*, for which we did not point out any candidate mutation. We observed a higher rate of molecular diagnosis in patients sequenced in trio, but no other difference in the diagnostic yield between other categories (Figure 21).

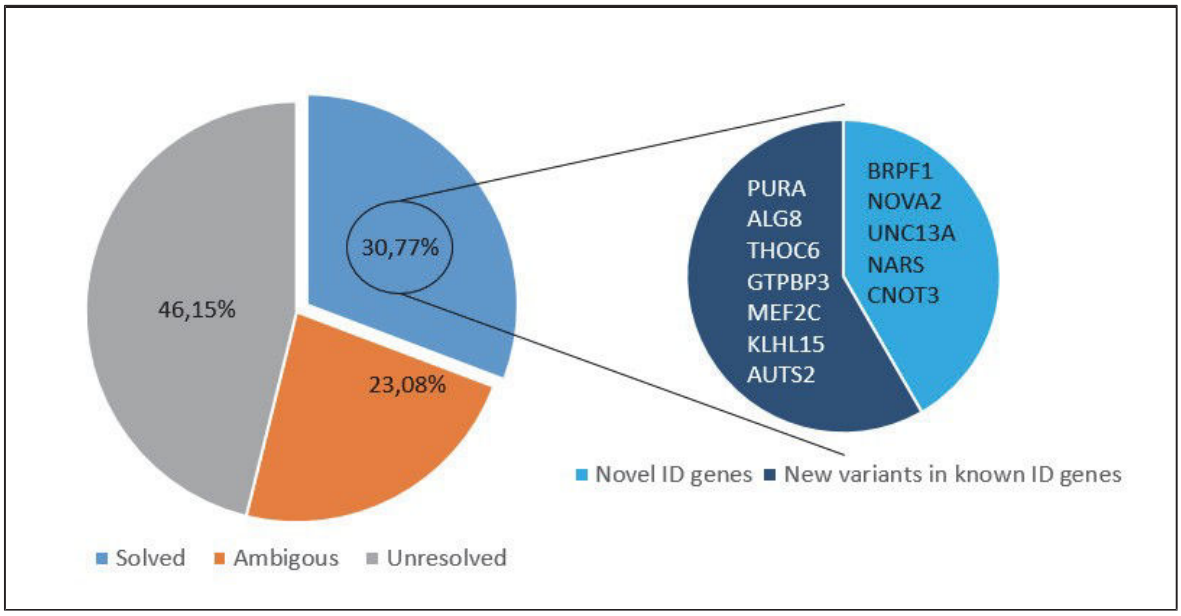


Figure 22: Percentage partitioning of variant classification in WES

A little bit more than half of the mutations was identified in genes already implicated in ID (7/12) (Figure 22). This is explained because either:

- the gene was recently implicated in ID (*i.e. PURA*, *THOC6*, *KLHL15, ALG8*), therefore it was not included in the gene panel used in TS;
- the patient was passed directly to WES (i.e. *GTPBP3*, *AUTS2*);
- the mutation did not retain attention in the first analysis by the targeted sequencing (*i.e.* mutations in the 5' UTR of *MEF2C*).

The remained identified pathogenic variants were in genes never been implicated in ID (*UNC13A*, *NARS*, *CNOT3*, *NOVA2* and *BRPF1*) (Figure 22).

The overall diagnostic yield is comparable to the one obtained from previous trio-WES studies. However, we combined two different NGS approaches; therefore, if we considered the overall NGS strategy used (TS + WES), we obtained a diagnostic yield of more than 40%.

Gender	Cohort (n=39)	Positive diagnosis			
		XLID	ADID	ARID	*Tot. variants*
Male	29	1	6	2	*9*
Female	10		1	2	*3*
Tot.	*39*	*1*	*7*	*4*	*12*

Table 13: Classification of the detected mutations in WES based on their inheritance mode

Among the positive diagnosis, we identified a majority of genes with an autosomal dominant inheritance (7 out of the 12 identified mutations) (Table 13), in line with the hypothesis that *de novo* mutations are the major cause of ID in outbreed populations (Vissers et al., 2010). In particular, three WES studies did not identify any autosomal recessive mutations (Hamdan et al., 2014; de Ligt et al., 2012; Rauch et al., 2012), but they were mainly focused on sporadic cases. Other studies including patients with a broader phenotype (including other neurodevelopmental disorders like epilepsy) still reported a majority of autosomal dominant mutations (ranging from 13.2% - 64% of the positive diagnosis) but also a low percentage of autosomal recessive ones (25.8% – 23% of the positive diagnosis) (Deciphering Developmental Disorders Study, 2015; Yang et al., 2013). In our cohort of patients, we identified a higher rate of autosomal recessive mutations (~33.3%; 4/12). This could be partially explained by the inclusion of some consanguineous families (4 + 2 suspected); however, we identified only two homozygous mutation in 2 of these families (1 known and 1 suspected), while we did not identify any deleterious variants for the remaining ones.

2.1.1 VALIDATION OF MUTATIONS IN KNOWN ID GENES

Many mutations identified by WES were in genes previously implicated in ID and their identification expands the clinical phenotype. For some of these I carried out some functional analysis to further delineate their pathophysiological mechanisms (see *Clinical And Functional Characterization Of Recurrent Missense Mutations Involved In THOC6-Related Intellectual Disability,* pg.124 and *Characterization Of Functional Consequences Of Truncating Mutations Affecting Long And Short AUTS2 Isoforms*, pg.126). Nevertheless, some of these variants were still classified as VUS as further analysis are required to prove their pathogenicity. Here I will describe an example of a validation analysis performed for a missense variant located in the 5'UTR of *MEF2C*.

NON-CODING VARIANT IN MEF2C

A girl with a Rett syndromic-like phenotype was previously tested for a Sanger sequencing of *MEF2C*, *FOXG1* and *MECP2* but no candidate pathogenic variants were identified; then a targeted-sequencing of 220 genes returned negative. The trio-WES pointed out a *de novo* candidate variant located in the non-coding 5'UTR region of *MEF2C*, a substitution located only 8 nucleotides upstream the initiation codon (NM_001193347.1: c.-8C>T). The nucleotide is well conserved (PhastCons: 1; PhyloP: 5.53) and

it is predicted to affect the Kozak sequence of the normal AUG, creating an alternative one in frame three amino-acid before (Figure 23) (http://dnafsminer.bic.nus.edu.sg/) (Liu and Wong, 2003). This variant is not present in gnomAD.

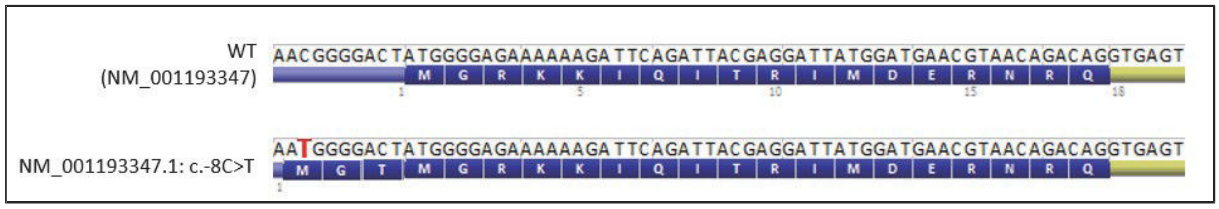


Figure 23: Schematic representation of the identified non-coding variant in MEF2C and its predicted effect

MEF2C belongs to the myocyte enhancer factor-2 family of transcription factor and is located in the 5q14.3 region, for which microdeletions have been reported in patients affected by a syndromic form of ID characterized by severe ID, epilepsy, muscular hypotonia and variable brain and other minor anomalies (Cardoso et al., 2009; Engels et al., 2009). Indeed, thanks also to the development of a conditional knock-out model, *MEF2C* was known to be implicated in the neuronal differentiation and regulation of excitatory synaptic number, suggesting a role in the synaptic plasticity, hence in learning and memory abilities as well as in seizure (Li et al., 2008). The identification of several patients with de novo truncating and missense variants in this gene eventually lead to the implication of this gene in the aetiology of this syndromic form of ID (Bienvenu et al., 2013; Le Meur et al., 2010; Zweier et al., 2010). Recently, several patients with a similar phenotype have been reported with balanced cytogenetic abnormalities breakpoints (BCAs) distal to *MEF2C,* disrupting the TAD containing the coding region of *MEF2C* thus leading to alterations in its expression (Redin et al., 2017).

Truncating mutations in *MEF2C* were shown to cause a decrease in the gene expression level (Redin et al., 2017; Zweier et al., 2010), while for the missense mutations either there was a significant increase (Patient 5, p.Leu38Gln from (Zweier et al., 2010)) or no difference with controls (Patient 8, p.Gly27Ala from (Zweier et al., 2010)) in the expression of *MEF2C.* We investigated the expression level of *MEF2C* in our patient by performing a RT-qPCR on cDNA obtained from reverse transcription using the SuperScript II (Invitrogen) on the extracted RNA from PAXgene blood tubes. We measured MEF2C expression level by using two different couple of primers; one that amplifies all the different isoforms (MEF2C_1: forward primer: ATCGACCTCCAAGTGCAGGTAACA; reverse primer: AGACCTGGTGAGTTTCGGGGATT), while a second one amplifies NM_00119335, NM_001308002 and NM_002397 transcripts (MEF2C_2: forward primer: GCCCTGAGTCTGAGGACAAG; reverse primer: AGTGAGCTGACAGGGTTGCT). The latter transcript is known to be highly expressed in brain (Zweier et al., 2010). qPCR was performed in triplicates and MEF2C mRNA level was quantified by the 2^{-DDCt}

method using GAPDH as a reference gene. Both couple of primers showed a significant increase in MEF2C expression in the patient compared to age- and sex- matched controls (n=4) (Figure 24).

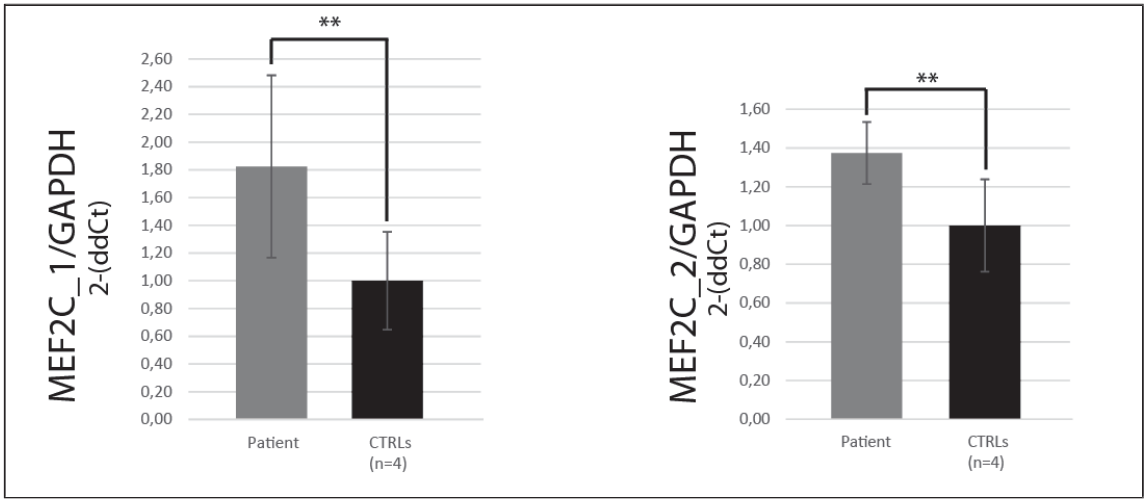


Figure 24: qPCR analysis of MEF2C expression level

On the left, the qPCR analysis done with the couple of primers amplifying all isoforms (MEF2C_1); on the right, the qPCR analysis performed with primers amplifying specific transcripts (MEF2C_2).

Patients with deleted or mutated *MEF2C* were reported to have a significant decrease in *MECP2* and *CDKL5* expression levels, potentially explaining the similar phenotype observed in patients with a mutation in one of these genes (Zweier et al., 2010). We investigated by qPCR the expression level of *CDKL5* (forward primer: GAAACACATGAAATTGTGGCG; reverse primer: GCTTGAGAGTCCGAAGCATT) and of *MECP2* (forward primer: ACTCCCCAGAATACACCTTGCTT; reverse primer: TGAGGCCCTGGAGGTCCT), but we did not observe any significant difference in the mRNA levels in both genes (data not shown).

The mutated Cytosine is located in a CpG and is predicted to be methylated (RoadMap Epigenome Browser v1.19). Therefore, we hypothesize that this variant may alter the methylation of this specific C. To answer this question, we are going to bisulfite-treat the genomic DNA of the patient, her parents and controls, amplify the region containing the variant by PCR and then Sanger sequence it. Thus, by observing the differences among the reference sequence we will be able to see if the nucleotide change affects its methylation status. However, it is unlikely that affecting the methylation of one Cytosine might directly lead to an increase of *MEF2C* expression. Other mechanisms might therefore be involved.

2.1.2 VALIDATION OF MUTATIONS IN NOVEL ID GENES

WES was confirmed to be an efficient approach for the identification of novel genes implicated in ID. Overall, I identified 5 mutations in genes never been implicated in ID before, so in about 12% of the total patients sequenced by WES.

We first identified two truncating mutations in two genes never been implicated in ID before (*BRPF1* and *NOVA2*). For these mutations, I performed some functional molecular analyses to characterize

their consequences and understand how they could lead to ID. They will be described in other sections (*Mutations in Histone Acetylase Modifier BRPF1 Cause an Autosomal-Dominant Form of Intellectual Disability with Associated Ptosis,* pg.115 and *De Novo Truncating Variants In The Neuronal Splicing Factor NOVA2 Cause A Syndromic Form Of Intellectual Disability With Angelman-Like Features,* pg.116*)*

NARS

We identified a nonsense variant in a boy affected by severe ID, epilepsy, spasticity, microcephaly, with no speech and walk. This *de novo* variant (NM_004539.3: c.1600C>T; p.Arg534*) is close to the end of *NARS*, a gene encoding for asparagine-tRNA synthetase, an enzyme necessary for translation. Though, *NARS* appears to be tolerant to LoF (ExAC pLI = 0.00 with numerous truncating variants reported) and this would exclude that haploinsufficiency of this gene could lead to a severe form of ID. The nonsense variant was then considered as VUS and submitted to GeneMatcher. Through this data exchange, we have been in contact with MD D. Koleen (Nijmegen) who identified 3 patients with the same *de novo* nonsense variant. The comparison of the phenotype showed a high similarity among the 4 individuals, with recurrent clinical features, such as severe ID, microcephaly, ataxia and/or spasticity and seizures, suggesting that this recurrent mutation is responsible for this phenotype. After this first match, we have been in contact with Dr. A. Manole and Dr. H. Houlden (University College of London) and GeneDx that identified several recessive cases with homozygous or compound heterozygous mutations in *NARS*. We have been also in contact with other teams who previously identified variants in *NARS* in sporadic cases and even in large families that were classified as variants of unknown significance. Overall, many patients have been detected with a mutation in *NARS*, for a total of 5 individuals with the same *de novo* nonsense mutation; 7 recessive cases (compound heterozygotes and homozygotes) and 11 homozygotes individuals from 4 unrelated families with the same missense variant (p.Arg545Cys). Due to the many truncating variants over the whole length of the protein up to a nonsense variant p.Arg522* reported in gnomAD, we hypothesized that the recurrent nonsense variant (p.Arg534*) does not act by haploinsufficiency, but may have a dominant negative or gain of function effect by removing the last 15 aa in the highly conserved C-terminal domain necessary for ATP-binding. The homozygous missense variant detected in the 4 distinct families might have a similar effect, since it affects the same C-terminal stretch.

aa change	Genotype	Allelic frequency in GnomAD (%)	SIFT	Mutation Taster	PP2 (HumVar)	GS	PhyloP
Arg11Pro	homo	0.0007218	Tolerated	Disease causing	Tolerated	103	2.3
Thr17Met	homo	0.001219	Deleterious	Polymorphism	Possibly damaging	81	3.03
Met34Leu	comp. ht. with Asn218Ser	0.01133	Tolerated	Disease causing	Benign	15	1.01
Met69Aspfs*4	Comp. ht with Asp356Ala	/	/	/	/	/	/
Asn218Ser	comp. ht. with Met34Leu or Arg322Leu	0.02634	Tolerated	Disease causing	benign	46	4.73
Arg322Leu	comp. ht. with Asn218Ser	/	Deleterious	Disease causing	Probably damaging	102	5.69
Leu350Pro	comp. ht with Ala422Thr	/	Deleterious	Disease causing	Probably damaging	98	4.64
Asp356Ala	comp. ht. with Met69Aspfs*4	0.09236	Tolerated	Disease causing	Benign	126	3.03
Ala422Thr	comp. ht with Leu350Pro	0.002033	Deleterious	Disease causing	Benign	58	5.61
Gly509Ser	homo	/	Tolerated	Disease causing	Probably damaging	56	5.69
Arg534*	*de novo* ht	/	/	/	/	/	/
Arg545Cys	homo	0.001807	Deleterious	Disease causing	Possibly damaging	180	5.86

Table 14: Mutations identified in NARS, their prediction scores and frequencies in the general population

Two missense variants found only at a compound heterozygous state (p.Asn218Ser and p.Asp356Ala) appear to have an allelic frequency that would be too high for a pathogenic allele associated to a severe recessive form of ID (Table 14). A missense variant predicted to be essential for enzymatic activity (p.Arg322Leu) (Pr. H. Becker, personal communication) is observed only at the compound heterozygous state (Table 14). Furthermore, many truncating mutations are present in the general population but only one has been observed in patients at the heterozygous state (p.Arg534*) (Table 14). We thus hypothesized that, given the vital function of NARS, two loss-of-function variants would be embryonic lethal while the combination of rare hypomorphic alleles can cause the disease. Functional analysis to prove this hypothesis are currently carrying out *in vivo* by Dr. Manole at the University College London, including yeast complementation tests and zebrafish modeling, and biochemically in vitro studies by Pr. H. Becker at the University of Strasbourg, comprising testing of the aminoacylation and charging capacity of the transcript and the kinetic of charging for $tRNA^{Asn}$.

CNOT3

We identified a *de novo* missense mutation in *CNOT3* (NM_014516.3: c.439G>A) in a boy affected by ID, epilepsy, nystagmus, stereotypical behaviours, growth retardation and facial dysmorphism. This variant was classified as VUS, since there was no obvious pathogenicity evidence, except that *CNOT3* in ExAC has a high constraint value for missense (z-score= 3.89) and it is intolerant to LoF variants (pLI=1). Recently, in a study of a large cohort of patients affected by neurodevelopmental disorders (Deciphering Developmental Disorders Study, 2017), several patients with a missense variant in this gene were reported. Furthermore, by submitting this gene to GeneMatcher, patients with a similar phenotype were identified. Overall, 17 patients have been detected with a de novo variant in *CNOT3*, among which 4 are missenses variants of the same amino acid, 4 other missense variants and 9 truncating mutations (Figure 25). At a general view, these patients share some recurrent clinical features, such as ASD, global developmental delay and macrocephaly. Curiously, all the missense mutations are located closer to the N-terminus, while the truncating variants are all over the length of the protein. It would be interesting to perform a genotype-phenotype comparison and to test functional effects to check if there are differences based on the type of mutation. The collection of clinical information is currently pursued by Dr. R. Martin at Newcastle University.

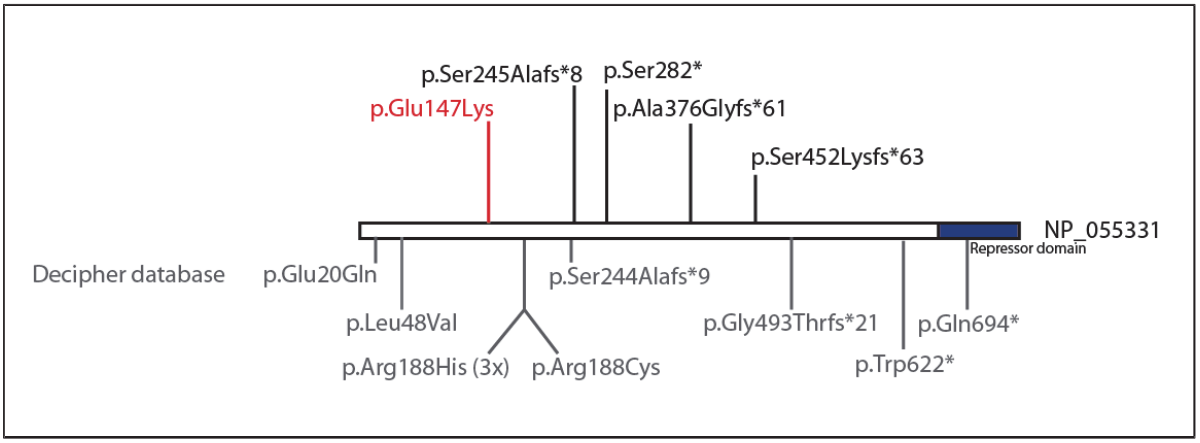


Figure 25: Identified mutations in CNOT3

UNC13A

In a girl affected by a severe encephalopathy, with severe ID, epilepsy, no speech and walk, macrocephaly and low weight, I identified a frameshift mutation and a de novo missense mutation, both not present in GnomAD, in the *UNC13A* gene. Segregation analyses showed that the frameshift mutation (NM_001080421.2:c.339_340insCAGGAAAC) is present in her two unaffected siblings and was transmitted from the mother, who does not show any clinical symptoms. On the other hand, the missense variant (NM_001080421.2: c.605G>A) arose *de novo* (Figure 26). An allele specific PCR (Forward primer: 5'-tgctgttgctcgtttcactgt-3'; reverse primer: 5'-tcatggcagacagtgagatctgtg-3') confirmed that the two mutations are in *trans*. The missense variant is in a well conserved nucleotide (PhyloP: 5.53; PhastCons: 1) and it is predicted to be deleterious (Mutation Taster and PolyPhen-2), even if the physical distance between the two amino acids is small (Grantham score: 29). In ExAC this gene has a high constraint metric with less observed missense variants than expected (z-score= 5.89) and it seems to be intolerant to loss-of-function mutations (pLI=1).

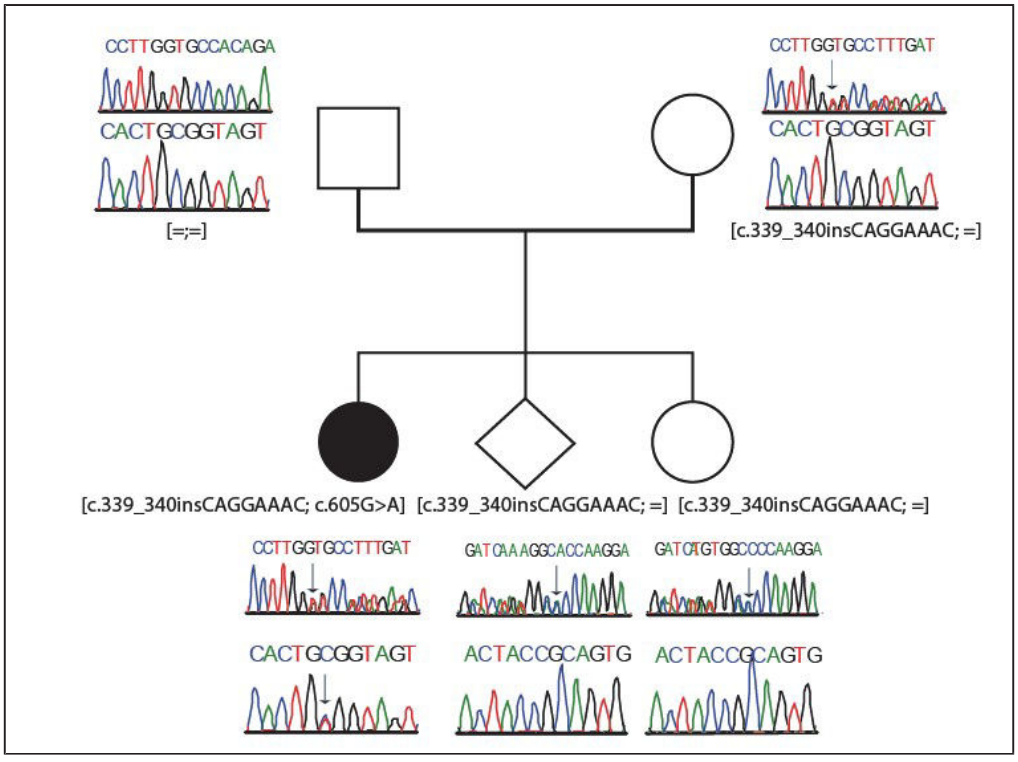


Figure 26: Pedigree of the family and mutations in UNC13A

UNC13A belongs to the UNC13 family proteins involved in the regulation of the neurotransmitter release at the nerve cell synapse. In particular, UNC13A is important for the priming of the synaptic vesicles, which leads to their fusion and release at the pre-synaptic plasma membrane. Knock-out mice for UNC13A died early after birth because of a severe paralysis and brains of the newborn showed an almost complete reduction of the readily releasable synaptic vesicle pool and a consequent reduction of spontaneous and action potential evoked glutamate release (Augustin et al., 1999).

Recently three mutations have been identified in *UNC13A*: a homozygous nonsense mutation (p.Gln102*), located at the beginning of the protein, in a patient affected by microcephaly and cortical hyperexcitability who died at 4 years of age for respiratory failure (Engel et al., 2016); a *de novo* missense variant p.(Pro814Leu) in an individual affected by a dyskinetic movement disorder, global developmental delay and autism (Lipstein et al., 2017); and a homozygous missense mutation p.(Glu52Lys) in a patient with global developmental delay, seizures, generalized hypotonia, myopathy and microcephaly (Lionel et al., 2017) (Figure 27). The nonsense mutation is located at the beginning of the protein and leads to a truncated protein lacking the syntaxin-1 (STX1A) binding site, through which it stabilizes the functional open conformation of STX1A. Therefore, the authors speculated that the clinical feature observed in the patient were caused by a persistent non-functional state of STX1A, which inhibited cholinergic transmission at the neuromuscular junction and glutamatergic transmission in the brain, as the observed clinical features in the patients well overlap with the ones observed in individuals with mutations in this gene (Engel et al., 2016). On the other hand, the missense variant p.(Pro814Leu) has a dominant gain of function that increases the fusion propensity of the synaptic vesicles, leading to a major synaptic vesicle release probability and abnormal short term plasticity, as demonstrated by functional analysis on electrophysiological studies on mouse neuronal cell cultures and *Caenorhabditis elegans* (Lipstein et al., 2017). No functional analyses were performed on the p.(Glu52Lys) variant.

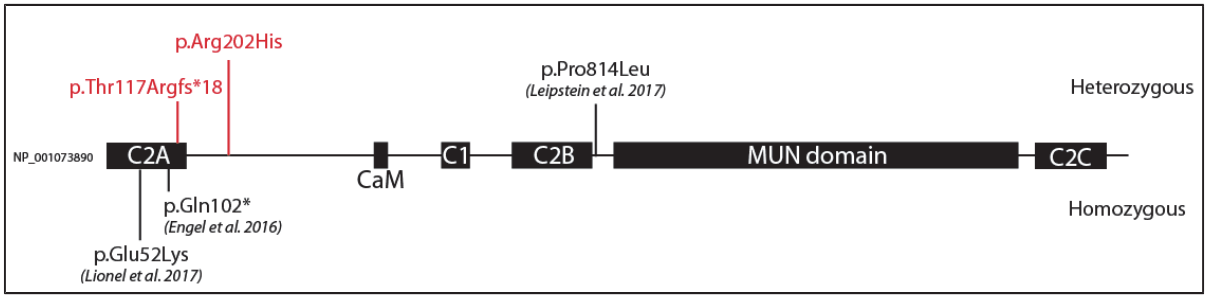


Figure 27: Mutations identified in UNC13A

Through GeneMatcher we have been able to identify several patients with a mutation in *UNC13A*, both *de novo* heterozygous or compound heterozygotes. Functional studies are currently performed by Dr. Lipstein at the Max-Planck Institute for experimental biology. She will use neurons from a double knock-out mouse for *Unc13a* and *Unc13b* and introduce the variants, by using lentiviruses. The expression of the different variants will then be analyzed either on a null background (especially for the missense mutation) or in a wild-type background to assess a dominant negative effect.

3. RNA-SEQUENCING IN PATIENTS WITH ID

Despite the high diagnostic yield obtained both with the TS and the WES, half of the patients did not receive a molecular diagnosis, underscoring the common limitations of these two strategies.

These unidentified cases could represent more complex forms of ID, caused by variants in several genes (oligogenism), epigenetic events or even non-genetic causes, but could also be from monogenic origin involving a mutation not identified by TS or WES, due to their technical limitations. These monogenic forms can be caused by: variations that are not located in CDS and could not have been identified by TES and WES, variations in CDS that have been missed by the techniques (small CNV affecting one or few exons of a gene, variations in poorly captured regions) or variations in CDS identified but whose effect at the mRNA level was misjudged. Some VUS and mutations in non-coding regions might affect RNA abundance or isoforms. Besides mutations affecting canonical splice sites or creating an exonic cryptic site with high prediction scores, the effect of SNV on splicing can be difficult to predict and in particular those which might affect exonic or intronic splicing enhancer sequences (ESE, ISE). Some informatic programs exist but predict a huge number of potential sequences, for which no real functional effects have been proven. Moreover, also intronic mutations located too far from exon/intron junctions and SNV in regulatory elements involved in transcription, such as promoters, 5'-UTR or distal enhancer are also not detected by TS and WES. Therefore, RNA-sequencing could be a useful complementary tool for variant identification and interpretation, as proved by recent studies on different patients' cohorts (Cummings et al., 2017; Kremer et al., 2017), but this was never tested in individuals with ID.

To evaluate the efficiency of RNA-sequencing as a complementary tool in the diagnosis of heterogeneous disorder, we sequenced the mRNA from individuals affected by ID and Bardet-Biedl syndrome (BBS). BBS is a disorder caused by defects in several genes mainly implicated in the formation and function of the primary cilium. BBS affects multiple organs, including the eye as most of the patients present retinis pigmentosa. We collected RNA from tissue available from patients but unfortunately non-relevant for the disorders, fibroblasts and blood. In this cohort we included some positive controls - *i.e.* individuals with a known mutation predicted to alter RNA levels – and patients with a VUS or with no pathogenic variant detected. We included 6 patients affected by BBS and 9 presenting ID, for a total of 15 patients. For the BBS individuals we only sequenced RNA from fibroblast, while for ID patients we sequenced RNA from fibroblasts for 3 patients and from whole blood for 6 others. For BBS we analysed 3 known pathogenic variants and 3 unknown cases, while for ID individuals we included 3 known mutations, 2 VUS, and 4 unknown cases (Table 15).

Ind.	Mutation	Disease	Tissue
Ind 1	*SDCCAG8;* c.836+356C>T	BBS	fibros
Ind 2	*BBS3;* deletion exon1-3	BBS	fibros
Ind 3	Unknown	BBS	fibros
Ind 4	*BBS1;* c.1168A>G + complex insertion in exon13	BBS	fibros
Ind 5	unknown	BBS	fibros
Ind 6	unknown	BBS	fibros
Ind 7	unknown	ID	fibros
Ind 8	unknown	ID	fibros
Ind 9	*DYRK1A;* c.951+1_951+4delGTAA	ID	fibros
Ind 10	*OPHN1;* duplication exon 4 to 5	ID	blood
Ind 11	*MEF2C,* c.-8C>T (VUS)	ID	blood
Ind 12	*CCDC101*; c.225-2dup (VUS)	ID	blood
Ind 13	unknown	ID	blood
Ind 14	unknown	ID	blood
Ind 15	*DYRK1A;* c.328-1G>T	ID	blood

Table 15: RNA-sequenced patients

3.1 RESULTS

On average 300 millions of sequences were generated per individuals. Before starting to analyse the obtained data, we checked the expression of the genes of interest in both tissue, by comparing the normalized number of reads using the reads per kilobase per million mapped reads (RPKM) (Mortazavi et al., 2008). We then analysed the expression of known genes implicated in ID and BBS in the two tissues, blood and fibroblasts.

The list of ID genes was retrieved from a combination of different European lists (Radboud UMC list, Nijmegen, https://issuu.com/radboudumc/docs/intellectual_disability_dg29?e=28355229/48848639; genes of the Genome England PanelApp https://panelapp.genomicsengland.co.uk/; list of genes from the SysID database: http://sysid.cmbi.umcn.nl/), for a total of 836 genes. Moreover, we specifically checked the expression level of the most recurrently mutated 40 genes, according to database and literature. Overall, the expression level of ID genes is higher in fibroblast than in blood cells (RPKM value ~20% more) (Figure 28A).

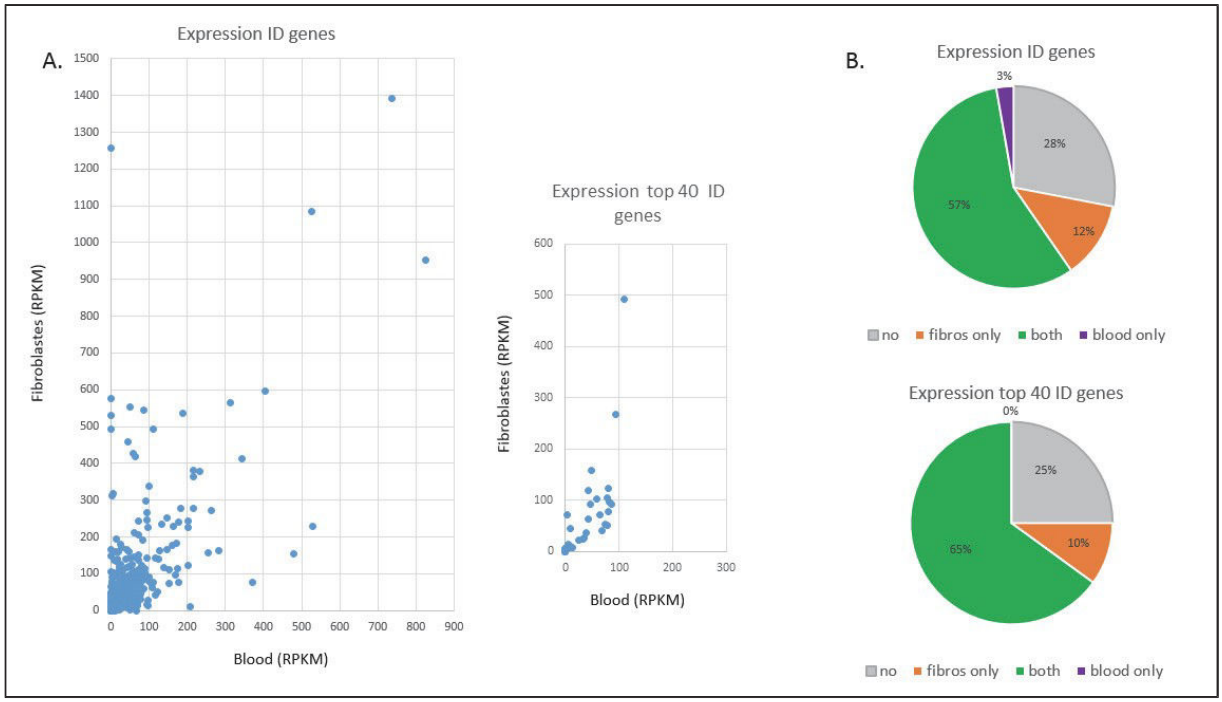


Figure 28: Expression of ID genes in fibroblast and blood cells

Overall, 57% of the ID genes are expressed (RPKM > 5) in both tissue, while 12% are expressed exclusively in fibroblasts and only 3% in blood (Figure 28B). However, a large portion of ID genes are not expressed neither in fibroblasts nor in blood (28%), among which 10 genes frequently reported to be mutated in ID. Indeed, some of these genes code for synaptic proteins (*KCNQ2, SCN2A, SCN1A, SYNGAP1, SHANK3, GRIN2B*), hence it was expected not to find them expressed in these tissues. However, some of the not expressed genes are involved in more common mechanism, such as transcription regulation or chromatin remodelling (*ASXL3, KAT6B, EP300, SATB2*). Generally speaking, the RNA-sequencing analysis seems more efficient from patients' fibroblasts than from blood cells, even if around a quarter of ID genes are not detected.

On the other hand, BBS genes are not highly expressed neither in fibroblasts nor in blood, with a general RPKM around 50 in both cells' types (A). Overall, about 36% of RP-genes are detected both in fibroblasts and blood cells, 12% are specifically expressed in fibroblasts while 3.6% in blood (Figure 30B).

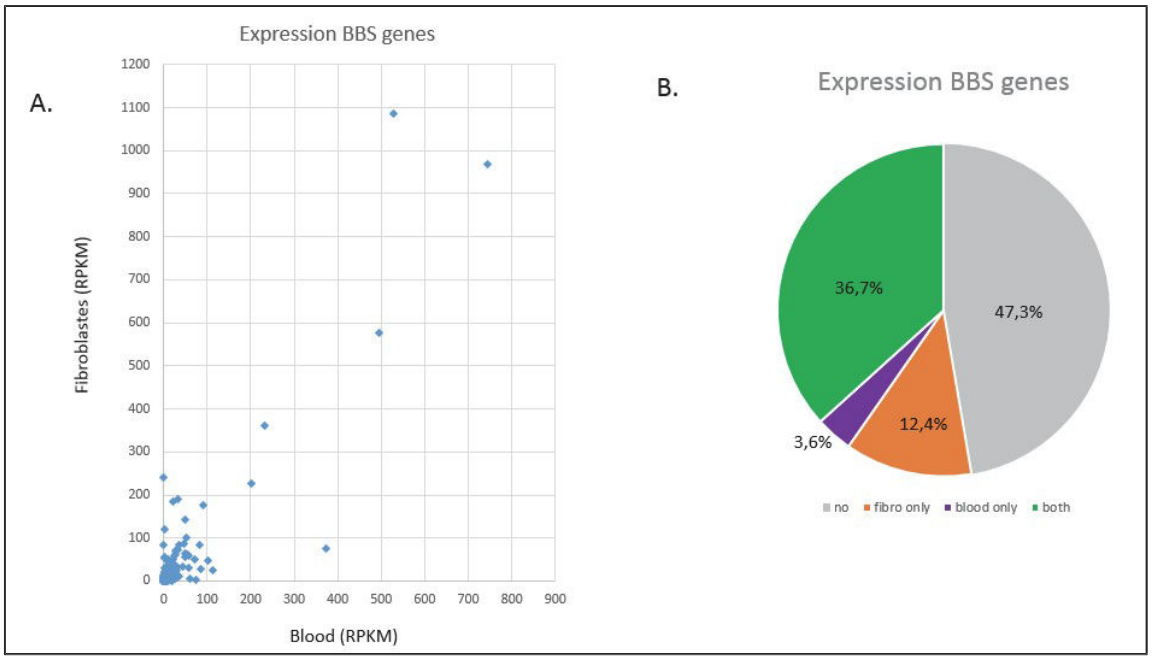


Figure 29: Expression of BBS genes in fibroblasts and cells:

This analysis allowed us to delineate the limit of the RNA-sequencing. We then analysed the data in order to identify the potential genetic causes of the disease of interest, which was the goal of this pilot study. Transcriptomic analyses were done by comparing one patient versus all the others, as the probability they have the same exact mutation is extremely low. The transcriptomic variation could be the result of different events, such as a mutation in the regulatory regions or in a splice-site. For instance, a mutation may lead to a dysregulation of gene expression, while a mutation in the splice-site causes an alteration in the alternative splicing. To this end, we investigated the variations in RNA abundance and sequences in three different situations: a change in the gene expression level; differences in the alternative-splicing; and allele-specific expression (Figure 29)

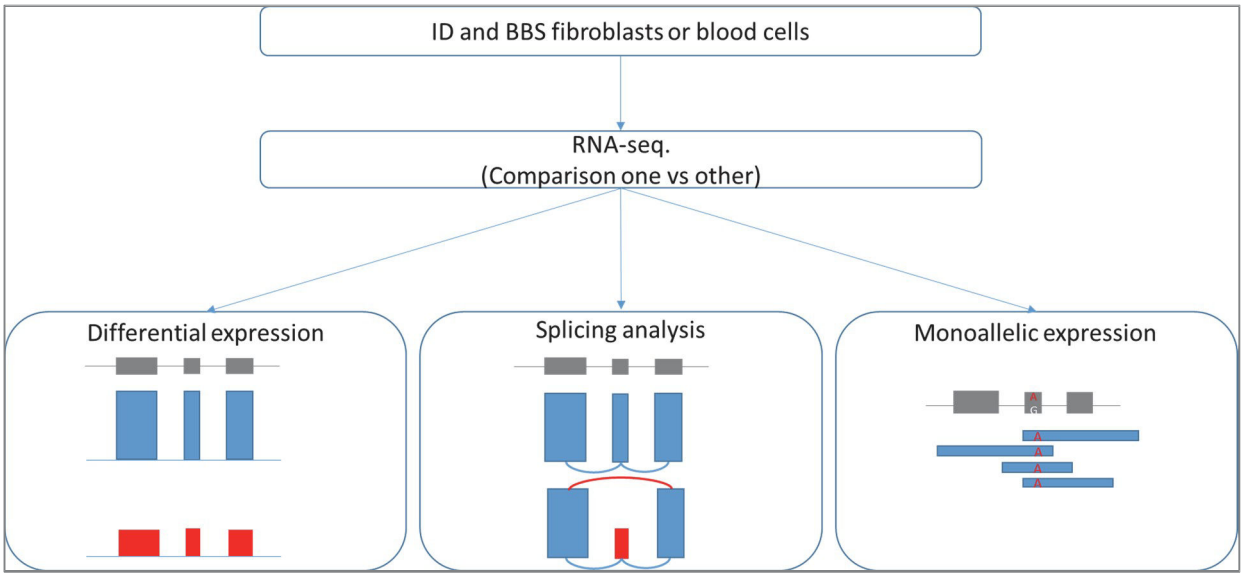


Figure 30: Workflow used for variant identification with RNA-sequencing

3.1.1 DIFFERENTIAL EXPRESSION

We analysed the differential expression of mRNA levels of each patient versus all the others. Differences in gene expression could be caused by mutations in regulatory regions (*e.g.* promoter, enhancers), by a NMD mechanism or by deletion/duplication of a gene. However, this analysis has several limitations: for instance, the differential expression could be the consequence of a mutation in a gene that regulate the expression of a subset of genes (*e.g.* transcription factor). Moreover, variants and mRNA located on the X-chromosome were not analysed separately.

I will not detail all the results, as the whole analysis is still ongoing, but I will show just an example.
In an individual with a known genetic cause of BBS (Ind 2), we detected a significant decrease in *BBS3* expression (log2FC= 8.42 and p-value= 7.98E-26) (Figure 31). Indeed, the known mutation is a homozygous deletion disrupting exon 1 to 3 of this gene and the differential expression analysis allowed us to detect the affected gene, showing the efficiency of this investigation.

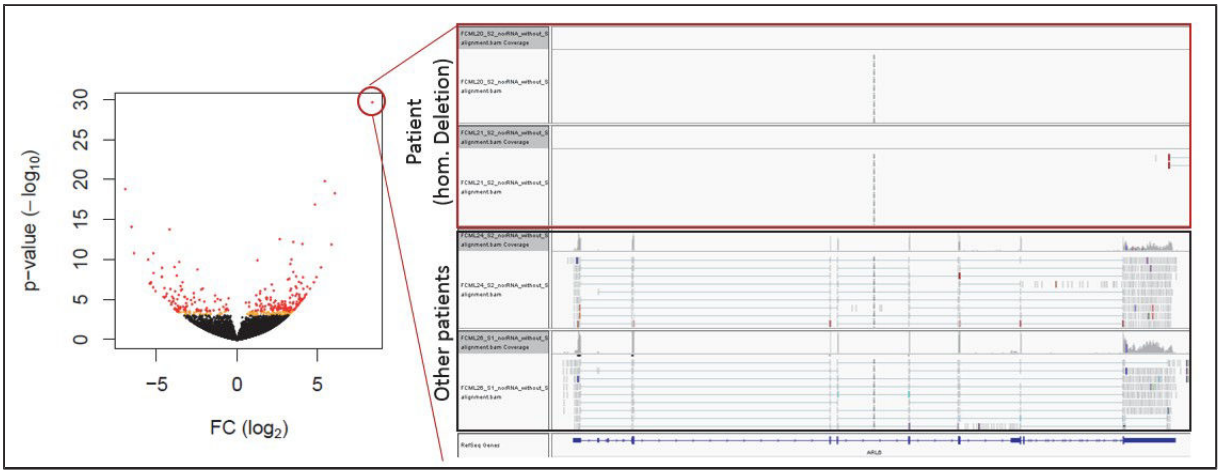


Figure 31: Volcano plot and RNA-sequencing reads of BBS3 in individual 2

3.1.2 SPLICING ANALYSIS

We checked for differential alternative splicing events in order to detect alterations possibly caused by mutations close to the exon-junction or in exonic/intronic splicing enhancer. Among the alternative splicing events there are the exon skipping, intron retention, mutually exclusive exons, alternative 5' and 3' splice sites, alternative first or last exons and more complex alternative splicing patterns. To detect splicing alterations, we used three different software: JunctionSeq, rMATS and LeafCutter.
We compared the number of splicing events detected by the three software (Table 16), using different filtering criteria. We noticed that, even if rMATS identifies the largest number of alternative splicing events, the filtering criteria must be extremely stringent (FDR < E-10) to avoid too many false positive

events. Moreover, rMATS returns separate files for each possible splicing event (skipped exon; retained intron; mutually exclusive exons; alternative 5' splice site; alternative 3' splice site) which increases the time of analysis. Junctionseq on the contrary restricts its analysis to skipped exons events. However, also with JunctionSeq we had to use stringent filters; for instance, to highlight potential differential splicing events, we had to consider a log2 Fold Change above 3 (Table 16), potentially missing some splicing events. On the other hand, with LeafCutter we did not use stringent filtering criteria on the fold change, but we observed a different number of splicing events when using a more stringent filter criterion on the adjusted p-value (Table 16).

Within the cohort, we included several individuals with a known mutation affecting the splicing. Among them, we could not detect any alteration in *OPHN1*, as the gene was poorly expressed in blood. Another known splice mutation (*DYRK1A*, NM_001396.3: c.328-1G>T) was not detected by all the three different software but we could not explain why, as the gene was expressed in the mRNA of the cell type. The second mutation in *DYRK1A* affecting the splicing (NM_001396.3:c.951+1_951+4delGTAA) was detected by both rMATS and LeafCutter, but it was filtered out when using a stringent filter (respectively for rMATS a FDR < E-10 and for LeafCutter a padj< 0.01), probably because the sample was sequenced in simplicate. The homozygous missense intronic variant in *SDCCAG8* (NM_001350248.1:c.836+356C>T) was detected in all the used software but, while with JunctionSeq and rMATS the analysis was more complicated due to the high number of splicing events detected, with LeafCutter we could easily retrieve this altered splicing event.

Ind.	Mutation	Disease	Tissue	rMATS					JunctionSeq		LeafCutter	
				FDR< E-10; |IncLevelDifference|> 0.05					padj< 0.01; dipersion< 0.05		|logef|> 0.6; |ΔΨ|> 0.2	
				SE	RI	MXE	A5SS	A3SS	|log2FC|> 3	|log2FC|> 0.6	padj< 0.05	padj< 0.01
Ind 1	*SDCCAG8*	BBS	fibros	83	0	14	6	5	39	168	58	35
Ind 2	*BBS3*	BBS	fibros	20	2	9	1	6	16	127	61	37
Ind 3	unknown	BBS	fibros	15	3	7	10	7	18	68	40	34
Ind 4	*BBS1*	BBS	fibros	11	4	6	5	3	0	6	33	12
Ind 5	unknown	BBS	fibros	29	7	16	3	5	43	329	113	69
Ind 6	unknown	BBS	fibros	28	1	5	2	6	29	139	114	56
Ind 7	unknown	ID	fibros	43	5	11	9	7	29	301	172	67
Ind 8	unknown	ID	fibros	132	23	30	22	19	312	1852	738	429
Ind 9	*DYRK1A*	ID	fibros	11	4	3	2	5	1	3	6	0
Ind 10	*OPHN1*	ID	blood	196	56	27	37	92	42	185	438	42
Ind 11	VUS (*MEF2C*)	ID	blood	303	65	60	64	130	73	279	1452	202
Ind 12	VUS (*CCDC101*)	ID	blood	244	61	41	78	79	164	754	1186	214
Ind 13	unknown	ID	blood	110	11	15	23	43	4	10	17	3
Ind 14	unknown	ID	blood	165	39	30	26	65	12	49	54	13
Ind 15	*DYRK1A*	ID	blood	53	4	8	12	9	0	0	4	2

Table 16: Number of splicing events detected by the three used sofwtare

FDR= false discovery rate ; |IncLevelDif.|= module of the difference among conditions between the percentage of the exon inclusion transcripts that splice from the upstream exon into the alternative exon and then into the downstream exon; SE = skipped exon; RI = retained intron, MXE= mutually exclusive exons; A5SS= alternative 5' splice site; A3SS= alternative 3' splice site; padj= adjusted p-value; log2FC= estimation of the Fold Change value in log2; |logef|= logarithm effect size; |ΔΨ|= module of the difference in usage proportion of each intron among the two conditions

3.1.3 MONOALLELIC EXPRESSION

The identification of a differential expression or splicing are limited to the detection of the consequences of one mutation, but not the detection of the mutation itself. It is possible to use the RNA-sequencing data to identify SNVs. Moreover, the comparison with DNA-sequencing data previously obtained could reveal a loss of expression of the mutant or the wild type allele (allele-specific expression). The complementation of these two NGS techniques helps in the detection of these events and eventually to suspect a second mutation in *trans* in the case of a recessive disorder. The comparison of the lists of variants (the vcf file) identified in DNA- and in RNA-sequencing is still ongoing. Then, thanks to the family barcode presents in varank, we are going to analyse variants present at the homozygous state in the RNA while they are at the heterozygous state in the DNA.

3.2 CONCLUSIONS AND PERSPECTIVES

We tested an RNA-sequencing approach in fibroblasts and blood cells of 15 patients affected by BBS or by ID, two heterogeneous disorders. We first checked the expression of the genes implicated in these disorders in these tissues; for the ID, more than half of them were expressed both in fibroblasts and blood cells with fibroblasts expressing the highest number of ID-genes compared to blood cells (Figure 28).

For the 6 BBS-patients we were able to detect in fibroblasts 3 variants: 2 positive controls and 1 unknown (Table 17). Among the 9 ID patients, we could not detect one known mutation as the gene of interest was not express (i.e. *OPHN1*), while we detected only in fibroblast one of the two known mutations altering splicing in *DYRK1A*, while the gene was expressed both in fibroblast and blood. However, we detected some candidate gene and are waiting for monoallelic expression to interpret these results. Overall, we analysed the efficiency of the RNA-sequencing for variant identification in heterogeneous disorders. We created an analysis pipeline that enables the identification of potential candidate mutations. In the next future, the monoallelic expression analysis will be carried out on the available patients, hopefully incrementing the number of detected mutations. Furthermore, the analysis of the monoallelic expression may also help in the implementation of the current workflow. For instance, it can be used as a first test to create a list of genes to focus for further analyses.

Ind.	Mutation	Disease	Tissue	Solved by RNA-seq.		
				DE	Splicing	MAE
Ind 1	*SDCCAG8* (c.836+356C>T)	BBS	fibros	-	+	Ongoing
Ind 2	*BBS3* (del. ex1-3)	BBS	fibros	+	-	
Ind 3	unknown	BBS	fibros	+	-	
Ind 4	*BBS1* (p.Met390Val + complex insertion)	BBS	fibros	-	-	
Ind 5	unknown	BBS	fibros	-	-	
Ind 6	unknown	BBS	fibros	-	-	
Ind 7	unknown	ID	fibros	-	-	
Ind 8	unknown	ID	fibros	-	-	
Ind 9	*DYRK1A* (c.951+1_951+4delGTAA)	ID	fibros	-	+	
Ind 10	*OPHN1* (dup. ex4-5)	ID	blood	-	-	
Ind 11	VUS (*MEF2C*, c.-8C>T)	ID	blood	-	?	
Ind 12	VUS (*CCDC101,* c.225-2dup)	ID	blood	-	?	
Ind 13	unknown	ID	blood	-	-	
Ind 14	unknown	ID	blood	-	-	
Ind 15	*DYRK1A* (c.328-1G>T)	ID	blood	-	-	

Table 17: Preliminary results of the RNA-sequencing analysis for variant identification

(DE= Differential Expression ; MAE= MonoAllelic Expression)

PART 2: DECIPHERING MOLECULAR MECHANISMS INVOLVED IN KNOWN AND NOVEL MONOGENIC FORMS OF ID

The NGS technologies used during my PhD project led to the identification of numerous candidate variants in known and novel ID-gene. These VUS/GUS require further steps of validation, which may include segregation analysis and functional studies, according to the gene function. Once the variant is proved to be deleterious, the next question concerns the specific role of the gene in this disorder. This issue is particularly important for novel ID-gene. In this section, I describe two novel ID-genes identified by trio-WES in patients with no mutation in known ID-genes. First, I identified a truncating mutation in a large family evocative for an autosomal dominant syndromic ID in *BRPF1,* a gene coding for a protein known to be a chromatin regulator. Investigation on the functional consequences of this mutation led to the better understanding of the role of this gene as well as its implication in ID (Mattioli et al., 2017). Similarly, I identified a de novo frameshift mutation in *NOVA2*, a RNA-binding protein involved in the alternative splicing of axon-guidance genes (Saito et al., 2016). Through data exchange we identified additional patients all sharing an Angelman-like ID. I am currently carrying out functional analyses to understand the pathogenicity of these mutations to explain the arising of such syndromic ID.

Even variants in known ID-gene require further steps of validation analysis to prove their pathogenicity and to better delineate the molecular mechanisms altered. For instance, I will describe the consequences of three homozygous missense variants in *THOC6*, a gene recently implicated in ID. These three variants are present at the same low frequencies in the general population (GnomAD data), suggesting they are in linkage. It is thus interesting to disentangle the contribution of each missense to the phenotype. Furthermore, by TS and WES we identified several truncating variants in *AUTS2*, a gene reported with a variable syndromic form of ID and ASD. Despite many studies on this gene, only few point mutations have been reported and the pathophysiological mechanisms involved are not yet clear, especially because the gene encodes two major and different isoforms. The characterization and identification of the molecular mechanisms altered in each monogenic form of ID is extremely important for a deeper understanding of the disorder. Moreover, the disentanglement of the concerned molecular pathways might lead to the development of therapeutic targets.

1. MUTATIONS IN HISTONE ACETYLASE MODIFIER *BRPF1* CAUSE AN AUTOSOMAL-DOMINANT FORM OF INTELLECTUAL DISABILITY WITH ASSOCIATED PTOSIS

By WES, I identified a truncating mutation in *BRPF1* in a large family evocative for an autosomal dominant syndromic ID. *BRPF1* encodes a protein known to be a chromatin regulator, controlling the histone acetyltransferase activity of the MYST family. At the transcript level, I observed that the mutated transcript was still expressed, indicating that it partially escaped the nonsense mediated decay, probably producing a truncated protein. I overexpressed human BRPF1 cDNA wild-type and with the frameshift mutation in HeLa cells and confirmed the existence of a truncated BPRF1 protein which also shows an aberrant cellular localization. Furthermore, I showed that the truncated protein loses certain protein interactors, specifically ING5 and MEAF6 that are important for the stabilization of the complex formation between BRPF1 and the histone acetyltransferases. Since BRPF1 is known to be a chromatin regulator, I analysed the histone modifications in patient's fibroblast, in particular at the level of the histone 3 (H3), which is a well-known target of this protein. I first studied the global acetylation level of H3, but I did not observe any difference, so I focused the analysis on the acetylation levels of specific lysines known to be preferentially acetylated by the complex of which BRPF1 is part and I observed a slight decrease for H3K23.

By exchanging data via Decipher and GeneMatcher we collected 6 additional cases with a mutations or deletion in *BRPF1*. Since *BRPF1* is located in the 3p15 region, which is implicated in a microdeletion syndrome causing ID, ptosis and growth delay for which a single gene was reported to be as the causative one (*SETD5*) (Grozeva et al., 2014), I carried out a phenotype-genotype comparison and showed that when *BRPF1* is disrupted all patients presenting ptosis.

Overall, I showed for the first time that *BRPF1* is a gene involved in ID with growth retardation, microcephaly and ptosis.

Mutations in Histone Acetylase Modifier *BRPF1* Cause an Autosomal-Dominant Form of Intellectual Disability with Associated Ptosis

Francesca Mattioli,[1,2,3,4,5] Elise Schaefer,[6] Alex Magee,[7] Paul Mark,[8] Grazia M. Mancini,[9] Klaus Dieterich,[10] Gretchen Von Allmen,[11] Marielle Alders,[12] Charles Coutton,[13] Marjon van Slegtenhorst,[9] Gaëlle Vieville,[13] Mark Engelen,[12] Jan Maarten Cobben,[12] Jane Juusola,[14] Aurora Pujol,[15,16,17] Jean-Louis Mandel,[1,2,3,4,5,18,19,*] and Amélie Piton[1,2,3,4,18,*]

Intellectual disability (ID) is a common neurodevelopmental disorder exhibiting extreme genetic heterogeneity, and more than 500 genes have been implicated in Mendelian forms of ID. We performed exome sequencing in a large family affected by an autosomal-dominant form of mild syndromic ID with ptosis, growth retardation, and hypotonia, and we identified an inherited 2 bp deletion causing a frameshift in *BRPF1* (c.1052_1053del) in five affected family members. *BRPF1* encodes a protein modifier of two histone acetyltransferases associated with ID: *KAT6A* (also known as *MOZ* or *MYST3*) and *KAT6B* (*MORF* or *MYST4*). The mRNA transcript was not significantly reduced in affected fibroblasts and most likely produces a truncated protein (p.Val351Glyfs*8). The protein variant shows an aberrant cellular location, loss of certain protein interactions, and decreased histone *H3K23* acetylation. We identified *BRPF1* deletions or point mutations in six additional individuals with a similar phenotype. Deletions of the 3p25 region, containing *BRPF1* and *SETD5*, cause a defined ID syndrome where most of the clinical features are attributed to *SETD5* deficiency. We compared the clinical symptoms of individuals carrying mutations or small deletions of *BRPF1* alone or *SETD5* alone with those of individuals with deletions encompassing both *BRPF1* and *SETD5*. We conclude that both genes contribute to the phenotypic severity of 3p25 deletion syndrome but that some specific features, such as ptosis and blepharophimosis, are mostly driven by *BRPF1* haploinsufficiency.

Intellectual disability (ID) characterizes a group of neurodevelopmental disorders that constitute a major public health, social, and educational problem because of the cumulated frequency and the heavy burden for affected individuals and families. ID is defined by significant limitations in both intellectual functioning and adaptive behavior associated with an intellectual quotient (IQ) below 70, and it affects about 2% of children or young adults. Moderate to severe forms of ID can be caused by chromosomal anomalies, including pathogenic deletions or duplications or single-gene defects with recessive, X-linked, or autosomal-dominant inheritance. More than 500 genes have been implicated in Mendelian forms of ID. Mutations can cause non-syndromic or syndromic ID with other associated clinical features. Additionally, a number of recurrent microdeletions also cause ID.

Terminal 3p and interstitial deletions of the 3p25–p26 region cause 3p deletion syndrome (MIM: 613792), characterized by mild to severe ID, growth retardation, microcephaly, and dysmorphic features, notably ptosis.[1] The terminal or interstitial deletions range from large deletions of several megabases to smaller deletions of fewer than 500 kb and do not always overlap, rendering it difficult to identify the genes associated with the phenotype. An increasing number of individuals harboring deletions of this region has advanced the understanding of the critical genes for this 3p25 region. Several individuals with a small 3p25.3 distal deletion present with a non-3p phenotype with ID, epilepsy, poor speech, ataxia, and stereotypic hand movements, and the two genes encoding GABA transporters, *SLC6A1* (MIM: 137165) and *SLC6A11* (MIM: 607952), were suspected to be involved.[2] For the more proximal deletions in 3p25, the most promising gene appears to be *SETD5* (MIM: 615743), encoding a putative histone methyltransferase. Indeed, variations in *SETD5* in individuals with ID and clinical features consistent with the 3p deletion syndrome have recently been reported.[3–5] However, some clinical features recurrent in 3p25 deletion syndrome, such as ptosis and

[1]Institut de Genetique et de Biologie Moleculaire et Cellulaire, 67400 Illkirch-Graffenstaden, France; [2]INSERM U964, 67400 Illkirch-Graffenstaden, France; [3]CNRS UMR 7104, 67400 Illkirch-Graffenstaden, France; [4]Université de Strasbourg, 67400 Illkirch, France; [5]Chaire de Génétique Humaine, Collège de France, 67400 Illkirch, France; [6]Service de Génétique Médicale, Hôpitaux Universitaires de Strasbourg, Institut de Génétique Médicale d'Alsace, 67000 Strasbourg, France; [7]Genetic Medicine, Belfast City Hospital, Belfast BT9 7AB, Ireland; [8]Spectrum Health Medical Group, Grand Rapids, MI 49544, USA; [9]Department of Clinical Genetics, Erasmus MC, Rotterdam 3015, the Netherlands; [10]Service de Génétique Clinique, Centre Hospitalier Universitaire de Grenoble site Nord, Hôpital Couple-Enfant, 38700 Grenoble, France; [11]Department of Pediatrics, McGovern Medical School, University of Texas in Houston, Houston, TX 77030, USA; [12]Department of Clinical Genetic, Academic Medical Center, Amsterdam 1100, the Netherlands; [13]INSERM 1209, CNRS UMR 5309, Laboratoire de Génétique Chromosomique, Centre Hospitalier Universitaire Grenoble Alpes, Institut Albert Bonniot, Université Grenoble Alpes, 38706 Grenoble, France; [14]GeneDx, Gaithersburg 20877, USA; [15]Neurometabolic Diseases Laboratory, Institute of Neuropathology, Institut d'Investigació Biomèdica de Bellvitge, 08908 Barcelona, Spain; [16]Center for Biomedical Research on Rare Diseases U759, L'Hospitalet de Llobregat, 08908 Barcelona, Spain; [17]Catalan Institution for Research and Advanced Studies, 08010 Barcelona, Spain; [18]Laboratoire de diagnostic génétique, Institut de Génétique Médicale d'Alsace, Hôpitaux Universitaires de Strasbourg, 67000 Strasbourg, France; [19]University of Strasbourg Institute for Advanced studies, 67000 Strasbourg, France

*Correspondence: jlmandel@igbmc.fr (J.-L.M.), piton@igbmc.fr (A.P.)

http://dx.doi.org/10.1016/j.ajhg.2016.11.010.

© 2017 American Society of Human Genetics.

blepharophimosis, are not consistently observed in individuals with *SETD5* mutations.

Here, we investigated the genetic origin of an autosomal-dominant syndromic form of mild ID associated with other features such as growth retardation, ptosis, and relative microcephaly, present in six affected relatives over three generations (Figure 1A). Ethical approval was obtained from the local ethics committees. The proband, III-2, was born at term with intrauterine growth restriction: weight 2,900 g (fifth percentile), height 46 cm (third percentile), and head circumference 32.5 cm (third percentile). Bilateral clubfeet were diagnosed during the pregnancy, and a karyotype was performed but was negative. At birth, edema of the back of the feet was noticed. He was hospitalized at the age of 1 month for the association of hypotonia and eating disorders without weight gain. The clinical examination found dysmorphic features with left ptosis, bilateral epicanthus, anteverted nostrils, a round face, a long philtrum, small and round ears, and unilateral cryptorchidism (Figure 3). Brachymetacarpia and clinodactyly of the toes were also noticed. Echocardiography, renal ultrasound, and cerebral echography found no anomaly. The cerebral computed tomography scan and hearing were normal. Gastroesophageal reflux was diagnosed. His development was significant for growth restriction and development of psychomotor delay. At 4 months old, the proband weighed 4,950 g (-1.5 SDs) and had a length of 54 cm (-3 SDs) and a head circumference of 39.5 cm (-1.5 SDs). At 4 years old, he weighed 14 kg (-1 SD) and had a length of 94 cm (-2 SDs) and a head circumference of 48.5 cm (-2 SDs). The boy sat at 16 months and walked at 30 months of age. He also presented with delayed language, and toilet training was acquired at 4 years of age. He had surgery for his ptosis and for cryptorchidism. His older brother (III-1) presented with no ID, growth disorder, or facial dysmorphism. However, his mother (II-2) presented with mild ID (permitting professional integration), short stature (150 cm), bilateral ptosis, facial dysmorphism similar to that of her son, and brachymetacarpia. Familial history revealed that her mother (deceased) and two of her sisters presented with the same phenotype. The phenotype is more severe for sister II-5, who had surgery twice for her ptosis with limited results and has had limited employment (Figure 3). She also presented with hypothyroidism. The other sister (II-3) also had surgery twice for her ptosis with limited results (Figure 3). She was 153 cm tall. Her daughter (III-4) presented with bilateral ptosis and mild ID with learning difficulties and concentration problems. Secondarily, the mother (II-2) had a new pregnancy: fetal echography showed a suspected anomaly of foot positioning, indicating possible clubfeet. At birth, the baby (III-3) had normal growth parameters: he weighed 3,560 g and had a length of 48 cm and a head circumference of 35 cm. He presented with pes varus, edema of the back of the feet, and the same facial dysmorphism as that of his brother. Progressively, the child presented with growth retardation, relative microcephaly, and developmental delay. At 19 months old, he could not walk. He weighed 8.2 kg (-3 SDs) and had a length of 75 cm (-2 SDs) and a head circumference of 45 cm (-2.5 SDs). His DNA was not available for testing. The child III-4, a cousin of the index individual, was born at term with short stature (47 cm), a normal weight (3,050 g), and a normal head circumference (33 cm). Bilateral ptosis was rapidly diagnosed and surgically repaired. Her motor development was within acceptable limits, given that she could sit at 8 months and walked at 18 months. Later, she presented with delayed language, difficulties at school, and behavioral disorders. Echocardiography, electroencephalogram, cerebral MRI, and a hearing test were normal. Unlike that of her cousins, her growth was normal: at 5.5 years, she weighed 22 kg and had a length of 114 cm and a head circumference of 50 cm. On clinical examination, the child presented with the same familial dysmorphism. Since then, the parents have had another child, who is in good health without developmental delay or facial dysmorphism.

The most severely affected individual (III-2) underwent multiple genetic tests before we decided to perform whole-exome sequencing (WES). In addition to karyotype, array comparative genomic hybridization, and fragile-X testing, many tests have been conducted, including evaluation of 22q11.2 (MIM: 611867) and 22q13.3 (MIM: 606232) deletion syndromes (by fluorescence in situ hybridization), as well as Prader-Willi (MIM: 176270) (15q11.2–q13 DNA methylation), DM1 myotonic dystrophy (MIM: 160900) (*DMPK* [MIM: 605377] expansion), Aarskog (MIM: 305400) (*FGD1* [MIM: 300546] sequencing), Noonan (MIM: 163950) (*PTPN11* [MIM: 176876], *SOS1* [MIM: 182530], *RAF1* [MIM: 164760], *SHOC2* [MIM: 602775] sequencing), and Saethre Chotzen (MIM: 101400) (*TWIST1* [MIM: 601622] and *FGFR3* [MIM: 134934] sequencing) syndromes.

Given that no pathogenic genetic event could be identified by these genetic investigations, we performed WES for individual III-2, his maternal cousin (III-4), and his maternal aunt (II-5). Libraries and captures from genomic blood DNA were done with the SureSelect XT Human All Exon V5 Kit (Agilent Technologies), and sequencing was performed on a 100 bp paired-end run on the HiSeq 2500 sequencer (Illumina). Reads were aligned and variants were called and annotated as previously described.[6,7] To identify a variant shared by the three affected individuals, we used the family barcode given by the VaRank ranking program.[6] Then, we filtered out the frequent mutations by using public databases and a large cohort of ID-affected individuals as previously described.[7] Applying these criteria, we identified four candidate variants: one loss-of-function (LoF) and three missense variants in the heterozygous state in all three affected members. The three missense variants, c.650G>A (p.Arg217His) (GenBank: NM_080668.3) in *CDCA5* (MIM: 609374), c.143C>T (p.Ser48Leu) (GenBank: NM_005199) in *CHRNG* (MIM: 100730), and c.1279C>T (p.Pro427Ser)

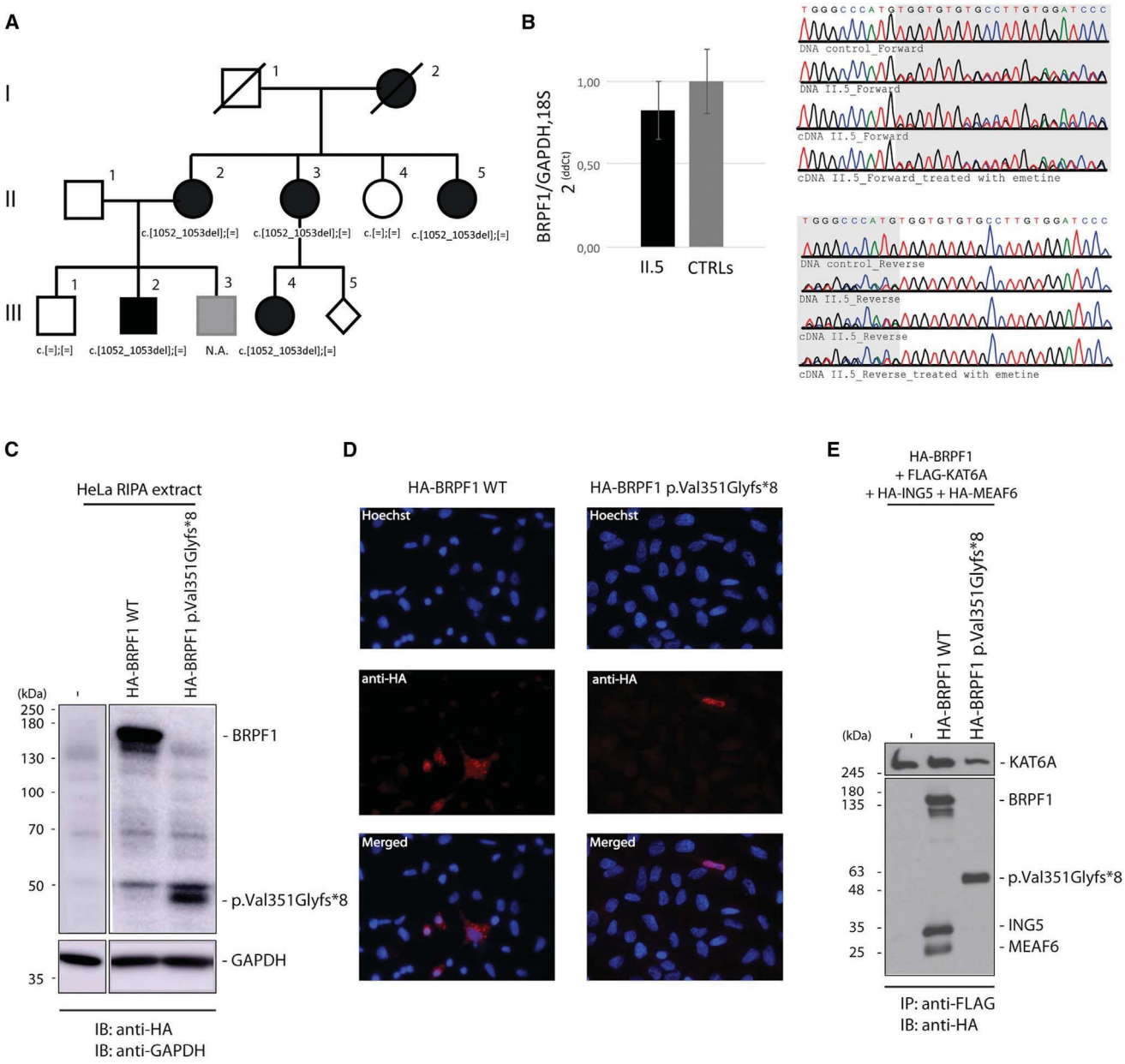


Figure 1. Identification of a Co-segregating 2 bp Deletion, c.1052_1053del, in *BRPF1* in a Family with Three Generations Affected by a Form of Mild ID Associated with Ptosis

(A) Pedigree of family A, which has three affected generations.

(B) The mutation partially escapes NMD. Quantitative real-time PCR was performed on RNA extracted (three extractions per individual) from fibroblasts of individual II-5 and three unrelated control individuals. The expression of *BRPF1* in relation to the average of two reference genes, *GAPDH* and *18S*, was calculated by the 2^{-DDCt} method. A t test was performed and showed no significant difference in the *BRPF1* mRNA level (error bars indicate the SD of three independent experiments). Sequences of blood DNA and fibroblast cDNA (treated or not with the NMD-blocker emetine) from individual II.5 are shown on the right.

(C) Expression of BRPF1 in HeLa cells. HeLa cells transfected with HA-tagged wild-type or p.Val351Glyfs*8 BRPF1 cDNA. Cells were harvested 36 hr after transfection. BRPF1 expression was analyzed by SDS-PAGE, and immunoblotting was performed with anti-HA antibody.

(D) HeLa cells were transfected with HA-tagged wild-type or p.Val351Glyfs*8 BRPF1 cDNA. BRPF1 localization was visualized by immunofluorescence with an anti-HA antibody. Nuclei were colored in blue by Hoechst staining.

(E) HA-tagged wild-type or mutant BRPF1 was transfected along with expression plasmids for FLAG-tagged KAT6A, HA-tagged ING5, and HA-tagged MEAF6 into HEK293 cells. HAT complexes were immunoprecipitated from protein extracts with anti-FLAG antibody to pull down KAT6A, and products of the complex were revealed by western blot using anti-HA antibody.

(GenBank: NM_198517) in *TBC1D10C* (MIM: 610831), were unlikely to be considered pathogenic for the syndromic ID phenotype (Table S1). The unique LoF variant identified was a 2 nt deletion, c.1052_1053del (GenBank: NM_001003694.1) in *BRPF1* (MIM: 602410), which encodes bromodomain and PHD finger-containing protein 1.

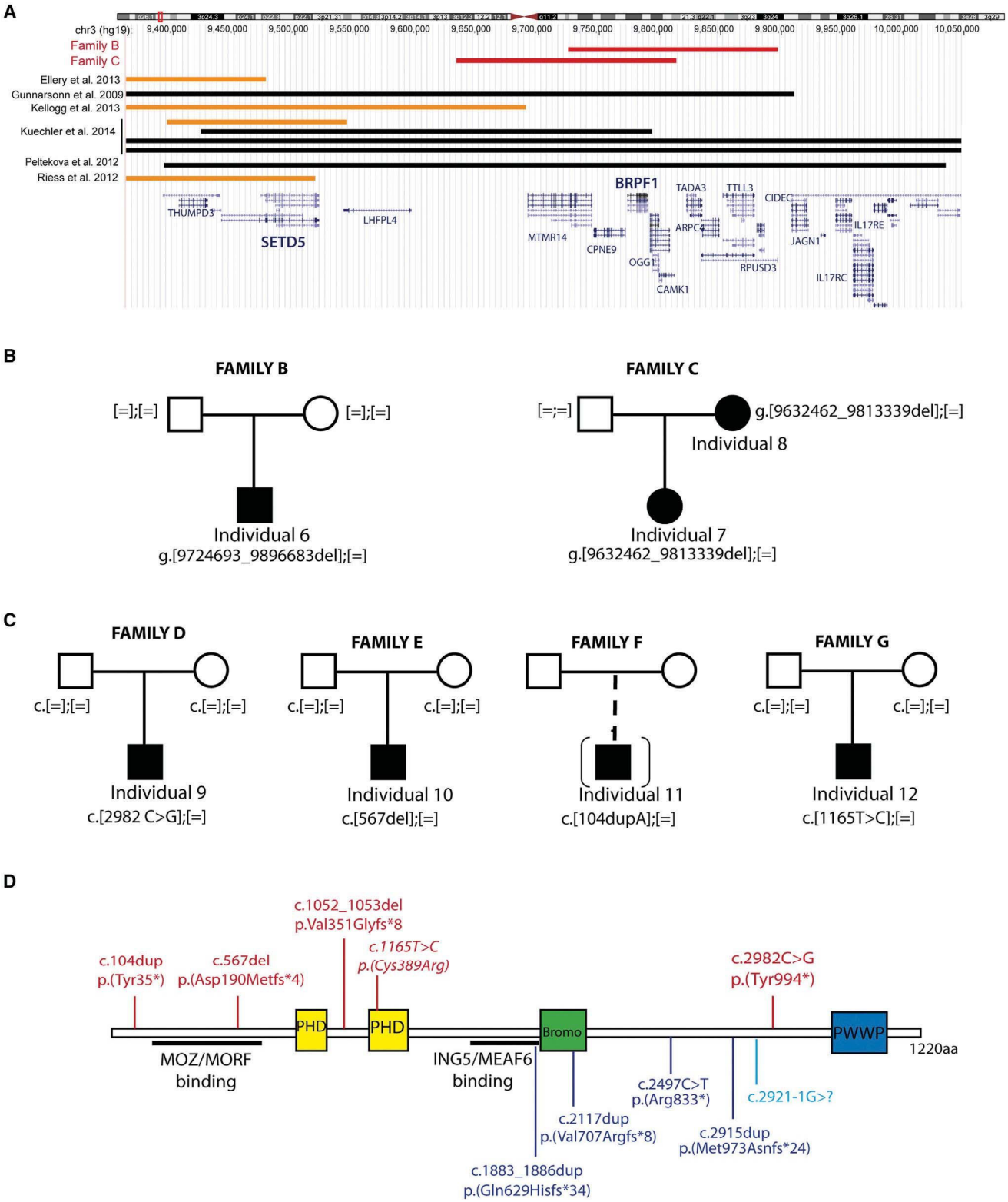


Figure 2. Mutations in *BRPF1* and Deletions in the 3p25 Region

(A) Overview of 3p25 deletions reported in the literature and in DECIPHER (from the UCSC Genome Browser). Black lines indicate a deletion encompassing both *SETD5* and *BRPF1*; orange lines represent the deletion containing *SETD5* but not *BRPF1*; and red lines indicate the two deletions including *BRPF1* but not *SETD5* reported in DECIPHER.

(B) Pedigree of the two additional families affected by 3p25 deletions encompassing *BRPF1* but not *SETD5*.

(C) Pedigree of four additional families with the *BRPF1* pathogenic variants shown in (D).

(legend continued on next page)

This deletion occurred in a well-conserved region, according to PhastCons and USCS Multiz alignment of 100 vertebrates and orthologs (from Ensembl), and was predicted to cause a frameshift leading to a premature stop codon eight amino acids downstream. Sanger sequencing in available family members confirmed that affected individuals carried deletion c.1052_1053del (Figure 1A), which has been added to ClinVar.

To evaluate whether *BRPF1* is tolerant of protein-truncating variants, we looked in the Exome Aggregation Consortium (ExAC) Browser, which contains 60,706 exomes from individuals unaffected by severe pediatric diseases. Here, we found five variants potentially leading to LoF in *BRPF1*: one nonsense and four splice variants. The nonsense variant is reported in one individual but present in only 21% of reads, suggesting a mosaic status (Table S2). The four splice variants are present in the heterozygous state. One of them is present in several (six) individuals but affects a known processed non-coding transcript (Ensembl: ENST00000469066.1). The three remaining variants are present in one individual each and affect canonical splice sites of exon 8 (324 nucleotides; might create an in-frame deletion of 108 amino acids), exon 9 (285 nucleotides; might create an in-frame deletion of 95 amino acids), or exon 11 (137 nucleotides; might create a frameshift). On the basis of gene length, 36 LoF variants in the *BRPF1* coding region could be expected for this gene; however, only five have been reported.[8] These data suggest that *BRPF1* is an extremely LoF-intolerant gene (probability of LoF intolerance ¼ 1).

To investigate whether the mutant *BRPF1* transcript undergoes nonsense-mediated decay (NMD), we obtained dermal fibroblasts from skin biopsy of individual II-5 and three unrelated control individuals. They were expanded as previously described.[9] Fibroblast RNA was extracted according to the TRI Reagent protocol (Molecular Research Center), treated with DNaseI (Roche Diagnostic), and reverse transcribed into cDNA with random hexamers and SuperScript II Reverse Transcriptase according to the manufacturer's recommendation. PCR was performed with specific primers (*BRPF1* 5^0-tgccagaacagcaatgtcatcctc-3^0 [forward] and 5^0-cgcacaggctccatcttcatgtaa-3^0 [reverse]). qPCR were performed in triplicate, and the *BRPF1* mRNA level was quantified by the 2^{-DDCt} method with an average of two reference genes, *GAPDH* and *18S*. A parametric Student's t test was performed to compare the relative *BRPF1* expression and revealed a slight but not significant decrease in *BRPF1* mRNA levels in individual II-5, suggesting that the mutated transcript partially escapes NMD. cDNA sequencing (GATC) from II-5 revealed the presence of both the wild-type and the mutant transcripts (Figure 1B), but peak heights were lower for the latter. A similar peak height could be restored when fibroblasts from individual II.5 were treated with emetine (100 mg/mL) to block NMD, confirming that the c.1052_1053del *BRPF1* transcript undergoes partial NMD.

The deletion leads to a frameshift with the appearance of a premature stop codon: p.Val351Glyfs*8. The truncated protein is predicted to contain 358 amino acids instead of 1,220 and lacks several essential functional domains, including the second PHD finger domain, the bromodomain, and the PWWP domains, which are involved in histone recognition and binding (Figure 2D). We were not able to detect wild-type BRPF1 by western blot in fibroblasts from control individuals with the anti-BRPF1 antibody (Peregrin N-16, sc-103110, Santa Cruz Biotechnology; PCRB-BRPF1-2A12, DSHB, University of Iowa). To evaluate the protein, we generated N-terminal HA-tagged wild-type and mutant BRPF1. HeLa cells were transfected (Lipofectamine 2000, Invitrogen), and total proteins were extracted after 36 hr. Western blot using anti-HA antibody revealed an ~50 kDA truncated BRPF1 that accumulated at a lower level than the wild-type, suggesting reduced stability (Figure 1C). Using fluorescence microscopy, we observed that wild-type BRPF1 localized to the cytoplasm with the formation of cytoplasmic puncta, as previously reported (Figure 1D).[12] By contrast, the mutant BRPF1 signal was weaker, and the truncated protein appeared to be more uniformly distributed in both the cytoplasm and nucleus.

BRPF1 is a chromatin regulator that promotes histone acetylation by bringing different histone acetyltransferases (HATs) of the MYST protein family (HBO1, KAT6A [also known as MOZ], and KAT6B [MORF]) into a complex with other regulator proteins, such as ING5 and MEAF6.[12,13] The truncated protein, p.Val351Glyfs*8, still contains the KAT6B and KAT6A interaction domains between amino acids 59 and 222.[12] A similarly truncated form of BRPF1 (DN-term1, truncated after amino acid 354) was still able to bind KAT6A.[13] However, the ING5-MEAF6 interaction is mediated by amino acids 540–640,[12] suggesting that p.Val351Glyfs*8 BRPF1 would not be able to bring these two proteins into the HAT complex. To test this, we transfected HA-tagged wild-type and p.Val351Glyfs*8 BRPF1, along with expression plasmids for FLAG-tagged KAT6A, HA-tagged ING5, and HA-tagged MEAF6, into HEK293 cells. The HAT complexes were immunoprecipitated from protein extracts with anti-FLAG antibody to pull down KAT6A, and products were analyzed by western blot using anti-HA antibody (Figure 1E). We observed that both wild-type and p.Val351Glyfs*8 BRPF1 were able to bind KAT6A. Whereas

(D) Top: schematic representation of BRPF1 and localization of the five different LoF and missense mutations. Bottom: the four de novo LoF variants described by the DDD project in individuals with neurodevelopmental conditions[10] (in dark blue) and the LoF variant identified in one boy with schizophrenia and mild ID.[11] Domains are colored as follows: yellow, PHD finger (PHD) domains; green, bromodomain (Bromo), involved in the recognition of acetylated lysine residues; blue, PWWP nucleosome-binding domain. Regions involved in binding with MOZ, MORF, ING5, and EAF6 are underlined.[12]

the wild-type was able to bind ING5 and MEAF6, the p.Val351Glyfs*8 variant failed to do so.

To investigate the effect of the *BRPF1* mutation on the global acetylation level of histone H3, we extracted histones from the fibroblasts of individual II-5 and three unrelated healthy control individuals. We used 2 mg of histone proteins to detect global histone H3 acetylation with the EpiQuik Global Histone H3 Acetylation Assay Kits (Epigentek) (Figure S1A). No significant difference in H3 acetylation levels was detected. To dissect more specifically the acetylation occurring at the different lysines known to be acetylated by the KAT6A-KAT6B HAT complex,[14,15] we performed western blot analysis on histone extractions with specific anti-H3K9 (ab4441, Abcam), anti-H3K14 (in house), and anti-H3K23 (9674, Cell Signaling) antibodies, and we normalized the intensities obtained to the intensity of global histone H3 (catalog no. 06755, lot 31949, Upstate). No change in acetylation levels was observed for H3K9 or K14 (Figure S1B); however, compared with control individuals, individual II-5 showed a slight but non-significant decrease in the acetylation level of H3K23. Histone H3 acetylation levels were also analyzed in histone extracts obtained from HeLa cells cotransfected with constructs encoding KAT6A, ING5, and MEAF6 with or without wild-type or p.Val351Glyfs*8 BRPF1. No difference was observed in the ability to stimulate K9 and K14 acetylation between wild-type and mutant BRPF1. However, unlike wild-type BRPF1, the p.Val351Glyfs*8 variant failed to stimulate K23 acetylation of histone H3 (Figure S1C).

BRPF1-KAT6A-KAT6B complexes are involved in the development of the forebrain and other organs in mice, and complete knockout causes embryonic lethality with vascular defects and abnormal neural tube closure.[16] Inactivation in mice and other animal models, including medaka fish, has demonstrated that BRPF1 acts through the regulation of *Hox* genes to effect skeletal development.[13,17] To determine whether BRPF1 also alters *HOX* expression in humans, we investigated the expression of human homologs of some *Hox* genes described as regulated by the murine BRPF1-KAT6A-KAT6B complex in individual II-5 fibroblasts. Results obtained for *HOXA7* (MIM: 142950) and *HOXC10* (MIM: 605560) were not interpretable as a result of variability in expression among control individuals (data not shown). However, low variability was observed in control individuals for the *HOXD8* (MIM: 142985) mRNA level, and we observed that the level of *HOXD8* mRNA was significantly higher in individual II-5 than in control individuals (Figure S2).

In order to confirm the association between *BRPF1* and ID, we performed data exchange to retrieve additional individuals carrying *BRPF1* mutations. We first queried DECIPHER to identify copy-number variants affecting *BRPF1* and identified two individuals with 3p25 deletions including *BRPF1* but not *SETD5* (a gene previously associated with ID) (Figure 2A). Clinical details of these two individuals are compared to the clinical symptoms of the first family (Tables 1 and 2; Table S3). The first individual has a de novo 172 kb deletion encompassing *BRPF1* and four other genes (family B individual 6; Figure 2B). The second has a 181 kb deletion including *BRPF1* and eight other genes (family C individual 7; Figure 2B; Figure 3); this was inherited from her mildly affected mother (individual 8). Both individuals have mild ID, ptosis or blepharophimosis, and a roundish face, clinical features that overlap those of members of the large family. Next, we used the GeneMatcher exchange database to search for ID-affected individuals with *BRPF1* mutations identified by WES analysis (where no other obvious candidate gene was present). We found three nonsense or frameshift variations—c.2982C>G (p. Tyr994*), c.567delT (p.Asp190Metfs*24), and c.104dupA (p.Tyr35*)—and one de novo missense variant, c.1165T>C (p.Cys389Arg) (Figure 2C). Two of the nonsense mutations occurred de novo, and one was from unknown inheritance (in an adopted boy with a family history in his biological family; no DNA was available for testing). The missense variant affects a well-conserved amino acid located in the second PHD domain and is predicted to be pathogenic (by SIFT and PolyPhen-2). These four individuals presented with mild to moderate ID, hand and feet anomalies, and similar facial appearances with the presence of ptosis (Tables 1 and 2; Table S3; Figure 3). In total, all individuals with *BRPF1* mutations or deletions have mild or moderate ID. Of the three individuals with moderate ID, two (individuals 10 and 11) carry the earliest truncating mutations, whose protein products would lack at least part of the interaction domain with KAT6A and KAT6B. This truncated protein product might increase the severity of the phenotype, but we cannot exclude other genetic or environmental modifiers in the variable expressivity of this disorder.

De novo truncating variants in *BRPF1* have also been recently reported in large studies: the Deciphering Developmental Disorders (DDD) study has reported four de novo LoF variations in BRPF1—c.1883_1886dup (p.Gln629Hisfs*34), c.2117dup (p.Val707Argfs*8), c.2497C>T (p.Arg833*), and c.2915dup (p.Met973Asnfs*24), identified in 4,293 UK individuals with neurodevelopmental disorders.[10] A de novo LoF variant was also reported in *BRPF1* in one male individual from a schizophrenia cohort.[11]

3p25 deletion syndrome is characterized by ID, growth retardation, microcephaly, hypotonia, and specific facial dysmorphism. The critical region contains *BRPF1* and *SETD5*, among other genes. Previous work has established that disruption of *SETD5* is involved in the cognitive phenotype of this 3p25 syndrome.[3–5] The identification of LoF mutations and deletions of *BRPF1* in individuals with ID led us to investigate the contribution of *BRPF1* in the 3p25 syndrome. We performed a genotype-phenotype comparison by using those individuals with mutations affecting either *BRPF1* (group 1) or *SETD5* (group 2) only as well as those with a 3p25 deletion including both *SETD5* and *BRPF1* (group 3) (Table 3; Table S3). For

Table 1. Clinical Features of Individuals Carrying *BRPF1* Mutations in Families A and B

	Family A					Family B
	Individual 1 (III.2)	Individual 2 (II.2)	Individual 3 (II.3)	Individual 4 (III.4)	Individual 5 (II.5)	Individual 6
Mutation (GenBank: NM_001003694.1)[a]	c.1052_1053del (p.Val351Glyfs*8)	c.1052_1053del (p.Val351Glyfs*8)	c.1052_1053del (p.Val351Glyfs*8)	c.1052_1053del (p.Val351Glyfs*8)	c.1052_1053del (p.Val351Glyfs*8)	deletion of chr3: 9,724,693–9,896,683
Mutation type	intragenic	intragenic	intragenic	intragenic	intragenic	NA, de novo
Sex	male	female	female	female	female	male
Age of examination	5 years, 9 months	32 years	34 years	6 years, 10 months	30 years	6 years, 6 months
Uneventful pregnancy	diagnosis of club feet	NA	NA	yes	NA	no (36.5 WoG)
Birth weight	$<5^{th}$ %	NA	NA	normal	NA	3^{rd} %
Birth length	$<3^{rd}$ %	NA	NA	$<5^{th}$ %	NA	NA
Birth OFC	$<3^{rd}$ %	NA	NA	$<5^{th}$ %	NA	NA
Neonatal hypotonia	yes	NA	NA	no	NA	no
Hypotonia	yes	NA	NA	yes	NA	no
Small stature	yes (104.5 cm; $<3^{rd}$ %)	yes (150 cm; $<3^{rd}$ %)	yes (153 cm; $<3^{rd}$ %)	no (122 cm)	yes (152 cm; $<3^{rd}$ %)	no (113 cm)
Low weight	yes (16 kg; $<3^{rd}$ %)	no (62 kg)	no (67 kg)	no (25 kg; $>90^{th}$ %)	NA	no (21 kg)
ID	mild	mild	mild	mild	mild	mild
Microcephaly	mild (50 cm; $<10^{th}$ %)	no (54.5 cm)	yes (53 cm; $<3^{rd}$ %)	no (50 cm)	mild (54 cm; $<10^{th}$ %)	mild (50.3 cm; $<10^{th}$ %)
Brain anomalies (MRI)	ACC (rostrum)	NA	NA	no	NA	NA
Seizures	no	no	no	no	no	no
Delay in walking	yes	NA	NA	yes	NA	no
Speech delay	yes	NA	NA	yes	NA	mild
Behavioral anomalies	no	no	no	hyperactivity	no	hyperactivity, shy, quiet
Vision or eye problems	strabismus, amblyopia	refraction problems	refraction problems	refraction problems	strabismus, amblyopia	NA
Ptosis and/or blepharophimosis	yes	yes	yes	yes	yes	yes (bilateral)
Hand anomalies	BM, BD	BM, BD	BM, BD	BM, BD	BM, BD	bilateral CD of fifth finger
Feet anomalies	clinodactyly, club feet	no	NA	no	no	syndactyly of the second and third toes

Abbreviations are as follows: %, percentile; ACC, agenesis of corpus callosum; BD, brachydactyly; BM, brachymetacarpia; CD, camptodactyly; MRI, magnetic resonance imaging; NA, information not available; OFC, occipital frontal circumference; and WoG, weeks of gestation.
[a]The mutation was absent from all of the available unaffected individuals in family A.

Table 2. Clinical Features of Individuals Carrying *BRPF1* Mutations or Deletions in Families C–G

	Family C		Family D	Family E	Family F	Family G
	Individual 7	Individual 8	Individual 9	Individual 10	Individual 11	Individual 12
Mutation (GenBank: NM_001003694.1)	deletion of chr3: 9,632,462–9,813,339		c.2982 C>G (p. Tyr994*)	c.567delT (p.Asp190Metfs*24)	c.104dupA (p.Tyr35*)	c.1165T>C (p.Cys389Arg)
Mutation type	NA, inherited from affected mother		intragenic, de novo	intragenic, de novo	intragenic, unknown	intragenic, de novo
Sex	female	female	male	male	male	male
Age of examination	3 years, 8 months	37 years	10 years	3 years	12 years	3 years, 9 months
Uneventful pregnancy	no (30 WoG)	NA	caesarean (37 WoG)	33 WoG	NA	yes
Birth weight	2,070 g	NA	normal	normal	NA	normal
Birth length	43 cm	NA	NA	43.2 cm (normal)	NA	NA
Birth OFC	30 cm	NA	NA	28cm ($<3^{rd}$ %)	NA	NA
Neonatal hypotonia	yes	NA	yes	no	NA	yes
Hypotonia	yes	NA	yes	no	NA	yes
Small stature	no	NA	141.5 cm	mild (91.4 cm; $<10^{th}$ %)	yes ($<3^{rd}$ %)	no
Low weight	no	NA	no (56.3 kg; $>97^{th}$ %)	no (17.7 kg; $>90^{th}$ %)	NA	no
ID	yes	mild	moderate	moderate	moderate	mild
Microcephaly	mild ($<10^{th}$ %)	NA	NA	mild (48 cm; $<10^{th}$ %)	yes ($<3^{rd}$ %)	mild ($<10^{th}$ %)
Brain anomalies (MRI)	NA	NA	yes[a]	NA	NA	no
Seizures	no	NA	yes	no	yes	no
Delay in walking	yes	yes	yes	yes	yes (mild)	yes
Speech delay	yes	NA	yes	no	yes (no words at 3 years)	yes (only few words)
Behavioral anomalies	impaired social interactions	shyness	NA	yes	hyperactivity, autism	very shy
Vision or eye problems	strabismus	NA	strabismus, refraction problems	NA	near sighted	strabismus
Ptosis and/or blepharophimosis	yes	NA	yes (bilateral)	yes (bilateral)	yes	yes
Hand anomalies	no	NA	CD (left second finger)	bilateral CD of fifth finger	bilateral CD of fifth finger	no
Feet anomalies	no	NA	long first toe	no	no	CD

Abbreviations are as follows: %, percentile; CD, camptodactyly; MRI, magnetic resonance imaging; NA, information not available; OFC, occipital frontal circumference; and WoG, weeks of gestation.
[a]Enlarged perivascular Virchow-Robin spaces.

Figure 3. Facial Characteristics of the Individuals with *BRPF1* Mutations
Pictures of individuals with *BRPF1* point mutations and deletions. Common features include a roundish face, blepharophimosis and ptosis, downslanted palpebral fissures, temporal narrowing, and a downturned mouth. Ethical approval was obtained from the local ethics committees. For all individuals included in this figure, families also gave consent for publication of the images.

individuals with *BRPF1* disruptions, only one index individual per family was taken into account. Clinical information for individuals with *SETD5* disruptions or 3p25 deletions of both *BRPF1* and *SETD5* was retrieved from the literature.[4,5,18,19] We observed that disruption of *SETD5* or *BRPF1* tends to lead to mild or moderate ID, whereas all of individuals with severe ID have disruptions of both *SETD5* and *BRPF1*. However, the degree of severity was evaluated by different clinical geneticists and lacked IQ testing for the individuals available, which could be biased. We performed a Fischer's exact test to compare clinical features between these groups. Although the majority of the individuals presented with delay in the acquisition of walking (86%, 83%, and 100% for groups 1, 2, and 3, respectively) and language (86%, 92%, and 100% for groups 1, 2, and 3, respectively), the severity is significantly increased in group 3. All individuals from group 3 acquired walking after 3 years of age (5/5 for group 3 versus 2/19 for groups 1 and 2, p value ¼ 0.0005) and presently have no language (5/5 for group 3 versus 1/19 for groups 1 and 2, p value ¼ 0.0001) (Table 3), suggesting that disruption of both *BRPF1* and *SETD5* contributes to the phenotype of 3p25 deletion syndrome. Interestingly, both genes encode proteins involved in histone modification and gene regulation, and they might have common targets. *SETD5* encodes a methyltransferase involved in the methylation of histones H3 and H4, whereas *BRPF1* binds methylated histone H3 and promotes its acetylation.

To investigate the contribution of *BRPF1* disruption to particular clinical features of the 3p25 microdeletion syndrome, we compared all individuals with disruptions in *BRPF1*, with or without *SETD5* disruptions (group 1 þ 3), with those individuals with only *SETD5* disruptions (group 2). A significant difference was observed between the two groups for the presence of microcephaly or borderline small head size (10/10 in group 1 þ 3 versus 1/13 in group 2, p value < 0.0001) and unilateral or bilateral ptosis and/or blepharophimosis (12/12 in group 1 þ 3 versus 1/14 in group 2, p value < 0.0001). These eye and/or eyelid anomalies were present in all individuals carrying a disruption of *BRPF1*. Other clinical features (small stature and strabismus) were enriched in individuals with *BRPF1* disruptions; however, these differences were not significant after Bonferroni correction for multiple testing (threshold p value < 0.0017). Better delineating other clinical features driven by *BRPF1* haploinsufficiency will require a larger cohort.

Recently, mutations in *KAT6B* (MIM: 605880) have been associated with syndromic ID, including Ohdo syndrome (MIM: 603736), genitopatellar syndrome (MIM: 606170), blepharophimosis-ptosis-epicanthus inversus syndrome, and even a Noonan-syndrome-like phenotype.[20–23] Mutations in *KAT6A* (MIM: 601408) are associated with ID with craniofacial dysmorphism, microcephaly or craniosynostosis, feeding difficulties, cardiac defects, and ocular anomalies (MIM: 616268).[24,25]

Zebrafish and mouse models of Brpf1 and BRPF1 disruption, respectively, are reported in the literature.[13,26] Zebrafish mutants show craniofacial defects, with shifts in segmental identities of craniofacial arches, as a result of a progressive loss of anterior *Hox* gene expression, indicating that Brpf1 plays a role in patterning the vertebrate head by mediating the expression of *Hox* genes. Mice with homozygous *Brpf1* deletion show embryonic lethality with different embryonic defects, including abnormal neural tube closure.[16,26]The forebrain-specific deletion of *Brpf1* results in early postnatal lethality and growth retardation. Viable mice show neocortical abnormalities, partial agenesis of the corpus callosum, and

Table 3. Dissection of *SETD5* and *BRPF1* Contributions to Clinical Features of 3p25 Deletion Syndrome

	Group 1 (*BRPF1* Only)		Group 2 (*SETD5* Only)		Group 3 (Both *SETD5* and *BRPF1*)	
	Percentage	Number	Percentage	Number	Percentage	Number
ID	100%	7/7	100%	14/14	100%	5/5
Mild or moderate ID	100%	6/6	100%	6/6	40%	2/5
Severe ID	0%	0/6	0%	0/6	60%	3/5[a]
General Characteristics						
Uneventful pregnancy (born at term)	33%	2/6	64%	9/14	40%	2/5
Low birth parameters	33%	2/6	8%	1/13	0%	0/4
Small stature	43%	3/7[b]	15%	2/13	100%	4/4[a]
Microcephaly or borderline small head size	100%	6/6[b,c]	8%	1/13	100%	4/4
Development						
Walking delay	86%	6/7	83%	10/12	100%	5/5
Severe walking delay (>3 years)	0%	0/7	17%	2/12	100%	5/5[a,c]
Speech delay	86%	6/7	92%	12/13	100%	5/5
No speech	0%	0/7	8%	1/12	100%	5/5[a,c]
Neurological Features						
Seizures	29%	2/7	21%	3/14	80%	4/5[a]
Hypotonia	67%	4/6	67%	4/6	100%	4/4
Brain anomalies (MRI)	67%	2/3	0%	0/4	25%	1/4
Behavioral anomalies	71%	5/7	77%	10/13	25%	1/4
Others Features						
Strabismus	80%	4/5[b]	36%	5/14	100%	4/4
Ptosis and/or blepharophimosis	100%	7/7[b,c]	7%	1/14	100%	5/5[a]
Hand anomalies	71%	5/7	50%	7/14	80%	4/5
Feet anomalies	57%	4/7	15%	2/13	40%	2/5
Congenital heart defect	0%	0/7	15%	2/13	40%	2/5

Clinical information for individuals with *SETD5* point mutations or deletions (group 2) and individuals with large 3p25 deletions encompassing *SETD5* and *BRPF1* was retrieved from the literature.[4,5,18,19] Clinical information for individuals with *BRPF1* point mutations or small 3p25 deletions reported in this publication (group 1) was retrieved from physicians attending the families. For the sake of avoiding artifacts, one member per family was considered. A 2 3 2 contingency table was made for analyzing the presence of each clinical sign, and because of the small sample size, a two-tailed Fisher's exact test was used to calculate the p value to highlight a statistically significant difference between groups.

[a]Clinical feature more prevalent when both genes are deleted (group 3) than when only one gene is deleted (groups 1 and 2) (p value < 0.05).

[b]Clinical feature significantly more associated with *BRPF1* disruption, with or without *SETD5* (group 1 and 3), than with *SETD5* disruption only (group 2) (p value < 0.05).

[c]Significant after Bonferroni correction for multiple testing (p value < 0.0017).

behavioral anomalies.[27] Interestingly, the investigators observed an alteration in the expression of several transcription factors involved in developmental processes and upregulation of *Hox* gene expression. These data indicate that *Brpf1* is involved in forebrain development and acts as both an activator and a repressor of gene expression.

Certain chromatin modifiers that are associated with ID when mutated in the germline are also associated with childhood cancer when mutated at the somatic level, for example, *SETBP1* (MIM: 611060) and *KMT2A* (MIM: 159555). Several somatic mutations affecting different regions of *BRPF1* have been reported in childhood leukemia[28] and adult medulloblastoma.[29]

In conclusion, we report here that LoF point mutations and small deletions affecting *BRPF1* are responsible for a syndromic form of ID associated with eye and/or eyelid phenotype, i.e., ptosis and/or blepharophimosis. *BRPF1* encodes the third member of the HAT KAT6A-KAT6B complex, which is involved in ID when functionally impaired.

We have therefore shown that *BRPF1*, together with *SETD5*, contributes to the severity of the 3p25 deletion syndrome phenotype and is responsible for some specific clinical features, such as ptosis and blepharophimosis.

Accession Numbers

The accession number for the data reported in this paper is ClinVar: SCV000328673.1.

Supplemental Data

Supplemental Data include two figures and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.11.010.

Acknowledgments

The authors thank the families for their participation in this study. The authors also thank Fondation Jerome Lejeune, Fondation Maladies Rares, and Association APLM for financial support. This study was also supported by grant ANR-10-LABX-0030-INRT, a French state fund managed by the Agence Nationale de la Recherche under the frame program Investissements d'Avenir ANR-10-IDEX-0002-02. The authors also thank all members of the Strasbourg Hospital molecular diagnostic lab, the Clinical Genetics Service of Prof. Hélène Dollfus, the Institut Génétique Biologie Moléculaire Cellulaire sequencing platform, and UMR_S 1112 (Bernard Jost, Stéphanie Le Gras, Mathieu Jung, Jean Muller, and Véronique Geoffroy) for their technical and bioinformatics support. We also thank Sylvain Daujat, Robert Schneider, Federica Evangelista, Tiago Baptista, and Lazlo Tora for histone H3 antibodies and technical advice; and Xiang-Jiao Yang for the cDNA of *BRPF1*, *KAT6A*, *ING5*, and *MEAF6* expression plasmids and technical help. J.J. is an employee of GeneDx, and A. Pujol is a consultant for GeneDx.

Received: August 16, 2016
Accepted: November 11, 2016
Published: December 8, 2016

Web Resources

ClinVar, http://www.ncbi.nlm.nih.gov/clinvar/
dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/
Decipher, https://decipher.sanger.ac.uk/
ExAC Browser, http://exac.broadinstitute.org/
GeneMatcher, https://genematcher.org/
Integrative Genomics Viewer (IGV), http://www.broadinstitute.org/igv/
Mutation Nomenclature, http://www.hgvs.org/mutnomen/recs.html
NHLBI Exome Sequencing Project (ESP) Exome Variant Server, http://evs.gs.washington.edu/EVS/
OMIM, http://www.omim/org/
RefSeq, http://www.ncbi.nlm.nih.gov/RefSeq
UCSC Genome Browser, http://genome.ucsc.edu/

References

1. Narahara, K., Kikkawa, K., Murakami, M., Hiramoto, K., Namba, H., Tsuji, K., Yokoyama, Y., and Kimoto, H. (1990). Loss of the 3p25.3 band is critical in the manifestation of del(3p) syndrome: karyotype-phenotype correlation in cases with deficiency of the distal portion of the short arm of chromosome 3. Am. J. Med. Genet. *35*, 269–273.
2. Dikow, N., Maas, B., Karch, S., Granzow, M., Janssen, J.W., Jauch, A., Hinderhofer, K., Sutter, C., Schubert-Bast, S., Anderlid, B.M., et al. (2014). 3p25.3 microdeletion of GABA transporters SLC6A1 and SLC6A11 results in intellectual disability, epilepsy and stereotypic behavior. Am. J. Med. Genet. A. *164A*, 3061–3068.
3. Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet *380*, 1674–1682.
4. Grozeva, D., Carss, K., Spasic-Boskovic, O., Parker, M.J., Archer, H., Firth, H.V., Park, S.M., Canham, N., Holder, S.E., Wilson, M., et al.; UK10K Consortium (2014). De novo loss-of-function mutations in SETD5, encoding a methyltransferase in a 3p25 microdeletion syndrome critical region, cause intellectual disability. Am. J. Hum. Genet. *94*, 618–624.
5. Kuechler, A., Zink, A.M., Wieland, T., Lüdecke, H.J., Cremer, K., Salviati, L., Magini, P., Najafi, K., Zweier, C., Czeschik, J.C., et al. (2015). Loss-of-function variants of SETD5 cause intellectual disability and the core phenotype of microdeletion 3p25.3 syndrome. Eur. J. Hum. Genet. *23*, 753–760.
6. Geoffroy, V., Pizot, C., Redin, C., Piton, A., Vasli, N., Stoetzel, C., Blavier, A., Laporte, J., and Muller, J. (2015). VaRank: a simple and powerful tool for ranking genetic variants. PeerJ *3*, e796.
7. Redin, C., Gérard, B., Lauer, J., Herenger, Y., Muller, J., Quartier, A., Masurel-Paulet, A., Willems, M., Lesca, G., El-Chehadeh, S., et al. (2014). Efficient strategy for the molecular diagnosis of intellectual disability using targeted high-throughput sequencing. J. Med. Genet. *51*, 724–736.
8. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
9. Lowry, W.E., Richter, L., Yachechko, R., Pyle, A.D., Tchieu, J., Sridharan, R., Clark, A.T., and Plath, K. (2008). Generation of human induced pluripotent stem cells from dermal fibroblasts. Proc. Natl. Acad. Sci. USA *105*, 2883–2888.
10. McRae, J.F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., Alvi, M., et al. (2016). Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation. bioRxiv. http://dx.doi.org/10.1101/049056.
11. Xu, B., Roos, J.L., Dexheimer, P., Boone, B., Plummer, B., Levy, S., Gogos, J.A., and Karayiorgou, M. (2011). Exome sequencing supports a de novo mutational paradigm for schizophrenia. Nat. Genet. *43*, 864–868.
12. Ullah, M., Pelletier, N., Xiao, L., Zhao, S.P., Wang, K., Degerny, C., Tahmasebi, S., Cayrou, C., Doyon, Y., Goh, S.L., et al. (2008). Molecular architecture of quartet MOZ/MORF histone acetyltransferase complexes. Mol. Cell. Biol. *28*, 6828–6843.
13. Laue, K., Daujat, S., Crump, J.G., Plaster, N., Roehl, H.H., Kimmel, C.B., Schneider, R., Hammerschmidt, M.; and Tübingen 2000 Screen Consortium (2008). The multidomain protein Brpf1 binds histones and is required for Hox gene expression and segmental identity. Development *135*, 1935–1946.
14. Doyon, Y., Cayrou, C., Ullah, M., Landry, A.J., Côté, V., Selleck, W., Lane, W.S., Tan, S., Yang, X.J., and Côté, J. (2006). ING tumor suppressor proteins are critical regulators of chromatin acetylation required for genome expression and perpetuation. Mol. Cell *21*, 51–64.

15. Voss, A.K., Collin, C., Dixon, M.P., and Thomas, T. (2009). Moz and retinoic acid coordinately regulate H3K9 acetylation, Hox gene expression, and segment identity. Dev. Cell *17*, 674–686.
16. You, L., Yan, K., Zou, J., Zhao, H., Bertos, N.R., Park, M., Wang, E., and Yang, X.J. (2015). The chromatin regulator Brpf1 regulates embryo development and cell proliferation. J. Biol. Chem. *290*, 11349–11364.
17. Hibiya, K., Katsumoto, T., Kondo, T., Kitabayashi, I., and Kudo, A. (2009). Brpf1, a subunit of the MOZ histone acetyl transferase complex, maintains expression of anterior and posterior Hox genes for proper patterning of craniofacial and caudal skeletons. Dev. Biol. *329*, 176–190.
18. Ellery, P.M., Ellis, R.J., and Holder, S.E. (2014). Interstitial 3p25 deletion in a patient with features of 3p deletion syndrome: further evidence for the role of SRGAP3 in mental retardation. Clin. Dysmorphol. *23*, 29–31.
19. Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., et al. (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am. J. Hum. Genet. *94*, 677–694.
20. Campeau, P.M., Lu, J.T., Dawson, B.C., Fokkema, I.F., Robertson, S.P., Gibbs, R.A., and Lee, B.H. (2012). The KAT6B-related disorders genitopatellar syndrome and Ohdo/SBBYS syndrome have distinct clinical features reflecting distinct molecular mechanisms. Hum. Mutat. *33*, 1520–1525.
21. Clayton-Smith, J., O'Sullivan, J., Daly, S., Bhaskar, S., Day, R., Anderson, B., Voss, A.K., Thomas, T., Biesecker, L.G., Smith, P., et al. (2011). Whole-exome-sequencing identifies mutations in histone acetyltransferase gene KAT6B in individuals with the Say-Barber-Biesecker variant of Ohdo syndrome. Am. J. Hum. Genet. *89*, 675–681.
22. Kraft, M., Cirstea, I.C., Voss, A.K., Thomas, T., Goehring, I., Sheikh, B.N., Gordon, L., Scott, H., Smyth, G.K., Ahmadian, M.R., et al. (2011). Disruption of the histone acetyltransferase MYST4 leads to a Noonan syndrome-like phenotype and hyperactivated MAPK signaling in humans and mice. J. Clin. Invest. *121*, 3479–3491.
23. Yu, H.C., Geiger, E.A., Medne, L., Zackai, E.H., and Shaikh, T.H. (2014). An individual with blepharophimosis-ptosis-epicanthus inversus syndrome (BPES) and additional features expands the phenotype associated with mutations in KAT6B. Am. J. Med. Genet. A. *164A*, 950–957.
24. Arboleda, V.A., Lee, H., Dorrani, N., Zadeh, N., Willis, M., Macmurdo, C.F., Manning, M.A., Kwan, A., Hudgins, L., Barthelemy, F., et al.; UCLA Clinical Genomics Center (2015). De novo nonsense mutations in KAT6A, a lysine acetyltransferase gene, cause a syndrome including microcephaly and global developmental delay. Am. J. Hum. Genet. *96*, 498–506.
25. Tham, E., Lindstrand, A., Santani, A., Malmgren, H., Nesbitt, A., Dubbs, H.A., Zackai, E.H., Parker, M.J., Millan, F., Rosenbaum, K., et al. (2015). Dominant mutations in KAT6A cause intellectual disability with recognizable syndromic features. Am. J. Hum. Genet. *96*, 507–513.
26. You, L., Zou, J., Zhao, H., Bertos, N.R., Park, M., Wang, E., and Yang, X.J. (2015). Deficiency of the chromatin regulator BRPF1 causes abnormal brain development. J. Biol. Chem. *290*, 7114–7129.
27. You, L., Yan, K., Zou, J., Zhao, H., Bertos, N.R., Park, M., Wang, E., and Yang, X.J. (2015). The lysine acetyltransferase activator Brpf1 governs dentate gyrus development through neural stem cells and progenitors. PLoS Genet. *11*, e1005034.
28. Huether, R., Dong, L., Chen, X., Wu, G., Parker, M., Wei, L., Ma, J., Edmonson, M.N., Hedlund, E.K., Rusch, M.C., et al. (2014). The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes. Nat. Commun. *5*, 3630.
29. Kool, M., Jones, D.T., Jäger, N., Northcott, P.A., Pugh, T.J., Hovestadt, V., Piro, R.M., Esparza, L.A., Markant, S.L., Remke, M., et al.; ICGC PedBrain Tumor Project (2014). Genome sequencing of SHH medulloblastoma predicts genotype-related response to smoothened inhibition. Cancer Cell *25*, 393–405.

The American Journal of Human Genetics, Volume *100*

Supplemental Data

Mutations in Histone Acetylase Modifier *BRPF1* Cause an Autosomal-Dominant Form of Intellectual Disability with Associated Ptosis

Francesca Mattioli, Elise Schaefer, Alex Magee, Paul Mark, Grazia M. Mancini, Klaus Dieterich, Gretchen Von Allmen, Marielle Alders, Charles Coutton, Marjon van Slegtenhorst, Gaëlle Vieville, Mark Engelen, Jan Maarten Cobben, Jane Juusola, Aurora Pujol, Jean-Louis Mandel, and Amélie Piton

Supplemental Figures

Figure S1.

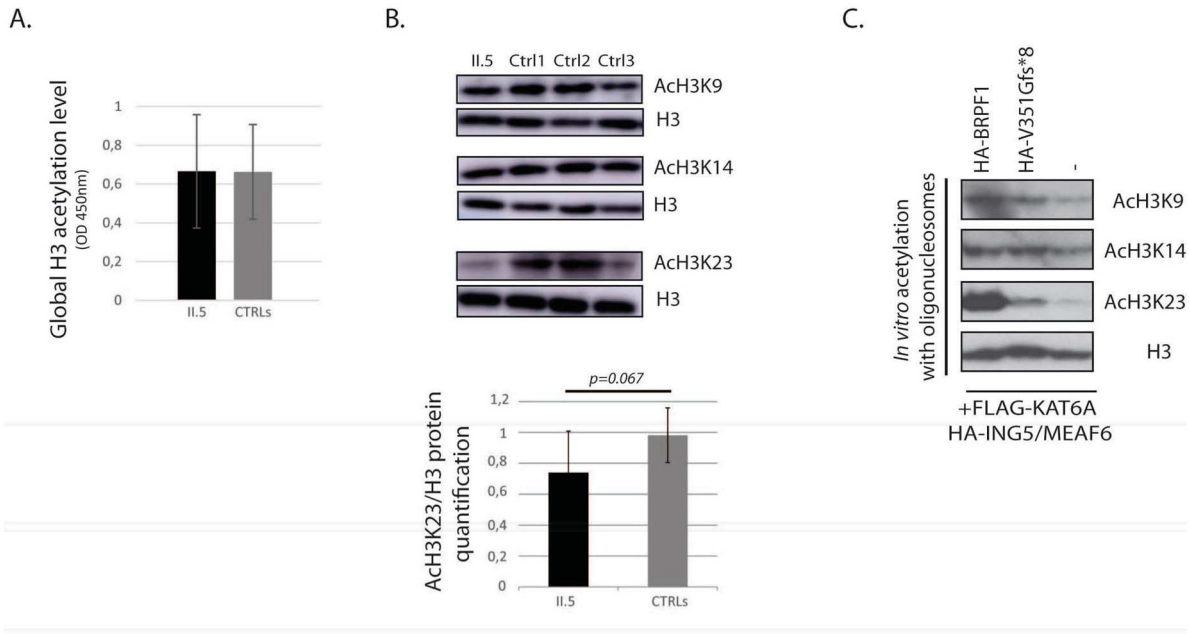


Figure S1. Effect of the c.1052_1053del variant on histone H3 acetylation level. (A) Global histone H3 acetylation level was compared between individual II-5 and three unrelated control individuals. Histone were extracted from fibroblasts and H3 acetylation level was measured (experiments performed in triplicate, error bars indicate the SD) using the EpiQuik Global Histone H3 kit. (B) Immunoblot performed on histone extracts from fibroblasts of individual II-5 and three unrelated control individuals (four extractions per individual) with specific anti-H3K9ac (Abcam, ab4441), anti-H3K14ac and anti-H3K23ac (cell signaling, #9674) antibodies and global H3 antibody (upstate, cat 06755, lot 31949). No difference was observed in H3K9 or H3K14 acetylation intensity between patient II-5 and controls when normalized by the intensities obtained for global H3. A non-significant decrease of H3K23 acetylation level was however observed for the patient II-5 compared to the three unrelated controls (experiments performed on four histone extractions per individuals) (C) HeLa cells were cotransfected with MOZ/KAT6A, ING5 and MEAF6 cDNAs with or without wild-type or mutant BRPF1 cDNA. Oligonuclesomes were extracted and *in vitro* histone acetylation assays were performed as previously described[7]. Acetylation levels were analyzed by immunoblotting using antibodies against histone H3 and its acetylated forms.

Figure S2.

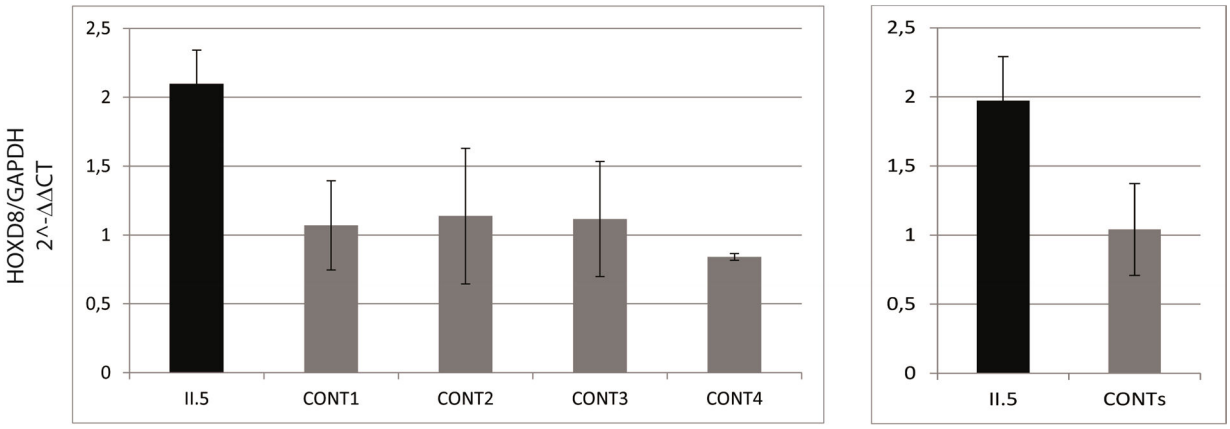


Figure S2. Significant increase in HOXD8 expression level was observed in individual II-5 cells. Quantitative real-time PCR was performed on reverse-transcribed mRNA extracted from fibroblasts of patient II-5 and of four unrelated controls (experiments performed on four different extractions per individual). The relative expression of HOXD8 *v.s*. GAPDH was calculated using the $2^{-(\Delta\Delta Ct)}$ method. Error bars indicate the SD of four independent experiments. A t-test was performed and showed a significant increase of *HOXD8* mRNA level in patient's fibroblasts.

Supplemental Tables

Table S1.

Gene	Variation	Status	Type of variation	Consequensis	At the protein level	dbSNP	ExAC	SIFT	Polyphen2
CDCA5	NM 080668.3:c.650G>A	heterozygous	substitution	missense	p.Arg217His	rs771200184	1 heterozygous	Deleterious (0)	POSSIBLY DAMAGING (0.828)
CHRNG	NM 005199: c.143C>T	heterozygous	substitution	missense	p.Ser48Leu	rs370022034		Deleterious (0.04)	BENIGN (0.108)
TBC1D10C	NM 198517: c.1279C>T	heterozygous	substitution	missense	p.Pro427Ser	rs756681324	15 heterozygous	Tolerated (0.12)	BENIGN (0.005)
BRPF1	NM 001003694: c.1052 1053del	heterozygous	deletion	frameshift	p.Val351Glyfs*8			NA	NA

Table S1. Rare non-synonymous variants identified by WES and common to the three affected family members sequenced.

Table S2.

Chrom	Position	Reference	Alternate	Protein Consequence	Transcript Consequence	Annotation	Allele Count	Allele Number	Number of homozygous	Population	Comments
3	9783787	C	T	p.Gln645Ter	ENST00000383829.2:c.1933C>T	stop gained	1	121 384	0	African	The variation is present in 14/65 reads. It could be a mosaic variant
3	9785260	A	C	p.?	ENST00000383829.2:c.2312 2A>C	splice acceptor	1	112 398	0	European (Non Finnish)	It affects the acceptor splice site of exon 8 (324 nts length), might create an in frame deletion of 108 a.a (Ala771 to Lys878)
3	9785263	A	G	p.?	ENST00000469066.1:n.217 2A>G	splice acceptor	6	113 094	0	LATINO	It affects a known processed transcript, not coding for a protein
3	9786193	G	A	p.?	ENST00000383829.2:c.2920+1G>A	splice donor	1	105 458	0	LATINO	It affects the donor splice site of exon 9 (285 nts length), might create an in frame deletion of 95 a.a (Gly879 to Met973). Rs191236303
3	9787255	AG	A	p.?	ENST00000383829.2:c.3069 1delG	splice acceptor	1	121 324	0	European (Non Finnish)	It affects 137 nts, might create a frameshift the acceptor splice site of exon 11 (137 nts length), might cause a frameshift

Table S2. Splice and nonsense variants identified in *BRPF1* in the ExAC general population

Table S3.

	Groupl	Group 2				Group 3
Reference	*this report*	*Grozeva et al. 2014*	*Pinto et al. 2014*	*Kuechler et al. 201S* (Patient 1, 2, 3, Riess et al. 201, Kellogg et al. 2013)	*Ellery et al. 2014*	*Kuechler et al. 201S* (Patient 4, 5, 6, Gunnarsson et al., Peltekova et al.)
Type of mutation	*BRPF1* Lof or deletions w/o *SETDS*	*SETDS* LoF	*SETDS* deletion	*SETDS* LoF and deletions w/ot *BRPF1*	*SETDS* deletion	3p25 deletions encompassing both *SETDS* and *BRPF1*
Sex	6M, 1F	7M	M	4F, 1M	M	4F, 1M
ID severity:	7/7	7/7	1	5/5	1	5/5
severity:						
mild	3/6		1	4/5		1/5
moderate	3/6			1/5		1/5
severe	0/6			0/5		3/5
General characteristics:						
Uneventful pregnancy (born at term)	2/6	4/7	1	4/5	0	2/5
low birth weight height/growth retardation	2/6	0/7	0	1/5	n.a.	0/4
Small stature	3/7	1/7	0	1/5		4/4
Microcephaly/smaller head size	6/6	1/7	0	0/5	n.a.	4/4
Development						
Mild walking delay (>18mo <3y)	6/7	6/7	0	4/4		0/5
Severe delay at walk (>3y)	0/7	2/7	0	0/4		5/5
speech delay	6/7	7/7	1	3/4	1	5/5

no speech	0/7	1/7	0	0/3	0	5/5
Neurological features						
Behavioral anomalies	5/7	5/7	1	4/5	n.a.	1/5
Brain anomalies (MRI)	2/3	0 (6n.a.)	n.a.	0/3	n.a.	1/4
Hypotonia	4/6	n.a.	n.a.	3/5	1	4/4
Seizures	2/7	0/7	0	2/5	1	4/5
Facial dysmorphisms						
Strabismus	4/5	1/7		4/5		3/5
Ptosis and/or blepharophimosis	7/7	1/7	0	0/5		5/5
Limbs						
Hand anomalies	5/7	2/7	0	4/5	1	4/5
Feet anomalies	4/7	1/7	0	1/4		2/5
Others features:						
Congenital heart defects	0/7	2/7	0	0/5		2/5

Table S3. Clinical features of patients reported in literature with mutations/deletions of *SETD5* (Group 2) or deletions encompassing both *SETD5* and *BRPF1* (Group 3)

Supplemental References

1. Geoffroy, V., Pizot, C., Redin, C., Piton, A., Vasli, N., Stoetzel, C., Blavier, A., Laporte, J., and Muller, J. (2015). VaRank: a simple and powerful tool for ranking genetic variants. PeerJ 3, e796.
2. Redin, C., Gerard, B., Lauer, J., Herenger, Y., Muller, J., Quartier, A., Masurel Paulet, A., Willems, M., Lesca, G., El Chehadeh, S., et al. (2014). Efficient strategy for the molecular diagnosis of intellectual disability using targeted high throughput sequencing. J Med Genet 51, 724 736.
3. Ellery, P.M., Ellis, R.J., and Holder, S.E. (2014). Interstitial 3p25 deletion in a patient with features of 3p deletion syndrome: further evidence for the role of SRGAP3 in mental retardation. Clin Dysmorphol 23, 29 31.
4. Grozeva, D., Carss, K., Spasic Boskovic, O., Parker, M.J., Archer, H., Firth, H.V., Park, S.M., Canham, N., Holder, S.E., Wilson, M., et al. (2014). De novo loss of function mutations in SETD5, encoding a methyltransferase in a 3p25 microdeletion syndrome critical region, cause intellectual disability. Am J Hum Genet 94, 618 624.
5. Kuechler, A., Zink, A.M., Wieland, T., Ludecke, H.J., Cremer, K., Salviati, L., Magini, P., Najafi, K., Zweier, C., Czeschik, J.C., et al. (2015). Loss of function variants of SETD5 cause intellectual disability and the core phenotype of microdeletion 3p25.3 syndrome. Eur J Hum Genet 23, 753 760.
6. Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., et al. (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet 94, 677 694.
7. Ullah, M., Pelletier, N., Xiao, L., Zhao, S.P., Wang, K., Degerny, C., Tahmasebi, S., Cayrou, C., Doyon, Y., Goh, S.L., et al. (2008). Molecular architecture of quartet MOZ/MORF histone acetyltransferase complexes. Mol Cell Biol 28, 6828 6843.

2. DE NOVO TRUNCATING VARIANTS IN THE NEURONAL SPLICING FACTOR *NOVA2* CAUSE A SYNDROMIC FORM OF INTELLECTUAL DISABILITY WITH ANGELMAN-LIKE FEATURES

Many genes that play important roles in the development of the nervous system undergo alternative splicing to generate protein variants with different functions. The use of alternative mRNA isoforms for the regulation of gene expression is a critical process during neuronal development, since specific proteins are required at different time and space. mRNA regulation is coordinated by different RNA-binding proteins (RBPs). NOVA (Neuro-Oncological Ventral Antigen) proteins NOVA1 and NOVA2 are two RBPs, which are involved in neuronal-specific alternative splicing. They have been first described as antigens in patients with a paraneoplastic neurologic syndrome (POMA), a neurological disorder characterized by ataxia with or without opsoclonus- myoclonus, with or without dementia, encephalopathy and cortical deficits along the other symptoms (Yang et al., 1998)(Darnell and Posner, 2003). The two proteins share three similar KH-domains through which they bind directly to YCAY motifs (where Y stands for a pyrimidine) in the RNA sequence (Buckanovich and Darnell, 1997; Jensen et al., 2000a; Lewis et al., 2000). According to the binding location on mRNA, they can either induce exon skipping or exon retention (Ule et al., 2006).

Both NOVA1 and NOVA2 are mainly expressed in the central nervous system. However, they are reciprocally present in specific brain areas in mouse. For instance, NOVA2 is majorly expressed in cortex and hippocampus, whereas NOVA1 is mainly present in midbrain and spinal cord (Saito et al., 2016; Yang et al., 1998). Both the two Nova null mice displayed growth retardation, a progressive motor dysfunction and they died shortly after birth but only Nova2$^{-/-}$ mice present agenesis of corpus callosum (Jensen et al., 2000b; Saito et al., 2016). This peculiar defect suggested that Nova1 and Nova2 control different set of RNA transcripts. As a matter of fact, Nova2 seems to be more associated to the splicing regulation of genes involved in axon guidance and axonal projection in development mouse cortex (E18.5). These splicing events were developmentally regulated between E12.5 and E18.5 in mouse cortex, highlighting an important role of Nova2 as an axonal pathfinder modifier during cortical development (Saito et al., 2016).

2.1 IDENTIFICATION OF MUTATIONS IN *NOVA2*

By WES, we identified a *de novo* frameshift mutation in *NOVA2* (NM_002516.3: c.782del) in a patient presenting with intellectual disability, growth retardation, microcephaly, epilepsy, subcortical atrophy and traits of pyramidal syndrome as main features. Through data exchanging with other French and international teams, we identified four additional patients with a *de novo* truncating mutations in this gene. All variants are predicted to remove the third and last KH domain (Figure 32), which is important for RNA recognition and binding. The five mutations cluster in a small GC- and repeat-rich domain, which is poorly covered in most of the WES (like in ExAC) and some mutations in this gene might thus

have been missed in large-scale sequencing projects. Nevertheless, only one LoF variant has been reported in GnomAD but it is predicted to affect the splice donor site of a non-canonical transcript. All the detected mutations are in the last and larger exon of the gene and they all lead to frameshift encoding a common C-terminal tail of 134 amino acids. The location of the different variants in the last exon of the gene suggests that the mutant transcripts would escape to nonsense-mediated decay (NMD), a hypothesis that we were not able to verify directly in patients, due to the low expression of NOVA2 in blood.

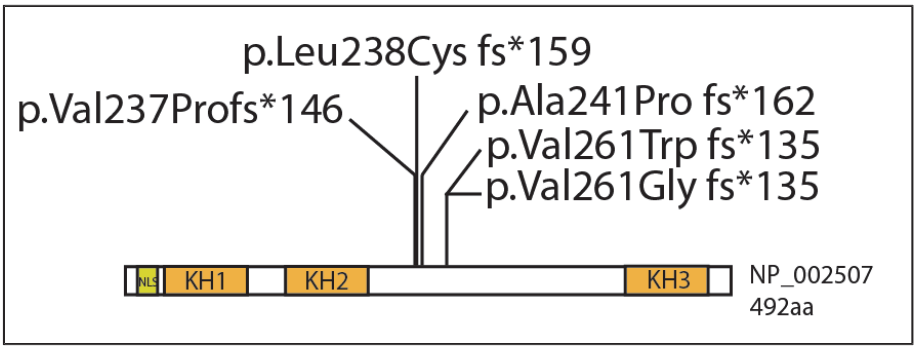


Figure 32: Schematic representation of the NOVA2 protein and the relative position of the mutations identified in ID patients

Patients share different clinical features in addition to ID (Table 18). All patients presented with a syndromic intellectual disability characterized by developmental, motor and speech delay. Most of the individuals show also abnormal behaviour (3/4), including also autistic traits (2), stereotypic movements with the hands (3/4) and frequent laughter (2/4). Most of them present hypotonia (3/4) and feeding difficulties (3/4). Spasticity and/or ataxic gait was reported in most of them (4/5). Brain malformations, reported in 3 out of 4 patients included cortical atrophy, Chiari Malformation and corpus callosum thinning. Two patients present with seizures. Several of these clinical features, such as the poor speech, the ataxic gait, the epilepsy, and the “jovial” behaviour as well as hand stereotypies may evoke Angelman syndrome (OMIM 105830). Interestingly, three patients were previously screened for it, including analysis of UBE3A methylation and sequencing.

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	n/tot.
cDNA mutation (NM_002516.3)	**c.782del**	**c.711_712insTG**	**c.701_720dup20**	**c.709_748del40**	**c.781del**	
Protein mutation (NP_002507.1)	**p.Val261Glyfs*135**	**p.Leu238Cysfs*159**	**p.Ala241Profs*162**	**p.Val237Profs*146**	**p.Val261Trpfs*135**	
Developmental delay	+	+	+	+	n.a.	4/4
ID	+	+	+	+	n.a.	4/4
Motor delay	+	+	+	+	n.a.	4/4
Speech delay	+	+	+	+	n.a.	4/4
Abnormal behavior	+	+	-	+	n.a.	3/4
Stereotypic movements	+	+	-	+	n.a.	3/4
Frequent laughter	+	-	-	+	n.a.	2/4
Feeding difficulties	+	+	-	+	n.a.	3/4
Hypotonia	-	+	+	+	n.a.	3/4
Epilepsy	+	-	-	-	+	2/5
Ataxie/Spasticity	+	+	+	+		4/4
Brain anomalies	+	-	+	+	n.a.	3/4
Previous genetic test	Angelman, ARX	Angelman, CDG	n.a.	Angelman, MECP2	n.a.	

Table 18: Main clinical phenotype of patients with a mutation in NOVA2

As NOVA2 is a gene that has never been implicated in ID, we performed some functional analysis to validate the pathogenicity of the variants identified, define their functional consequences and decipher the pathophysiological mechanisms.

2.2 FUNCTIONAL CONSEQUENCES OF MUTATIONS IN *NOVA2*

Since NOVA2 is specifically expressed in brain, we could not have access to tissue expressing it. To overcome this problem, we generated N-terminal FLAG-tagged wild-type and the four mutated NOVA2 cDNAs (NM_002516: c.782del; c.711_712insTG; c.701_720dup20; c.709_748del40). These constructs were then overexpressed in HeLa cells, where NOVA2 is physiologically not expressed. Cells were transfected with either the wild-type or one of the four mutant NOVA2 cDNA (Lipofectamine 2000, Invitrogen) and total proteins were extracted after 36 hours. Western blot analysis using an anti-FLAG antibody revealed that the four truncated NOVA2 proteins are expressed at a similar level to the wild-type, suggesting that the truncated proteins are stable (Figure 33).

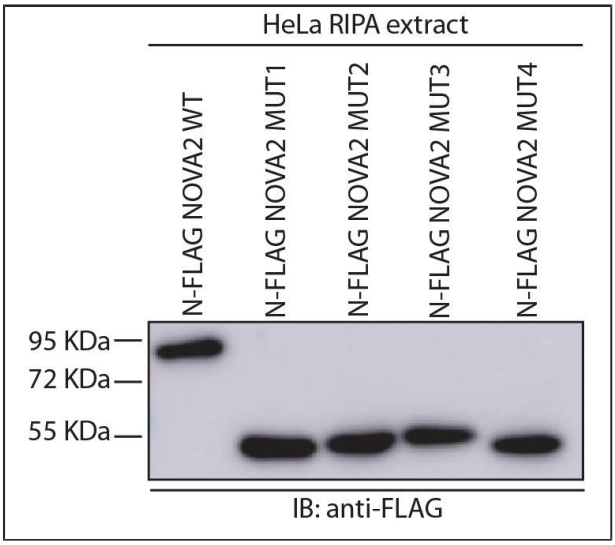


Figure 33: Western blot on N-FLAG NOVA2 wild-type and mutated overexpressed in HeLa cells

As the truncated NOVA2 proteins are expressed and since protein localization should not be altered as the frameshift variants are located downstream of the nuclear localization signals, we hypothesized that the protein function of NOVA2 is altered. NOVA2 is a neuronal splicing factor that regulates the splicing of axon guidance genes; altered splicing events in these genes were reported in *Nova2* knock-out mouse model (Saito et al., 2016). We retrieved the list of alternative splicing events depending from Nova2 in mice and we tested two of these alternative splicing events in our human cells overexpressing NOVA2. RNA was extracted after 36hours and a reverse transcription was performed (SuperScript IV, Invitrogen), then the obtained cDNA was amplified using specific primers designed to surround the splicing event to test: 1) if the spliced mRNA of interest is expressed in HeLa and 2) if NOVA2 is regulating the splicing. Among the analysed genes, we found that overexpression of human NOVA2 alters the splicing of exon 26 of *NEO1* and exon 14 of *APLP2*. *NEO1* and *APLP2* codes for two

different transmembrane receptors, well expressed in human brain tissues (Human Protein Atlas). Neogenin1 (NEO1) is a transmembrane receptor well expressed in differentiating neurons (Wilson and Key, 2007). NOVA2 has been shown to temporally regulate the splicing of a specific intracellular domain of NEO1. The presence or absence of these amino acids lead to the activation of different signalling pathways, which may play important roles during development. Amyloid Beta A4 Precursor-like protein 2 (APLP2) belongs to the conserved amyloid precursor protein gene family, which are important for the formation, maintenance and plasticity of synapses (Han et al., 2017). We observed that overexpression of NOVA2 leads to an increase of skipping of *NEO1* exon 26 compared to non-transfected cells (respectively, 61.15% vs 8.4%; p-value < 0.01) (Figure 34 on the left). On the other hand, it increases exon 14 retention in *APLP2* mRNA (10.10% in transfected vs 3.58% in non-transfected HeLa; p-value < 0.01) (Figure 34 on the right). We repeated the experiment and transfected HeLa cells with the four mutant NOVA2 and tested if these splicing events were affected. We observed that mutant NOVA2 proteins are less efficient than the wild-type to regulate these alternative splicing events, as we detected a significant difference in the average concentration ratio of the two isoform of *NEO1* (skipped/retained exon 26) in transfected HeLa with NOVA2 mutated compared to non-transfected (0.83 vs 0.09; p-value < 0.01) and also compared to HeLa overexpressing in NOVA2 wild-type (0.83 vs 1.66; p-value < 0.01) (Figure 34, on the left); similarly, the average ratio of the two isoform of APLP2 (skipped/retained intron 14) between the transfected HeLa with the mutations differs from the one of the non-transfected cells (respectively, 20.05 vs 27.15; p-value < 0.01) and from the one of the HeLa transfected with NOVA2 wild-type (20.05 vs 8.9; p-value < 0.01) (Figure 34 on the right).

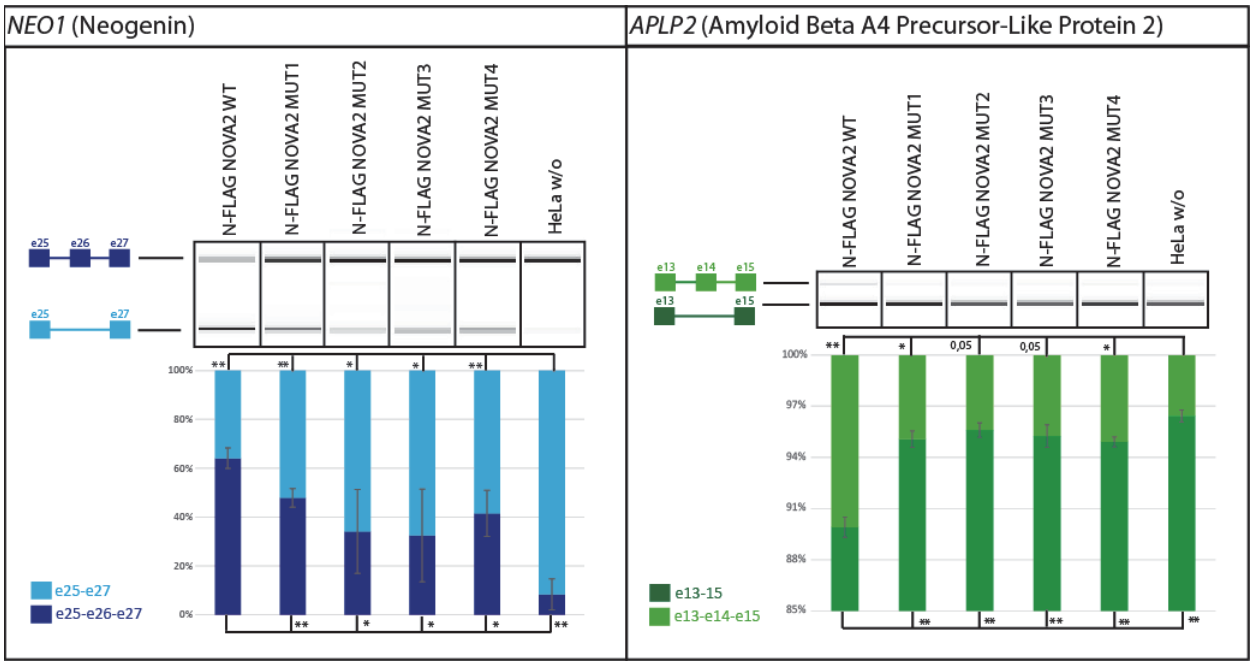


Figure 34: NEO1 and APLP2 splicing in HeLa cells

(**: p-value < 0.01; *: p-value < 0.05; 0.05: p-value = 0.05)

To confirm the role of human NOVA2 on the regulation of NEO1 and APLP2 splicing in a neuronal cell model, we investigated the effects of the inactivation of *NOVA2* in human neural stem cells (hNSCs), where NOVA2 is physiologically expressed. hNSCs were transfected (Interferin, PolyPlus) with siRNA targeting human *NOVA2* (ON-TARGETplus siRNA HUMAN NOVA2, GE Healthcare Dharmacon). siRNA efficiency was checked after two and four days of transfection by RT-qPCR and we confirmed a decrease of *NOVA2* mRNA level respectively of about 50% and 60% (Figure 35A). Expression of the paralog *NOVA1* was also assessed to exclude an eventual compensatory mechanism and to check the specificity of the inactivation (Figure 35B). Splicing events in *NEO1* and *APLP2* were then tested in extracted RNA from siRNA hNSCs. Only one *APLP2* isoform (the one excluding exon 14) was detected in hNSCs, with or without *NOVA2*. However, partial inactivation of *NOVA2* affects the splicing of NEO1. We observed an increase of the NEO1 isoform containing exon 26, meaning that inactivation of NOVA2 lead to a reduction of exon 26 skipping (Figure 35C). This result is consistent with the increase of exon 26 skipping observed when NOVA2 is overexpressed (in HeLa cells). A similar result is obtained in hNSCs after four days of siRNA transfection. RNA-sequencing experiments are on-going on RNA extracts from these cells to detect other altered alternative splicing events related to *NOVA2* inactivation.

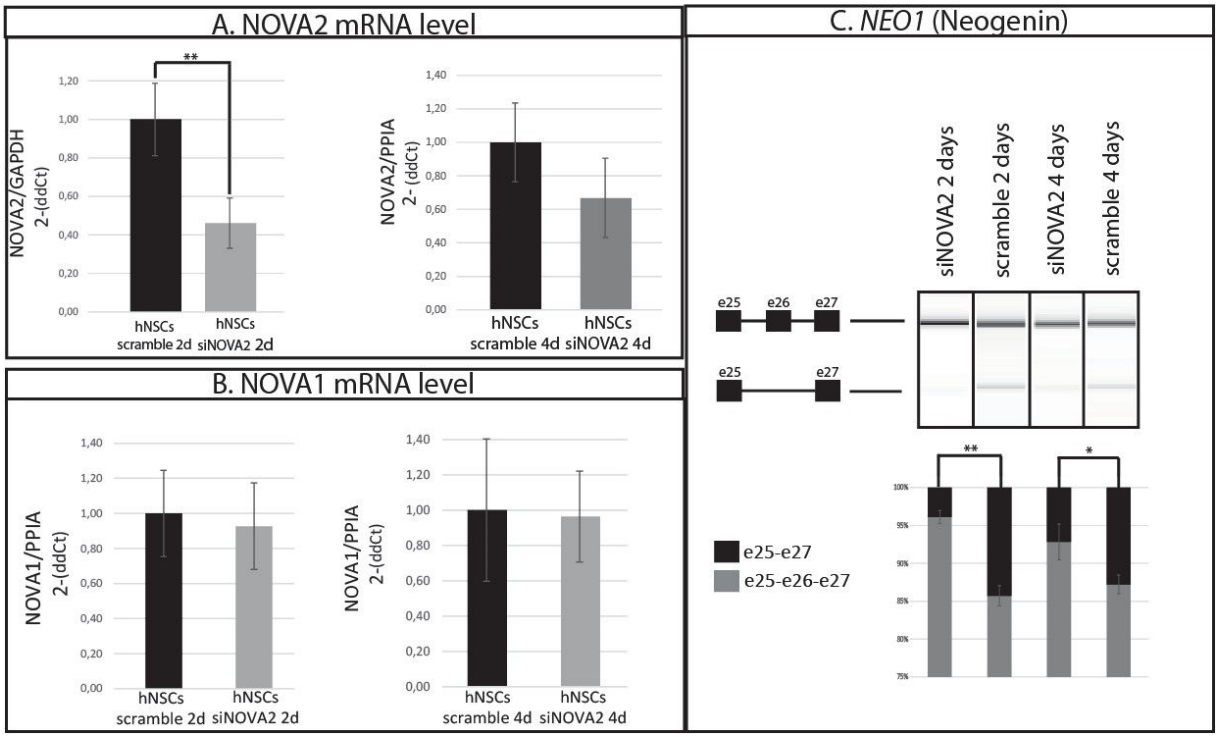


Figure 35: NOVA2 and NOVA1 expression and NEO1 alternative splicing in hNSCs

2.3 CONLUSIONS AND PERSPECTIVES

We identified five patients with a *de novo* frameshift mutation in *NOVA2*, a gene coding for a neuronal splicing factor. These mutations all cluster in a small, GC- and repeat-rich region and they are all predicted to lead to a truncated NOVA2 protein lacking the last KH domain. Interestingly, all 5 mutations gave the same frame, leading to the insertion of a common C-terminal amino acid sequence.

Patients show an overlapping phenotype and they are mainly affected by ID, developmental, motor and speech delay and some of them present also behavioural anomalies. Most of the reported individuals were previously tested for Angelman-syndrome, as their clinical manifestations are highly evocative of this syndrome. As all the individuals present an Angelman-like phenotype, we are currently investigating an additional cohort of patients with ID who were previously suspected to have Angelman-syndrome and for who *UBE3A* methylation or sequencing was performed as previous genetic screening, in order to identify additional mutations (Strasbourg University Hospital).

The frameshift mutations lead to stably expressed truncated NOVA2 proteins. As all the frameshift variants lead to a truncated protein lacking the last KH domain, we are currently investigating the ability of truncated NOVA2 proteins to recognize and bind the RNA (Collaboration with Nicolas Charlet-Berguerand, IGBMC). We showed that the overexpression of NOVA2 in HeLa cells – where NOVA2 is physiologically absent – specifically regulates the alternative splicing events of *NEO1* and *APLP2*, encoding two transmembrane proteins playing important role in brain development. The overexpression of the mutant NOVA2 in HeLa cells lead to a partial dysregulation of these alternative splicing. The reduction of NOVA2 mRNA level in hNSCs confirmed the role of NOVA2 on regulation of NEO1 splicing. Overall, these data suggest that frameshift mutations in *NOVA2* result in at least a partial loss-of-function, altering alternative splicing events occurring in NOVA2 target genes. These splicing events, naturally occurring during neurodevelopment, are important for neuronal migration and differentiation. We are currently testing the consequences of NOVA2 inactivation or overexpression on neurite outgrowth in neuronal N2A cells. After a treatment of few days with retinoic acid to induce their differentiation into neurons, the number and length of neurites formed by N2A will be measured to see if inactivation of NOVA2 leads to defects in neurite outgrowth and if this can be rescued by overexpressing wild-type or mutant NOVA2 proteins. If on one hand our data suggest a partial loss of function effect, it is quite unusual that all five frameshift mutations are in the same frame. The shared C-terminal amino acid sequence (Figure 36) might contain some novel putative protein-protein interaction domains, suggesting they could have a novel molecular function of the mutant proteins. The experiments performed in N2A, as well as the use of a zebrafish model overexpressing wild-type and mutant NOVA2 (Collaboration with Gaëlle Hayot and Christelle Golzio, IGBMC) will allow to test the gain and the loss of function behaviour of the mutations identified.

Overall, these preliminary results show for the first time that *NOVA2* is a novel gene implicated in an Angelman-like syndromic ID.

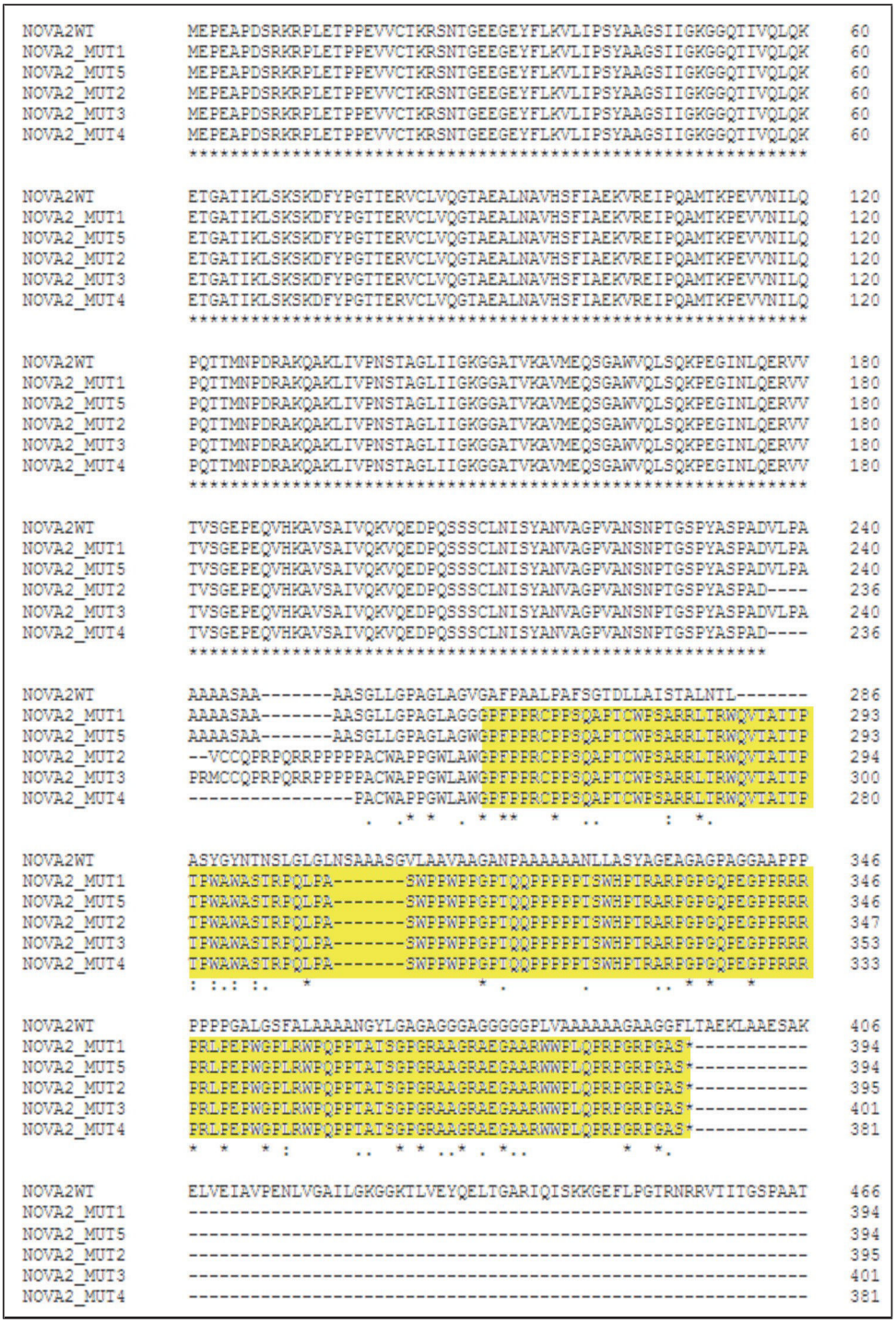


Figure 36: Mutated and wild-type NOVA2 protein alignment (Clustal Omega, EMBL-EBI)

In yellow are highlighted the shared C-terminal amino acid sequence of the mutated proteins

3. CLINICAL AND FUNCTIONAL CHARACTERIZATION OF RECURRENT MISSENSE MUTATIONS INVOLVED IN *THOC6*-RELATED INTELLECTUAL DISABILITY

THOC6 is part of the THO complex, which is part of the larger TREX complex, known to be involved in mRNA processing and transport. *THOC6* has been already implicated in a recessive syndromic form of ID in a large consanguineous Hutterite family (Beaulieu et al., 2013) and additional patients with homozygous and compound heterozygote mutations have been reported since then. By trio-WES I identified three homozygous variants in THOC6 (p.(Trp100Arg, Val234Leu, Gly275Asp)) in a boy affected by ID. Curiously, these variants are present at the heterozygous state in the mother, suggesting a maternal uniparental disomy of chromosome 16 (where *THOC6* lies). We have been in contact with a clinician following a girl who is compound heterozygote for this three missense changes and a previously reported missense variant (*THOC6*: p.Gly190Glu). Interestingly, these three missense variants are reported at the heterozygous state in different database (GnomAD, 1000Genomes, etc), at the same minor allele frequency (MAF) in each different subpopulation, suggesting they are in linkage disequilibrium. Due to the relative recurrence of this haplotype, it was crucial to delineate the consequences of these three missense variants and to understand if it is one specific amino acid change or the combination of the three that cause the disease. To do that, I overexpressed human THOC6 cDNA carrying each missense variant alone and combined, plus two other mutations reported in at least two patients (p.Arg87* and p.Gly190Glu) in HeLa or HEK293T cells to check if they affect its stability and its cellular localization. While the protein expression was not altered by the overexpression of the haplotype, the physiologically subcellular localization was altered, and it was driven by only one missense variant (p.Trp100Arg) of the haplotype. Alteration in the physiologically subcellular localization were noticed also in the additional two mutations. We hypothesized that the subcellular localization could be the consequence of a disrupted interaction between other members of the THO complex. Thus, I overexpressed THOC6 cDNA wild-type and mutated in HEK293T cells, immunoprecipitated them and detected the potential interactors by western-blot. Among the tested interactors, I observed a decrease in THOC1 and THOC5 in the three-variant-haplotype and the known mutation. However, inside the haplotype, it is a different missense change (p.Gly275Asp) that causes this phenotype. During these functional analysis, another patient carrying the same three missense variants was reported (Casey et al., 2016), underscoring the relative recurrence of this haplotype. Overall, we demonstrated that the three mutations tested, the truncating p.(Arg87*), the missense p.(Gly190Glu) variant and the haplotype of three variants, alter THOC6 physiological cellular localization and interaction with other members of the THO complex, THOC1 and THOC5. However, it

does not seem to exist a direct link between the two types of alterations, because within the haplotype it is two different variants that drive these alterations.

Clinical and functional characterization of recurrent missense variants implicated in *THOC6*-related intellectual disability

Mattioli Francesca[1,2,3,4], Isidor Bertrand[5], Abdul-Rahman Omar[6], Andrew Gunter[6], Raman Kumar[7], Beaulieu Chandree[8], Jozef Gecz[7], Boycott Kim[8], Innes Micheil[9], Mandel Jean-Louis[1,2,3,4,10,11], Piton Amelie[1,2,3,4,10]

[1] *Institut de Génétique et de Biologie Moléculaire et Cellulaire, 67 400, Illkirch, France*

[2] *Centre National de la Recherche Scientifique, UMR7104, 67 400, Illkirch, France*

[3] *Institut National de la Santé et de la Recherche Médicale, U964, 67 400, Illkirch, France*

[4] *Université de Strasbourg, 67 400 Illkirch, France*

[5] *Service de Génétique Médicale, CHU de Nantes, Nantes, France*

[6] *University of Mississippi Medical Center, Jackson, Mississippi, USA*

7Adelaide Medical School and Robinson Research Institute, University of Adelaide, Adelaide, SA5000, Australia

[8] *Children's Hospital of Eastern Ontario Research Institute, University of Ottawa, Canada*

[9] *Department of Medical Genetics, University of Calgary, Calgary, Canada*

[10] *Molecular Genetics Unit, Hôpitaux Universitaires de Strasbourg, Strasbourg, France*

[11] *University of Strasbourg Institute of Advanced Studies*

The authors declare no conflict of interest.

Address for correspondence:
Amélie Piton, PhD
Laboratoire "Mécanismes génétiques des maladies neurodéveloppementales",
IGBMC
1, rue Laurent Fries,
67400, Illkirch, France
Tel: +33369551652
E-mail: piton@igbmc.fr

ABSTRACT

THOC6 encodes a subunit of the THO complex that is part of a highly-conserved TREX complex, known to perform roles in mRNA processing and export. Few homozygous or compound heterozygote variants have been identified in the *THOC6* gene in patients with a syndromic form of intellectual disability (ID) (Beaulieu-Boycott-Innes syndrome, BBIS MIM# 613680). Here we report two additional individuals affected with BBIS originating from the north of Europe and who share an haplotype carrying three very rare missense changes in *THOC6*: p.(Trp100Arg, Val234Leu, Gly275Asp). The first affected individual is a boy who is homozygous for the three-variant haplotype, due to a maternal uniparental disomy event. The second is a girl, who is compound heterozygote for this haplotype and a previously reported p.(Gly190Glu) missense variant. We analyzed impact of these different amino acid changes identified on THOC6 protein stability, cellular localization, and interaction with the other THO complex subunits. We show that the different *THOC6* variants alter its physiological nuclear localization and interaction with at least two THO subunits, THOC1 and THOC5. Two amino acid changes of the three-variant-haplotype have alone specific effects and might contribute to the pathogenicity of the haplotype. Overall, we expanded the cohort of currently known BBIS affected individuals by reporting two individuals carrying the same recurrent European haplotype composed of three amino acid changes affecting THOC6 localization and interaction with THO protein partners.

INTRODUCTION

Intellectual disability (ID) is one of the most frequent neurodevelopmental disorders (NDDs) and affects about 2% of children or young adults. It is characterized by significant limitations in both intellectual functioning and adaptive behavior and it is associated with an intellectual quotient below 70 before the age of 18. It is estimated that a genetic anomaly is the cause of ID in about 60% of cases (Gilissen et al., 2014). The genetic origins of ID can be due to chromosomal abnormalities, copy number variations (CNVs) and point mutations or small insertion/deletions affecting a single gene. For the latter category, more than 700 genes have been identified so far (Vissers et al., 2016). Around 300 genes have been implicated in autosomal-recessive ID, the majority of them related to syndromic form of ID, that is to say that combine other symptoms in addition to ID. One of these genes, *THOC6,* codes for THOC6 protein that is component of a highly-conserved TREX mRNA export complex, and was implicated in the Beaulieu-Boycott-Innes syndrome (BBIS, MIM: #613680), an autosomal-recessive syndromic form of ID associated to various cardiac and renal malformations and peculiar facial dysmorphism (Beaulieu et al., 2013; Boycott et al., 2010). This syndrome was first identified in two related Hutterite families, followed by discovery of other nonsense and missense variants in patients with consistent BBIS clinical manifestations. In total, nine affected individuals from seven unrelated families have been reported with deleterious variants in *THOC6* (Amos et al., 2017; Anazi et al., 2016; Beaulieu et al., 2013; Casey et al., 2016).

THOC6 encodes a subunit of the THO complex that is involved in mRNA processing and mRNA export. The THO complex was first identified in yeast that is composed of three subunits. In humans, the THO complex is composed of homologous subunits of the yeast Hpr1 (THOC1), Tho2 (THOC2) and Tex1 (THOC3) and three additional subunits that do not have a yeast counterpart: THOC5, THOC6 and THOC7. Even though their names are akin, the THO

complex proteins are not orthologues and contain different protein domains (Masuda et al., 2005). The THO complex is part of the larger Transcription and export complex (TREX), which is recruited to the 5'end of the mRNA in a splicing- and cap-dependent way (Cheng et al., 2006; Masuda et al., 2005). In *Drosophila*, the THO complex is involved in the extracellular stimuli mediated signal transduction (Rehwinkel et al., 2004). A similar role is performed by human THOC5 (Tran et al., 2014a, 2014b), suggesting the contribution of the THO complex in cellular proliferation, differentiation and stress response. Indeed, THOC6 is shown to have a role in apoptosis regulation (Beaulieu et al., 2013). Identification of four *THOC2* missense variants in 4 unrelated large families with syndromic X-linked ID (Kumar et al., 2015) and a *de novo* *PTK2-THOC2* gene fusion in a patient with psychomotor retardation and congenital cerebellar hypoplasia (Di Gregorio et al J Med Genet 2013;50: 543–551) further suggests the important role of the THO complex in neuronal development.

Here we report two individuals with clinical features consistent with BBIS, carrying the same rare haplotype composed of these three amino acid changes p.(Trp100Arg, Val234Leu, Gly275Asp): a boy, homozygous for this haplotype due to a maternal uniparental disomy event and a girl, compound heterozygote for this haplotype together with a previously reported missense p.(Gly190Glu) variant. We investigated the consequences of these amino acid changes on THOC6 expression, localization and interactions with known protein partners, compared to the effects of the first reported nonsense variant p.(Arg87*) (Anazi et al., 2016), and show that the different variants lead to an altered subcellular localization of THOC6 proteins and disruption of its interaction with members of the THO complex. While our studies were in progress, a BBIS patient homozygous for the same haplotype p.(Trp100Arg,

Val234Leu, Gly275Asp) inherited from unrelated parents was reported (Casey et al., 2016), confirming that this haplotype is recurrently involved in BBIS in the European population.

RESULTS

Identification of two affected individuals with missense variants in THOC6

Patient 1 (**Figure 1A**) underwent multiple genetic testing and no candidate variant was detected by targeted-sequencing. By whole-exome sequencing (WES) we identified only one *de novo* loss-of-function nonsense variant in the *PSPN* gene (NM_004158.2: c.436C>T; p.Gln146*). This variant, coding for a 10 amino acid truncated protein, was considered to be non-pathogenic as *PSPN* is largely tolerant to loss-of-function (LoF) variants (probability of LoF intolerance, pLI = 0.01 with 3 truncating variants reported in the general population ExAC). Surprisingly, our exome sequencing data identified 58 rare (MAF< 0.0045) homozygous variants on chromosome 16; all heterozygous in the mother and absent from the father's DNA. Among these homozygous variants, three missense changes (NM_024339.3: c.298T>A, p.Trp100Arg; c.700G>C, p.Val234Leu; c.824G>A, p.Gly275Asp) were located in *THOC6*, a gene previously implicated in ID. The missense change p.Trp100Arg is predicted to be the most deleterious according to prediction programs (**Table 2**). We hypothesized that these homozygous variants originated from a maternal uniparental disomy of chromosome 16. A SNP-array analysis was performed on the proband and confirmed the existence of two different regions of homozygosity, with the following approximate coordinates: chr16:1-16,227,147 and chr16: 84,453,753-90,354,753. *THOC6* is located in the first region. Patient 1 was born at 35 weeks of pregnancy, characterized by the identification of an intrauterine growth restriction as well as a micropenis and a short corpus callossum. Birth measurements were the following: 2.060 kg, 46 cm and occipital-frontal head circumference (OFC) of 32 cm. He presented with neonatal hypotonia and feeding difficulties At 3.5 years, his measurements were: 13kg, 95.5cm and OFC 47cm. He has severe psychomotor retardation with severe speech delay (no word) and walk

acquired at 24 months. He has upper limbs stereotypies and autistic behavior. He presents facial dysmorphism (tall forehead, short palpebral fissures, long noise, retrognathia) consistent with the other BBSI patients previously reported. Genetic and molecular investigations performed in Patient 2 (**Figure 1B**) include high-resolution karyotype, SNP microarray, plasma amino acids, urine organic acids, and urine amino acids. All of these studies were normal with the exception of urine organic acids and plasma amino acids, which were abnormal in non-diagnostic patterns. Whole exome sequencing analysis revealed with parental samples used for segregation analysis. She was born at 34 weeks due to concerns for fetal movement; birth measurements were 1.92kg, 43.5cm length, and OFC of 33.3cm. Prenatal concerns included ventriculomegaly and dichorionic-diamniotic twin gestation; her brother is healthy and developmentally appropriate for his age. After birth, a cleft palate, micrognathia, choanal atresia, and a congenital heart defect were noted, in addition to hydrocephalus. She had hypotonia. Patient 2 had a ventriculoperitoneal shunt placed and patent ductus arteriosus ligation shortly after birth and currently has a stable atrial septal defect (ASD); her cleft palate was repaired at age 21 months. Patient 2 (now 6 years old) has severe global developmental delays: she has never developed words and is non-ambulatory (although she can roll over). Despite intensive therapies, Patient 2 cycles through periods of achievement followed by regression, usually around one month following skill acquisition (for example, patient 2 stopped standing with minimal support and no longer attempts to stand). She has oral aversion and takes all feeds through a G tube. Dysmorphic features associated with BBIS include: microcephaly, prominent forehead, short palpebral fissures with epicanthal folds, low-hanging columella, abnormally shaped dentition with malocclusion. She also presents cupped ears, small anteverted nares, down-turned corners of the mouth with a tented upper lip, maxillary hypoplasia, prominent fetal pads on hands and feet, and bilateral overlapping toes. Patient 2 has additional medical complications that include alternating exotropia, nystagmus, hyperopia,

bilateral sensorineural hearing loss, seizures, chronic lung disease, and pulmonary hypertension.

The haplotype p.(Trp100Arg, Val234Leu, Gly275Asp) is recurrent in the European population

The haplotype composed of the three missense variants p.(Trp100Arg, Val234Leu, Gly275Asp) identified in our two affected individuals was recently described at the homozygous state in another individual presenting a BBIS form of ID (Casey et al., 2016). These three missense variants are also reported at the heterozygous state in the GnomAD database at the same minor allele frequency (MAF) in the different subpopulations suggesting they are in total linkage disequilibrium (**Supplementary Table 1**). The highest frequency is obtained in the European population. In 1000 Genomes database, the variants are present in only one European individual of a British origin. Recent data from UK10K indicate a high allele frequency at 0.001 for these variants in the population from United Kingdom (n=3,577 individuals). Together the data suggest that this haplotype may have origin in this specific region. The three BBIS individuals with this haplotype have also a Northern European origin: Patient1 is from North of France, Patient2 with likely Northern European descent (surname sounding English) and the patient reported by Casey et al. (2016) is of Irish traveler origin.

We compared the clinical manifestations present in Patients 1 and 2 to the individual reported by Casey et al and the other patients reported with different *THOC6* variants (**Table 1, Figure 1C**). Overall, a similar phenotype was observed among patients with a variant in *THOC6* and no obvious difference between missense and nonsense variants was noted. We do not see any clinical features specific to carriers of the triple variants haplotype. Due to the relatively high recurrence of this haplotype (3 out of 11 BBIS cases), we found it critical to delineate the consequences of these three missense changes with an ultimate aim of understanding if one or more of the variants contributed to pathogenicity. To answer this question, we analyzed the

effect of each of the missense variants alone or combined together. We also included in our analysis, the missense c.596 G>A, p.(Gly190Glu) variant identified in Patient 2 and a patient reported previously (Amos et al., 2017), and a recurrent nonsense c.259 C>T, p.(Arg87*) variant (Amos et al., 2017; Anazi et al., 2016) (Figure 1C).

Gly275Asp alone or the truncating variant Arg87 reduce the THOC6 protein stability*

At first, we investigated if THOC6 variants affect stability of the proteins in HEK293T cells. HEK293T cells were transfected with plasmids expressing human FLAG-tagged wild type or mutant forms with: Trp100Arg, Val234Leu or Gly275Asp alone, the combination of the three missense variants ("triple mutant"), Gly190Glu or Arg87* THOC6. FLAG-THOC6 proteins were analyzed by immunoblotting using an anti-FLAG antibody. Quantification of THOC6 protein level (normalized to housekeeping protein TUB2A2) showed reduced levels of Arg87* or Gly275Asp but not the triple mutant THOC6 protein (**Figure 2, Table 2**).

p.(Trp100Arg), p.(Gly190Glu) and p.(Arg87) variants alter localization of normally-nuclear THOC6 protein*

All the THO complex subunit proteins, including THOC6 are localized in the nucleus, specifically co-localizing with the splicing factors in the nuclear speckle domains (Dias et al., 2010; Masuda et al., 2005) (Beaulieu et al., 2013). The first homozygous Gly46Arg missense variant implicated in BBIS was shown to affect the natural nuclear localization of THOC6 in HeLa cells (Beaulieu et al., 2013). We performed immunostaining on HeLa cells transfected with plasmids expressing FLAG-tagged human wild-type or variant *THOC6* cDNAs using anti-FLAG antibody and found that whereas wild-type showed a normal nuclear localization, the triple mutant, Gly190Asp and Arg87* THOC6 showed an abnormal cytosolic localization

(**Figure 3, Table 2**). Among the amino acid changes composing the haplotype, only Trp100Arg affects the nuclear localization of THOC6. This missense variant was predicted to be the most pathogenic of the haplotype by the different programs (**Table 2**) and changes a tryptophan located in a WD rich domain, which are usually implicated in multiprotein assembly and interaction.

Interaction with protein partners from the THO complex

As no nuclear localization signal (NLS) was identified in THOC6, we hypothesized that its import into the nucleus is facilitated by other protein/s. As THOC6 belongs to the THO complex, we wondered if the variants leading to abnormal THOC6 localization might affect their interactions with other proteins of the THO complex. We performed immunoprecipitation experiments using an anti-FLAG antibody on HEK293T cells transfected with human wild-type or variant THOC6 expression plasmids. A Coomassie Blue staining revealed the presence, among the immunoprecipitated proteins, of FLAG-THOC6 along with other proteins at an estimated size of 75 and 28 KDa (data not shown). The immunopreciptated proteins were immunoblotted for THOC1, THOC2, THOC3, THOC4/ALY, THOC5, CIP29 and CBP80 TREX subunits. Among these, only THOC1 and THOC5 were detected in the immunoprecipitated proteins (**Figure 4**). Interactions with THOC1 and THOC5 was abolished or decreased in the case of Gly275Asp, the triple mutant, Gly190Glu and to a lesser extent Arg87* (**Figure 4**).

DISCUSSION

Here we describe two affected individuals carrying the same three missense variants p.(Trp100Arg; p.Val234Leu; Gly275Asp) in the *THOC6* gene presenting clinical manifestations consistent with BBIS. These three missense variants were also recently reported in an another BBSI affected individual (Casey et al., 2016). Patient 1 haplotype composed of

the three missense variants was inherited from his mother. This haplotype is heterozygous in the mother and homozygous in the affected child, resulting from a maternal uniparental disomy in chromosome 16. Patient 2 inherited the haplotype from his mother and a single missense variant p.(Gly190Glu) reported previously (Amos et al., 2017) from his father. Overall, three distinct Northern European BBIS affected individuals carry this haplotype most likely with its origin in the United Kingdom region. To better understand the pathogenicity of this haplotype recurrently associated with BBIS in Europe, we analyzed the three missense variants individually or in combination to determine if they affected the THOC6 protein levels, localization or interactions with the known THO protein partners. We asked if pathogenicity in the affected individuals was caused by one or more of these THOC6 variant proteins. We also included in our analyses two other recurrent variants: the second missense variant identified in trans in Patient 2 p.(Gly190Glu) as well as a nonsense variant p.(Arg87*) identified in several individuals. We confirmed the nuclear localization of the THOC6 protein and identified an interaction between the overexpressed THOC6 protein and two other members of the THO complex, THOC1 and THOC5, without determining whether it is direct or indirect interactions, in HEK293 cells. A direct interaction between THOC5 and THOC6 was already reported (katahira et al 2013) but a previous study did not find any interaction between THOC6 and THOC5/1 (El Bounkari et al., 2009). Thoc1 and Thoc5 knockout mice were embryonically lethal, indicating their important role during development (Li et al., 2005; Mancini et al., 2010). We showed that the three-variant haplotype, p.Gly190Glu and p.Arg87* lead to a mislocalization of THOC6 in the cytosplasm, and to a loss or a decrease of its interactions with THOC1 and THOC5. Therefore, the variants might make THOC6 unable to carry out its normal function and impact mRNA export, leading to clinical outcomes. Our results combined to the previous ones indicate that compound heterozygosity for the two most common pathogenic missense variants (the haplotype and Gly190Glu found in Patient2) appears, on the basis of single patient descriptions, as severe as homozygosity for the haplotype (Patient 1, and Patient

3 described by Casey et al.), homozygosity for a null mutation (Patient 8 described by Anazi et al. and Patient 9 described by Amos et al.) and by combination of a null and a missense pathogenic variant (Patients 10 described by Amos et al.). Thus about all the known alleles appear to play equivalent roles as a null allele. Combining the frequency in Europeans of the known mutations to the cumulative frequency of other LoF variants in this population (9 variants in 10 individuals, see **Sup Table 2**), one can estimate the minimum frequency of pathogenic alleles in Europeans is 7.03e-4, corresponding to an incidence of affected individuals of 0.5 per million.

We have speculated at the beginning that it could have existed a link between the alterations in THOC6 binding to THOC5 and/or THOC1 and the THOC6 mislocalization in the cytoplasm. It is the case for another member of the THO complex deprived of NLS, THOC7. El Bounkari et al. previously showed that THOC5 directly interacts with THOC7 and that this interaction is responsible for bringing THOC7 to the nucleus. However, in our case, the amino acid change leading to an abnormal cytoplasmic THOC6 localization, Trp100Arg, does not affect interactions with THOC5 and THOC1. Therefore, disruptions of these interactions might not be the primary cause of the cellular mis-localization of THOC6. We can hypothesize that the cytosolic abnormal localization would be the result of a loss of interaction with another protein, which we were not able to identify in our study.

Previously published immunoprecipitation experiments with THOC2 reported a direct interaction between THOC2 and THOC6 (Cheng et al., 2006; Masuda et al., 2005). However, we did not detect THOC2 in our immunoprecipitation experiments, may be because the protein is large (~182 KDa). THOC2 is involved in the export of polyA$^+$ RNA (Chi et al., 2013). Missense variants in *THOC2* were reported in patients with an X-linked ID characterized by elevated BMI, speech delay, short stature and seizures a phenotype that do not well overlap with BBIS (Kumar et al., 2015). By comparing the few available common clinical features

among the two cohorts of patients, we did not notice a strong overlap between clinical manifestations of patients with THOC6 and THOC2 variants. Facial dysmorphy observed in BBIS seems to be specific only to THOC6 variants. It has been proposed that specific THO/TREX subunits are differentially recruited to specific subsets of mRNA (Heath et al., 2016), which could explain that consequences are different when they are altered by genetic variants. The microcephaly is significantly more pronounced in these patients than in patients with THOC2 mutations, which is consistent with a role of THOC6 in apoptotic processes, which was previously shown (Beaulieu et al., 2013). Interestingly, THOC1 has a death domain through which it regulates cell-cycle and induce p53-independent apoptosis (Gasparri et al., 2004) and THOC5 is known to play a role in cell differentiation and proliferation (Mancini et al., 2010).

In conclusion, we have expanded the cohort of BBIS affected individuals by reporting two additional European patients, both carrying the same haplotype, originating from the north of Europe and composed of three missense variants. This highlights its relative high frequency, as up to now it is one of the most frequent BBIS variant present in more than about a quarter of the patients (3/11). We did not observe any obvious clinical feature specific to this haplotype. We demonstrated that this haplotype, as well as two other recurrent variants identified in non-consanguineous European population, the truncating p.(Arg87*) and the missense p.(Gly190Glu) variants, alter THOC6 physiological nuclear localization and its interaction with other members of the THO complex, THOC1 and THOC5. However, it seems that there is no direct link between these two alterations, because they are driven by two different missense changes within the haplotype. The pathogenicity of the haplotype results therefore of a combined effect of at least two of the three missense changes.

MATERIALS AND METHODS

Patient recruitment and genomic analysis

Patient1 underwent multiple genetic testing that included karyotyping, array comparative genomic hybridization, fragile X-test and targeted-sequencing of more than 400 genes implicated in ID. As no clear pathogenic variant was detected, a trio-whole exome sequencing was then performed. Libraries and captures from genomic blood DNA were prepared with the SureSelect XT Human All Exon V5 Kit (Agilent Technologies), and 100 bp paired-end sequencing was performed on the HiSeq2500 sequencer (Illumina). Reads were aligned and variants called and annotated as described previously (Geoffroy et al., 2015; Redin et al., 2014). Potential pathogenic variants were identified using VaRank ranking program (Geoffroy et al., 2015). Variants were filtered according to different inheritance scenarios by using public databases and a large cohort of ID-affected individuals as previously described (Redin et al., 2014). A SNP array (Infinium HumanCytoSNP-12 v2.1 BeadChip, Illumina) containing 300 000 SNPs was performed to study the unidisomy event. The potential functional effects of the amino-acid changes on the protein was assessed using several bioinformatics programs including SIFT (Ng and Henikoff, 2003), PolyPhen2(Adzhubei et al., 2010) and Mutation Taster. (Schwarz et al., 2010)

Site-directed mutagenesis

The variant FLAG-THOC6 expression plasmids were generated by site-directed mutagenesis of thepcDNA3.1-FLAG-THOC6 construct reported previously (Beaulieu et al., 2013) using the following specific primers: c.298T>A_For 5'-ATGGGGAGGTGAAGGCCaGGCTTTGGG-3' and c.298T>A_Rev 5'-GAGCATCTCCGCCCAAAGCCtGGCCTTCACC-3', c.700G>C_For 5'-AACTGATTCCGACTGGATGcTCTGTGGAGG-3' and c.700G>C_Rev 5'-TGGGCCCCCTCCACAGAgCATCCAGTC-3', c.824G>A_For 5'-

GACCTGATTCTGTCAGCTGaCCAGGGCCG-3' and c.824G>A_Rev 5'-TTGACGCAGCGGCCCTGGtCAGCTGACAG-3', c.256C>T_For 5'-ATAGCATGGTTTCCACCGATtGACATCTGC-3'and c.256C>T_Rev 5'-CAGCACTAAGCAGATGTCaATCGGTGGAAAC-3', c.569G>A_For 5'-CTGTCAGGTGGCGAGGATGaAGCTGTTCGAC-3' and c.569G>A_Rev 5'-GTCCCAAAGTCGAACAGCTtCATCCTCGCC-3'. The variant FLAG-THOC6 plasmids were confirmed by sequencing (GATC, Germany).

Western Blot analysis

HEK293T cells were transiently transfected with the different plasmids expressing FLAG-tagged wild-type or variant forms of *THOC6* using Lipofectamine2000 (Invitrogen) and harvested 36 hours after transfection. For immunoblotting analysis, proteins were lysed in RIPA buffer combined with protease inhibitors (Roche). THOC6 expression was analyzed by SDS-PAGE, and immunoblotting was performed with anti-FLAG (1:1000, F1804, SIGMA) and anti-TUB2A2 (in house, 1:4000) antibodies. Protein semi-quantification was done by measuring the band intensity using ImageJ software and then calculating a ratio between FLAG and TUB2A2 intensities. The experiments were performed in quadruplicates.

Immunoprecipitation experiments

For immunoprecipitation studies, proteins were extracted using a NP40 buffer combined with anti-proteases (Roche) (25mM Tris-HCl pH8, 150mM NaCl, 10% glycerol, 1% NP-40, 2mM EDTA). FLAG-THOC6 was immunoprecipitated with Dynabeads Protein A (Invitrogen) using 2µg of mouse anti-FLAG antibody (F1804, SIGMA) or 2µg of a negative control (AR-441, SantaCruz). After washing steps, proteins were eluted with SDS sample buffer. Coomassie

staining revealed the presence of the immunoprecipitated proteins. The presence of proteins of the THO/TREX complex were investigated by immunoblotting with THOC1 (1:2000; Bethyl A302-839A), THOC2 (1:300, ProteinTech, 55178-1-AP) THOC3 (1:2500; Sigma-Aldrich HPA044009), THOC4/ALY (1:4000; Bethyl A302-892A), THOC5 (1:5000; Bethyl A302-120A), CIP29 (1:2000; Thermo Scientific PA5-21783) and CBP80 (1:5000; Bethyl A301-794A). The experiments were performed in duplicates.

Analysis of cellular localization by immunofluorescence

Wild-type and mutant FLAG-THOC6 localization was performed by immunofluorescence detection. HeLa cells were transfected with the plasmids described above using Lipofectamine 2000 (Invitrogen) and after 36 hours cells were fixed with 4% paraformaldehyde and permeabilized with 1×-TBS containing 0,2% TritonX-100, 10% fetal calf serum and 1% bovine serum albumin. Cells were incubated overnight with anti-FLAG antibody (1:500, F1804, SIGMA) and then with the secondary antibody Alexa-594 conjugated goat anti-mouse (1:1000, Invitrogen). Cells were stained with Hoechst and mounted onto microscope slides. Images were obtained with an upright motorized fluorescent microscope (Leica Microsystems).

ACKNOWLEDGEMENTS

The authors thank the families for their participation in this study. The authors also thank all members of the Strasbourg Hospital Molecular Diagnostic and Cytogenetic Laboratories (Christel Depienne), the Institut Genetique Biologie Moleculaire Cellulaire sequencing platform (Bernard Jost, Stephanie Le Gras, Mathieu Jung) for performing sequencing and UMR_S 1112 (Jean Muller, and Veronique) for developing and maintaining the ranking program Varank.

FUNDINGS

This work was funded by Fondation Jerome Lejeune, Fondation Maladies Rares and Association APLM. This study was also supported by grant ANR-10-LABX-0030-INRT, a French state fund managed by the Agence Nationale de la Recherche under the frame program Investissements d'Avenir ANR-10-IDEX-0002-02.

REFERENCES

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Amos, J.S., Huang, L., Thevenon, J., Kariminedjad, A., Beaulieu, C.L., Masurel-Paulet, A., Najmabadi, H., Fattahi, Z., Beheshtian, M., Tonekaboni, S.H., et al. (2017). Autosomal recessive mutations in THOC6 cause intellectual disability: syndrome delineation requiring forward and reverse phenotyping. Clin. Genet. *91*, 92–99.

Anazi, S., Alshammari, M., Moneis, D., Abouelhoda, M., Ibrahim, N., and Alkuraya, F.S. (2016). Confirming the candidacy of THOC6 in the etiology of intellectual disability. Am. J. Med. Genet. A. *170A*, 1367–1369.

Beaulieu, C.L., Huang, L., Innes, A.M., Akimenko, M.-A., Puffenberger, E.G., Schwartz, C., Jerry, P., Ober, C., Hegele, R.A., McLeod, D.R., et al. (2013). Intellectual disability associated with a homozygous missense mutation in THOC6. Orphanet J. Rare Dis. *8*, 62.

Boycott, K.M., Beaulieu, C., Puffenberger, E.G., McLeod, D.R., Parboosingh, J.S., and Innes, A.M. (2010). A novel autosomal recessive malformation syndrome associated with developmental delay and distinctive facies maps to 16ptel in the Hutterite population. Am. J. Med. Genet. A. *152A*, 1349–1356.

Casey, J., Jenkinson, A., Magee, A., Ennis, S., Monavari, A., Green, A., Lynch, S.A., Crushell, E., and Hughes, J. (2016). Beaulieu-Boycott-Innes syndrome: an intellectual disability syndrome with characteristic facies. Clin. Dysmorphol. *25*, 146–151.

Cheng, H., Dufu, K., Lee, C.-S., Hsu, J.L., Dias, A., and Reed, R. (2006). Human mRNA export machinery recruited to the 5' end of mRNA. Cell *127*, 1389–1400.

Chi, B., Wang, Q., Wu, G., Tan, M., Wang, L., Shi, M., Chang, X., and Cheng, H. (2013). Aly and THO are required for assembly of the human TREX complex and association of TREX components with the spliced mRNA. Nucleic Acids Res. *41*, 1294–1306.

Dias, A.P., Dufu, K., Lei, H., and Reed, R. (2010). A role for TREX components in the release of spliced mRNA from nuclear speckle domains. Nat. Commun. *1*, 97.

El Bounkari, O., Guria, A., Klebba-Faerber, S., Claussen, M., Pieler, T., Griffiths, J.R., Whetton, A.D., Koch, A., and Tamura, T. (2009). Nuclear localization of the pre-mRNA associating protein THOC7 depends upon its direct interaction with Fms tyrosine kinase interacting protein (FMIP). FEBS Lett. *583*, 13–18.

Gasparri, F., Sola, F., Locatelli, G., and Muzio, M. (2004). The death domain protein p84N5, but not the short isoform p84N5s, is cell cycle-regulated and shuttles between the nucleus and the cytoplasm. FEBS Lett. *574*, 13–19.

Geoffroy, V., Pizot, C., Redin, C., Piton, A., Vasli, N., Stoetzel, C., Blavier, A., Laporte, J., and Muller, J. (2015). VaRank: a simple and powerful tool for ranking genetic variants. PeerJ *3*, e796.

Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W.M., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature *511*, 344–347.

Heath, C.G., Viphakone, N., and Wilson, S.A. (2016). The role of TREX in gene expression and disease. Biochem. J. *473*, 2911–2935.

Kumar, R., Corbett, M.A., van Bon, B.W.M., Woenig, J.A., Weir, L., Douglas, E., Friend, K.L., Gardner, A., Shaw, M., Jolly, L.A., et al. (2015). THOC2 Mutations Implicate mRNA-Export Pathway in X-Linked Intellectual Disability. Am. J. Hum. Genet. *97*, 302–310.

Li, Y., Wang, X., Zhang, X., and Goodrich, D.W. (2005). Human hHpr1/p84/Thoc1 regulates transcriptional elongation and physically links RNA polymerase II and RNA processing factors. Mol. Cell. Biol. *25*, 4023–4033.

Mancini, A., Niemann-Seyde, S.C., Pankow, R., El Bounkari, O., Klebba-Färber, S., Koch, A., Jaworska, E., Spooncer, E., Gruber, A.D., Whetton, A.D., et al. (2010). THOC5/FMIP, an mRNA export TREX complex protein, is essential for hematopoietic primitive cell survival in vivo. BMC Biol. *8*, 1.

Masuda, S., Das, R., Cheng, H., Hurt, E., Dorman, N., and Reed, R. (2005). Recruitment of the human TREX complex to mRNA during splicing. Genes Dev. *19*, 1512–1517.

Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.

Redin, C., Gérard, B., Lauer, J., Herenger, Y., Muller, J., Quartier, A., Masurel-Paulet, A., Willems, M., Lesca, G., El-Chehadeh, S., et al. (2014). Efficient strategy for the molecular diagnosis of intellectual disability using targeted high-throughput sequencing. J. Med. Genet. *51*, 724–736.

Rehwinkel, J., Herold, A., Gari, K., Köcher, T., Rode, M., Ciccarelli, F.L., Wilm, M., and Izaurralde, E. (2004). Genome-wide analysis of mRNAs regulated by the THO complex in Drosophila melanogaster. Nat. Struct. Mol. Biol. *11*, 558–566.

Schwarz, J.M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods *7*, 575–576.

Tran, D.D.H., Koch, A., and Tamura, T. (2014a). THOC5, a member of the mRNA export complex: a novel link between mRNA export machinery and signal transduction pathways in cell proliferation and differentiation. Cell Commun. Signal. CCS *12*, 3.

Tran, D.D.H., Saran, S., Williamson, A.J.K., Pierce, A., Dittrich-Breiholz, O., Wiehlmann, L., Koch, A., Whetton, A.D., and Tamura, T. (2014b). THOC5 controls 3'end-processing of immediate early genes via interaction with polyadenylation specific factor 100 (CPSF100). Nucleic Acids Res. *42*, 12249–12260.

Vissers, L.E.L.M., Gilissen, C., and Veltman, J.A. (2016). Genetic studies in intellectual disability and related disorders. Nat. Rev. Genet. *17*, 9–18.

FIGURE AND TABLE LEGEND

Figure 1. Two previously unreported individuals with syndromic form of ID carrying *THOC6* missense variants.

A, B. Pedigrees of the patients (Patient 1, **A**; Patient 2, **B**) described in this study**. C.** THOC6 protein showing the variants reported previously and the missense variants identified in the two patients reported here (highlighted in red).

Figure 2. **THOC6 variant protein stability is reduced in HEK293T cells.**

HEK293 cells transfected with plasmids expressing FLAG-tagged wild-type or variant THOC6 proteins were harvested 36 hr after transfection and the THOC6 levels were analyzed by SDS-PAGE and immunoblotting with anti-FLAG antibody.

Figure 3. Cellular localization of FLAG-tagged wild-type or THOC6 variant proteins in HEK293T cells.

HEK293T cells were transiently transfected with plasmids expressing FLAG-tagged wild-type or variant THOC6 proteins. Red fluorescence (Alexa-594) (left panels) indicates the localization of the FLAG-tagged THOC6 proteins. Middle panels, Hoescht indicates the position of the nuclei. Right panels, merged Hoescht and FLAG-THOC6 showing nuclear (wild-type, Val234Leu and Gly275Asp) and cytoplasmic (triple variant, Trp100Arg, Gly190Glu and Arg87*) THOC6 localization.

Figure 4. Interactions between wild-type or THOC6 variant proteins with the other known THO complex subunits.

HEK293T cells were transiently transfected with plasmids expressing FLAG-tagged wild-type or THOC6 variant proteins. FLAG-THOC6 was immunoprecipitated from the cell lysates with anti-FLAG antibody and analysed by Western blot analysis using anti-FLAG, anti-THOC1 and anti-THOC5 antibodies . THOC1 and THOC5 proteins are present in immunoprecipitates of the wild-type, Trp100Arg and Val234Leu THOC6 proteins but at low level or absent for the triple haplotype, Gly190Glu and Arg87*.

Table 1. Clinical manifestations observed in patients carrying the p.(Trp100Arg, Val234Leu, Gly275Asp) haplotype compared to the other BBIS patients.

Mutations are indicated according to NM_024339.3. [a] 3 var: c.[298T>A, 700G>C, 824G>A]; [b] 3 vear : p.[(Trp100Arg,p.Val234Leu,p.Gly275Asp)]

Table 2. *In silico* predictions and functional consequences of the different *THOC6* mutations

Variants are indicated according to NM_024339.3. HumVar prediction scores are indicated for Polyphen2 [a] truncated protein; [b]: results described in Beaulieu et al., 2013; n.a.: not applicable; n.t.: not tested

References	*this report*	*this report*	*Casey et al. 2016*	Other BBIS patients
Individual	Patient 1	Patient 2		*n.a.*
Sex	M	F	F	*n.a.*
Ethinicity		North European	Irish Traveller	*n.a.*
Mutation (NM_024339.3)	c.[3 var[a]];[3 var[a]]	c.[3 var[a]];[596G>A]	c.[3 var[a]];[3 var[a]]	*n.a.*
Consequences	p.[3 var[b]];[3 var[b]]	p.[3 var[b]];[Gly190Glu]	p.[3 var[b]];[3 var[b]]	*n.a.*
Age	3.5 years	5.3 years		*n.a.*
Height (centile)	95.5cm (10th)	96.5cm (<2nd)	0,4th	*n.a.*
Weight	13kg (10th)	15.4kg (7th)	na	*n.a.*
OFC (centile)	47.5cm (<3rd)	42cm (<2nd)	2nd	*n.a.*
mild to moderate microcephaly	+	+	+	8/8
facial dysmorphy	+	+	+	8/8
Tall forehead	+	+	+	
Deep set eyes	+	+	+	
Long nose	+	+	+	
Epicanthus		+	+	
Low hanging columnella		+		
Flat philtrum	+	-	+	
Retrognathia	+	+	+	
Dental problems (malocclusion/caries)		+	+	6/8
ID	severe	severe	severe	8/8 (severe 2/8)
Speech delay	+	+	+	7/8
Brain anomalies	+		+	2/4
ventricular dilatation	-	+ (hydrocephalus)	+	
corpus callosum dysgenesis	+	+	-	
Cardiac anomalies	-	+	+	4/8
Renal anomalies	-	-	+	2/8
Genitourinary problems	micropenis, hypospadias	-	-	5/8

482

Table 1

Mutation		**triple mutant**	c.298 T>A, p.Trp100Arg	c.700 G>C, p.Val234Leu	c.824 G>A, p.Gly275Asp	**c.569G>A, p.Gly190Glu**	**c.259 C>T, p.Arg87***	**c.136G>A, p.Gly46Arg**
References		Casey et al., this report				this report, Amos	Amos, Anazi	Beaulieu
***In silico* predictions**	SIFT	n.a	Deleterious (0,01)	Tolerated (0,79)	Tolerated (0,09)	**Deleterious (0)**	**n.a**	**Deleterious (0)**
	Polyphen2	n.a	**Proba. damaging (0,990)**	Benign (0,029)	Benign (0,164)	**Proba. damaging (1)**	**n.a**	**Proba. damaging (0,999)**
	Mutation T@ster	n.a	**Disease causing (1)**	**Disease causing (0,995)**	**Disease causing (1)**	**Disease causing (1)**	**n.a**	**Disease causing (1)**
	Grantham Score	n.a	**101**	32	94	**98**	**n.a**	**125**
Functional studies	Protein expression	**Normal**	Normal	Normal	**Decreased**	**Normal**	**Decreased**[a]	n.t.
	Nuclear localization	**Abnormal**	**Abnormal**	Normal	Normal	**Abnormal**	**Abnormal**	**Abnormal**[b]
	THOC1/5 protein interaction	**Decreased**	Yes	Yes	**Decreased**	**Decreased**	**Decreased**	n.t.

Table 2

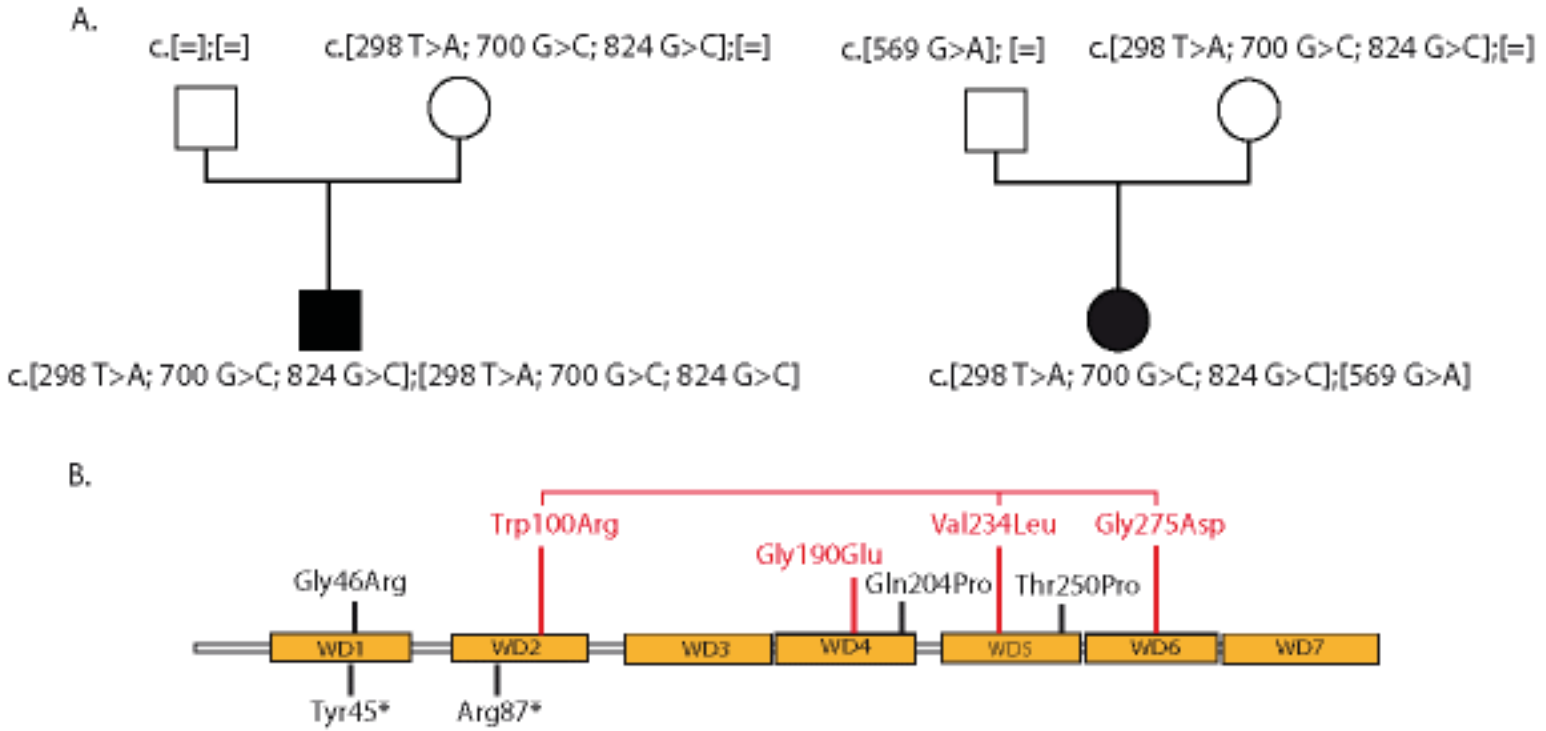

A.
c.[=];[=]
c.[298 T>A; 700 G>C; 824 G>C];[=]
c.[298 T>A; 700 G>C; 824 G>C];[298 T>A; 700 G>C; 824 G>C]
c.[569 G>A]; [=]
c.[298 T>A; 700 G>C; 824 G>C];[=]
c.[298 T>A; 700 G>C; 824 G>C];[569 G>A]
B.
Trp100Arg
Val234Leu
Gly275Asp
Gly190Glu
Gln204Pro
Thr250Pro
Gly46Arg
WD1
WD2
WD3
WD4
WD5
WD6
WD7
Tyr45*
Arg87*

Figure 1

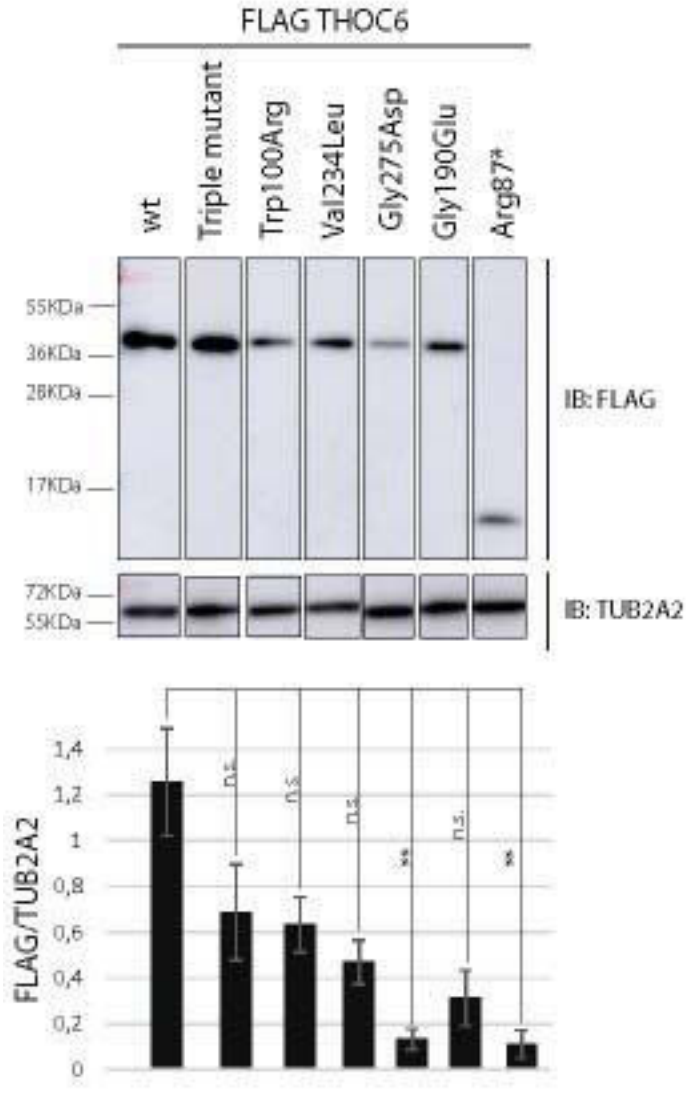

FLAG THOC6
wt
Triple mutant
Trp100Arg
Val234Leu
Gly275Asp
Gly190Glu
Arg87*
55KDa
36KDa
28KDa
17KDa
IB: FLAG
72KDa
55KDa
IB: TUB2A2
FLAG/TUB2A2
1,4
1,2
1
0,8
0,6
0,4
0,2
0
n.s.
n.s.
n.s.
**
n.s.
*

Figure 2

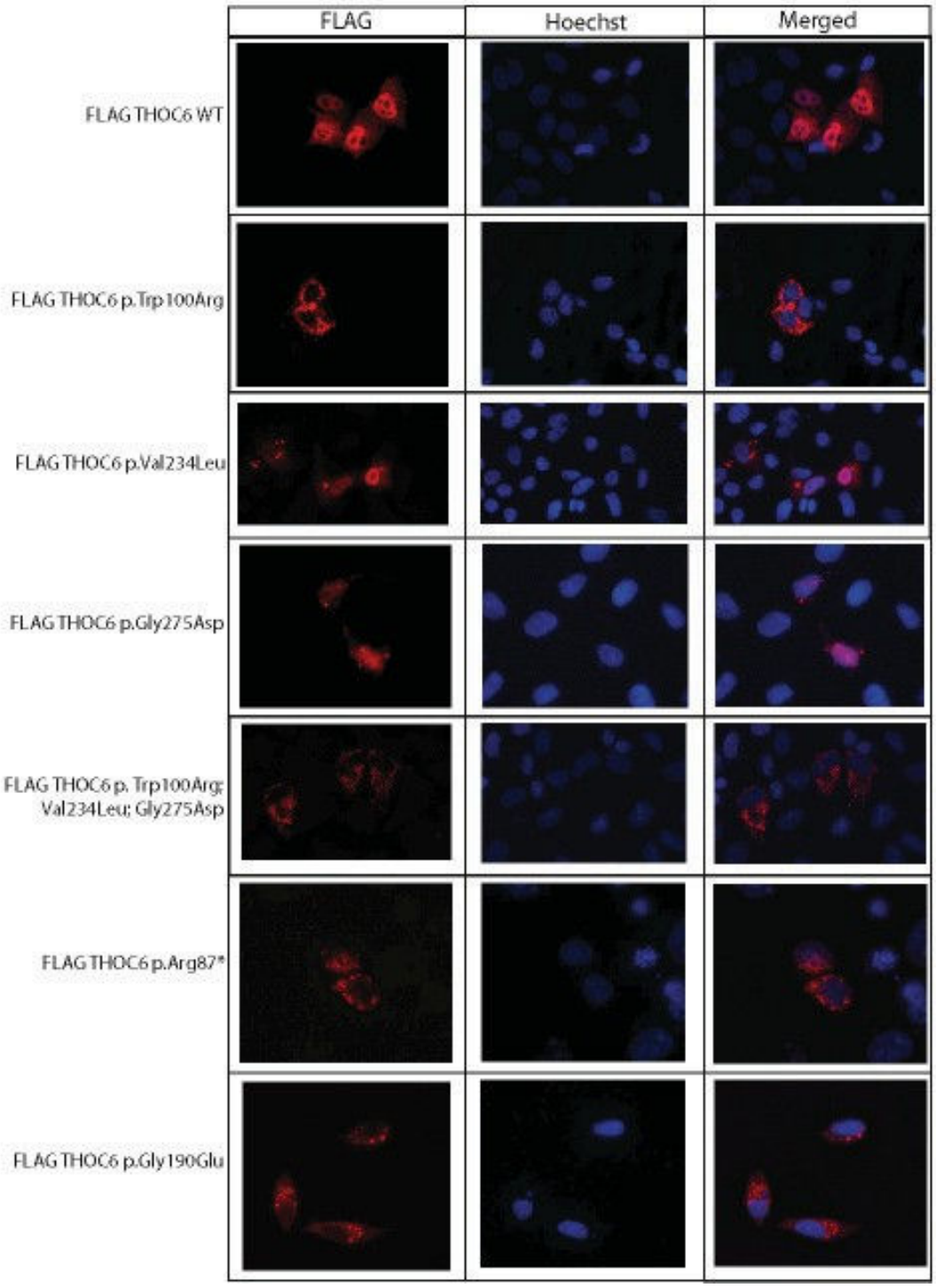

FLAG
Hoechst
Merged
FLAG THOC6 WT
FLAG THOC6 p.Trp100Arg
FLAG THOC6 p.Val234Leu
FLAG THOC6 p.Gly275Asp
FLAG THOC6 p. Trp100Arg; Val234Leu; Gly275Asp
FLAG THOC6 p.Arg87*
FLAG THOC6 p.Gly190Glu

Figure 3

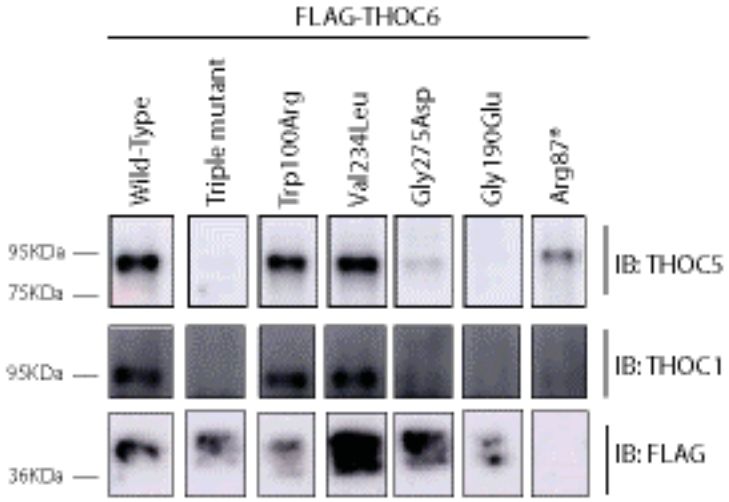

FLAG-THOC6
Wild-Type
Triple mutant
Trp100Arg
Val234Leu
Gly275Asp
Gly190Glu
Arg87*
95KDa
75KDa
95KDa
36KDa
IB: THOC5
IB: THOC1
IB: FLAG

Figure 4

		c.135C>A, p.(Tyr45*)	c.136G>A, p.Gly46Arg	c.259 C>T, p.Arg87*	c.298 T>A, p.Trp100Arg	c.700 G>C, p.Val234Leu	c.824 G>A, p.Gly275Asp	c.569G>A, p.Gly190Glu	c.611 A>C, p.(Gln204Pro)	c.784 A>C, p.(Thr250Pro)
	Reported	Amos et al., 2017	Beaulieu et al., 2013	Amos et al. 2017 Anazi et al., 2016	this report, Casey et al., 2016			this report, Amos et al., 2017	Amos et al., 2017	Amos et al., 2017
	Number of patients	1	4*	2	3			2	1	1
	Patients	P9	P4, P5, P6, P7	P8, P10	P1, P2, P3			P2, P9	P11	P10
Allelic frequency in BBIS patients										
	European/European American	1		1	5			2		1
	Middle East			2					2	
	Hutterite		8							
Frequency in the general population										
GnomAD	**European (Non-Finnnish)**	1 (8.95e-6)	1 (8.95e-6)	1 (3,29e-5)	36 (2.87e-4)			29 (2.30e-4)	0	0
	African	0	0	0	2 (8.33e-5)			1 (4.18e-5)	0	0
	Ashkenazi Jewish	0	0	0	0			0	0	0
	East Asian	0	0	0	0			0	0	0
	European (Finnish)	0	0	0	3 (1.17e-4)			0	0	0
	Latino	0	0	0	4 (1.17e-4)			1 (2.91e-5)	0	0
	South Asian	0	0	0	1 (3.30e-5)			0	0	0
	Other	0	0	0	1 (1.56e-4)			0	0	0
	Total count (MAF)	1 (4,06e-6)	1 (4,06e-6)	1 (4,1e-6)	47 (1.71e-4)			31 (1.12e-4)	0	0
1000Genomes	**All (British)**	0	0	0	1 (2.00e-4)			0,00E+00	0	*0*
UK10K	**ALSPAC (Count)**				2			1		
	ALSPAC (Freq)				1,00E-03			2.59e-4		
	TWINSUK (Count)				2			1		
	TWINSUK (Freq)				1,00E-03			2.80e-4		

Supplementary Table 1: Allelic frequencies reported for each variant in BBIS patients and in the general population

P1, P2 (This report) ; P3 (Casey et al. 2016); P4, P5, P6, P7 (Beaulieu et al. 2013) ; P8 (Anazi et al. 2016); P9, P10, P11 (Amos et al. 2017)

4. CHARACTERIZATION OF FUNCTIONAL CONSEQUENCES OF TRUNCATING MUTATIONS AFFECTING LONG AND SHORT *AUTS2* ISOFORMS

AUTS2 is one of the largest genes in mammals and contains 19 exons, among which the first six are separated by large introns whereas the remaining ones have clustered introns at the 3'end. *AUTS2* has been shown to have an important role in human-specific evolution as its first half region displays the strongest statistical signal in a genomic screen differentiating modern humans from Neanderthals. Other regions found to be statistically different included genes involved in cognitive and social interactions such as *DYRK1A* and *NRG3*, suggesting that specific changes of *AUTS2* occurred in modern humans probably led to cognitive traits specific to human (Green et al., 2010). On the other hand, the 3' end region is well conserved.

AUTS2 was described for the first time in a pair of twins affected by ASD and growth retardation who had an identical balanced translocation interrupting *AUTS2* (Sultana et al., 2002). Since then, many inter- and intra-genic deletion have been identified in patients with a variable phenotype, ranging from neurodevelopmental disorders to other neurological phenotypes (Amarillo et al., 2014; Liu et al., 2015; Nagamani et al., 2013). Nevertheless, some clinical features are more recurrent than others, such as developmental delay, ASD, growth retardation with or without microcephaly and facial dysmorphisms, helping in the delineation of the phenotype caused by this gene. The study of a relative large cohort of patients with a CNV disrupting at least one exon of *AUTS2*, led to the observation that individuals with a deletion encompassing the 3' UTR region have a more severe phenotype compared to the others (Beunders et al., 2013); a rapid amplification of the 5' cDNA ends (5'-RACE) showed that this part of the gene encodes a short C-terminal isoform, possibly explaining the difference in the severity of the phenotype (Beunders et al., 2013). The inactivation of *auts2* in zebrafish leads to microcephaly that could be rescued either by injection of the full-length transcript or by just the C-terminal isoform, suggesting that the severity of the phenotype could be related to the position of the deletion thus if it affects or not the short isoforms located at the 3'UTR region (Beunders et al., 2013). The authors showed that this short-transcript is well expressed in brain and has a transcription start site (TSS) located in exon 9 of the long isoform NM_015570 (starting from Met555, Figure 38). However, in the last genome annotation (Hg38/GRC38), the short isoform is reported to start from exon 6 and has as initiation codon a proline (Pro249, Figure 38). The short isoforms has been also characterized in mouse but two transcripts were identified: one starting from exon 7 (with an initiation codon in exon 8) and the second one starting from exon 9 (Hori et al., 2014). A recent study showed the high transcriptional complexity of *auts2* in zebrafish, reporting numerous different isoforms mediated by alternative splicing as well as alternative promoter usage. Moreover, the expression of the different transcripts is

spatially and temporally regulated (Kondrychyn et al., 2017). Therefore, the characterization of this short isoform is not yet clear.

Several mouse models have been generated to understand the role of *Auts2*: a constitutive knockout affecting the long and short isoforms and one in which only exon 8 is removed, and a conditional knockout with a nestin promoter-driven disrupting exon 7. For the latter, the full homozygous and the heterozygous *Auts2* knockout were characterized and in both models a growth retardation was observed, with heterozygotes showing an intermediate phenotype between wild-type and homozygotes, indicating a gene-dosage dependent effect. Defects in righting reflex and in emitted ultrasonic vocalizations were noticed (Gao et al., 2014). On the other hand, both the constitutive homozygotes were neonatally lethal. Examinations of embryonic brains did not point out any morphological or histological differences between wild-type and mutant mice but defects in cortical neuronal migration were observed. In the knockout mouse lacking the exon 8 the full-length isoform was eliminated while the expression of a short isoform starting at exon 9 was alternatively increased, suggesting a compensatory mechanism. Further analysis in *Auts2*$^{de8l/+}$ and *Auts2*$^{de8l/del8}$ mice showed defects in neuronal cortical migration and axonal elongation in a gene dosage-dependent way, indicating that the long isoform is majorly involved in these processes, since the short isoform was still expressed in these knockout models. As a matter of fact, these phenotypes were rescued by co-electroporation of the full-length *AUTS2* isoform (Hori et al., 2014). Behavioural analyses on the constitutive heterozygous knockout mice for both isoforms showed neurocognitive, recognition and associative memory defects, suggesting that *AUTS2* is implicated in emotional control as well as in learning and memory formation (Hori et al., 2015).

4.1 ROLE OF AUTS2

In mouse, *Auts2* is well expressed in several brain regions, among which cortex, hippocampus and cerebellum, from early neurodevelopmental stages at the embryonic day 12 and being continuously present postnatally, even if the expression level is lower after birth (Bedogni et al., 2010a). *Auts2* is well expressed in the developing cerebral cortex where is majorly expressed in the prefrontal region. *Auts2* has been reported to be exclusively present in nuclei during development and, as the neurodevelopment advances, it emerges also in the cytoplasm and dendrites and axons of the differentiated neurons (Hori et al., 2014).

It has been shown that the two *AUTS2* isoforms are differentially expressed: the long transcript is present both in the cell nuclei and cytoplasm and it is expressed during development and after birth; conversely, the short one is nuclear and it disappears shortly after birth (Hori et al., 2014). These differential expressions could be explained by a distinctive role of AUTS2. For instance, it has been demonstrated that the nuclear AUTS2 acts as a transcriptional regulator for neuronal development. A combined ChIP analysis and RNA-sequencing on mouse embryonic forebrain showed that AUTS2 binds

to promoters of genes highly expressed in the developing forebrain (Oksenberg et al., 2014). Furthermore, AUTS2 interacts with the polycomb repressive complex 1 (PRC1) known to repress transcription. When AUTS2 binds to the PRC1, it inhibits its repressive activity by recruiting the caseine kinase 2 (CK2) and activating gene transcription (Gao et al., 2014). On the other hand, AUTS2 in the cytoplasm is involved in cytoskeletal regulation from one side by activating via the proline-rich domain 1 (PR1) Rac1, which is a Rho family small GTPase that regulates polymerization and microtubule dynamics and promoting lamellipodia formation and neurite extension and, on the other side by inhibiting Cdc42, another Rho family GTPase, leading to repression of filopodia formation in neurites and cell bodies of neurons. Knockdown of *Auts2* in neurons of embryonic mouse brains resulted in neuronal migration defects that were rescued only by the introduction of the full-length AUTS2, indicating that the long isoform is involved in neuronal migration (Hori et al., 2014).

4.2 IDENTIFICATION OF PATIENTS WITH POINT MUTATION IN *AUTS2*

Despite the large number of reported patients with a deletion encompassing *AUTS2*, only two point mutations were reported in literature in individuals affected by moderate ID, microcephaly and ASD, but none of them was located in this short isoform (Beunders et al., 2015, 2016). Later on, SNVs have been reported in the Decipher database and still the majority of them affects both isoforms (Deciphering Developmental Disorders Study, 2017). By targeted-sequencing and WES, we identified four truncating mutations: two of them are *de novo*, one in a boy mainly affected by growth retardation and moderate ID and another one in a girl with mild ID and ASD (Figure 37). The other two SNVs were

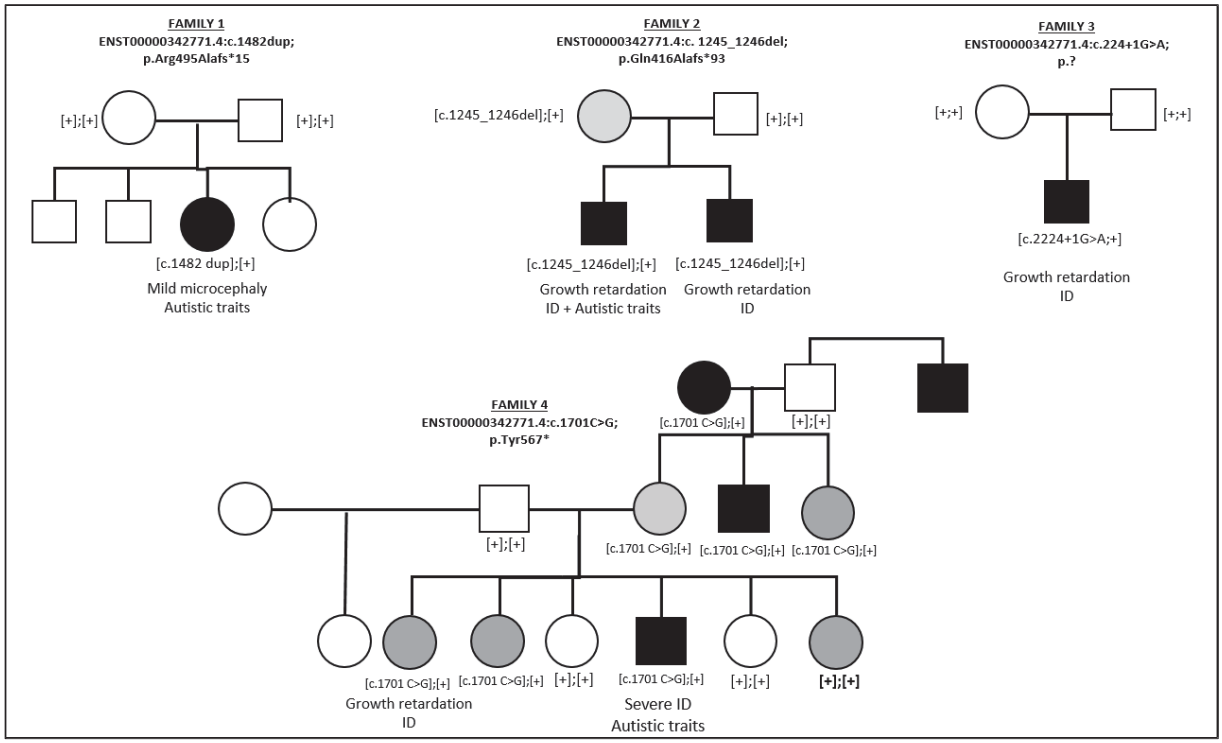


Figure 37: Pedigree of the identified four families with a point mutation in AUTS2

identified in two unrelated families with a variable degree of ID evocating an autosomal dominant inheritance: in one family the variant is transmitted by the mildly affected mother to the two more severely affected child, while the second one includes at least three generations affected by a variable syndromic form of ID and ASD (Figure 37).

Moreover, we have been in contact with other research teams that detected inter- and intra-genic deletions of *AUTS2* and 2 additional patients with a truncating SNV in this gene, located upstream of the putative short transcript identified by Beunders et al. (2013) (Figure 38).

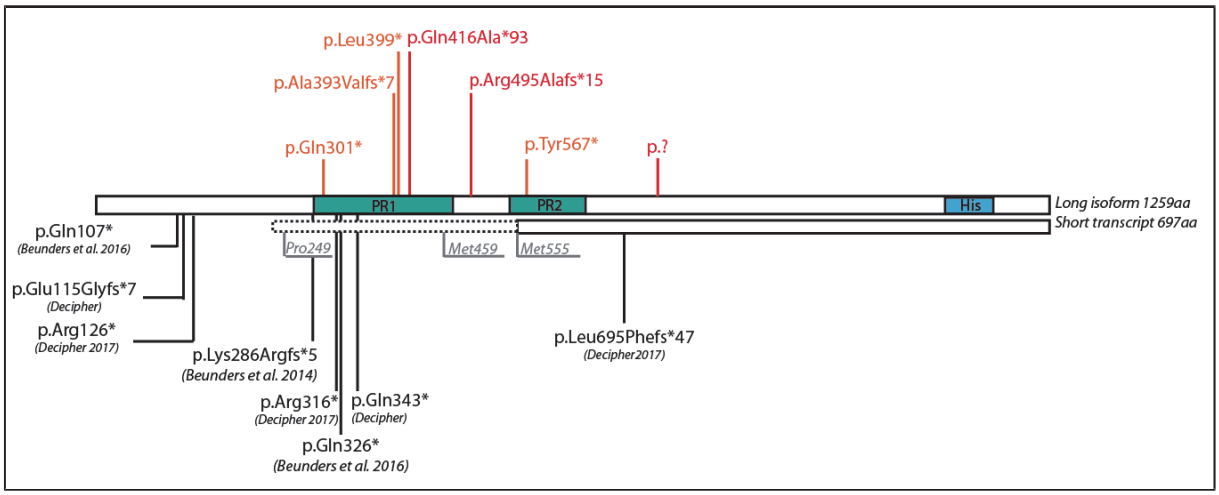


Figure 38: AUTS2 protein and the relative position of the identified SNVs

In red are shown the variants of patients whose fibroblast were available; in orange are reported the additional variants identified; in black the mutations reported either in literature or in the Decipher database. PR stands for proline-rich and His stands for histinde-rich domain.

Overall, several studies revealed the complexity of AUTS2, having different roles as a transcriptional activator as well as cytoskeletal regulator. However, it is still not clear how human mutations – and in particular SNVs – lead to such variable neurodevelopmental disorder, comprising ID and ASD, thus the affected molecular pathways are still not identified. It has been demonstrated that deletions of this gene encompassing the 3' region lead to a more severe phenotype, implicating the contribution of the short isoform majorly in giving rise to the phenotype. Nevertheless, different short transcripts have been described in animal models.

Therefore, the objectives of this project are to:

- characterize the *AUTS2* isoforms in different tissues;
- analyse the functional consequences of *AUTS2* mutations and investigate their effect in gene expression regulation;
- Delineate the clinical phenotype of AUTS2 patients

4.3 CHARACTERIZATION OF *AUTS2* ISOFORMS

In a RNA-sequencing previously performed on human neuronal stem cells (hNSCs) (Quartier et al., 2018), we observed that the short transcript is majorly present (~90%) in this type of cells. In the RNA-

sequencing of human fibroblasts we still noticed the presence of the short isoform, but the ratio is more balanced (~60% short: 40% long). Furthermore, we identified reads in the RNA-sequencing that led us to suspect the existence of two supplementary exons in the short transcript upstream of exon 6, which have named 5b and 5c, where 5c is closer to exon 6 than 5b (Figure 39). These two exons are present in an exclusive manner, *i.e.* either only one exon is included so they are not both present in the same transcript.

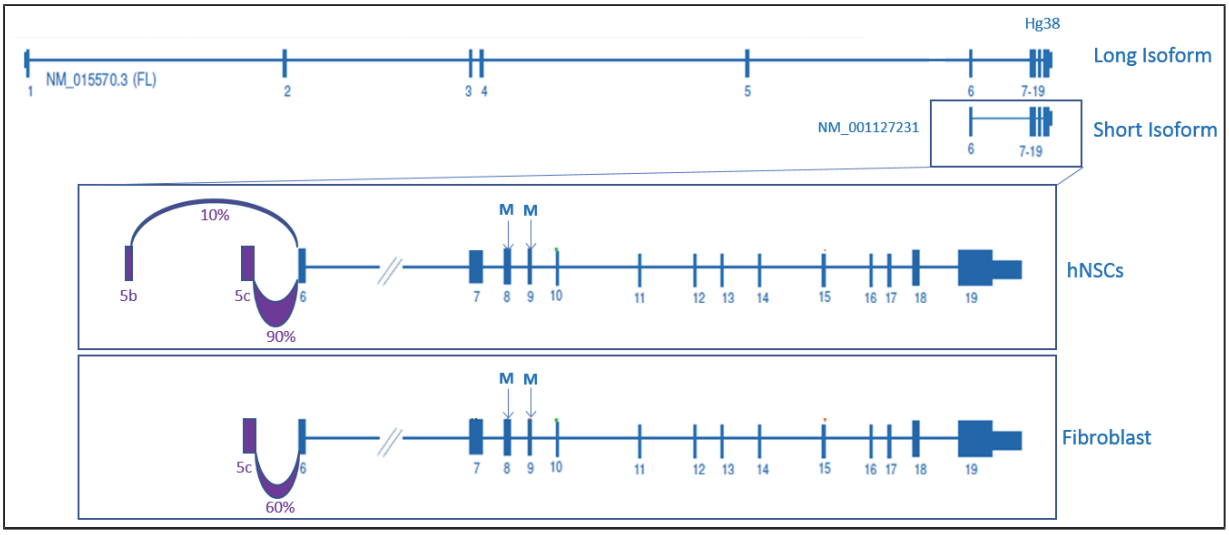


Figure 39: Scheme of the AUTS2 long and short isoform in the last genome version (hg38) and additional identified exon

In the upper part, a schematic representation of the long and short isoform in the last genome version. In the bottom part, a scheme of the identified supplementary exon and their relative ratio in hNSCs and human fibroblasts.

Exon 5c is annotated in Ensembl (ENST00000489774.1) but in an isoform with only 2 exons. Our RNA-sequencing data and RT-PCR confirms that 5c is included in the short transcript both in human fibroblasts and in the 85% of the short isoform of the hNSCs, while the remaining 15% contains exon 5b. On the other hand, 5b is not present in human fibroblasts and the exon 5c is included in the 65% of the short transcript (Figure 39). The exon 5c is also found in mouse (ENSMUST00000161374.7) and its presence is confirmed by RNA-sequencing on brain mouse available online (https://web.stanford.edu/group/barres_lab/brain_rnaseq.html) (Zhang et al., 2014). On the other hand, neither in human nor in mouse there is an in-phase methionine before the exon 8 (Met459), hence it is not possible to exclude the presence of another initiation codon, such as the Pro249.

4.4 FUNCTIONAL CONSEQUENCES OF *AUTS2* MUTATIONS

To investigate the functional consequences of the identified SNVs and to prove their pathogenicity, we analysed the RNA from available patients' fibroblast.

First, we performed a qPCR analysis to check if mutations may undergo a NMD mechanism, using two different couple of primers, one encompassing the long and the short isoform, while the second one amplifies from exon 5c to exon 7. Overall, we did not observe a significant decrease of AUTS2 mRNA

level in patients' RNA compared to RNA from 4 matched controls, except from the patient with a splicing mutation.

To analyse if there is a mutation effect on the gene expression regulation, we performed a transcriptomic analysis on the available patients' fibroblast. A total of 4 patients (of which 2 females and 2 males) from family 1, 2 and 3 were RNA-sequenced in parallel with 4 controls that were sex- and age-matched. Bioinformatic and statistical analysis was then performed as described in *Differential Expression* (pg. 83).

By RNA-sequencing we have been able to validate the spliced mutation identified in family 3, as the mRNA sequence showed a retained intron (Figure 40), validating the *in silico* predictions.

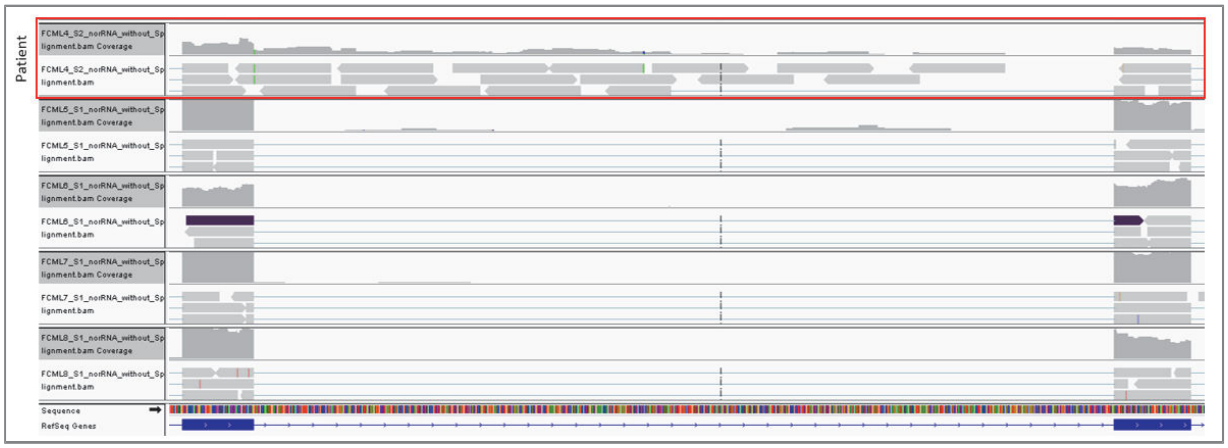


Figure 40: Retained AUTS2 intron in patient from family 3

Overall, about 500 genes were differentially expressed in patients. Two lists of genes were obtained after the filtering-step: the up- and the down-regulated genes. Among the up-regulated genes, we noticed in *AUTS2* patients a significant increase of *RELN* expression at different levels among boys and girls; *RELN* is known to be differentially expressed in males and females (Lintas and Persico, 2010) and it is involved in cerebrocortical development. *RELN* encodes a glycoprotein produced by specific cell types within the developing brain and it activates a signalling pathway that regulates different molecular mechanisms (*e.g.* actin dynamics) that are required for proper neuronal migration and lamination, dendrite and spine development and synaptic function. For instance, homozygous mutations in this gene have been reported in patients affected by lissencephaly, a cortical development malformation. Curiously, *RELN* and *AUTS2* are activated by the same transcription factor *TBR1*, which is a brain-specific T-box transcription factor that regulates laminar identity of postmitotic cortical neurons, axonal pathfinding and neuronal migration (Bedogni et al., 2010b). *RELN* increase in *AUTS2* patients was further validated by qPCR in three different RNA extractions.

The up- and down-regulated gene lists were then analysed on the functional annotation tool of the Database for Annotation, Visualization and Integrated Discovery v6.8 (DAVID) and on the Ingenuity Pathways Analysis (IPA®, Qiagen®). Only modules with a Bonferroni corrected p-value below 0.01 were considered. While we did not observe the presence of any significant annotation cluster in the

upregulated genes, we identified a significant enrichment of clusters that included terms such as cell cycle, mitosis, cell division; centromere, kinetochore, chromosome; kinesin, microtubule (Figure 41). Particularly, we noticed a reduction in genes coding for kinesins (*e.g. KIF2C, KIF4A*), centrosomal proteins (*e.g. CENPE, CENPJ*) and for all the subunits of the Ndc80 complex (*NDC80*, *NUF2*, *SPC24* and *SPC25*), which is involved in microtubule-kinetochore attachment.

Cytoskeletal organization is crucial during neurodevelopment, as it is implicated in a large variety of molecular mechanisms, ranging from the regulation of cell division and migration, to the growth of extensive dendritic arbores and axonal branches, to the transport of cargos along those fibres. The cytoskeletal components microtubules and microtubule-associated proteins are known to play an essential role in the different phases of brain development, such as neurogenesis, neuronal migration, axon growth and synapse formation. Indeed, mutations in human genes coding for several tubulin isoforms (*e.g. TUBG1, TUBB*) or microtubule motor proteins (*e.g. KIF2A, KIF4A*) have been associated to various neurodevelopmental disorders with or without cortical developmental malformations, which are usually caused by defects in neuronal migration or proliferation.

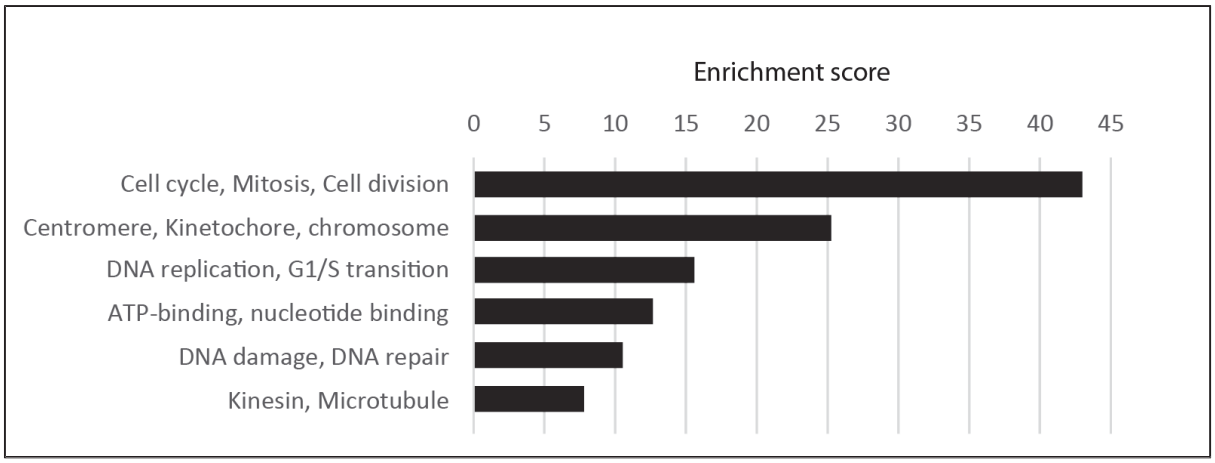


Figure 41: Enrichments scores of the significant enriched functional annotation clusters in DAVID

Centrosome is the major microtubule-organizing centre that controls several cellular processes, among which cell cycle progression, DNA damage response, mitotic spindle formation and genome stability (Barbelanne and Tsang, 2014). The identification of mutations in genes encoding for centrosomal protein in patients affected by primary microcephaly or cortical malformations further places the centrosome as a key regulator in cortical development, especially in neuronal division, migration and proliferation.

The RNA-sequencing in *AUTS2* patients showed a decrease in the expression level of different microtubule-associated proteins, among which centrosomal proteins that have been implicated in primary microcephaly, in detail: *STIL, WDR62, ASPM, CENPE, CENPJ* and *KNL1*. These genes are known to encode for proteins which are involved in mitotic spindle assembly and cell division. Furthermore, expressions of different genes encoding for kinesin proteins, which are molecular motors that uses

microtubules to actively transport cargos along, were reduced in *AUTS2* patients, specifically, 10 kinesin out of the 45 total mammalian kinesin (~22%) (Hirokawa et al., 2009). Other genes implicated in microtubule processes have a decreased expression in *AUTS2* patients, such as all the genes encoding for the subunit for the Ndc80 complex. This complex has a role in assembling the kinetochore itself and its ability to bind and link spindle microtubules to mitotic chromosomes, hence it is an important component for chromosome segregation.

As the RNA-sequencing data pointed to an alteration of genes encoding for proteins involved in the spindle apparatus, we tested if patients' fibroblasts have any defects in the mitotic spindle orientation, by checking the mitotic spindle angle. Control and patients' fibroblast were treated with R03306 (Roche) – an inhibitor of CDK1 - at a final concentration of 10µM for 19 hours to synchronize cells in the G2 phase in order to release them into the mitotic phase. Fibroblasts were then fixed with 4% paraformaldehyde, blocked in 10% foetal calf serum in 0.2% triton, 0.1% bovine serum albumin-TBS 1X and immunostained for tubulin (Abcam, AB6160, 1:1000) and pericentrine (Millipore, AB859, 1:500) overnight at 4°C. Alexa-coupled secondary antibody were used at 1:1000, followed by Hoescht staining. Images were then acquired using a confocal microscope (Leica). At first, we also checked for chromosome lagging in patients' fibroblast but we did not observe it. Spindle angle measurements were then made using the Macro spindle orientation cells plugins of ImageJ. As schematically represented in Figure 42, spindle angles measurement derived from the measures of spindle pole positions that were taken from fixed and immunostained adherent cells (Mannen et al., 2016).

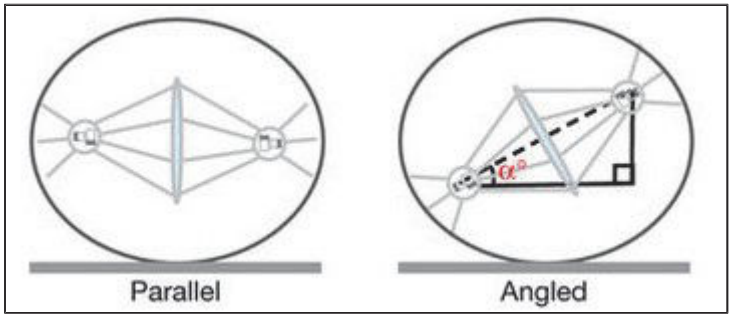


Figure 42: Schematic representation of the mitotic spindle angles measurements (Adapted from (Decarreau et al., 2017))

4 patient cell's lines, including 3 individuals whose RNA was sequenced plus an additional patient with a deletion encompassing *AUTS2*, were compared to 4 different control's fibroblasts. Overall, we observed higher mitotic spindle angles in patients than in controls (Figure 43), indicating that mutations in AUTS2 lead to defects in the spindle apparatus.

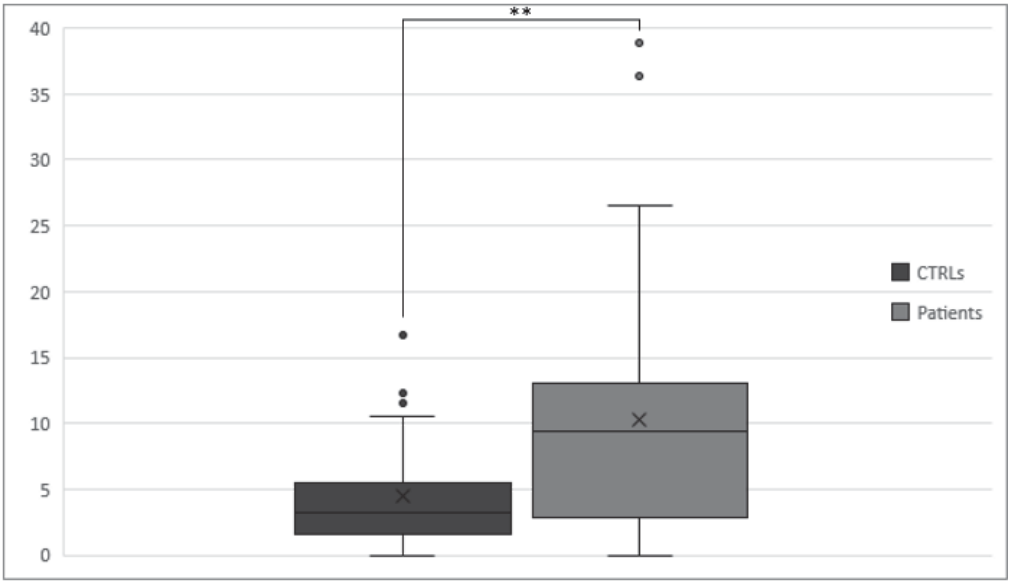


Figure 43: Spindle angle measurements

(** = p-value < 0.01)

The RNA-sequencing data points to a role of AUTS2 in microtubule regulation, with a special focus on microtubule-associated proteins involved in cell division and chromosome segregation, as showed also by the mitotic spindle angle measurements, as well as in cargo transport of the protein kinesin family. The full-length isoform and cytosolic AUTS2 was already described to regulate actin dynamics via the regulation of two Rho family GTPases, thus it is involved in neurite outgrowth and branch formation as well as neuronal migration via Rac1 signalling pathway (Hori et al., 2014). Our data showed that truncating mutations leads to a decrease in expression levels of genes involved in microtubule regulation, particularly during cell division. We observed a significant difference in the mitotic spindle angles between controls' and patients' fibroblasts, suggesting a defect in the spindle apparatus probably due to a dysregulation in microtubules and microtubules-associated proteins.

4.5 CONCLUSIONS AND PERSPECTIVES

Overall, we identified four rare point mutations in four unrelated families: two of these mutations are de novo, while the other two are present transmitted in an autosomal dominant manner and they give rise to a variable phenotype, even among relatives. The identification of either a deletion or a SNVs are found in a relative high percentage of patients affected by ID/ASD (0.3-0.4%), revealing *AUTS2* as one of the most mutated gene implicated in these neurodevelopmental disorders.

Moreover, we characterized two novel exons of the short isoforms of AUTS2 in hNSCs and fibroblasts. However, it is difficult to understand if exon 5c, 6 and 7 are coding in the short isoform, hence to understand if mutations located in those sites are affecting this short transcript.

We investigated on the functional consequences of these SNVs in four available patients' fibroblasts from family 1, 2, and 3. We observed a NMD decay for only one patient (family 3), while we did not

see it in the other individuals. A RNA-sequencing analysis pointed out a decrease in the expression levels of genes involved in cell cycle, mitosis, cell division; centromere, kinetochore, chromosome; kinesin, microtubule. Analyses on the spindle angles from *AUTS2* patients further confirmed a dysregulation in the mitotic apparatus spindle, probably due to transcriptional regulation alteration of genes coding for proteins related to microtubules.

To further confirm that *AUTS2* mutations alters the transcriptional regulation of genes coding for microtubule-related proteins (especially centrosome and kinesin) additional studies should be carried out, especially in different cell type. For instance, it would be interesting to inactivate *AUTS2* in hNSCs by using siRNA to verify if the expression of the identified downregulated genes is directly affected by a decreased level of *AUTS2* or it is rather a long-term effect. Furthermore, it would be interesting to separately inactivate the short and the long transcripts to better understand the contribution of each isoforms: this would be possible thanks to the identification of additional exons specific of the short isoforms. Moreover, it would be useful for the identification of the translational initiation codon, so to check if the identified SNVs affect the coding sequence of the short isoform. The better characterization of alterations of the centrosome (*i.e.* structure and protein localization) and of the microtubular cellular transport could help in the better characterization and understanding of the SNVs effect.

Another goal of this project is to better delineate the *AUTS2* syndrome. Indeed, individuals with a mutation in this gene have a variable phenotype, which may include ID with or without ASD as well as other clinical features. This variability could be explained by the localization of the mutation, as previously showed by the comparison of a severity score between patients with a disruption affecting the N and the C terminus of *AUTS2* (Beunders et al., 2013). On the other hand, even individuals with the same mutation have a variable phenotype (see family 2 and 4), suggesting the presence of other factors involved (*e.g.* mutations in other gene(s)). Therefore, it is extremely difficult to predict a correlation between the genotype and the phenotype; it is hence essential to collect as much clinical information as possible to calculate a comparable clinical severity score. This will also help to test if effectively males are more severely affected than females, as we observed in our cohort of patients. For instance, we observed in *AUTS2* patients a significant increase of *RELN* expression, a gene known to be differentially expressed in males and females. Interestingly, *AUTS2* and *RELN* are activated by the same transcription factor TBR1, a brain-specific T-box transcription factor that has been showed to regulate several other transcription factors involved in ASD, implicating it in the activation of a transcriptional cascade probably implicated in the autism pathogenesis (Chuang et al., 2015). Interestingly, in *Tbr1*$^{-/-}$ mice the expression level of both *RELN* and *AUTS2* were reduced. Additional studies showed that TBR1 directly binds to the promoter region of both genes and regulates their transcription (Bedogni et al., 2010b; Chuang et al., 2015). Furthermore, it has been shown that many

genes associated with *Tbr1* in mouse are also shared by *Ar* coding for the androgen receptor (*i.e. Ovos2*, *Cldn3, Nrgn* and *Cd44*) (Chuang et al., 2015). Therefore, mutations in *AUTS2* might lead to a dysregulation of this pathway, leading to alteration in *RELN* expression. However, the molecular mechanism is not clear (does AUTS2 directly regulates RELN or it does so indirectly, for example, by altering PRC1 regulation?), hence further investigations are required (*i.e.* ChIP, protein analysis).

GENERAL DISCUSSION

Before the introduction of NGS technologies, genetic testing in ID patients was previously limited to fragile-X testing, array-CGH and direct sequencing of genes associated to specific syndromes evoked by the patient's phenotype, limiting the diagnostic yield, especially if we consider that more than 700 genes are implicated in ID. This high genetic heterogeneity hinders patients' diagnosis that is necessary for genetic counselling and to give parents an explanation, and may be useful for medical prognosis of the patient as well as his management and healthcare. NGS technologies greatly improved the diagnostic yield and they are now routinely used in clinical laboratories. For instance, when I arrived in the lab, a TS on an initial cohort comprising more than 200 patients gave a molecular yield of about 25% (Redin et al., 2014). Despite the TS gave a relatively high diagnostic yield, most of the patients remained without a molecular diagnosis. This could be explained by limitations of the TS approach itself that could be overcome by the WES. To increase this diagnosis yield, we sequenced mainly by trio-WES a small cohort of patients that did not receive a molecular diagnosis via TS as well as individuals from multiplex families and with a highly evocative syndromic ID. This strategy increased the general diagnostic yield: overall, WES gave a diagnostic yield of about 30% in the cohort, and more than 40% as a cumulative estimate of our combined NGS strategy (TS + WES).

Conversely to our approach, WES is frequently used in clinical laboratory as a first-intention genetic testing. It has been argued that early testing by WES could save money and time, especially for children with a severe phenotype (Soden et al., 2014). It has been estimated that the cost of prior negative test is 19 100 $ per family, while an exome sequencing is no more than 7640 $ per family (and 2996 $ per individual). Beside the economic factor, the diagnosis could have been made years in advance, ending the diagnostic odyssey that many families are experiencing. A comparison study was also performed between a panel with 500 genes, WES and WGS and concluded that the most optimal approach for diagnosis of ID is the WES, since it well covers the genes present in the panel but it is not as expensive and labour as the WGS (Sun et al., 2015). However, the WES has to reach an adequate coverage in order to limit false positive and negative results. Furthermore, WES data analysis is fast and exhaustive if patients are sequenced in parallel with their parents, significantly increasing the cost, otherwise - in my limited experience - the number of variants is too high to be carefully analysed. One solution could be to focus only on the known ID-genes but this limits the power of the WES approach, which is the detection also of unknown ID genes. The augmented use of WES has raised some ethical issues concerning the delivery of the results of the so-called *incidental* or *secondary findings* that, up to now, is still a matter of debate that is eventually becoming addressed by the scientific community. According to the American College of Medical Genetic and Genomic (ACMG), it is better to refer to them as secondary findings, since they also need to be analysed to conclude on their pathogenicity. Another issue is whether to return all of these detected variants or just the ones in genes for which is possible to prevent and manage a disease, referred to as *actionable* genes (*e.g. BRCA1, TP53*). In the recent

guidelines provided by the ACMG, 59 medically actionable genes were recommended for return in clinical genomic sequences (Kalia et al., 2017). However, there are still ethical concerns on whether to report a known pathogenic variant in a gene with incomplete penetrance, so with an uncertain outcome, or in a gene causing a disease for which actions cannot be taken.

The routinely use of NGS technologies has led to a significant increase in novel mutation and genes implicated in ID but, in parallel, there is also a growing number of variants whose significance is ambiguous (VUS). With the expanding use in recent years of WES in clinical laboratories, a related term to VUS came out indicating genes whose implication in ID is not proven or not clear (GUS). The validation of these VUS and GUS is partially facilitated by the increasing number of patients sequenced, as the chances to identify several or clustered mutations in the same gene are higher. The development of tools promoting such data exchange (*i.e.* MatchMaker) is indeed facilitating variant interpretation, as demonstrated by their important contribution in the identification of a large number of novel ID genes. These tools are an effective method for validating VUS and GUS pathogenicity, especially in clinical laboratories, where the advent of routine WES led not only to variant interpretation but also to novel ID gene identification. For instance, thanks to data exchanging, we have been able to re-evaluate several VUS and GUS (*i.e. NARS, CNOT3*) in a relative short time. However, the comparison of the phenotype can be problematic if the clinical manifestation is variable or not specific. Therefore, it is not always easy to conclude on the pathogenicity of the variant and functional analysis are still required but, at present, the number of VUS or GUS generated by NGS technologies are more that can be validated by just one clinical laboratory. One solution is the collaboration between clinical and research laboratories, but research studies take a long time that is not compatible with the need of a fast diagnosis in hospital. Another possibility is to refer to a laboratory that already published on this gene, so that has an already established protocol and analysis, as we did for the identified variants in *UNC13A.* However, it is difficult to obtain the lab to test for variants after the initial publication or without the inclusion of the patient in a paper.

As the molecular characterization of novel and recurrent ID-genes is progressing, molecular signatures are identified, such as for example the genome-wide methylation signature identified in patient with a mutation in *NSD1*, a gene coding for a methyltransferase and implicated in a syndromic ID (Choufani et al., 2015). Our investigations on deciphering the molecular mechanisms involved in monogenic forms of ID in *BRPF1*, *NOVA2*, *THOC6* and *AUTS2* helped in the identification of potential molecular signatures but future works have still to be addressed in this direction. The characterized molecular signatures could then be used for a rapid screening to test the pathogenicity of a VUS, even in a clinical laboratory.

The identification of genes implicated in ID is extremely important for dissecting ID, delineating the involved molecular pathway and for understanding human brain development. By using a combined NGS strategies (TS+WES), we have been able to identify a novel genetic cause of ID in 13% of individuals in our small cohort of patients.

Interestingly, most of the identified mutations were in genes involved in regulation of gene expression, highlighting the importance of the fine-tuning gene expression during neurodevelopment. This molecular pathway is becoming more and more investigated, rising several questions. For example, many chromatin-related genes identified in ID with a germline mutation have been reported in cancer when mutated at the somatic level (*e.g. KMT2A*, *SETBP1, BRPF1*) but in most of cases germline mutations in these genes do not (or only slightly) increase the risk of tumour formation (*e.g. EP300*). Another issue is to understand how a change in this global regulation could cause such a specific phenotype. This raises the current question of how germline mutations in such ubiquitous proteins could lead to specific defects in brain development and synaptic functions. An explanation might be that neural cells are more sensitive than other cell types, in particular during their differentiation into neurons. Interestingly, most of the mutations in chromatin-related genes associated to neurodevelopmental disorders seem to be gene-dosage sensitive (Berdasco and Esteller, 2013) and the majority of them give rise to a phenotype with a haploinsufficiency mechanism, as we also observed in *BRPF1*. A well-known example of the importance of gene dosage is showed by the X-linked *MECP2* gene. Mutations in this gene are implicated in the Rett-syndrome (RTT) (OMIM: 312750) in females, while a lethal encephalopathy was observed in males whom mother was either an asymptomatic or mildly affected carrier (Zeev et al., 2002), but mutations leading to a truncated protein with partial function result in a syndromic ID (Couvert et al., 2001; Meloni et al., 2000). On the other hand, duplication of this gene have been reported to cause a syndromic ID with high penetrance in males but not in females, who are usually asymptomatic (Van Esch et al., 2005). Studies on the olfactory receptor neurons of RTT patients (Matarazzo et al., 2004) and also in the mouse model (Kishi and Macklis, 2004; Palmer et al., 2008) showed defects in neuronal maturation. Similarly, analyses on induced pluripotent stem cells (iPSCs) showed a reduced number of dendritic spines and synapses (Landucci et al., 2018; Marchetto et al., 2010). Studies on human embryonic stem cells (hESCs), developing to neuronal precursor cells and then to neurons, reported that MECP2 acts as a transcriptional activator in neurons but not in neuronal precursors (Li et al., 2013). Overall, these studies indicate a major role of MECP2 in neuronal maturation, functioning and maintenance, rather than neurogenesis that could explain why the RTT phenotype is observed starting from 6-18 months of age, which is indeed a timing not related to neurogenesis. As a matter of fact, also the temporal regulation could lead to the rising of specific defects in brain. Among the various regulatory mechanisms, there is the alternative splicing, which regulates the presence of isoforms specifically

expressed in brain at particular time point, a mechanism mainly orchestrated by RBPs. *NOVA2* is a RBP regulating such alternative splicing events and the identification of mutations in patients affected by a syndromic ID leading to alterations in splicing events, further support the important role of RBPs in brain development regulation. Interestingly, NOVA2 and its paralogous NOVA1 have been implicated in the paraneoplastic neurologic syndrome (POMA), a neurological disorder characterized by ataxia with or without opsoclonus-myoclonus. POMA is caused by the production of antibodies directed toward a malignant tumor that damage another normal tissue. NOVA2 was first identified as an autoantigen in a subset of POMA patients that developed also cognitive impairment and in some cases encephalopathy (Yang et al., 1998): this neurological disorder is thus caused by the absence of NOVA2, as it is sequestered by the antibody. Defects in RBP have been implicated in a broad spectrum of human diseases, including neurological disorders, encompassing neurodevelopmental and neurodegenerative ones. The difference in the phenotype could be explained by a different type of mutations leading to a different molecular mechanisms. For instance, deletions and severe missenses in *PUM1* have been identified in patients affected by a global developmental delay syndrome, including speech delay and ID as well as ataxia and seizure, while a rare missense variant was identified in a family with several members affected by an adult-onset ataxia with incomplete penetrance (Gennarino et al., 2018). Functional investigations on these variants revealed that the degree of severity of these two disorders is linked to the amount of PUM1 protein, which is a RBP involved in the translation repression. Among its targets, it has been described *ATXN1*, for which mutations have been implicated in spinocerebellar ataxia-1 (Gennarino et al., 2015). The fragile-X syndrome protein FMRP has been also implicated in a late onset adult-onset neurodegenerative disease: when CGG expansion in the 5'UTR of *FMR1* has an intermediate number of repeats, it is not sufficient to cause Fragile-X but it causes the Fragile-X Tremor Ataxia Syndrome (FXTAS). Two main molecular mechanisms have been proposed for this disorder: one in which the produced expanded CGG-repeats in the 5'UTR of *FMR1* mRNA sequester several proteins resulting in neuronal dysfunction (Iwahashi et al., 2006), and a second one in which FXTAS is caused by repeat-associated non-AUG (RAN) translation of the expanded repeats (Sellier et al., 2017), producing small toxic peptides. Overall, defects in RBPs causing both neurodevelopment and neurodegenerative disorders highlight the importance of gene expression regulation in neuron survival and function and provide a link between these two categories of brain disorders.

Even if more patients than before received a molecular diagnosis, analysis of coding sequences of the genome cannot explain all the genetic cases of ID. The role of non-coding regions - corresponding to 98% of the genome - is now emerging as more and more important. Indeed, non-coding regions are particularly important in gene regulation, as they contain many regulatory elements (*i.e.* promoter, alternative splicing and enhancer), and they are also crucial for proper chromatin and epigenetic

organization. It has been shown that regulatory regions are among the most constrained regions of the human genome (Iulio et al., 2018), in line also with the described contribution of *de novo* mutations in regulatory elements to severe neurodevelopmental disorders (Short et al., 2018). As its cost is constantly decreasing, WGS approach is becoming more frequently used, as it is the most comprehensive tool, since it enables the detection of all SNVs; moreover, due to its uniform coverage, it allows the identification of balanced and unbalanced structural variants, which is not possible with the WES. Therefore, as it has been estimated that a high percentage of genetic anomalies in a coding region is a cause of ID (Gilissen et al., 2014), we can reason that WGS is going to solve a large portion of patients, not only those with variants in non-coding regions.

Despite the technical inconvenient of the WGS (*e.g.* data elaboration and storage), the big limitation of WGS is that we still do not have a complete and full understanding of non-coding regions, thus resulting in poor variant annotation and consequently in poor variant prioritization. As in an individual there are millions of variants and many of them are unique, we are thus back to the classical *"needle in the stack"*. Even if tools for understanding variants in non-coding regions are emerging, most of them are restricted to specific functional categories and we are still missing pieces of information. For example, the interpretation of a rare *de novo* variant in the 5'UTR of *MEF2C* was hindered by the lack of integrated interpretation tools. Nonetheless, specific international consortia (*e.g.* ENCODE, Roadmap Epigenomics) are progressing our understanding as well as the development of integrated map with all the regulatory elements of the human genome. Furthermore, it is also emerging the importance of the 3D genome: for instance, topologically associated domains (TADs) are extremely important for bringing together the proper regulatory elements and their disruption has been shown to cause gene expression dysregulation leading to disease, even by disruption of distal elements (Lupiáñez et al., 2015; Redin et al., 2017). It is thus important to consider the 3D genomic architecture and tools for the predictions of regulatory perturbations and interactions within the genome are currently emerging (Bianco et al., 2018; Stadhouders, 2018). The ultimate goal would be a comprehensive database including gene annotations of non-coding regions, functional and constraint data as well as allelic topological conformations in different cell-types at different times, in order to predict the effects of variants in non-coding regions integrating all our knowledge on regulatory elements, gene expression, chromatin folding and modifications.

Many mutations in non-coding regions might affect the mRNA level; to this end, RNA-sequencing can be a useful tool for directly visualizing the mutation effect, as recently demonstrated (Cummings et al., 2017; Kremer et al., 2017). Even if our preliminary results on a small cohort of ID patients are not so encouraging and much work has to be done to implement this analysis, RNA-sequencing could be a good complementary tool to WES and WGS to identify variants in non-coding region and understand their consequences at the same time, obviating the need of prediction tools for variant interpretation.

Indeed, its big and obvious limitation in neurodevelopmental disorders like ID is the inaccessibility of the relevant tissue, so relevant genes that are tissue-specifically expressed are missed; a solution to overcome this issue could be the RNA-sequencing on iPSCs derived from patients' cells. The Genotype-Tissue Expression (GTEx) project proved the efficiency of the RNA-sequencing to identify the association between genetic variation and gene expression levels, by characterizing eQTL (both local and distal) across multiple tissues (GTEx Consortium et al., 2017). Overall, these information are important for assessing the functional properties and consequences of genetic variants and to verify their effect in a specific tissue of interests and could thus explain the many genetic variants previously identified by GWAS that have been associated with human complex disorders. For instance, this could also help in the better understanding of ASD for which - even in our small cohort - its genetic seems to be more complex. This could open the way to the identification of several risk factors with a moderate or variable risk rather than the full penetrance of a single genetic cause, which identification is limited by our current approaches.

Nevertheless, even with the future advent of WGS, we can predict that we will not identify a single pathogenic mutation in all ID-patients. There might exist some more complex genetic forms of neurodevelopmental disorders for which our current methods of analyses are still limited. For instance, while prioritizing the variants, we filter out alleles that are present in the unaffected parents. Therefore, the identification of ADID with incomplete penetrance is still limited, even if the clinical relevance for inherited or common variants was previously shown, for example, in different CNVs, such as in the 15q13.3 and 16p11.2 regions (van Bon et al., 2009; Zufferey et al., 2012). In some cases, the re-evaluation of the parents may reveal a milder phenotype that was previously not noticed (as we observed in two *AUTS2* families), underscoring the importance of clinical evaluations.

Moreover, polymorphisms and variants with an allelic frequency too high in the general population are usually filtered out; however, some of these may be pathogenic in combination with rare variants. For example, the thrombocytopenia–absent radius (TAR) syndrome (characterized by hypomegakaryocytic thrombocytopenia and bilateral radial aplasia) has been associated with a microdeletion on chromosome 1q21.1 but - as an AR inheritance was evocated - other additional modifiers were suspected to contribute to the disorder (Klopocki et al., 2007). It has been only years later that two low-frequent SNPs in two non-coding regions were identified as causative implicated alleles (Albers et al., 2012), postulating a hypomorphic mechanism in which one allele is null and the expression of the second one is reduced, due to the SNP. Due to their relatively high MAF, these SNPs would have probably been filtered out in a WGS.

A variable phenotype could also be explained by a more complex inheritance pattern, such as the combination of multiple genetic variants; however, these studies are limited by the need of a large number of individuals to prove the effect at the cohort level, even for a digenic event. For instance, it has been showed that patients with 16p12.1 microdeletions have significantly more often a second large CNV compared to controls and some of them have a more severe phenotype, whereas carrier parents present more neurological or neuropsychiatry anomalies more often than the non-carrier parents (Girirajan et al., 2010). Another study reported a significant enrichment of variants in *SHANK2* affecting conserved amino acids in ASD patients compared to controls (Leblond et al., 2012), which were also associated to alterations in the synapse density in neuronal cell cultures. Moreover, the authors also reported that ASD probands carry intragenic de novo deletions in *SHANK2* along with inherited CNVs at the 15q11-q13 locus affecting either *CHRNA7* or *CYFPIP1* (Leblond et al., 2012). These results supported the presence of putative modifiers genes (that could be either protective or risk genetic factors), in line with a multiple hit hypothesis. Indeed, a multifactorial model has been proposed in neurodevelopmental disorders, particularly in ASD, as several studies demonstrated that *de novo* CNVs and LoF SNVs are more present in children with ASD than in controls (Iossifov et al., 2014; Sanders et al., 2012; Schaaf et al., 2011) and recent studies reported that common variants are largely implicated in the risk of autism (Gaugler et al., 2014; Klei et al., 2012). Nevertheless, multifactorial models with an interplay between genetic and environmental factors have been also proposed, as also environmental factors could play a role (*i.e.* the prenatal exposure to androgens (Quartier et al., 2018)).

In conclusions, in recent years there has been a great improvement in NGS technologies that allowed us to identify and to better understand novel genetic causes of ID, paving the way toward personalized medicine and potential therapeutic targets. In the next-future, one can imagine there is going to be a similar increasing in the understanding of variants in non-coding regions, providing a molecular diagnosis for some of the current unexplained cases of ID. Nevertheless, not all of them can be explained by a monogenic and highly penetrant model and new methodologies of analysis have to be developed to disentangle more complex inheritance patterns.

BIBLIOGRAPHY

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Abbasi-Moheb, L., Mertel, S., Gonsior, M., Nouri-Vahid, L., Kahrizi, K., Cirak, S., Wieczorek, D., Motazacker, M.M., Esmaeeli-Nieh, S., Cremer, K., et al. (2012). Mutations in NSUN2 cause autosomal-recessive intellectual disability. Am. J. Hum. Genet. *90*, 847–855.

Abu-Elneel, K., Liu, T., Gazzaniga, F.S., Nishimura, Y., Wall, D.P., Geschwind, D.H., Lao, K., and Kosik, K.S. (2008). Heterogeneous dysregulation of microRNAs across the autism spectrum. Neurogenetics *9*, 153–161.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Albers, C.A., Paul, D.S., Schulze, H., Freson, K., Stephens, J.C., Smethurst, P.A., Jolley, J.D., Cvejic, A., Kostadima, M., Bertone, P., et al. (2012). Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. Nat. Genet. *44*, 435–439, S1-2.

Alexander-Bloch, A.F., McDougle, C.J., Ullman, Z., and Sweetser, D.A. (2016). IQSEC2 and X-linked syndromal intellectual disability. Psychiatr. Genet. *26*, 101–108.

Alrahbeni, T., Sartor, F., Anderson, J., Miedzybrodzka, Z., McCaig, C., and Müller, B. (2015). Full UPF3B function is critical for neuronal differentiation of neural stem cells. Mol. Brain *8*, 33.

Amarillo, I.E., Li, W.L., Li, X., Vilain, E., and Kantarci, S. (2014). De novo single exon deletion of AUTS2 in a patient with speech and language disorder: a review of disrupted AUTS2 and further evidence for its role in neurodevelopmental disorders. Am. J. Med. Genet. A. *164A*, 958–965.

Amos, J.S., Huang, L., Thevenon, J., Kariminedjad, A., Beaulieu, C.L., Masurel-Paulet, A., Najmabadi, H., Fattahi, Z., Beheshtian, M., Tonekaboni, S.H., et al. (2017). Autosomal recessive mutations in THOC6 cause intellectual disability: syndrome delineation requiring forward and reverse phenotyping. Clin. Genet. *91*, 92–99.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinforma. Oxf. Engl. *31*, 166–169.

Aranda, P.S., LaJoie, D.M., and Jorcyk, C.L. (2012). Bleach gel: a simple agarose gel for analyzing RNA quality. Electrophoresis *33*, 366–369.

Au, P.Y.B., You, J., Caluseriu, O., Schwartzentruber, J., Majewski, J., Bernier, F.P., Ferguson, M., Care for Rare Canada Consortium, Valle, D., Parboosingh, J.S., et al. (2015). GeneMatcher aids in the identification of a new malformation syndrome with intellectual disability, unique facial dysmorphisms, and skeletal and connective tissue abnormalities caused by de novo variants in HNRNPK. Hum. Mutat. *36*, 1009–1014.

Augustin, I., Rosenmund, C., Südhof, T.C., and Brose, N. (1999). Munc13-1 is essential for fusion competence of glutamatergic synaptic vesicles. Nature *400*, 457–461.

Bain, J.M., Cho, M.T., Telegrafi, A., Wilson, A., Brooks, S., Botti, C., Gowans, G., Autullo, L.A., Krishnamurthy, V., Willing, M.C., et al. (2016). Variants in HNRNPH2 on the X Chromosome Are Associated with a Neurodevelopmental Disorder in Females. Am. J. Hum. Genet. *99*, 728–734.

Bannister, A.J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Res. *21*, 381–395.

Barbelanne, M., and Tsang, W.Y. (2014). Molecular and cellular basis of autosomal recessive primary microcephaly. BioMed Res. Int. *2014*, 547986.

Beaulieu, C.L., Huang, L., Innes, A.M., Akimenko, M.-A., Puffenberger, E.G., Schwartz, C., Jerry, P., Ober, C., Hegele, R.A., McLeod, D.R., et al. (2013). Intellectual disability associated with a homozygous missense mutation in THOC6. Orphanet J. Rare Dis. *8*, 62.

Bedogni, F., Hodge, R.D., Nelson, B.R., Frederick, E.A., Shiba, N., Daza, R.A., and Hevner, R.F. (2010a). Autism susceptibility candidate 2 (Auts2) encodes a nuclear protein expressed in developing brain regions implicated in autism neuropathology. Gene Expr. Patterns GEP *10*, 9–15.

Bedogni, F., Hodge, R.D., Elsen, G.E., Nelson, B.R., Daza, R.A.M., Beyer, R.P., Bammler, T.K., Rubenstein, J.L.R., and Hevner, R.F. (2010b). Tbr1 regulates regional and laminar identity of postmitotic neurons in developing neocortex. Proc. Natl. Acad. Sci. U. S. A. *107*, 13129–13134.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300.

Berdasco, M., and Esteller, M. (2013). Genetic syndromes caused by mutations in epigenetic genes. Hum. Genet. *132*, 359–383.

Bernier, R., Golzio, C., Xiong, B., Stessman, H.A., Coe, B.P., Penn, O., Witherspoon, K., Gerdts, J., Baker, C., Vulto-van Silfhout, A.T., et al. (2014). Disruptive CHD8 mutations define a subtype of autism early in development. Cell *158*, 263–276.

Beunders, G., Voorhoeve, E., Golzio, C., Pardo, L.M., Rosenfeld, J.A., Talkowski, M.E., Simonic, I., Lionel, A.C., Vergult, S., Pyatt, R.E., et al. (2013). Exonic deletions in AUTS2 cause a syndromic form of intellectual disability and suggest a critical role for the C terminus. Am. J. Hum. Genet. *92*, 210–220.

Beunders, G., de Munnik, S.A., Van der Aa, N., Ceulemans, B., Voorhoeve, E., Groffen, A.J., Nillesen, W.M., Meijers-Heijboer, E.J., Frank Kooy, R., Yntema, H.G., et al. (2015). Two male adults with pathogenic AUTS2 variants, including a two-base pair deletion, further delineate the AUTS2 syndrome. Eur. J. Hum. Genet. EJHG *23*, 803–807.

Beunders, G., van de Kamp, J., Vasudevan, P., Morton, J., Smets, K., Kleefstra, T., de Munnik, S.A., Schuurs-Hoeijmakers, J., Ceulemans, B., Zollino, M., et al. (2016). A detailed clinical analysis of 13 patients with AUTS2 syndrome further delineates the phenotypic spectrum and underscores the behavioural phenotype. J. Med. Genet. *53*, 523–532.

Bhalla, K., Phillips, H.A., Crawford, J., McKenzie, O.L.D., Mulley, J.C., Eyre, H., Gardner, A.E., Kremmidiotis, G., and Callen, D.F. (2004). The de novo chromosome 16 translocations of two patients with abnormal phenotypes (mental retardation and epilepsy) disrupt the A2BP1 gene. J. Hum. Genet. *49*, 308–311.

Bianco, S., Lupiáñez, D.G., Chiariello, A.M., Annunziatella, C., Kraft, K., Schöpflin, R., Wittler, L., Andrey, G., Vingron, M., Pombo, A., et al. (2018). Polymer physics predicts the effects of structural variants on chromatin architecture. Nat. Genet. *50*, 662–667.

Bienvenu, T., Diebold, B., Chelly, J., and Isidor, B. (2013). Refining the phenotype associated with MEF2C point mutations. Neurogenetics *14*, 71–75.

Boczonadi, V., Müller, J.S., Pyle, A., Munkley, J., Dor, T., Quartararo, J., Ferrero, I., Karcagi, V., Giunta, M., Polvikoski, T., et al. (2014). EXOSC8 mutations alter mRNA metabolism and cause hypomyelination with spinal muscular atrophy and cerebellar hypoplasia. Nat. Commun. *5*, 4287.

van Bokhoven, H. (2011). Genetic and epigenetic networks in intellectual disabilities. Annu. Rev. Genet. *45*, 81–104.

van Bon, B.W.M., Mefford, H.C., Menten, B., Koolen, D.A., Sharp, A.J., Nillesen, W.M., Innis, J.W., de Ravel, T.J.L., Mercer, C.L., Fichera, M., et al. (2009). Further delineation of the 15q13 microdeletion and duplication syndromes: a clinical spectrum varying from non-pathogenic to a severe outcome. J. Med. Genet. *46*, 511–523.

Buckanovich, R.J., and Darnell, R.B. (1997). The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. Mol. Cell. Biol. *17*, 3194–3201.

Cardoso, C., Boys, A., Parrini, E., Mignon-Ravix, C., McMahon, J.M., Khantane, S., Bertini, E., Pallesi, E., Missirian, C., Zuffardi, O., et al. (2009). Periventricular heterotopia, mental retardation, and epilepsy associated with 5q14.3-q15 deletion. Neurology *72*, 784–792.

Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q., and Krainer, A.R. (2003). ESEfinder: a web resource to identify exonic splicing enhancers. Nucleic Acids Res. *31*, 3568–3571.

Carvill, G.L., Heavin, S.B., Yendle, S.C., McMahon, J.M., O'Roak, B.J., Cook, J., Khan, A., Dorschner, M.O., Weaver, M., Calvert, S., et al. (2013). Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. Nat. Genet. *45*, 825–830.

Casey, J., Jenkinson, A., Magee, A., Ennis, S., Monavari, A., Green, A., Lynch, S.A., Crushell, E., and Hughes, J. (2016). Beaulieu-Boycott-Innes syndrome: an intellectual disability syndrome with characteristic facies. Clin. Dysmorphol. *25*, 146–151.

Chamberlain, S.J. (2013). RNAs of the human chromosome 15q11-q13 imprinted region. Wiley Interdiscip. Rev. RNA *4*, 155–166.

Chelly, J., Khelfaoui, M., Francis, F., Chérif, B., and Bienvenu, T. (2006). Genetics and pathophysiology of mental retardation. Eur. J. Hum. Genet. EJHG *14*, 701–713.

Chen, L.-F., Zhou, A.S., and West, A.E. (2017). Transcribing the connectome: roles for transcription factors and chromatin regulators in activity-dependent synapse development. J. Neurophysiol. *118*, 755–770.

Chérot, E., Keren, B., Dubourg, C., Carré, W., Fradin, M., Lavillaureix, A., Afenjar, A., Burglen, L., Whalen, S., Charles, P., et al. (2018). Using medical exome sequencing to identify the causes of neurodevelopmental disorders: Experience of 2 clinical units and 216 patients. Clin. Genet. *93*, 567–576.

Choufani, S., Cytrynbaum, C., Chung, B.H.Y., Turinsky, A.L., Grafodatskaya, D., Chen, Y.A., Cohen, A.S.A., Dupuis, L., Butcher, D.T., Siu, M.T., et al. (2015). NSD1 mutations generate a genome-wide DNA methylation signature. Nat. Commun. *6*, 10207.

Chuang, H.-C., Huang, T.-N., and Hsueh, Y.-P. (2015). T-Brain-1--A Potential Master Regulator in Autism Spectrum Disorders. Autism Res. Off. J. Int. Soc. Autism Res. *8*, 412–426.

Coffee, B., Keith, K., Albizua, I., Malone, T., Mowrey, J., Sherman, S.L., and Warren, S.T. (2009). Incidence of fragile X syndrome by newborn screening for methylated FMR1 DNA. Am. J. Hum. Genet. *85*, 503–514.

Cooper, G.M., Coe, B.P., Girirajan, S., Rosenfeld, J.A., Vu, T.H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., et al. (2011). A copy number variation morbidity map of developmental delay. Nat. Genet. *43*, 838–846.

Couvert, P., Bienvenu, T., Aquaviva, C., Poirier, K., Moraine, C., Gendrot, C., Verloes, A., Andrès, C., Le Fevre, A.C., Souville, I., et al. (2001). MECP2 is highly mutated in X-linked mental retardation. Hum. Mol. Genet. *10*, 941–946.

Crow, Y.J., Leitch, A., Hayward, B.E., Garner, A., Parmar, R., Griffith, E., Ali, M., Semple, C., Aicardi, J., Babul-Hirji, R., et al. (2006). Mutations in genes encoding ribonuclease H2 subunits cause Aicardi-Goutières syndrome and mimic congenital viral brain infection. Nat. Genet. *38*, 910–916.

Cukier, H.N., Perez, A.M., Collins, A.L., Zhou, Z., Zoghbi, H.Y., and Botas, J. (2008). Genetic modifiers of MeCP2 function in Drosophila. PLoS Genet. *4*, e1000179.

Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O'Grady, G.L., et al. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. *9*.

Darnell, R.B., and Posner, J.B. (2003). Paraneoplastic syndromes involving the nervous system. N. Engl. J. Med. *349*, 1543–1554.

Dauber, A., Golzio, C., Guenot, C., Jodelka, F.M., Kibaek, M., Kjaergaard, S., Leheup, B., Martinet, D., Nowaczyk, M.J.M., Rosenfeld, J.A., et al. (2013). SCRIB and PUF60 are primary drivers of the multisystemic phenotypes of the 8q24.3 copy-number variant. Am. J. Hum. Genet. *93*, 798–811.

De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209–215.

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

Deciphering Developmental Disorders Study (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature *542*, 433–438.

Decker, C.J., and Parker, R. (2012). P-bodies and stress granules: possible roles in the control of translation and mRNA degradation. Cold Spring Harb. Perspect. Biol. *4*, a012286.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. *37*, e67.

Di Donato, N., Neuhann, T., Kahlert, A.-K., Klink, B., Hackmann, K., Neuhann, I., Novotna, B., Schallner, J., Krause, C., Glass, I.A., et al. (2016). Mutations in EXOSC2 are associated with a novel syndrome characterised by retinitis pigmentosa, progressive hearing loss, premature ageing, short stature, mild intellectual disability and distinctive gestalt. J. Med. Genet. *53*, 419–425.

Dias, C., Estruch, S.B., Graham, S.A., McRae, J., Sawiak, S.J., Hurst, J.A., Joss, S.K., Holder, S.E., Morton, J.E.V., Turner, C., et al. (2016). BCL11A Haploinsufficiency Causes an Intellectual Disability Syndrome and Dysregulates Transcription. Am. J. Hum. Genet. *99*, 253–274.

Edvardson, S., Shaag, A., Kolesnikova, O., Gomori, J.M., Tarassov, I., Einbinder, T., Saada, A., and Elpeleg, O. (2007). Deleterious mutation in the mitochondrial arginyl-transfer RNA synthetase gene is associated with pontocerebellar hypoplasia. Am. J. Hum. Genet. *81*, 857–862.

Engel, A.G., Selcen, D., Shen, X.-M., Milone, M., and Harper, C.M. (2016). Loss of MUNC13-1 function causes microcephaly, cortical hyperexcitability, and fatal myasthenia. Neurol. Genet. *2*, e105.

Engels, H., Wohlleber, E., Zink, A., Hoyer, J., Ludwig, K.U., Brockschmidt, F.F., Wieczorek, D., Moog, U., Hellmann-Mersch, B., Weber, R.G., et al. (2009). A novel microdeletion syndrome involving 5q14.3-q15: clinical and molecular cytogenetic characterization of three patients. Eur. J. Hum. Genet. EJHG *17*, 1592–1599.

Esteller, M. (2011). Non-coding RNAs in human disease. Nat. Rev. Genet. *12*, 861–874.

Fairbrother, W.G., Yeo, G.W., Yeh, R., Goldstein, P., Mawson, M., Sharp, P.A., and Burge, C.B. (2004). RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res. *32*, W187–W190.

Favaro, F.P., Alvizi, L., Zechi-Ceide, R.M., Bertola, D., Felix, T.M., de Souza, J., Raskin, S., Twigg, S.R.F., Weiner, A.M.J., Armas, P., et al. (2014). A noncoding expansion in EIF4A3 causes Richieri-Costa-Pereira syndrome, a craniofacial disorder associated with limb defects. Am. J. Hum. Genet. *94*, 120–128.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806–811.

Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am. J. Hum. Genet. *84*, 524–533.

Fogel, B.L., Wexler, E., Wahnich, A., Friedrich, T., Vijayendran, C., Gao, F., Parikshak, N., Konopka, G., and Geschwind, D.H. (2012). RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. Hum. Mol. Genet. *21*, 4171–4186.

Forsberg, S.L., Ilieva, M., and Maria Michel, T. (2018). Epigenetics and cerebral organoids: promising directions in autism spectrum disorders. Transl. Psychiatry *8*, 14.

Frints, S.G.M., Ozanturk, A., Rodríguez Criado, G., Grasshoff, U., de Hoon, B., Field, M., Manouvrier-Hanu, S., E Hickey, S., Kammoun, M., Gripp, K.W., et al. (2018). Pathogenic variants in E3 ubiquitin ligase RLIM/RNF12 lead to a syndromic X-linked intellectual disability and behavior disorder. Mol. Psychiatry.

Gao, Z., Lee, P., Stafford, J.M., von Schimmelmann, M., Schaefer, A., and Reinberg, D. (2014). An AUTS2-Polycomb complex activates gene expression in the CNS. Nature *516*, 349–354.

Gaugler, T., Klei, L., Sanders, S.J., Bodea, C.A., Goldberg, A.P., Lee, A.B., Mahajan, M., Manaa, D., Pawitan, Y., Reichert, J., et al. (2014). Most genetic risk for autism resides with common variation. Nat. Genet. *46*, 881–885.

Gecz, J., Gedeon, A.K., Sutherland, G.R., and Mulley, J.C. (1996). Identification of the gene FMR2, associated with FRAXE mental retardation. Nat. Genet. *13*, 105–108.

Gécz, J., Shoubridge, C., and Corbett, M. (2009). The genetic landscape of intellectual disability arising from chromosome X. Trends Genet. TIG *25*, 308–316.

Geisheker, M.R., Heymann, G., Wang, T., Coe, B.P., Turner, T.N., Stessman, H.A.F., Hoekzema, K., Kvarnung, M., Shaw, M., Friend, K., et al. (2017). Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. Nat. Neurosci. *20*, 1043–1051.

Gennarino, V.A., Singh, R.K., White, J.J., De Maio, A., Han, K., Kim, J.-Y., Jafar-Nejad, P., di Ronza, A., Kang, H., Sayegh, L.S., et al. (2015). Pumilio1 haploinsufficiency leads to SCA1-like neurodegeneration by increasing wild-type Ataxin1 levels. Cell *160*, 1087–1098.

Gennarino, V.A., Palmer, E.E., McDonell, L.M., Wang, L., Adamski, C.J., Koire, A., See, L., Chen, C.-A., Schaaf, C.P., Rosenfeld, J.A., et al. (2018). A Mild PUM1 Mutation Is Associated with Adult-Onset Ataxia, whereas Haploinsufficiency Causes Developmental Delay and Seizures. Cell *172*, 924-936.e11.

Geoffroy, V., Pizot, C., Redin, C., Piton, A., Vasli, N., Stoetzel, C., Blavier, A., Laporte, J., and Muller, J. (2015). VaRank: a simple and powerful tool for ranking genetic variants. PeerJ *3*, e796.

Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W.M., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature *511*, 344–347.

Gilman, S.R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M., and Vitkup, D. (2011). Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron *70*, 898–907.

Girirajan, S., Rosenfeld, J.A., Cooper, G.M., Antonacci, F., Siswara, P., Itsara, A., Vives, L., Walsh, T., McCarthy, S.E., Baker, C., et al. (2010). A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. Nat. Genet. *42*, 203–209.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710–722.

Grozeva, D., Carss, K., Spasic-Boskovic, O., Parker, M.J., Archer, H., Firth, H.V., Park, S.-M., Canham, N., Holder, S.E., Wilson, M., et al. (2014). De novo loss-of-function mutations in SETD5, encoding a methyltransferase in a 3p25 microdeletion syndrome critical region, cause intellectual disability. Am. J. Hum. Genet. *94*, 618–624.

Grozeva, D., Carss, K., Spasic-Boskovic, O., Tejada, M.-I., Gecz, J., Shaw, M., Corbett, M., Haan, E., Thompson, E., Friend, K., et al. (2015). Targeted Next-Generation Sequencing Analysis of 1,000 Individuals with Intellectual Disability. Hum. Mutat. *36*, 1197–1204.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Guilmatre, A., Dubourg, C., Mosca, A.-L., Legallic, S., Goldenberg, A., Drouin-Garraud, V., Layet, V., Rosier, A., Briault, S., Bonnet-Brilhault, F., et al. (2009). Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. Arch. Gen. Psychiatry *66*, 947–956.

Halevy, A., Lerer, I., Cohen, R., Kornreich, L., Shuper, A., Gamliel, M., Zimerman, B.-E., Korabi, I., Meiner, V., Straussberg, R., et al. (2014). Novel EXOSC3 mutation causes complicated hereditary spastic paraplegia. J. Neurol. *261*, 2165–2169.

Hamdan, F.F., Piton, A., Gauthier, J., Lortie, A., Dubeau, F., Dobrzeniecka, S., Spiegelman, D., Noreau, A., Pellerin, S., Côté, M., et al. (2009a). De novo STXBP1 mutations in mental retardation and nonsyndromic epilepsy. Ann. Neurol. *65*, 748–753.

Hamdan, F.F., Gauthier, J., Spiegelman, D., Noreau, A., Yang, Y., Pellerin, S., Dobrzeniecka, S., Côté, M., Perreau-Linck, E., Perreault-Linck, E., et al. (2009b). Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. N. Engl. J. Med. *360*, 599–605.

Hamdan, F.F., Daoud, H., Rochefort, D., Piton, A., Gauthier, J., Langlois, M., Foomani, G., Dobrzeniecka, S., Krebs, M.-O., Joober, R., et al. (2010). De novo mutations in FOXP1 in cases with intellectual disability, autism, and language impairment. Am. J. Hum. Genet. *87*, 671–678.

Hamdan, F.F., Srour, M., Capo-Chichi, J.-M., Daoud, H., Nassif, C., Patry, L., Massicotte, C., Ambalavanan, A., Spiegelman, D., Diallo, O., et al. (2014). De novo mutations in moderate or severe intellectual disability. PLoS Genet. *10*, e1004772.

Han, K., Müller, U.C., and Hülsmann, S. (2017). Amyloid-precursor Like Proteins APLP1 and APLP2 Are Dispensable for Normal Development of the Neonatal Respiratory Network. Front. Mol. Neurosci. *10*, 189.

Haraksingh, R.R., and Snyder, M.P. (2013). Impacts of Variation in the Human Genome on Gene Regulation. J. Mol. Biol. *425*, 3970–3977.

Harripaul, R., Noor, A., Ayub, M., and Vincent, J.B. (2017). The Use of Next-Generation Sequencing for Research and Diagnostics for Intellectual Disability. Cold Spring Harb. Perspect. Med. *7*.

Hartley, S.W., and Mullikin, J.C. (2016). Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq. Nucleic Acids Res. *44*, e127.

Hartman, J.L., Garvik, B., and Hartwell, L. (2001). Principles for the buffering of genetic variation. Science *291*, 1001–1004.

Hirokawa, N., Noda, Y., Tanaka, Y., and Niwa, S. (2009). Kinesin superfamily motor proteins and intracellular transport. Nat. Rev. Mol. Cell Biol. *10*, 682–696.

Hoischen, A., Krumm, N., and Eichler, E.E. (2014). Prioritization of neurodevelopmental disease genes by discovery of new mutations. Nat. Neurosci. *17*, 764–772.

Hori, K., Nagai, T., Shan, W., Sakamoto, A., Taya, S., Hashimoto, R., Hayashi, T., Abe, M., Yamazaki, M., Nakao, K., et al. (2014). Cytoskeletal regulation by AUTS2 in neuronal migration and neuritogenesis. Cell Rep. *9*, 2166–2179.

Hori, K., Nagai, T., Shan, W., Sakamoto, A., Abe, M., Yamazaki, M., Sakimura, K., Yamada, K., and Hoshino, M. (2015). Heterozygous Disruption of Autism susceptibility candidate 2 Causes Impaired Emotional Control and Cognitive Memory. PLOS ONE *10*, e0145979.

Hu, H., Haas, S.A., Chelly, J., Van Esch, H., Raynaud, M., de Brouwer, A.P.M., Weinert, S., Froyen, G., Frints, S.G.M., Laumonnier, F., et al. (2016). X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes. Mol. Psychiatry *21*, 133–148.

Hu, H., Kahrizi, K., Musante, L., Fattahi, Z., Herwig, R., Hosseini, M., Oppitz, C., Abedini, S.S., Suckow, V., Larti, F., et al. (2018). Genetics of intellectual disability in consanguineous families. Mol. Psychiatry.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Huguet, G., Ey, E., and Bourgeron, T. (2013). The genetic landscapes of autism spectrum disorders. Annu. Rev. Genomics Hum. Genet. *14*, 191–213.

Huisman, S.A., Redeker, E.J.W., Maas, S.M., Mannens, M.M., and Hennekam, R.C.M. (2013). High rate of mosaicism in individuals with Cornelia de Lange syndrome. J. Med. Genet. *50*, 339–344.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.-H., Narzisi, G., Leotta, A., et al. (2012). De novo gene disruptions in children on the autistic spectrum. Neuron *74*, 285–299.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.

Iulio, J. di, Bartha, I., Wong, E.H.M., Yu, H.-C., Lavrenko, V., Yang, D., Jung, I., Hicks, M.A., Shah, N., Kirkness, E.F., et al. (2018). The human noncoding genome defined by genetic diversity. Nat. Genet. *50*, 333–337.

Iwahashi, C.K., Yasui, D.H., An, H.-J., Greco, C.M., Tassone, F., Nannen, K., Babineau, B., Lebrilla, C.B., Hagerman, R.J., and Hagerman, P.J. (2006). Protein composition of the intranuclear inclusions of FXTAS. Brain J. Neurol. *129*, 256–271.

Jansen, S., Geuer, S., Pfundt, R., Brough, R., Ghongane, P., Herkert, J.C., Marco, E.J., Willemsen, M.H., Kleefstra, T., Hannibal, M., et al. (2017). De Novo Truncating Mutations in the Last and Penultimate Exons of PPM1D Cause an Intellectual Disability Syndrome. Am. J. Hum. Genet. *100*, 650–658.

Jensen, K.B., Musunuru, K., Lewis, H.A., Burley, S.K., and Darnell, R.B. (2000a). The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain. Proc. Natl. Acad. Sci. U. S. A. *97*, 5740–5745.

Jensen, K.B., Dredge, B.K., Stefani, G., Zhong, R., Buckanovich, R.J., Okano, H.J., Yang, Y.Y., and Darnell, R.B. (2000b). Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. Neuron *25*, 359–371.

Jin, P., Alisch, R.S., and Warren, S.T. (2004). RNA and microRNAs in fragile X mental retardation. Nat. Cell Biol. *6*, 1048–1053.

Kalia, S.S., Adelman, K., Bale, S.J., Chung, W.K., Eng, C., Evans, J.P., Herman, G.E., Hufnagel, S.B., Klein, T.E., Korf, B.R., et al. (2017). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet. Med. Off. J. Am. Coll. Med. Genet. *19*, 249–255.

Kalscheuer, V.M., Freude, K., Musante, L., Jensen, L.R., Yntema, H.G., Gécz, J., Sefiani, A., Hoffmann, K., Moser, B., Haas, S., et al. (2003). Mutations in the polyglutamine binding protein 1 gene cause X-linked mental retardation. Nat. Genet. *35*, 313–315.

Kanai, Y., Dohmae, N., and Hirokawa, N. (2004). Kinesin transports RNA: isolation and characterization of an RNA-transporting granule. Neuron *43*, 513–525.

Kaufman, L., Ayub, M., and Vincent, J.B. (2010). The genetic basis of non-syndromic intellectual disability: a review. J. Neurodev. Disord. *2*, 182–209.

Ke, R., Mignardi, M., Hauling, T., and Nilsson, M. (2016). Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences. Hum. Mutat. *37*, 1363–1367.

Kiebler, M.A., and Bassell, G.J. (2006). Neuronal RNA granules: movers and makers. Neuron *51*, 685–690.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Kirchner, S., and Ignatova, Z. (2015). Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. Nat. Rev. Genet. *16*, 98–112.

Kishi, N., and Macklis, J.D. (2004). MECP2 is progressively expressed in post-migratory neurons and is involved in neuronal maturation rather than cell fate decisions. Mol. Cell. Neurosci. *27*, 306–321.

Kleefstra, T., Schenck, A., Kramer, J.M., and van Bokhoven, H. (2014). The genetics of cognitive epigenetics. Neuropharmacology *80*, 83–94.

Klei, L., Sanders, S.J., Murtha, M.T., Hus, V., Lowe, J.K., Willsey, A.J., Moreno-De-Luca, D., Yu, T.W., Fombonne, E., Geschwind, D., et al. (2012). Common genetic variants, acting additively, are a major source of risk for autism. Mol. Autism *3*, 9.

Klopocki, E., Schulze, H., Strauss, G., Ott, C.-E., Hall, J., Trotier, F., Fleischhauer, S., Greenhalgh, L., Newbury-Ecob, R.A., Neumann, L.M., et al. (2007). Complex inheritance pattern resembling autosomal recessive inheritance involving a microdeletion in thrombocytopenia-absent radius syndrome. Am. J. Hum. Genet. *80*, 232–240.

Kochinke, K., Zweier, C., Nijhof, B., Fenckova, M., Cizek, P., Honti, F., Keerthikumar, S., Oortveld, M.A.W., Kleefstra, T., Kramer, J.M., et al. (2016). Systematic Phenomics Analysis Deconvolutes Genes Mutated in Intellectual Disability into Biologically Coherent Modules. Am. J. Hum. Genet. *98*, 149–164.

Kondrychyn, I., Robra, L., and Thirumalai, V. (2017). Transcriptional Complexity and Distinct Expression Patterns ofauts2Paralogs inDanio rerio. G3 Bethesda Md *7*, 2577–2593.

Kopajtich, R., Murayama, K., Janecke, A.R., Haack, T.B., Breuer, M., Knisely, A.S., Harting, I., Ohashi, T., Okazaki, Y., Watanabe, D., et al. (2016). Biallelic IARS Mutations Cause Growth Retardation with Prenatal Onset, Intellectual Disability, Muscular Hypotonia, and Infantile Hepatopathy. Am. J. Hum. Genet. *99*, 414–422.

de Kovel, C.G.F., Brilstra, E.H., van Kempen, M.J.A., Van't Slot, R., Nijman, I.J., Afawi, Z., De Jonghe, P., Djémié, T., Guerrini, R., Hardies, K., et al. (2016). Targeted sequencing of 351 candidate genes for epileptic encephalopathy in a large cohort of patients. Mol. Genet. Genomic Med. *4*, 568–580.

Kozol, R.A., Abrams, A.J., James, D.M., Buglo, E., Yan, Q., and Dallman, J.E. (2016). Function Over Form: Modeling Groups of Inherited Neurological Conditions in Zebrafish. Front. Mol. Neurosci. *9*, 55.

Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., et al. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat. Commun. *8*, 15824.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.

Kumar, R., Corbett, M.A., van Bon, B.W.M., Woenig, J.A., Weir, L., Douglas, E., Friend, K.L., Gardner, A., Shaw, M., Jolly, L.A., et al. (2015). THOC2 Mutations Implicate mRNA-Export Pathway in X-Linked Intellectual Disability. Am. J. Hum. Genet. *97*, 302–310.

Küry, S., van Woerden, G.M., Besnard, T., Proietti Onori, M., Latypova, X., Towne, M.C., Cho, M.T., Prescott, T.E., Ploeg, M.A., Sanders, S., et al. (2017). De Novo Mutations in Protein Kinase Genes CAMK2A and CAMK2B Cause Intellectual Disability. Am. J. Hum. Genet. *101*, 768–788.

Kuss, A.W., Garshasbi, M., Kahrizi, K., Tzschach, A., Behjati, F., Darvish, H., Abbasi-Moheb, L., Puettmann, L., Zecha, A., Weissmann, R., et al. (2011). Autosomal recessive mental retardation: homozygosity mapping identifies 27 single linkage intervals, at least 14 novel loci and several mutation hotspots. Hum. Genet. *129*, 141–148.

Lal, D., Pernhorst, K., Klein, K.M., Reif, P., Tozzi, R., Toliat, M.R., Winterer, G., Neubauer, B., Nürnberg, P., Rosenow, F., et al. (2015). Extending the phenotypic spectrum of RBFOX1 deletions: Sporadic focal epilepsy. Epilepsia *56*, e129-133.

Landucci, E., Brindisi, M., Bianciardi, L., Catania, L.M., Daga, S., Croci, S., Frullanti, E., Fallerini, C., Butini, S., Brogi, S., et al. (2018). iPSC-derived neurons profiling reveals GABAergic circuit disruption and acetylated α-tubulin defect which improves after iHDAC6 treatment in Rett syndrome. Exp. Cell Res.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lardelli, R.M., Schaffer, A.E., Eggens, V.R.C., Zaki, M.S., Grainger, S., Sathe, S., Van Nostrand, E.L., Schlachetzki, Z., Rosti, B., Akizu, N., et al. (2017). Biallelic mutations in the 3' exonuclease TOE1 cause pontocerebellar hypoplasia and uncover a role in snRNA processing. Nat. Genet. *49*, 457–464.

Le Meur, N., Holder-Espinasse, M., Jaillard, S., Goldenberg, A., Joriot, S., Amati-Bonneau, P., Guichet, A., Barth, M., Charollais, A., Journel, H., et al. (2010). MEF2C haploinsufficiency caused by either microdeletion of the 5q14.3 region or mutation is responsible for severe mental retardation with stereotypic movements, epilepsy and/or cerebral malformations. J. Med. Genet. *47*, 22–29.

Leblond, C.S., Heinrich, J., Delorme, R., Proepper, C., Betancur, C., Huguet, G., Konyukh, M., Chaste, P., Ey, E., Rastam, M., et al. (2012). Genetic and functional analyses of SHANK2 mutations suggest a multiple hit model of autism spectrum disorders. PLoS Genet. *8*, e1002521.

Lee, T.-Y., Lin, Z.-Q., Hsieh, S.-J., Bretaña, N.A., and Lu, C.-T. (2011). Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. Bioinforma. Oxf. Engl. *27*, 1780–1787.

Lejeune, J., Turpin, R., and Gautier, M. (1959). [Chromosomic diagnosis of mongolism]. Arch. Fr. Pediatr. *16*, 962–963.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Lemke, J.R., Riesch, E., Scheurenbrand, T., Schubach, M., Wilhelm, C., Steiner, I., Hansen, J., Courage, C., Gallati, S., Bürki, S., et al. (2012). Targeted next generation sequencing as a diagnostic tool in epileptic disorders. Epilepsia *53*, 1387–1398.

Lessel, D., Schob, C., Küry, S., Reijnders, M.R.F., Harel, T., Eldomery, M.K., Coban-Akdemir, Z., Denecke, J., Edvardson, S., Colin, E., et al. (2017). De Novo Missense Mutations in DHX30 Impair Global Translation and Cause a Neurodevelopmental Disorder. Am. J. Hum. Genet. *101*, 716–724.

Lévy, J., Coussement, A., Dupont, C., Guimiot, F., Baumann, C., Viot, G., Passemard, S., Capri, Y., Drunat, S., Verloes, A., et al. (2017). Molecular and clinical delineation of 2p15p16.1 microdeletion syndrome. Am. J. Med. Genet. A. *173*, 2081–2087.

Lewis, H.A., Musunuru, K., Jensen, K.B., Edo, C., Chen, H., Darnell, R.B., and Burley, S.K. (2000). Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. Cell *100*, 323–332.

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

Li, H., Radford, J.C., Ragusa, M.J., Shea, K.L., McKercher, S.R., Zaremba, J.D., Soussou, W., Nie, Z., Kang, Y.-J., Nakanishi, N., et al. (2008). Transcription factor MEF2C influences neural stem/progenitor cell differentiation and maturation in vivo. Proc. Natl. Acad. Sci. U. S. A. *105*, 9397–9402.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. *25*, 2078–2079.

Li, J.J., Bickel, P.J., and Biggin, M.D. (2014). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ *2*, e270.

Li, Y., Wang, H., Muffat, J., Cheng, A.W., Orlando, D.A., Lovén, J., Kwok, S.-M., Feldman, D.A., Bateup, H.S., Gao, Q., et al. (2013). Global transcriptional and translational repression in human-embryonic-stem-cell-derived Rett syndrome neurons. Cell Stem Cell *13*, 446–458.

Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. Nat. Genet. *50*, 151–158.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.

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

Lines, M.A., Huang, L., Schwartzentruber, J., Douglas, S.L., Lynch, D.C., Beaulieu, C., Guion-Almeida, M.L., Zechi-Ceide, R.M., Gener, B., Gillessen-Kaesbach, G., et al. (2012). Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly. Am. J. Hum. Genet. *90*, 369–377.

Lintas, C., and Persico, A.M. (2010). Neocortical RELN promoter methylation increases significantly after puberty. Neuroreport *21*, 114–118.

Lionel, A.C., Costain, G., Monfared, N., Walker, S., Reuter, M.S., Hosseini, S.M., Thiruvahindrapuram, B., Merico, D., Jobling, R., Nalpathamkalam, T., et al. (2017). Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet. Med. Off. J. Am. Coll. Med. Genet.

Lipstein, N., Verhoeven-Duif, N.M., Michelassi, F.E., Calloway, N., van Hasselt, P.M., Pienkowska, K., van Haaften, G., van Haelst, M.M., van Empelen, R., Cuppen, I., et al. (2017). Synaptic UNC13A protein variant causes increased neurotransmission and dyskinetic movement disorder. J. Clin. Invest. *127*, 1005–1018.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013). Global epigenomic reconfiguration during mammalian brain development. Science *341*, 1237905.

Liu, H., and Wong, L. (2003). Data mining tools for biological sequences. J. Bioinform. Comput. Biol. *1*, 139–167.

Liu, Y., Zhao, D., Dong, R., Yang, X., Zhang, Y., Tammimies, K., Uddin, M., Scherer, S.W., and Gai, Z. (2015). De novo exon 1 deletion of AUTS2 gene in a patient with autism spectrum disorder and developmental delay: a case report and a brief literature review. Am. J. Med. Genet. A. *167*, 1381–1385.

Lou, C.H., Shao, A., Shum, E.Y., Espinoza, J.L., Huang, L., Karam, R., and Wilkinson, M.F. (2014). Posttranscriptional control of the stem cell and neurogenic programs by the nonsense-mediated RNA decay pathway. Cell Rep. *6*, 748–764.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Loviglio, M.N., Leleu, M., Männik, K., Passeggeri, M., Giannuzzi, G., van der Werf, I., Waszak, S.M., Zazhytska, M., Roberts-Caldeira, I., Gheldof, N., et al. (2017). Chromosomal contacts connect loci associated with autism, BMI and head circumference phenotypes. Mol. Psychiatry *22*, 836–849.

Lubs, H.A. (1969). A marker X chromosome. Am. J. Hum. Genet. *21*, 231–244.

Lubs, H.A., Stevenson, R.E., and Schwartz, C.E. (2012). Fragile X and X-linked intellectual disability: four decades of discovery. Am. J. Hum. Genet. *90*, 579–590.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Lupiáñez, D.G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. Trends Genet. TIG *32*, 225–237.

Lupski, J.R. (1998). Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet. TIG *14*, 417–422.

Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D.C.Y., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D.A., et al. (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N. Engl. J. Med. *362*, 1181–1191.

Mannen, T., Yamashita, S., Tomita, K., Goshima, N., and Hirose, T. (2016). The Sam68 nuclear body is composed of two RNase-sensitive substructures joined by the adaptor HNRNPL. J. Cell Biol. *214*, 45–59.

Marchetto, M.C.N., Carromeu, C., Acab, A., Yu, D., Yeo, G.W., Mu, Y., Chen, G., Gage, F.H., and Muotri, A.R. (2010). A model for neural development and treatment of Rett syndrome using human induced pluripotent stem cells. Cell *143*, 527–539.

Martin, C.L., Duvall, J.A., Ilkin, Y., Simon, J.S., Arreaza, M.G., Wilkes, K., Alvarez-Retuerto, A., Whichello, A., Powell, C.M., Rao, K., et al. (2007). Cytogenetic and molecular characterization of A2BP1/FOX1 as a candidate gene for autism. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet. *144B*, 869–876.

Martínez, F., Caro-Llopis, A., Roselló, M., Oltra, S., Mayo, S., Monfort, S., and Orellana, C. (2017). High diagnostic yield of syndromic intellectual disability by targeted next-generation sequencing. J. Med. Genet. *54*, 87–92.

Mary, L., Piton, A., Schaefer, E., Mattioli, F., Nourisson, E., Feger, C., Redin, C., Barth, M., El Chehadeh, S., Colin, E., et al. (2018). Disease-causing variants in TCF4 are a frequent cause of intellectual disability: lessons from large-scale sequencing approaches in diagnosis. Eur. J. Hum. Genet. EJHG.

Matarazzo, V., Cohen, D., Palmer, A.M., Simpson, P.J., Khokhar, B., Pan, S.-J., and Ronnett, G.V. (2004). The transcriptional repressor Mecp2 regulates terminal neuronal differentiation. Mol. Cell. Neurosci. *27*, 44–58.

Mathe, E., Olivier, M., Kato, S., Ishioka, C., Hainaut, P., and Tavtigian, S.V. (2006). Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. Nucleic Acids Res. *34*, 1317–1325.

Mattioli, F., Piton, A., Gérard, B., Superti-Furga, A., Mandel, J.-L., and Unger, S. (2016). Novel de novo mutations in ZBTB20 in Primrose syndrome with congenital hypothyroidism. Am. J. Med. Genet. A. *170*, 1626–1629.

Mattioli, F., Schaefer, E., Magee, A., Mark, P., Mancini, G.M., Dieterich, K., Von Allmen, G., Alders, M., Coutton, C., van Slegtenhorst, M., et al. (2017). Mutations in Histone Acetylase Modifier BRPF1 Cause an Autosomal-Dominant Form of Intellectual Disability with Associated Ptosis. Am. J. Hum. Genet. *100*, 105–116.

Maulik, P.K., Mascarenhas, M.N., Mathers, C.D., Dua, T., and Saxena, S. (2011). Prevalence of intellectual disability: a meta-analysis of population-based studies. Res. Dev. Disabil. *32*, 419–436.

McLaren, J., and Bryson, S.E. (1987). Review of recent epidemiological studies of mental retardation: prevalence, associated disorders, and etiology. Am. J. Ment. Retard. AJMR *92*, 243–254.

Meloni, I., Bruttini, M., Longo, I., Mari, F., Rizzolio, F., D'Adamo, P., Denvriendt, K., Fryns, J.P., Toniolo, D., and Renieri, A. (2000). A mutation in the rett syndrome gene, MECP2, causes X-linked mental retardation and progressive spasticity in males. Am. J. Hum. Genet. *67*, 982–985.

Meyer-Schuman, R., and Antonellis, A. (2017). Emerging mechanisms of aminoacyl-tRNA synthetase mutations in recessive and dominant human disease. Hum. Mol. Genet. *26*, R114–R127.

Mircsof, D., Langouët, M., Rio, M., Moutton, S., Siquier-Pernet, K., Bole-Feysot, C., Cagnard, N., Nitschke, P., Gaspar, L., Žnidarič, M., et al. (2015). Mutations in NONO lead to syndromic intellectual disability and inhibitory synaptic defects. Nat. Neurosci. *18*, 1731–1736.

Mirzaa, G.M., Enyedi, L., Parsons, G., Collins, S., Medne, L., Adams, C., Ward, T., Davitt, B., Bicknese, A., Zackai, E., et al. (2014). Congenital microcephaly and chorioretinopathy due to de novo heterozygous KIF11 mutations: five novel mutations and review of the literature. Am. J. Med. Genet. A. *164A*, 2879–2886.

Moore, M.J. (2005). From birth to death: the complex lives of eukaryotic mRNAs. Science *309*, 1514–1518.

Morita, M., Oike, Y., Nagashima, T., Kadomatsu, T., Tabata, M., Suzuki, T., Nakamura, T., Yoshida, N., Okada, M., and Yamamoto, T. (2011). Obesity resistance and increased hepatic expression of catabolism-related mRNAs in Cnot3+/- mice. EMBO J. *30*, 4678–4691.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Morton, D.J., Kuiper, E.G., Jones, S.K., Leung, S.W., Corbett, A.H., and Fasken, M.B. (2018). The RNA exosome and RNA exosome-linked disease. RNA N. Y. N *24*, 127–142.

Musante, L., and Ropers, H.H. (2014a). Genetics of recessive cognitive disorders. Trends Genet. *30*, 32–39.

Musante, L., and Ropers, H.H. (2014b). Genetics of recessive cognitive disorders. Trends Genet. TIG *30*, 32–39.

Nachtergaele, S., and He, C. (2017). The emerging biology of RNA post-transcriptional modifications. RNA Biol. *14*, 156–163.

Nagamani, S.C.S., Erez, A., Ben-Zeev, B., Frydman, M., Winter, S., Zeller, R., El-Khechen, D., Escobar, L., Stankiewicz, P., Patel, A., et al. (2013). Detection of copy-number variation in AUTS2 gene by targeted exonic array CGH in patients with developmental delay and autistic spectrum disorders. Eur. J. Hum. Genet. EJHG *21*, 343–346.

Najmabadi, H., Hu, H., Garshasbi, M., Zemojtel, T., Abedini, S.S., Chen, W., Hosseini, M., Behjati, F., Haas, S., Jamali, P., et al. (2011). Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature *478*, 57–63.

Nakajima, J., Okamoto, N., Tohyama, J., Kato, M., Arai, H., Funahashi, O., Tsurusaki, Y., Nakashima, M., Kawashima, H., Saitsu, H., et al. (2015). De novo EEF1A2 mutations in patients with characteristic facial features, intellectual disability, autistic behaviors and epilepsy. Clin. Genet. *87*, 356–361.

Napoli, I., Mercaldo, V., Boyl, P.P., Eleuteri, B., Zalfa, F., De Rubeis, S., Di Marino, D., Mohr, E., Massimi, M., Falconi, M., et al. (2008). The fragile X syndrome protein represses activity-dependent translation through CYFIP1, a new 4E-BP. Cell *134*, 1042–1054.

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242–245.

Need, A.C., Shashi, V., Hitomi, Y., Schoch, K., Shianna, K.V., McDonald, M.T., Meisler, M.H., and Goldstein, D.B. (2012). Clinical application of exome sequencing in undiagnosed genetic conditions. J. Med. Genet. *49*, 353–361.

Nesbitt, A., Bhoj, E.J., McDonald Gibson, K., Yu, Z., Denenberg, E., Sarmady, M., Tischler, T., Cao, K., Dubbs, H., Zackai, E.H., et al. (2015). Exome sequencing expands the mechanism of SOX5-associated intellectual disability: A case presentation with review of sox-related disorders. Am. J. Med. Genet. A. *167A*, 2548–2554.

Neves-Pereira, M., Müller, B., Massie, D., Williams, J.H.G., O'Brien, P.C.M., Hughes, A., Shen, S.-B., Clair, D.S., and Miedzybrodzka, Z. (2009). Deregulation of EIF4E: a novel mechanism for autism. J. Med. Genet. *46*, 759–765.

Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.

Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010a). Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. *42*, 30–35.

Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., et al. (2010b). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat. Genet. *42*, 790–793.

Nguyen, L.S., Kim, H.-G., Rosenfeld, J.A., Shen, Y., Gusella, J.F., Lacassie, Y., Layman, L.C., Shaffer, L.G., and Gécz, J. (2013). Contribution of copy number variants involving nonsense-mediated mRNA decay pathway genes to neuro-developmental disorders. Hum. Mol. Genet. *22*, 1816–1825.

Notwell, J.H., Heavner, W.E., Darbandi, S.F., Katzman, S., McKenna, W.L., Ortiz-Londono, C.F., Tastad, D., Eckler, M.J., Rubenstein, J.L.R., McConnell, S.K., et al. (2016). TBR1 regulates autism risk genes in the developing neocortex. Genome Res. *26*, 1013–1022.

Nousbeck, J., Spiegel, R., Ishida-Yamamoto, A., Indelman, M., Shani-Adir, A., Adir, N., Lipkin, E., Bercovici, S., Geiger, D., van Steensel, M.A., et al. (2008). Alopecia, neurological defects, and endocrinopathy syndrome caused by decreased expression of RBM28, a nucleolar protein associated with ribosome biogenesis. Am. J. Hum. Genet. *82*, 1114–1121.

Oberlé, I., Vincent, A., Abbadi, N., Rousseau, F., Hupkes, P.E., Hors-Cayla, M.C., Gilgenkrantz, S., Oostra, B.A., and Mandel, J.L. (1991). New polymorphism and a new chromosome breakpoint establish the physical and genetic mapping of DXS369 in the DXS98-FRAXA interval. Am. J. Med. Genet. *38*, 336–342.

Oksenberg, N., Haliburton, G.D.E., Eckalbar, W.L., Oren, I., Nishizaki, S., Murphy, K., Pollard, K.S., Birnbaum, R.Y., and Ahituv, N. (2014). Genome-wide distribution of Auts2 binding localizes with active neurodevelopmental genes. Transl. Psychiatry *4*, e431.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246–250.

O'Roak, B.J., Stessman, H.A., Boyle, E.A., Witherspoon, K.T., Martin, B., Lee, C., Vives, L., Baker, C., Hiatt, J.B., Nickerson, D.A., et al. (2014). Recurrent de novo mutations implicate novel genes underlying simplex autism risk. Nat. Commun. *5*, 5595.

Ostergaard, P., Simpson, M.A., Mendola, A., Vasudevan, P., Connell, F.C., van Impel, A., Moore, A.T., Loeys, B.L., Ghalamkarpour, A., Onoufriadis, A., et al. (2012). Mutations in KIF11 cause autosomal-dominant microcephaly variably associated with congenital lymphedema and chorioretinopathy. Am. J. Hum. Genet. *90*, 356–362.

Pak, C., Garshasbi, M., Kahrizi, K., Gross, C., Apponi, L.H., Noto, J.J., Kelly, S.M., Leung, S.W., Tzschach, A., Behjati, F., et al. (2011). Mutation of the conserved polyadenosine RNA binding protein, ZC3H14/dNab2, impairs neural function in Drosophila and humans. Proc. Natl. Acad. Sci. U. S. A. *108*, 12390–12395.

Palmer, A., Qayumi, J., and Ronnett, G. (2008). MeCP2 mutation causes distinguishable phases of acute and chronic defects in synaptogenesis and maintenance, respectively. Mol. Cell. Neurosci. *37*, 794–807.

Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. Am. J. Hum. Genet. *102*, 11–26.

Pertea, M., Lin, X., and Salzberg, S.L. (2001). GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res. *29*, 1185–1190.

Petterson, B., Bourke, J., Leonard, H., Jacoby, P., and Bower, C. (2007). Co-occurrence of birth defects and intellectual disability. Paediatr. Perinat. Epidemiol. *21*, 65–75.

Philippakis, A.A., Azzariti, D.R., Beltran, S., Brookes, A.J., Brownstein, C.A., Brudno, M., Brunner, H.G., Buske, O.J., Carey, K., Doll, C., et al. (2015). The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. Hum. Mutat. *36*, 915–921.

Pober, B.R. (2010). Williams-Beuren syndrome. N. Engl. J. Med. *362*, 239–252.

Poirier, K., Lebrun, N., Broix, L., Tian, G., Saillour, Y., Boscheron, C., Parrini, E., Valence, S., Pierre, B.S., Oger, M., et al. (2013). Mutations in TUBG1, DYNC1H1, KIF5C and KIF2A cause malformations of cortical development and microcephaly. Nat. Genet. *45*, 639–647.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

Popp, B., Støve, S.I., Endele, S., Myklebust, L.M., Hoyer, J., Sticht, H., Azzarello-Burri, S., Rauch, A., Arnesen, T., and Reis, A. (2015). De novo missense mutations in the NAA10 gene cause severe non-syndromic developmental delay in males and females. Eur. J. Hum. Genet. EJHG *23*, 602–609.

Puffenberger, E.G., Jinks, R.N., Sougnez, C., Cibulskis, K., Willert, R.A., Achilly, N.P., Cassidy, R.P., Fiorentini, C.J., Heiken, K.F., Lawrence, J.J., et al. (2012). Genetic mapping and exome sequencing identify variants associated with five novel diseases. PloS One *7*, e28936.

Quartier, A., Chatrousse, L., Redin, C., Keime, C., Haumesser, N., Maglott-Roth, A., Brino, L., Le Gras, S., Benchoua, A., Mandel, J.-L., et al. (2018). Genes and Pathways Regulated by Androgens in Human Neural Cells, Potential Candidates for the Male Excess in Autism Spectrum Disorder. Biol. Psychiatry.

Rauch, A., Hoyer, J., Guth, S., Zweier, C., Kraus, C., Becker, C., Zenker, M., Hüffmeier, U., Thiel, C., Rüschendorf, F., et al. (2006). Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. Am. J. Med. Genet. A. *140*, 2063–2074.

Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet Lond. Engl. *380*, 1674–1682.

Redin, C., Gérard, B., Lauer, J., Herenger, Y., Muller, J., Quartier, A., Masurel-Paulet, A., Willems, M., Lesca, G., El-Chehadeh, S., et al. (2014). Efficient strategy for the molecular diagnosis of intellectual disability using targeted high-throughput sequencing. J. Med. Genet. *51*, 724–736.

Redin, C., Brand, H., Collins, R.L., Kammin, T., Mitchell, E., Hodge, J.C., Hanscom, C., Pillalamarri, V., Seabra, C.M., Abbott, M.-A., et al. (2017). The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. Nat. Genet. *49*, 36.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.

Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. J. Comput. Biol. J. Comput. Mol. Cell Biol. *4*, 311–323.

Regan, R., and Willatt, L. (2010). Mental Retardation: Definition, Classification and Etiology. *18*, 16–30.

Riazuddin, S., Hussain, M., Razzaq, A., Iqbal, Z., Shahzad, M., Polla, D.L., Song, Y., van Beusekom, E., Khan, A.A., Tomas-Roca, L., et al. (2016). Exome sequencing of Pakistani consanguineous families identifies 30 novel candidate genes for recessive intellectual disability. Mol. Psychiatry.

Rice, G., Patrick, T., Parmar, R., Taylor, C.F., Aeby, A., Aicardi, J., Artuch, R., Montalto, S.A., Bacino, C.A., Barroso, B., et al. (2007). Clinical and molecular phenotype of Aicardi-Goutieres syndrome. Am. J. Hum. Genet. *81*, 713–725.

Rice, G.I., Bond, J., Asipu, A., Brunette, R.L., Manfield, I.W., Carr, I.M., Fuller, J.C., Jackson, R.M., Lamb, T., Briggs, T.A., et al. (2009). Mutations involved in Aicardi-Goutières syndrome implicate SAMHD1 as regulator of the innate immune response. Nat. Genet. *41*, 829–832.

Rice, G.I., Kasher, P.R., Forte, G.M.A., Mannion, N.M., Greenwood, S.M., Szynkiewicz, M., Dickerson, J.E., Bhaskar, S.S., Zampini, M., Briggs, T.A., et al. (2012). Mutations in ADAR1 cause Aicardi-Goutières syndrome associated with a type I interferon signature. Nat. Genet. *44*, 1243–1248.

Ronan, J.L., Wu, W., and Crabtree, G.R. (2013). From neural development to cognition: unexpected roles for chromatin. Nat. Rev. Genet. *14*, 347–359.

Ronemus, M., Iossifov, I., Levy, D., and Wigler, M. (2014). The role of de novo mutations in the genetics of autism spectrum disorders. Nat. Rev. Genet. *15*, 133–141.

Ropers, H.H. (2010). Genetics of early onset cognitive impairment. Annu. Rev. Genomics Hum. Genet. *11*, 161–187.

Rutherford, S.L. (2000). From genotype to phenotype: buffering mechanisms and the storage of genetic information. BioEssays News Rev. Mol. Cell. Dev. Biol. *22*, 1095–1105.

Saito, Y., Miranda-Rottmann, S., Ruggiu, M., Park, C.Y., Fak, J.J., Zhong, R., Duncan, J.S., Fabella, B.A., Junge, H.J., Chen, Z., et al. (2016). NOVA2-mediated RNA regulation is required for axonal pathfinding during development. ELife *5*.

Sanchis, A., Cerveró, L., Bataller, A., Tortajada, J.L., Huguet, J., Crow, Y.J., Ali, M., Higuet, L.J., and Martínez-Frías, M.L. (2005). Genetic syndromes mimic congenital infections. J. Pediatr. *146*, 701–705.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241.

Sarachana, T., Zhou, R., Chen, G., Manji, H.K., and Hu, V.W. (2010). Investigation of post-transcriptional gene regulatory networks associated with autism spectrum disorders by microRNA expression profiling of lymphoblastoid cell lines. Genome Med. *2*, 23.

Sartor, F., Anderson, J., McCaig, C., Miedzybrodzka, Z., and Müller, B. (2015). Mutation of genes controlling mRNA metabolism and protein synthesis predisposes to neurodevelopmental disorders. Biochem. Soc. Trans. *43*, 1259–1265.

Schaaf, C.P., Sabo, A., Sakai, Y., Crosby, J., Muzny, D., Hawes, A., Lewis, L., Akbar, H., Varghese, R., Boerwinkle, E., et al. (2011). Oligogenic heterozygosity in individuals with high-functioning autism spectrum disorders. Hum. Mol. Genet. *20*, 3366–3375.

Schenck, A., Bardoni, B., Langmann, C., Harden, N., Mandel, J.L., and Giangrande, A. (2003). CYFIP/Sra-1 controls neuronal connectivity in Drosophila and links the Rac1 GTPase pathway to the fragile X protein. Neuron *38*, 887–898.

Schmitt, A.D., Hu, M., and Ren, B. (2016). Genome-wide mapping and analysis of chromosome architecture. Nat. Rev. Mol. Cell Biol. *17*, 743–755.

Schuurs-Hoeijmakers, J.H.M., Vulto-van Silfhout, A.T., Vissers, L.E.L.M., van de Vondervoort, I.I.G.M., van Bon, B.W.M., de Ligt, J., Gilissen, C., Hehir-Kwa, J.Y., Neveling, K., del Rosario, M., et al. (2013). Identification of pathogenic gene variants in small families with intellectually disabled siblings by exome sequencing. J. Med. Genet. *50*, 802–811.

Schwarz, J.M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods *7*, 575–576.

Scott, D.A., Hernandez-Garcia, A., Azamian, M.S., Jordan, V.K., Kim, B.J., Starkovich, M., Zhang, J., Wong, L.-J., Darilek, S.A., Breman, A.M., et al. (2017). Congenital heart defects and left ventricular non-compaction in males with loss-of-function variants in NONO. J. Med. Genet. *54*, 47–53.

Sellier, C., Buijsen, R.A.M., He, F., Natla, S., Jung, L., Tropel, P., Gaucherot, A., Jacobs, H., Meziane, H., Vincent, A., et al. (2017). Translation of Expanded CGG Repeats into FMRpolyG Is Pathogenic and May Contribute to Fragile X Tremor Ataxia Syndrome. Neuron *93*, 331–347.

Shaheen, R., Han, L., Faqeih, E., Ewida, N., Alobeid, E., Phizicky, E.M., and Alkuraya, F.S. (2016). A homozygous truncating mutation in PUS3 expands the role of tRNA modification in normal cognition. Hum. Genet. *135*, 707–713.

Shashi, V., Xie, P., Schoch, K., Goldstein, D.B., Howard, T.D., Berry, M.N., Schwartz, C.E., Cronin, K., Sliwa, S., Allen, A., et al. (2015). The RBMX gene as a candidate for the Shashi X-linked intellectual disability syndrome. Clin. Genet. *88*, 386–390.

Shen, S., Park, J.W., Lu, Z., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. U. S. A. *111*, E5593-5601.

Short, P.J., McRae, J.F., Gallone, G., Sifrim, A., Won, H., Geschwind, D.H., Wright, C.F., Firth, H.V., FitzPatrick, D.R., Barrett, J.C., et al. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. Nature.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

Simons, C., Griffin, L.B., Helman, G., Golas, G., Pizzino, A., Bloom, M., Murphy, J.L.P., Crawford, J., Evans, S.H., Topper, S., et al. (2015). Loss-of-function alanyl-tRNA synthetase mutations cause an autosomal-recessive early-onset epileptic encephalopathy with persistent myelination defect. Am. J. Hum. Genet. *96*, 675–681.

Skraban, C.M., Wells, C.F., Markose, P., Cho, M.T., Nesbitt, A.I., Au, P.Y.B., Begtrup, A., Bernat, J.A., Bird, L.M., Cao, K., et al. (2017). WDR26 Haploinsufficiency Causes a Recognizable Syndrome of Intellectual Disability, Seizures, Abnormal Gait, and Distinctive Facial Features. Am. J. Hum. Genet. *101*, 139–148.

Snijders Blok, L., Madsen, E., Juusola, J., Gilissen, C., Baralle, D., Reijnders, M.R.F., Venselaar, H., Helsmoortel, C., Cho, M.T., Hoischen, A., et al. (2015). Mutations in DDX3X Are a Common Cause of Unexplained Intellectual Disability with Gender-Specific Effects on Wnt Signaling. Am. J. Hum. Genet. *97*, 343–352.

Snijders Blok, L., Hiatt, S.M., Bowling, K.M., Prokop, J.W., Engel, K.L., Cochran, J.N., Bebin, E.M., Bijlsma, E.K., Ruivenkamp, C.A.L., Terhal, P., et al. (2018). De novo mutations in MED13, a component of the Mediator complex, are associated with a novel neurodevelopmental disorder. Hum. Genet.

Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene. Hum. Mutat. *36*, 928–930.

Sobreira, N.L.M., Cirulli, E.T., Avramopoulos, D., Wohler, E., Oswald, G.L., Stevens, E.L., Ge, D., Shianna, K.V., Smith, J.P., Maia, J.M., et al. (2010). Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. PLoS Genet. *6*, e1000991.

Soden, S.E., Saunders, C.J., Willig, L.K., Farrow, E.G., Smith, L.D., Petrikin, J.E., LePichon, J.-B., Miller, N.A., Thiffault, I., Dinwiddie, D.L., et al. (2014). Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. Sci. Transl. Med. *6*, 265ra168.

Srivastava, A.K., and Schwartz, C.E. (2014). Intellectual disability and autism spectrum disorders: causal genes and molecular mechanisms. Neurosci. Biobehav. Rev. *46 Pt 2*, 161–174.

Srivastava, A., McGrath, B., and Bielas, S.L. (2017). Histone H2A Monoubiquitination in Neurodevelopmental Disorders. Trends Genet. TIG *33*, 566–578.

Stadhouders, R. (2018). Expanding the toolbox for 3D genomics. Nat. Genet. *50*, 634–635.

Stessman, H.A.F., Xiong, B., Coe, B.P., Wang, T., Hoekzema, K., Fenckova, M., Kvarnung, M., Gerdts, J., Trinh, S., Cosemans, N., et al. (2017). Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. Nat. Genet. *49*, 515–526.

Stettner, G.M., Shoukier, M., Höger, C., Brockmann, K., and Auber, B. (2011). Familial intellectual disability and autistic behavior caused by a small FMR2 gene deletion. Am. J. Med. Genet. A. *155A*, 2003–2007.

Straub, J., Konrad, E.D.H., Grüner, J., Toutain, A., Bok, L.A., Cho, M.T., Crawford, H.P., Dubbs, H., Douglas, G., Jobling, R., et al. (2018). Missense Variants in RHOBTB2 Cause a Developmental and Epileptic Encephalopathy in Humans, and Altered Levels Cause Neurological Defects in Drosophila. Am. J. Hum. Genet. *102*, 44–57.

Sultana, R., Yu, C.-E., Yu, J., Munson, J., Chen, D., Hua, W., Estes, A., Cortes, F., de la Barra, F., Yu, D., et al. (2002). Identification of a novel gene on chromosome 7q11.2 interrupted by a translocation breakpoint in a pair of autistic twins. Genomics *80*, 129–134.

Sun, E., and Shi, Y. (2015). MicroRNAs: Small molecules with big roles in neurodevelopment and diseases. Exp. Neurol. *268*, 46–53.

Sun, Y., Ruivenkamp, C.A.L., Hoffer, M.J.V., Vrijenhoek, T., Kriek, M., van Asperen, C.J., den Dunnen, J.T., and Santen, G.W.E. (2015). Next-generation diagnostics: gene panel, exome, or whole genome? Hum. Mutat. *36*, 648–655.

Tabet, R., Moutin, E., Becker, J.A.J., Heintz, D., Fouillen, L., Flatter, E., Krężel, W., Alunni, V., Koebel, P., Dembélé, D., et al. (2016). Fragile X Mental Retardation Protein (FMRP) controls diacylglycerol kinase activity in neurons. Proc. Natl. Acad. Sci. U. S. A. *113*, E3619-3628.

Taft, R.J., Vanderver, A., Leventer, R.J., Damiani, S.A., Simons, C., Grimmond, S.M., Miller, D., Schmidt, J., Lockhart, P.J., Pope, K., et al. (2013). Mutations in DARS cause hypomyelination with brain stem and spinal cord involvement and leg spasticity. Am. J. Hum. Genet. *92*, 774–780.

Talebizadeh, Z., Butler, M.G., and Theodoro, M.F. (2008). Feasibility and relevance of examining lymphoblastoid cell lines to study role of microRNAs in autism. Autism Res. Off. J. Int. Soc. Autism Res. *1*, 240–250.

Tammimies, K., Marshall, C.R., Walker, S., Kaur, G., Thiruvahindrapuram, B., Lionel, A.C., Yuen, R.K.C., Uddin, M., Roberts, W., Weksberg, R., et al. (2015). Molecular Diagnostic Yield of Chromosomal Microarray Analysis and Whole-Exome Sequencing in Children With Autism Spectrum Disorder. JAMA *314*, 895–903.

Tan, C.A., Topper, S., Del Gaudio, D., Nelakuditi, V., Shchelochkov, O., Nowaczyk, M.J.M., Zeesman, S., Brady, L., Russell, L., Meeks, N., et al. (2015). Characterization of patients referred for non-specific intellectual disability testing: the importance of autosomal genes for diagnosis. Clin. Genet.

Tarpey, P.S., Raymond, F.L., Nguyen, L.S., Rodriguez, J., Hackett, A., Vandeleur, L., Smith, R., Shoubridge, C., Edkins, S., Stevens, C., et al. (2007). Mutations in UPF3B, a member of the nonsense-mediated mRNA decay complex, cause syndromic and nonsyndromic mental retardation. Nat. Genet. *39*, 1127–1133.

Tarpey, P.S., Smith, R., Pleasance, E., Whibley, A., Edkins, S., Hardy, C., O'Meara, S., Latimer, C., Dicks, E., Menzies, A., et al. (2009). A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nat. Genet. *41*, 535–543.

Tatton-Brown, K., and Rahman, N. (2013). The NSD1 and EZH2 overgrowth genes, similarities and differences. Am. J. Med. Genet. C Semin. Med. Genet. *163C*, 86–91.

Tatton-Brown, K., Seal, S., Ruark, E., Harmer, J., Ramsay, E., Del Vecchio Duarte, S., Zachariou, A., Hanks, S., O'Brien, E., Aksglaede, L., et al. (2014). Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability. Nat. Genet. *46*, 385–388.

Ten Kate, L.P., Teeuw, M., Henneman, L., and Cornel, M.C. (2010). Autosomal recessive disease in children of consanguineous parents: inferences from the proportion of compound heterozygotes. J. Community Genet. *1*, 37–40.

Tzschach, A., Grasshoff, U., Beck-Woedl, S., Dufke, C., Bauer, C., Kehrer, M., Evers, C., Moog, U., Oehl-Jaschkowitz, B., Di Donato, N., et al. (2015). Next-generation sequencing in X-linked intellectual disability. Eur. J. Hum. Genet. EJHG *23*, 1513–1518.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B.J., and Darnell, R.B. (2006). An RNA map predicting Nova-dependent splicing regulation. Nature *444*, 580–586.

Van Esch, H., Bauters, M., Ignatius, J., Jansen, M., Raynaud, M., Hollanders, K., Lugtenberg, D., Bienvenu, T., Jensen, L.R., Gecz, J., et al. (2005). Duplication of the MECP2 region is a frequent cause of severe mental retardation and progressive neurological symptoms in males. Am. J. Hum. Genet. *77*, 442–453.

Vasileiou, G., Vergarajauregui, S., Endele, S., Popp, B., Büttner, C., Ekici, A.B., Gerard, M., Bramswig, N.C., Albrecht, B., Clayton-Smith, J., et al. (2018). Mutations in the BAF-Complex Subunit DPF2 Are Associated with Coffin-Siris Syndrome. Am. J. Hum. Genet. *102*, 468–479.

Vissers, L.E.L.M., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M., et al. (2010). A de novo paradigm for mental retardation. Nat. Genet. *42*, 1109–1112.

Vissers, L.E.L.M., Gilissen, C., and Veltman, J.A. (2016). Genetic studies in intellectual disability and related disorders. Nat. Rev. Genet. *17*, 9–18.

van der Voet, M., Nijhof, B., Oortveld, M.A.W., and Schenck, A. (2014). Drosophila models of early onset cognitive disorders and their clinical applications. Neurosci. Biobehav. Rev. *46 Pt 2*, 326–342.

Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., Tian, Y., Horvath, S., Mill, J., Cantor, R.M., Blencowe, B.J., and Geschwind, D.H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature *474*, 380–384.

de Vries, B.B.A., Pfundt, R., Leisink, M., Koolen, D.A., Vissers, L.E.L.M., Janssen, I.M., Reijmersdal, S. van, Nillesen, W.M., Huys, E.H.L.P.G., Leeuw, N. de, et al. (2005). Diagnostic genome profiling in mental retardation. Am. J. Hum. Genet. *77*, 606–616.

Wahle, E., and Winkler, G.S. (2013). RNA decay machines: deadenylation by the Ccr4-not and Pan2-Pan3 complexes. Biochim. Biophys. Acta *1829*, 561–570.

Wan, J., Yourshaw, M., Mamsa, H., Rudnik-Schöneborn, S., Menezes, M.P., Hong, J.E., Leong, D.W., Senderek, J., Salman, M.S., Chitayat, D., et al. (2012). Mutations in the RNA exosome component gene EXOSC3 cause pontocerebellar hypoplasia and spinal motor neuron degeneration. Nat. Genet. *44*, 704–708.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wickramasinghe, V.O., and Laskey, R.A. (2015). Control of mammalian gene expression by selective mRNA export. Nat. Rev. Mol. Cell Biol. *16*, 431–442.

Will, C.L., and Lührmann, R. (2011). Spliceosome structure and function. Cold Spring Harb. Perspect. Biol. *3*.

Willemsen, M.H., Ba, W., Wissink-Lindhout, W.M., de Brouwer, A.P.M., Haas, S.A., Bienek, M., Hu, H., Vissers, L.E.L.M., van Bokhoven, H., Kalscheuer, V., et al. (2014). Involvement of the kinesin family members KIF4A and KIF5C in intellectual disability and synaptic function. J. Med. Genet. *51*, 487–494.

Wilson, N.H., and Key, B. (2007). Neogenin: one receptor, many functions. Int. J. Biochem. Cell Biol. *39*, 874–878.

Windpassinger, C., Piard, J., Bonnard, C., Alfadhel, M., Lim, S., Bisteau, X., Blouin, S., Ali, N.B., Ng, A.Y.J., Lu, H., et al. (2017). CDK10 Mutations in Humans and Mice Cause Severe Growth Retardation, Spine Malformations, and Developmental Delays. Am. J. Hum. Genet. *101*, 391–403.

Winkelmann, J., Lin, L., Schormair, B., Kornum, B.R., Faraco, J., Plazzi, G., Melberg, A., Cornelio, F., Urban, A.E., Pizza, F., et al. (2012). Mutations in DNMT1 cause autosomal dominant cerebellar ataxia, deafness and narcolepsy. Hum. Mol. Genet. *21*, 2205–2210.

Winter, J., Jung, S., Keller, S., Gregory, R.I., and Diederichs, S. (2009). Many roads to maturity: microRNA biogenesis pathways and their regulation. Nat. Cell Biol. *11*, 228–234.

Xu, G.L., Bestor, T.H., Bourc'his, D., Hsieh, C.L., Tommerup, N., Bugge, M., Hulten, M., Qu, X., Russo, J.J., and Viegas-Péquignot, E. (1999). Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. Nature *402*, 187–191.

Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N. Engl. J. Med. *369*, 1502–1511.

Yang, Y.Y., Yin, G.L., and Darnell, R.B. (1998). The neuronal RNA-binding protein Nova-2 is implicated as the autoantigen targeted in POMA patients with dementia. Proc. Natl. Acad. Sci. U. S. A. *95*, 13254–13259.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. J. Comput. Mol. Cell Biol. *11*, 377–394.

Ylikallio, E., Woldegebriel, R., Tumiati, M., Isohanni, P., Ryan, M.M., Stark, Z., Walsh, M., Sawyer, S.L., Bell, K.M., Oshlack, A., et al. (2017). MCM3AP in recessive Charcot-Marie-Tooth neuropathy and mild intellectual disability. Brain J. Neurol. *140*, 2093–2103.

Zeev, B.B., Yaron, Y., Schanen, N.C., Wolf, H., Brandt, N., Ginot, N., Shomrat, R., and Orr-Urtreger, A. (2002). Rett syndrome: clinical manifestations in males with MECP2 mutations. J. Child Neurol. *17*, 20–24.

Zhang, P., Casaday-Potts, R., Precht, P., Jiang, H., Liu, Y., Pazin, M.J., and Mattson, M.P. (2011). Nontelomeric splice variant of telomere repeat-binding factor 2 maintains neuronal traits by sequestering repressor element 1-silencing transcription factor. Proc. Natl. Acad. Sci. U. S. A. *108*, 16434–16439.

Zhang, Y., Chen, K., Sloan, S.A., Bennett, M.L., Scholze, A.R., O'Keeffe, S., Phatnani, H.P., Guarnieri, P., Caneda, C., Ruderisch, N., et al. (2014). An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. J. Neurosci. Off. J. Soc. Neurosci. *34*, 11929–11947.

Zheng, G., Dahl, J.A., Niu, Y., Fu, Y., Klungland, A., Yang, Y.-G., and He, C. (2013). Sprouts of RNA epigenetics: the discovery of mammalian RNA demethylases. RNA Biol. *10*, 915–918.

Zheng, X., Yang, P., Lackford, B., Bennett, B.D., Wang, L., Li, H., Wang, Y., Miao, Y., Foley, J.F., Fargo, D.C., et al. (2016). CNOT3-Dependent mRNA Deadenylation Safeguards the Pluripotent State. Stem Cell Rep. *7*, 897–910.

Ziats, M.N., and Rennert, O.M. (2016). The Evolving Diagnostic and Genetic Landscapes of Autism Spectrum Disorder. Front. Genet. *7*, 65.

Zufferey, F., Sherr, E.H., Beckmann, N.D., Hanson, E., Maillard, A.M., Hippolyte, L., Macé, A., Ferrari, C., Kutalik, Z., Andrieux, J., et al. (2012). A 600 kb deletion syndrome at 16p11.2 leads to energy imbalance and neuropsychiatric disorders. J. Med. Genet. *49*, 660–668.

Zweier, M., Gregor, A., Zweier, C., Engels, H., Sticht, H., Wohlleber, E., Bijlsma, E.K., Holder, S.E., Zenker, M., Rossier, E., et al. (2010). Mutations in MEF2C from the 5q14.3q15 microdeletion syndrome region are a frequent cause of severe mental retardation and diminish MECP2 and CDKL5 expression. Hum. Mutat. *31*, 722–733.

APPENDIX 1:
Novel de novo mutations in *ZBTB20* in Primrose syndrome with congenital hypothyroidism

We identified 2 *de novo* missense mutations that have never been reported in *ZBTB20* – a gene recently implicated in Primrose syndrome (*Cordeddu et al. 2014*) – in a patient whose phenotype well resembles to the one described. These two variants are close to the previously identified mutations – which are located in the zinc-finger and linker region –, confirming the implication of the gene in this syndromic form of ID and the important role of this region of the protein in the syndrome etiology. Reverse phenotyping showed that this patient presents with classic features of Primrose syndrome (dysmorphic facies, macrocephaly, hearing loss, hypotonia, hypoplasia of the corpus callosum) and, in addition, congenital hypothyroidism. Review of the literature reveals another Primrose syndrome patient with hypothyroidism and thus this may represent an under recognized component that should be investigated in other patients.

CLINICAL REPORT

AMERICAN JOURNAL OF medical genetics PART A

Novel De Novo Mutations in *ZBTB20* in Primrose Syndrome With Congenital Hypothyroidism

Francesca Mattioli,[1] Amelie Piton,[1,2*] Bénédicte Gérard,[2] Andrea Superti-Furga,[3] Jean-Louis Mandel,[1,2,4] and Sheila Unger[5**]

[1]Department of Translational Medicine and Neurogenetics, IGBMC, Illkirch, France

[2]Laboratoire de Diagnostic Génétique, Hôpitaux Universitaires de Strasbourg, Strasbourg, France

[3]Centre for Molecular Diseases and Department of Pediatrics, Lausanne University Hospital (CHUV), University of Lausanne, Lausanne, Switzerland

[4]Chaire de Génétique Humaine, Collège de France, Paris, France

[5]Service of Medical Genetics, Lausanne University Hospital (CHUV), Lausanne, Switzerland

Manuscript Received: 17 December 2015; Manuscript Accepted: 16 March 2016

The cardinal features of Primrose syndrome (MIM 259050) are dysmorphic facial features, macrocephaly, and intellectual disability, as well as large body size, height and weight, and calcified pinnae. A variety of neurological signs and symptoms have been reported including hearing loss, autism, behavioral abormalities, hypotonia, cerebral calcifications, and hypoplasia of the corpus callosum. Recently, heterozygous de novo missense mutations in *ZBTB20*, coding for a zing finger protein, have been identified in Primrose syndrome patients. We report a boy with intellectual disability carrying two de novo missense mutations in the last exon of *ZBTB20* (Ser616Phe and Gly741Arg; both previously unreported). One of them, Ser616Phe, affects an amino acid located in one of the C2H2 zing-fingers involved in DNA-binding and close to other missense mutations already described. Reverse phenotyping showed that this patient presents with classic features of Primrose syndrome (dysmorphic facies, macrocephaly, hearing loss, hypotonia, hypoplasia of the corpus callosum) and, in addition, congenital hypothyroidism. Review of the literature reveals another Primrose syndrome patient with hypothyroidism and thus, this may represent an under recognized component that should be investigated in other patients. © 2016 Wiley Periodicals, Inc.

Key words: Primrose syndrome; congenital hypothyroidism; *ZBTB20*; macrocephaly

How to Cite this Article:
Mattioli F, Piton A, Gérard B, Superti-Furga A, Mandel J-L, Unger S. 2016. Novel de novo mutations in *ZBTB20* in Primrose syndrome with congenital hypothyroidism. Am J Med Genet Part A 9999A:1–4.

INTRODUCTION

In 1982, Dr. D.A. Primrose reported a man with severe intellectual disabilities, calcified pinnae, and muscle wasting. The patient also had cataracts and probably torus palatinus (« a hard mass filling in the cavity of the hard palate ») [Primrose, 1982]. The second report also described an adult male with intellectual disabilites, calcified pinnae, cataracts, and palatal mass [Collacott et al., 1986]. This second patient is noted as having a head circumference on the 98th percentile and thus relative macrocephaly [Collacott et al., 1986]. Primrose syndrome has subsequently been reported in a handful of individuals, mostly diagnosed in adulthood [Lindor et al., 1996; Battisti et al., 2002; Mathijssen et al., 2006; Dalal et al., 2010; Posmyk et al., 2011]. The large calcified external ears have been labelled a « telltale sign » but it is possible that they calcify only in adulthood and thus may not be a reliable pediatric indicator and their absence should not preclude the diagnosis [Dalal et al., 2010]. Later case reports have stressed the differential diagnosis with overgrowth syndromes [Cordeddu et al., 2014]. Whole exome sequencing and candidate gene sequencing, performed on patients with Primrose syndrome, identified eight heterozygote de novo missense mutations in the gene encoding the transcription factor ZBTB20.

Conflict of interest: none.
Grant sponsor: Agence de Biomedicine, Fondation Jerome Lejeune, CREGEMES.
*Correspondence to:
Amelie Piton, Department of Translational Medicine and Neurogenetics, IGBMC, Illkirch, France.
E-mail: piton@igbmc.fr
**Correspondence to:
Sheila Unger, Service of Medical Genetics, Lausanne University Hospital (CHUV), Lausanne 1011, Switzerland.
E-mail: sheila.unger@chuv.ch
Article first published online in Wiley Online Library (wileyonlinelibrary.com): 00 Month 2016
DOI 10.1002/ajmg.a.37645

© 2016 Wiley Periodicals, Inc.

ZBTB20 is one of the genes implicated in the 3q13.31 microdeletion syndrome that has important phenotypic overlap with Primrose syndrome (developmental delay, muscular hypotonia, postnatal overgrowth, and other features). All of the missense mutations reported in Primrose syndrome are located in the first two C2H2 zing fingers domains at the C-terminal part of the protein, region involved in DNA-binding. We report here a new *ZBTB20* mutation, identified during a targeted sequencing approach of candidate genes ina cohort of individuals with mild to severe intellectual disability, in a patient with features of Primrose syndrome.

CLINICAL REPORT

The proband is the second child of a healthy non-consanguineous couple. His older brother is healthy, developing normally, and has a normal body size. The proband was born following an uneventful pregnancy with a birthweight of 4.3 kg (95th centile), birth length of 54 cm (75th centile), and head circumference of 34 cm (10th centile). The perinatal period was remarkable for the discovery of congenital hypothyroidism after a newborn screening test revealed a TSH of 32 mU/L (normal range: 0.2–3.5 mU/L). Follow-up testing revealed a TSH of 162 mU/L, and a free T4 <5 pmol/L (normal range: 9–19 pmol/L). A Tc-99m (pertechnetate) thyroid scan showed a large volume thyroid gland with homogeneous uptake suggesting a problem in thyroid hormone synthesis. Hormone replacement was initiated and values normalized. He was also found to have bilateral cryptorchidism that was surgically repaired at age 2 years.

Newborn screening tests for audition were also abnormal. Formal testing detected a 70 dB loss on the right and a 50–60 dB loss on the left for middle and high frequencies. He was treated with hearing aids. A cranial MRI was performed at 1 year of age and showed partial agenesis of the corpus callosum and wide pericerebral spaces but normal cochlea and no anatomical explanation for the neurosensory hearing loss. An EEG showed non-specific slowing.

He was first seen in genetics at 14 months of age for overgrowth and hypotonia. His growth parameters at that age were: weight 13.5 kg (97th centile), length 85 cm (>97th centile), and head circumference 53.5 cm (>97th centile). He had a prominent forehead, right sided convergent strabismus, ptosis, broad face, large ears, and a depressed nasal bridge (Fig. 1). The pinnae were supple to palpation with no evidence of calcification. He had axial and peripheral hypotonia. The proband was also followed in neurodevelopmental clinic for significant global developmental delay and behavioral problems. Despite intensive physiotherapy, he was unable to stand unaided at age 4 years and was unstable in sitting position. He did not speak. He had low tolerance for frustration and exhibited recurrent temper tantrums in which he would cry and throw himself backwards. These symptoms improved partially with treatment by risperidone (selective monoaminergic antagonist). He also presented self-injuries (biting).

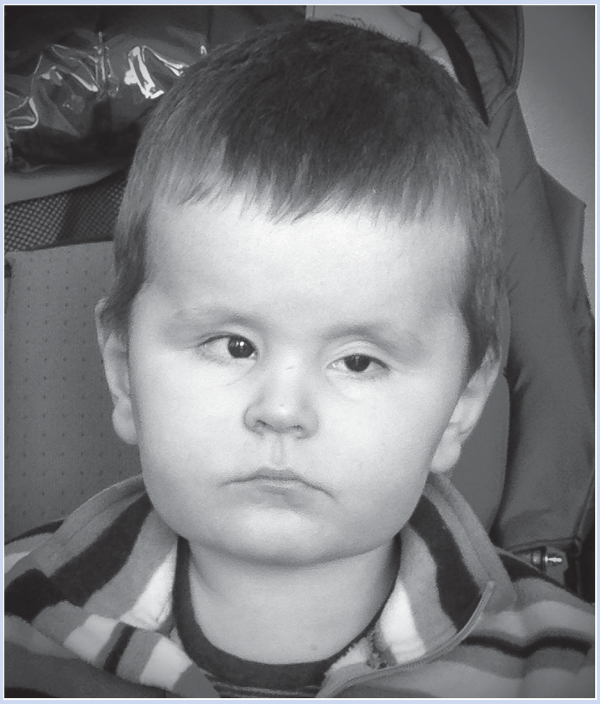


FIG. 1. Clinical photo of the proband. He has a broad prominent forehead, small deep set eyes with ptosis (more marked on the left), and epicanthal folds. The mouth is small with downturned corners but the jaw is large. The ears are also large.

MOLECULAR INVESTIGATIONS

An array-CGH (Agilent oligoNT 180 K) was performed and did not reveal any pathogenic CNV. A targeted exome sequencing was then considered. A sequencing library was prepared from the patient's DNA as previously described [Redin et al., 2014], including a capture enrichment reaction with specific baits corresponding to the coding regions of 275 genes certainly or potentially involved in intellectual disability (SureSelect, Agilent, Santa Clara, CA). Pair-end sequencing (2 × 101-bp on Illumina HiSeq2500 sequencer), read mapping, variant calling, annotation, filtration, and ranking were performed as previously described [Redin et al., 2014; Geoffroy et al., 2015]. Targeted sequencing led to the identification of two new non-synonymous substitutions in the coding sequence of *ZBTB20* (NM_001164342.1). Both variants, c.1847 C>T and c.2221 G>A, lead to missense changes (p.Ser616Phe and p.Gly741Arg) affecting two amino-acids located in the last coding-exon of *ZBTB20*. PCR amplification and Sanger sequencing of this exon in parental DNA samples surprisingly revealed that these variations, separated by more than 400 bp, both arose de novo (Fig. 2A). Trio compatibility was checked by using polymorphic microsatellite markers (PowerPlex 16HS system, Promega, France). Allele-specific amplification with the following primers (5^0-CAAGCCTCCGG AAATGTAAT-3^0 with 5^0-ATCTGTTGGCGCTCCTTCTT-3^0) for the specific amplification of c.1847 C>T followed by Sanger sequencing demonstrated that these two events occurred on the same chromosome, with no obvious molecular explanation as there is no homolog sequence for this exon in the human genome.

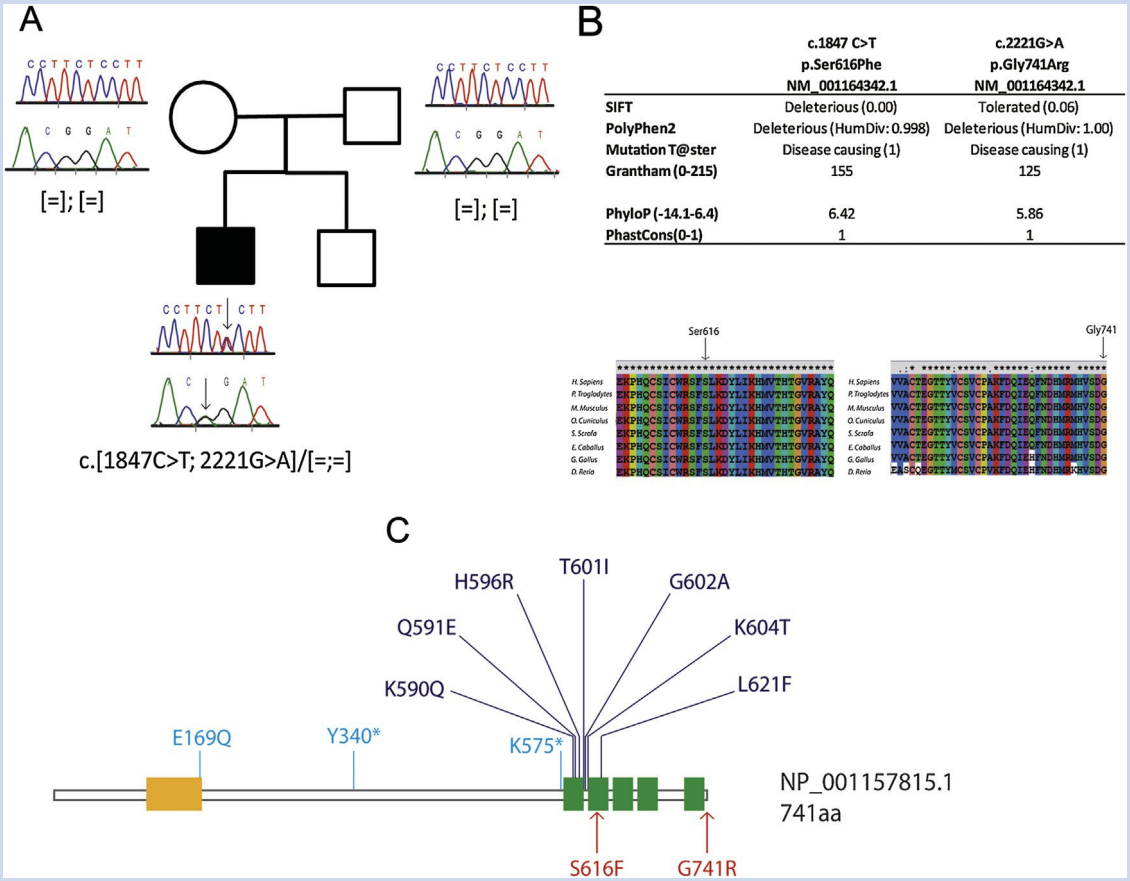


FIG. 2. Mutations in *ZBTB20*. A: Two de novo missense variants are identified in the proband. B: Predicted pathogenicity of the two de novo missense changes and nucleotide conservation scores for the two positions of the single nucleotide substitutions and alignment of the ZBTB20 protein orthologs using ClustalX. C: ZBTB20 protein structure (yellow box- BTB domain; green boxes-zinc-finger domains) with localization of previously reported mutations (dark blue) [Cordeddu et al., 2014], the mutations described here (red) and variants of unknown significance reported in the Deciphering Developmental Disorders Study (light blue). [Color figure can be seen in the online version of this article, available at http://wileyonlinelibrary.com/journal/ajmga].

DISCUSSION

In the original report, Primrose described an adult male with severe developmental delay, muscle wasting, cataracts, and calcified pinnae [Primrose, 1982]. The patient had unremarkable height and weight (177 cm and 53 kg, respectively), but head circumference was not recorded. The case report mentions "moderate hydrocephaly" but the photos do not show clear macrocephaly. A second patient was reported 4 years later [Collacott et al., 1986]. This case report also concerned an adult male with severe developmental delay and calcified pinnae. Subsequently, Primrose syndrome has been described in a series of case reports [Lindor et al., 1996; Battisti et al., 2002; Mathijssen et al., 2006; Dalal et al., 2010; Carvalho and Speck-Martins, 2011; Posmyk et al., 2011]. The gene discovery paper included an additional 3 previously unreported patients thus, bringing the total to 11 known patients (8 mutations proven) prior to our case report [Cordeddu et al., 2014]. The general phenotype is that of an overgrowth syndrome with macrocephaly and severe developmental delay/behavioral disturbances. Some features are variably present such as hearing loss, hypoplasia of the corpus callosum, cataracts, cryptorchidism, and torus palatinus. Several other features of the disorder become apparent only in adulthood: diabetes, calcified pinnae, sparse body hair, and distal muscle wasting [Cordeddu et al., 2014].

We present a 4-year-old boy carrying two missense mutations in the same exon of *ZBTB20*: Ser616Phe and Gly741Arg. Investigation of parental DNA showed that both mutations arose de novo. We also confirmed that they were in cis by using allele specific amplification. These substitutions affect two well-conserved nucleotide positions (phyloP: 6.42 and 5.86; PhastCons: 1 and 1). The corresponding amino acid changes are important (Grantham score: 155 and 125) and affect highly conserved residues in the C2H2 zinc-finger domain, which is important for DNA-binding (Fig. 2B). The p.Ser616Phe is predicted to be damaging for the protein function according to the different prediction programs (SIFT, Polyphen, Mutation T@ster) (Fig. 2B), and it is located in

the same region (AA 590–621) where the other missense mutations involved in Primrose syndrome were recently identified [Cordeddu et al., 2014] (Fig. 2C).

Cordeddu et al. [2014] reported that all the mutated proteins showed a strong reduction in ZBTB20 DNA binding and in its ability to repress transcription, indicating an important role for the zinc-finger, and the linker region. The p.Ser616Phe missense change might therefore be solely responsible for the patient's phenotype, which is consistent with Primrose syndrome. However, it is not possible to conclude if the second missense change, p. Gly741Arg, predicted slightly less damaging (Fig. 2B), and previously reported at the heterozygous state in one African individual from the EXAC database, does not also participate to some extent in the clinical features presented by the patient.

Our patient has many features of Primrose syndrome including overgrowth, macrocephaly, hypotonia, hypoplasia of the corpus callosum, hearing loss, behavioral abnormalities (self-injurious behaviors), and cryptorchidism. However, he did not have torus palatinus, calcified pinnae, or cataracts (slit lamp examination was not performed). It is possible that these features and the distal muscle wasting may appear later in life. The calcified pinnae in particular have only been observed in adults. While all adults with Primrose have had calcified pinnae, the two cases reported in childhood (4 and 8 years of age) did not [Primrose, 1982; Collacott et al., 1986; Lindor et al., 1996; Cordeddu et al., 2014].

The lone atypical feature in our patient was the congenital hypothyroidism that led us to consider a differential diagnosis of Pendred syndrome (MIM 274600) and the congenital disorders of glycosylation. On review of the case reports, at least one other Primrose patient had hypothyroidism [Dalal et al., 2010]. The number of cases with proven mutations is too small (2/9) to determine if the hypothyroidism is associated with *ZBTB20* mutations or if this is a coincidence. However, it would seem reasonable to propose thyroid hormone evaluations in children with the diagnosis of Primrose syndrome. Intriguingly, *ZBTB20* expression was shown to be regulated by thyroid hormone (T3) treatment in murine neural progenitor cell lines [Chatonnet et al., 2013].

In retrospect, our patient has findings typical of Primrose syndrome and the molecular diagnosis should have come as no surprise. However, the inclusion of congenital hypothyroidism as a sign oriented our differential diagnosis towards other conditions. The use of a multi-gene panel for patients with developmental delay enabled the diagnosis of Primrose syndrome and only through diagnosis of further patients can we clarify if there is an increased risk of hypothyroidism in this syndrome.

ACKNOWLEDGMENTS

We would like to thank the family for their gracious participation in this project. We thank also the members of the IGBMC sequencing platform. This work was funded by Agence de Biomedecine, Fondation Jerome Lejeune, and CREGEMES.

REFERENCES

Battisti C, Dotti MT, Cerase A, Rufa A, Sicurelli F, Scarpini C, Federico A. 2002. The Primrose syndrome with progressive neurological involvement and cerebral calcification. J Neurol 249:1466–1468.

Carvalho DR, Speck-Martins CE. 2011. Additional features of unique Primrose syndrome phenotype. Am J Med Genet Part A 155A: 1379–1383.

Chatonnet F, Guyot R, Benoît G, Flamant F. 2013. Genome-wide analysis of thyroid hormone receptors shared and specific functions in neural cells. Proc Natl Acad Sci USA 110:E766–E775.

Collacott RA, O'Malley BP, Young ID. 1986. The syndrome of mental handicap, cataracts, muscle wasting and skeletal abnormalities: Report of a second case. J Ment Defic Res 30:301–308.

Cordeddu V, Redeker B, Stellacci E, Jongejan A, Fragale A, Bradley TE, Anselmi M, Ciolfi A, Cecchetti S, Muto V, Bernardini L, Azage M, Carvalho DR, Espay AJ, Male A, Molin AM, Posmyk R, Battisti C, Casertano A, Melis D, van Kampen A, Baas F, Mannens MM, Bocchinfuso G, Stella L, Tartaglia M, Hennekam RC. 2014. Mutations in ZBTB20 cause Primrose syndrome. Nat Genet 46:815–817.

Dalal P, Leslie ND, Lindor NM, Gilbert DL, Espay AJ. 2010. Motor tics, stereotypies, and self-flagellation in primrose syndrome. Neurology 75:284–286.

Geoffroy V, Pizot C, Redin C, Piton A, Vasli N, Stoetzel C, Blavier A, Laporte J, Muller J. 2015. VaRank: A simple and powerful tool for ranking genetic variants. PeerJ 3:e796.

Lindor NM, Hoffman AD, Primrose DA. 1996. A neuropsychiatric disorder associated with dense calcification of the external ears and distal muscle wasting: 'Primrose syndrome'. Clin Dysmorphol 5:27–34.

Mathijssen IB, van Hasselt-van der Velde J, Hennekam RC. 2006. Testicular cancer in a patient with Primrose syndrome. Eur J Med Genet 49:127–133.

Posmyk R, Leśniewicz R, Chorąży M, Wołczyński S. 2011. New case of Primrose syndrome with mild intellectual disability. Am J Med Genet Part A 155A:2838–2840.

Primrose DA. 1982. A slowly progressive degenerative condition characterized by mental deficiency, wasting of limb musculature and bone abnormalities, including ossification of the pinnae. J Ment Defic Res 26:101–106.

Redin C, Gerard B, Lauer J, Herenger Y, Muller J, Quartier A, Masurel-Paulet A, Willems M, Lesca G, El-Chehadeh S, Le Gras S, Vicaire S, Philipps M, Dumas M, Geoffroy V, Feger C, Haumesser N, Alembik Y, Barth M, Bonneau D, Colin E, Dollfus H, Doray B, Delrue MA, Drouin-Garraud V, Flori E, Fradin M, Francannet C, Goldenberg A, Lumbroso S, Mathieu-Dramard M, Martin-Coignard D, Lacombe D, Morin G, Polge A, Sukno S, Thauvin-Robinet C, Thevenon J, Doco-Fenzy M, Genevieve D, Sarda P, Edery P, Isidor B, Jost B, Olivier-Faivre L, Mandel JL, Piton A. 2014. Efficient strategy for the molecular diagnosis of intellectual disability using targeted high-throughput sequencing. J Med Genet 51:724–736.

APPENDIX 2:

Disease-causing variants in *TCF4* are a frequent cause of intellectual disability: lessons from large-scale sequencing approaches in diagnosis

We present here the clinical reevaluation of ten patients, identified by high throughput sequencing (HTS) of several hundred of genes as carrying a pathogenic mutation in the *TCF4* gene, among a cohort of 1,000 patients with intellectual disability (ID) for which no Pitt Hopkins syndrome (PTHS) was suspected. We showed that if three of them present clinical features a posteriori consistent with a PTHS, three were only moderately evocative and two others not evocative of this syndrome. In parallel, we summarized all the mutations identified during HTS studies previously published by other teams (exome sequencing or targeted sequencing of large panels). With this study we demonstrate that mutations in *TCF4* can cause a broad spectrum of ID forms, from PTHS to nonsyndromic ID, without any correlation between the nature or the location of the mutation. We also show that *TCF4* mutations represent a frequent cause of ID (16/2,230 = 0.7%), and should be considered for routine genetic screening in case of ID (meaning that it should be included in all HTS panels used for the genetic diagnosis of nonsyndromic ID).

European Journal of Human Genetics
https://doi.org/10.1038/s41431-018-0096-4

ARTICLE

Disease-causing variants in *TCF4* are a frequent cause of intellectual disability: lessons from large-scale sequencing approaches in diagnosis

Laura Mary[1] · Amélie Piton[1,2] · Elise Schaefer[3] · Francesca Mattioli[2,4] · Elsa Nourisson[1] · Claire Feger[1] · Claire Redin[2,4] · Magali Barth[5] · Salima El Chehadeh[3,6] · Estelle Colin[5] · Christine Coubes[7] · Laurence Faivre[6] · Elisabeth Flori[3] · David Geneviève[7] · Yline Capri[8] · Laurence Perrin[8] · Jennifer Fabre-Teste[8] · Dana Timbolschi[3] · Alain Verloes[8] · Robert Olaso[9] · Anne Boland[9] · Jean-François Deleuze[9] · Jean-Louis Mandel[1,2,4] · Bénédicte Gerard[1] · Irina Giurgea[10,11]

Received: 8 June 2017 / Revised: 11 December 2017 / Accepted: 23 December 2017
© European Society of Human Genetics 2018

Abstract

High-throughput sequencing (HTS) of human genome coding regions allows the simultaneous screen of a large number of genes, significantly improving the diagnosis of non-syndromic intellectual disabilities (ID). HTS studies permit the redefinition of the phenotypical spectrum of known disease-causing genes, escaping the clinical inclusion bias of gene-by-gene Sanger sequencing. We studied a cohort of 903 patients with ID not reminiscent of a well-known syndrome, using an ID-targeted HTS of several hundred genes and found de novo heterozygous variants in *TCF4* (transcription factor 4) in eight novel patients. Piecing together the patients from this study and those from previous large-scale unbiased HTS studies, we estimated the rate of individuals with ID carrying a disease-causing *TCF4* mutation to 0.7%. So far, *TCF4* molecular abnormalities were known to cause a syndromic form of ID, Pitt–Hopkins syndrome (PTHS), which combines severe ID, developmental delay, absence of speech, behavioral and ventilation disorders, and a distinctive facial gestalt. Therefore, we reevaluated ten patients carrying a pathogenic or likely pathogenic variant in *TCF4* (eight patients included in this study and two from our previous ID-HTS study) for PTHS criteria defined by Whalen and Marangi. *A posteriori*, five patients had a score highly evocative of PTHS, three were possibly consistent with this diagnosis, and two had a score below the defined PTHS threshold. In conclusion, these results highlight *TCF4* as a frequent cause of moderate to profound ID and broaden the clinical spectrum associated to *TCF4* mutations to nonspecific ID.

Introduction

Since a few years, the development of new high throughput sequencing technologies (HTS) permitted the study of either a large number of genes or the entire exome/genome in

These authors contributed equally: Laura Mary and Amélie Piton.

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41431-018-0096-4) contains supplementary material, which is available to authorized users.

✉ Amélie Piton
piton@igbmc.fr

✉ Irina Giurgea
irina.giurgea@inserm.fr

1 Laboratoire de Diagnostic Génétique, Hôpitaux Universitaires de Strasbourg, Strasbourg, France

2 Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), CNRS UMR-7104, Inserm U964, Université de Strasbourg, Strasbourg, France

3 Département de Génétique Médicale, CHU de Hautepierre, Strasbourg, France

4 Chaire de Génétique Humaine, Collège de France, Paris, France

5 Département de Biochimie et de Génétique, CHU d'Angers, Angers, France

6 Centre de Génétique et Centre de Référence Anomalies du développement et Syndromes malformatifs, Hôpital d'Enfants, CHU de Dijon, Dijon, France

7 Département de Génétique médicale, Maladies rares et Médecine personnalisée, Hôpital Arnaud de Villeneuve, Chu Montpellier, France

8 AP-HP, Department of Genetics, Hôpital Robert Debré, Paris, France

9 Centre National de Recherche en Génomique Humaine, Institut de Biologie François Jacob, CEA, Evry, France

10 U.F. de Génétique moléculaire, Hôpital Armand Trousseau, Assistance Publique—Hôpitaux de Paris, Paris 75012, France

11 INSERM UMR S933, Sorbonne Université, Paris 75012, France

Published online: 26 April 2018

patients with non-syndromic ID. These allowed the identification of disease-causing variants in genes involved in syndromic forms of intellectual disability (ID) in patients whom clinical manifestations were not typical of the corresponding disorders. In our previous targeted sequencing (TES) study performed in 106 individuals with unexplained ID using a panel of 217-ID genes, only four genes were found to be mutated in more than one family. Three mutations were identified in *MECP2* (MIM *300005, involved in Rett syndrome #312750), two de novo point mutations in two girls and one maternally inherited complex rearrangement in exon 4 of the gene in one boy removing 60 amino acids inherited from his mother (speech delay) [1]. Two disease-causing variants were identified in another X-linked gene, *KDM5C* (MIM *314690), and in two autosomal genes *DYRK1A* (MIM *600855) and *TCF4* (MIM *602272). *TCF4* (transcription factor 4) is located in 18q21, and encodes a class I basic helix-loop-helix transcription factor binding to E-boxes on DNA after dimerization, which is involved in cell signaling, cell survival and neurodevelopment [5]. So far, *TCF4* is the single gene involved in Pitt-Hopkins Syndrome (PTHS, MIM #610954) [2–4], a rare, well-characterized, neurodevelopmental disorder usually presenting with severe intellectual disability associated with distinctive facial features, various neurological and behavioral impairment and gastro-intestinal dysfunction, hypotonia, ataxia, breathing abnormalities, and seizures [6]. This provided a rationale for *TCF4* Sanger-sequencing in patients with syndromic ID after ruling out differential diagnoses by PTHS clinical scores [6, 7]. Since implementation of HTS in ID screening, we and others have suggested *TCF4* implication in isolated ID [1, 8–10].

To assess the frequency of *TCF4* molecular abnormalities in non-syndromic ID patients, we studied 903 novel patients with mild to severe ID and reviewed the previous published targeted, exome or genome sequencing studies [1, 11–15]. To better delineate the phenotype related to *TCF4* mutations we re-analyzed *a posteriori* the phenotype of all the patients carrying a pathogenic or likely pathogenic variant in this gene (as defined by the American College of Medical Genetics and Genomics), but for whom PTHS diagnostic was not clinically suspected.

Materials and methods

Patients

DNA samples (from peripheral blood or saliva) of the 903 patients were referred to the laboratory of genetic diagnosis. Patients presented with non-specific intellectual disability and no major congenital anomalies. The cohort includes patients with mild ID or ID of unknown severity (around 25%), moderate (around 40%), or severe to profound (around 35%) ID, based on clinician's appreciations. The most current causes of cognitive impairment were dismissed by fragile-X test, array-CGH, and metabolic explorations (in 90% of patients or more). Among the more recurrent tests, *UBE3A* (MIM *601623) sequencing or methylation analysis were performed in <20% of the patients, and *MECP2, ARX* (MIM *300382) or *DMPK* (MIM *605377) in around 12%. Clinical data were recorded before inclusion following a standardized clinical questionnaire highlighting prenatal history, developmental milestones, neurological, and behavioral disorders. ID severity was assessed by medical geneticists upon clinical evaluation and was not a discriminating inclusion criterion. However, the cohort was enriched in severe and moderate forms of ID compared to the distribution in ID population. After obtaining the molecular diagnosis, the patient was reevaluated by the clinical geneticist. All the clinical data were re-collected, with a specific attention to PTHS clinical signs. This study was approved by the local Ethics Committee of the Strasbourg University Hospital (Comité Consultatif de Protection des Personnes dans la Recherche Biomédicale - CCPPRB). For all patients, a written informed consent for genetic testing was obtained from their legal representative.

Targeted genes and capture design

DNA samples were extracted from peripheral blood or saliva. HTS targeted libraries were prepared, as previously described [1] with individual in-solution SureSelect capture reaction for each DNA sample (custom design for genes known to be involved in ID, Agilent, Santa Clara, California, USA). Capture experiments were performed using probes corresponding to a panel of 275 (in 207 patients), 451 (in 66 patients) or 456 (in 630 patients) ID genes. Paired-end sequencing (2 × 101-bp) was performed on an Illumina HiSeq 2500, multiplexing in average 32 samples per sequencing lane. Read mapping, variant calling and annotation were performed, as previously described [1]. Detected variants, short indels and single nucleotide variants (SNVs), were annotated and ranked by VaRank software [16].

Sanger sequencing confirmation

TCF4 pathogenic or likely pathogenic variants identified by HTS were confirmed in patients and the de novo status was checked in their parents by Sanger sequencing. Pedigree (parents-child) concordance was confirmed by checking the segregation of several highly polymorphic microsatellite markers (PowerPlex 16 HS System, Promega, Madison, WI, USA) or frequent variants (when TES was also performed for parental DNA). We reported the variants

identified in *TCF4* in a specific database (https://databases.lovd.nl/shared/genes/TCF4).

PTHS clinical scoring

To facilitate the clinical diagnosis of PTHS two scoring tests have been developed in 2012. The first one, established by Whalen et al., was based on the scoring of the following criteria: facial gestalt (8 points), severe motor delay (2 points), absent language (2 points), stereotypic movements (2 points), hyperventilation (1 point), anxiety (1 point), hypotonia (1 point), smiling appearance (1 point), ataxic gait (1 point), and strabismus (1 point). This score was validated in patients evocative of PTHS with ($n = 33$) or without ($n = 100$) pathogenic variant identified in *TCF4*. A threshold of 15/20 was considered as a good indicator of *TCF4*. A score between 10 and 15 could also be suggestive of this diagnosis, especially for young patients [6]. The second scoring, established by Marangi et al. scored the following symptoms: typical/partial facial features (4 points/2 points), moderate/severe intellectual disability (2 points), poor/absent language (1 point/2 point), normal growth parameters at birth (1 point), microcephaly (1 point), epilepsy/EEG abnormalities (1 point), ataxic gait (1 point), hyperventilation (1 point), constipation (1 point), brain MRI abnormalities (1 point) and strabismus or ophthalmologic abnormalities (1 point) [7]. These criteria were evaluated in patients evocative of PTHS with ($n = 18$) or without ($n = 60$) pathogenic variants in *TCF4* and a score above 10/16 was recommended for a molecular study of *TCF4*. Whalen and Marangi's scores were calculated after a clinical reexamination (*a posteriori* after obtaining the molecular diagnosis) for the patients described in this paper plus the two we previously described [1].

Results

Pathogenic or likely pathogenic *TCF4* variants in undiagnosed ID patients

Through HTS targeted sequencing of several hundred of ID genes in 903 patients with undiagnosed ID, we identified eight pathogenic or likely pathogenic *TCF4* variants among which four were novel (Table 1, Fig. 1). All these variants occurred de novo, were not reported in ExAC general population database and affected amino acids included in all the isoforms of the gene. Named here according to the NM_001083962.1, we identified four nonsense or frameshift variants c.873C>A p.(Tyr291*), c.1662del p.(Asp554Glufs*4), c.1726C>T p.(Arg576*) and c.1927G>T p.(Glu643*), three missense variants affecting conserved amino acid located in the bHLH domain of the protein and predicted to be damaging by in silico tools (SIFT, Polyphen2): c.1705C>T p.(Arg569Trp), c.1733G>A p.(Arg578His) and c.1841C>T p.(Ala614Val), and one silent variant altering the last nucleotide of exon 12 (according to NG_011716) and predicted to modify the donor splice site (c.990G>A, p.?). In addition, two variants affecting only one alternative isoform (NM_001243231.1: c.7G>T p.(Glu3*) and c.2T>C, p.(Met1?)) have been identified, both inherited from an unaffected parent and were therefore classified as likely benign.

TCF4 mutation rate is of 0.7% (16/2239) in individuals with undiagnosed ID

Piecing together the 8 patients out of the 903 of this study with the two out of 106 patients that we have previously reported [1], the frequency of *TCF4* disease-causing variants is of 1% (10/1009) in our cohort of individuals with ID undiagnosed by a geneticist. Furthermore, we reviewed data from other large scale studies, including TES of ID genes [12, 15], and WES performed in patients with non-specific ID [11, 13, 14] and calculate the *TCF4* mutation rate in patients with non-syndromic ID (Tables 1 and 2). Altogether with our results, 16 individuals with pathogenic or likely pathogenic *TCF4* variants were identified during the large-scale sequencing studies performed in 2230 patients with nonspecific ID, providing a *TCF4* mutation rate of 0.7% (Table 2, Fig. 1).

TCF4 mutations can cause ID poorly suggestive of PTHS

A posteriori clinical reevaluation was performed for the 10 patients (eight novel patients included in this study and two from our previous ID-HTS study) carrying a *TCF4* disease-causing variant (Table 3, Fig. 2). All probands, except MMPN166, were born from unrelated healthy parents, with irrelevant family history. According to Whalen and Marangi scores, five patients (MMPN166, MMPN68, APN-214, B00H4MR, and B00H4U1) had features reminiscent of PTHS (>12/20 Whalen's and 10/16 Marangi's score), three individuals (B00H4R8, APN-210, and APN-41) were slightly evocative of PTHS (only one of the scores was upper to the threshold) and two patients (APN-149 and APN-117) were not consistent with PTHS (both scoring were below the threshold). To widely asses the phenotype of patients with a *TCF4* pathogenic variant identified through TES or WES, we further evaluate the phenotype of the patients reported by other groups [11–13, 15] (Table 4). Clinical data were available for four out of the six reported patients. The phenotype could be evocative of a PTHS for three of the patients, but not in the last one who had only

Table 1 Pathogenic or likely pathogenic variants in *TCF4* identified in patients with intellectual disability (ID) by large-scale sequencing approaches

Reference	Variant nomenclature		Inheritance	dbSNP	Previously reported in Clinvar	Previously described in PTHS screening	*TCF4*-LOVD Individual ID	SIFT	PP2	Individual	Gender
This report	c.873C>A	p.(Tyr291*)	de novo	NA	NA	no	000100	NA	NA	MMPN166	F
This report Tan et al. [15]	c.990G>A	p.?	de novo, de novo	rs587784469	RCV000147730.1 (Likely. pathogenic)	no	000098	NA	NA	- APN149 -Patient 6	M F
This report	c.1662del	p.(Asp554Glufs*4)	de novo	NA	NA	no	000101	NA	NA	B00H4MR	F
This report	c.1705C>T	p.(Arg569Trp)	de novo	NA	NA	yes [6]	000026	Deleterious	Prob. Damaging	MMPN68	M
This report	c.1726C>T	p.(Arg576*)	de novo	NA	NA	no	000102	NA	NA	APN214	M
This report	c.1733G>A	p.(Arg578His)	de novo	rs121909123	RCV000079458.4|RCV000189738.1 (Pathogenic)	yes [22]	000029	Deleterious	Poss. Damaging	APN210	F
This report	c.1841C>T	p.(Ala614Val)	de novo	NA	NA	yes [32]	000021	Deleterious	Prob. Damaging	B00H4R8	F
This report	c.1927G>T	p.(Glu643*)	de novo	NA	NA	no	000103	NA	NA	B00H4U1	F
Redin et al. [1]	c.514_517del	p.(Lys172Phefs*61)	de novo	rs398123561	RCV000079461.4 (Pathogenic)	yes [6]	000047	NA	NA	APN41	M
Redin et al. [1]	c.520C>T	p.(Arg174*)	de novo	NA	RCV000224478.1 (Pathogenic)	yes [6]	000023	NA	NA	APN117	F
Grozeva et al. [12]	c.505C>T	p.(Gln169*)	NA	NA	NA	no	000104	NA	NA	5410771	M
Grozeva et al. [12]	c.550−1G>A g.52946888 C>T	p.?	NA	NA	NA	no	000105	NA	NA	5411380	M
Tan et al. [15]	c.991−2A>G g.52927260 T>C	p.?	NA	rs587784470	RCV000147731.1 (Pathogenic)	no	000106	NA	NA	Patient 5	F
Hamdan et al. [13]	c.1153C>T	p.(Arg385*)	de novo	rs121909122	RCV000007797.4 (Pathogenic)	yes [4, 33]	000003	NA	NA	Case 045.400	M
De Ligt et al. [11]	c.1727G>A	p.(Arg576Gln)	de novo	NA	RCV000431775.1 (Pathogenic)	yes [33]	000027	Deleterious	Poss. Damaging	Trio 15	F

All the c. positions were given according to NM_001083962.1 isoform (and NG_011716.2 isoform for intronic variants). *PP2* Polyphen2 SIFT scores. Variants are classified following recommendations from the American College of Medical Genetics and Genomics

Table 2 Pathogenic or likely pathogenic variants identified in *TCF4* during targeted sequencing (TES), whole exome sequencing (WES) or whole genome sequencing (WGS) in patients with intellectual disability (ID)

Cohort	Reference	Approach	Number of patients	TCF4 mutations
ID (mild to severe)	this study	TES (275–456 genes)	903	8
ID (mild to severe)	Redin et al. [1]	TES (217 genes)	106	2
ID (moderate to severe)	Grozeva et al. [12]	TES (575 genes)	986	2
ID	Tan et al. [15]	TES (90 genes)	52	2
ID (severe)	Rauch et al. [14]	WES	51	(1*)
ID (moderate to severe)	de Ligt et al. [11]	WES	100	1
ID (moderate to severe)	Hamdan et al. [13]	WES	41	1
Total			2239	16 (0.7%)

* de novo missense variant predicted to be benign, not included in the statistics

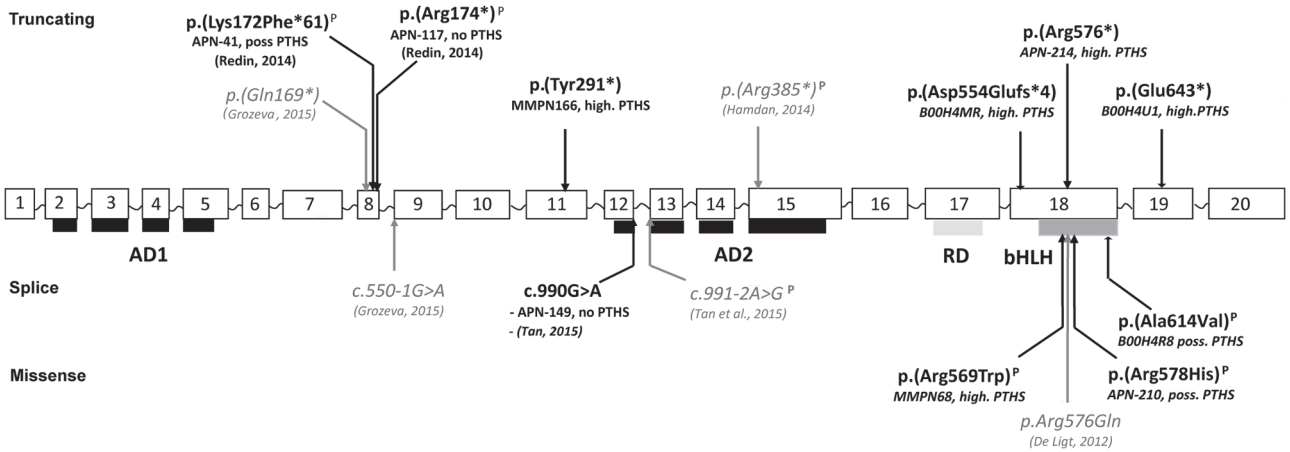


Fig. 1 Schematic representation of disease-causing variants identified in *TCF4*. *AD1, AD2* transactivation domains; *RD* repressor domain; *bHLH* DNA-binding domain. In bold: variants identified by TES in our cohort. Patient number is indicated as well as the severity of his PTHS phenotype: no not evocative of PTHS, poss. possibly evocative of PTHS, high. highly evocative of PTHS), in italic: variants identified in other HTS studies [P]: variants previously described in PTHS patients

a mild ID. Taken together, in nearly half of the patients (6/13) studied by HTS and carrying a disease-causing *TCF4* variant, clinical features were poorly or not evocative of PTHS.

Discussion

Targeted or whole exome HTS used in routine diagnosis have demonstrated their efficiency in the diagnosis of isolated ID [1, 11, 14]. Unexpected rates of pathogenic variants in genes implicated in syndromic cognitive impairment were found with these clinically unbiased approaches. We studied 903 patients with undiagnosed ID by targeted HTS of ID known genes, and identified eight novel patients carrying a pathogenic or likely pathogenic variant in *TCF4*. We also analyzed data from previous HTS studies, and found eight additional patients carrying a disease-causing variant in *TCF4*, including two patients reported by our group [1]. Taken together, we count 16 patients carrying a *TCF4* disase-causing variant (of which 15 distinct variants) among 2239 ID patients and we obtained a *TCF4* mutation rate of 0.7% in non-specific ID (Table 2). This mutation rate is close to those of the most frequent causes of ID such as *FMR1* expansions [17, 18] or *ARID1B* mutations [19] in Fragile-X and Coffin-Siris syndromes. Otherwise, *TCF4* mutation rate gets down to 0.3% (13/4293) in studies including patients with developmental disorders in which ID is not a mandatory sign, such as the Deciphering Developmental Disorder (DDD) project [20]. Indeed, a very recent study reported ID in 100% (47/47) of patients carrying a disease-causing variant in *TCF4*, collected though a web-based database [21]. However, in the DDD data, *TCF4* still appears in the top-twenty of the most frequently mutated genes in with developmental disorders.

The patients included in our TES study were referred by a geneticist after several biological, radiological and

Table 3 A posteriori reevaluation of PTHS clinical signs in seven patients carrying a pathogenic mutation in *TCF4*

	Patient MMPN166	Patient APN149	Patient B00H4MR	Patient MMPN68	Patient APN214	Patient APN210	Patient B00H4R8	Patient B00H4U1	Patient *(Redin et al., 2014)* APN41	Patient *(Redin et al., 2014)* APN117
TCF4 variant	c.873 C>A, p.(Tyr291*)	[a]c.990 G > A, p.?	c.1662del, p. (Asp554Glufs4*)	[a]c.1705C>T, p.(Arg569Trp)	c.1726C>T, p. (Arg576)	[a]c.1733G>A, p.(Arg578His)	[a]c.1841C>T p. (Ala614Val)	c.1927G>T p. (Glu643*)	[a]c.514 517del p.(Lys172Phefs*61)	[a]c.520 C>T, p.(Arg174*)
Other variants/CNV	dup 22q11	no	no	no	no	no	no	no	no	no
Patient	Female	Male	Female	Male	Male	Female	Female	Female	Male	Female
Age of inclusion	4 y-o	5 y-o	4 y-o	5 y-o	18 y-o	6 y-o	7 y-o	2 y-o	3 y-o	10 y-o
Age of reexamination	6 y-o	9 y-o	5y 8 m-o	8 y-o	20 y-o	9 y-o	7 y-o	4 y-o	6 y-o	13 y-o
Classical PTHS symptoms										
Facial gestalt	typical	mild	typical	mild	typical	mild	not typical[b]	typical	mild	mild
Growth (statural/ ponderal)	M/M	M/M	−1/−2	M/ + 5 SD	−3.5 SD/M	M/ + 3 SD	−3 SD/M	M/M	−2SD/−2SD	M/M
Head circumference	−1 DS	M	−2.5	M	M	+ 0,5 SD	+ 1.5 SD	M	M	−1.8 SD
Cognitive impairment	profound	moderate	severe	profound	profound	moderate	severe	severe	severe	moderate
Walking	absent	16 mo	NA	5 y	3 y	25 mo	5 y	absent	3 y	28 mo
Absent speech	yes	yes	no (few words)	yes	yes	yes	yes	yes	yes	no
Hyperventilation/ apneas	no	no	no	no	yes	yes	no	yes	no	no
Happy appearance	yes	yes	no	no	yes	yes	NA	no	yes	yes
Sleep disturbance	no	no	yes	no	yes	yes	no	no	no	yes
Behavior problems	no	yes (Self-aggress.)	no	no	yes (Self-aggress.)	no	yes, severe	no?	yes (poor interactions)	no
Stereotypic behavior	yes	no	yes	yes	yes	no	yes	yes	yes	no
Seizures	yes (3–6 mo)	no	no	yes	yes	no	no	no	yes	no
Hypotonia	yes	no	yes	yes	yes	no	no	yes	yes	yes
Ataxic gait	no	no	yes	yes (mild)	yes	no	instable	no	yes	yes
Ophthalmologic anomalies	yes strabismus	no	yes strabismus	yes	yes strabismus, astigmatism	yes Duane anomaly	No (ptosis)	yes strabismus hyperopia	no	no
Constipation	no	yes	yes	yes	yes	yes	no	no	yes	yes
Gastro-esophageal reflux	no	yes	no	no	yes	no	NA	yes (infancy)	yes	no
Fetal pads	no	no	yes	yes	no	no	no	no	yes	yes
Cerebral MRI	normal	–	abormal (1)	abnormal	abnormal	normal	abnormal (2)	normal	normal	normal

Table 3 (continued)

	Patient MMPN166	Patient APN149	Patient B00H4MR	Patient MMPN68	Patient APN214	Patient APN210	Patient B00H4R8	Patient B00H4U1	Patient (*Redin et al., 2014*) APN41	Patient (*Redin et al., 2014*) APN117
Other signs	no	no	cervical syringomyelia	no	cryptorchidism, abolition of osteo-tendinous reflexes	headaches	no	heterotaxy, bifid uvula, long thumbs, labia minora hypoplasia	chronic otitis	no
Whalen's score (> 12 or 15/20)	16	7	15	15	17	13	13	15	12	9
Marangi's score (> 10/16)	12	7	13	14	13	10	10	12	12	7
Conclusion PTHS	highly	no	highly	highly	highly	possibly	possibly	highly	possibly	no

PTHS Pitt-Hopkins syndrome, *y-o* year-old; mo: months, *SD* standard deviation, *M* value in normal range, *MRI* Magnetic resonance imaging

[a] mutation previously described in databases or literature

[b] Rubinstein-Taybi facial features suggested. MRI anomalies 1: posterior atrophy of corpus callosum; 2: Hypersignal of the subcortical white matter in temporal lobes. The cDNA nomenclature given according to NM_001083962.1 isoform for all the variants.

molecular tests which did not allow a diagnosis. The *a posteriori* analysis of the clinical features of the ten patients carrying a *TCF4* disease-causing variant showed that PTHS could have been suspected in five patients. However, even if the diagnosis would have been possible in three additional cases, by using Whalen and Marangi clinical scores, facial gestalt of those patients was not typical of PTHS. Furthermore, for two patients, Whalen and Marangi clinical scores were low and PTHS could not have been suspected clinically. Indeed, in absence of distinctive signs of PTHS, such as a typical facial gestalt (4/10) (Fig. 2) or hyperventilation (3/10) which can appear later in childhood [21, 22], the clinical diagnosis remains challenging, especially for the patients with moderate ID. In contrast, absence of speech (8/10), noticeable delay in walking (after 3 years of age, if acquired) (8/10), seizures (4/10), behavior problems (self-aggressiveness, poor social interactions) (4/10), smiling appearance (6/10), strabismus (4/10) and constipation (7/10) were observed in our patients, but were not sufficiently discriminatory signs of PTHS, as they can be also found in non-specific ID. This study suggests that even if some were *a posteriori* evocative of PTHS, other ones presenting nonspecific ID and only few PTHS features could not be diagnosed clinically showing that phenotypic spectrum associated to a *TCF4* disease-causing variant is wider than we used to think.

The main differential diagnoses described for PTHS are Angelman and Rett syndromes [23]. Consistent with that, previous genetic tests performed in the patients, before identification of a disease-causing variant in *TCF4,* were *UBE3A* methylation testing or point mutation screening (64% of the patients), and *MECP2* sequencing (36%). A third known differential diagnosis, the Mowat–Wilson syndrome, was suspected in one patient. This later syndrome is associated with cardiac and urogenital malformations and Hirschsprung disease, which are features more discriminative for clinical diagnosis. Surprisingly, a Steinert syndrome was suspected in four patients, maybe due to hypotonia observed in those patients. Analysis of the 17p11.2 deletion (Smith–Magenis syndrome) and of *ARX* coding sequences were also performed in two patients. Taken together, these explorations assess the difficulty to evoke clinically PTHS when the patient only presents with severe delay of psychomotor acquisitions with mild dysmorphic features.

Most of the disease-causing *TCF4* variants previously associated to PTHS are truncating mutations localized between the exons 7 and 18 and are probably responsible of haploinsufficiency. Missense variants mainly concern the bHLH domain of the protein including the arginine residues 578 and 580, spots of recurrent mutations [6]. In in vitro functional studies, Sepp et al. highlighted the variation in expression, patterning, dimerization and DNA binding of

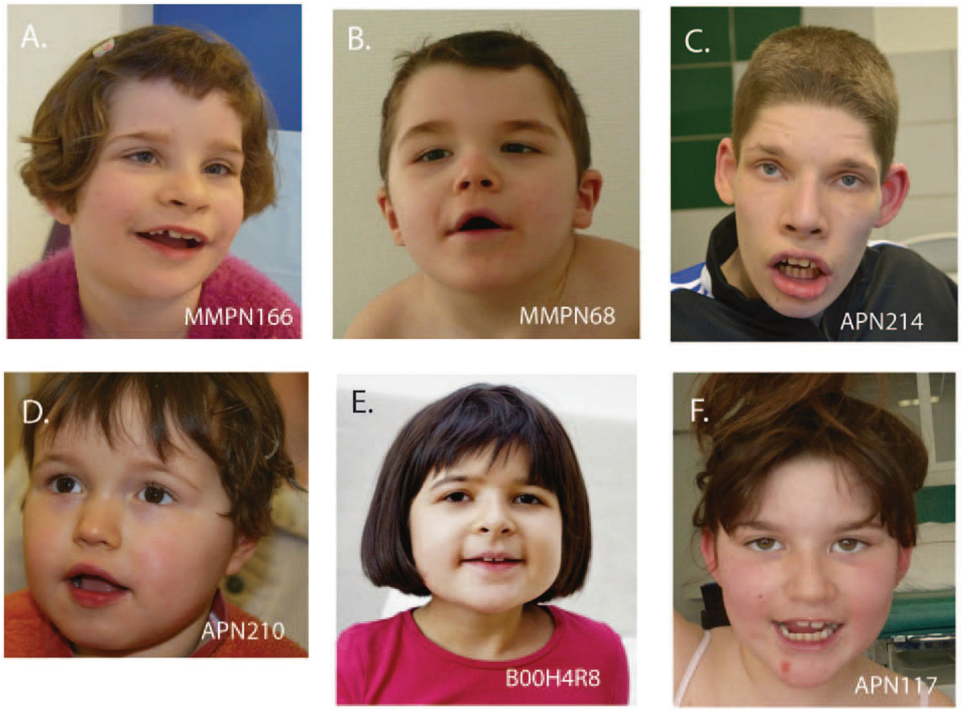


Fig. 2 Pictures of Patients carrying de novo heterozygous disease-causing variants in *TCF4*. a Patient MMPN166, b Patient MMPN68, c Patient APN214, d Patient APN210, e Patient B00H4R8, and f Patient APN117

different TCF4 mutants comparing to WT proteins, suggesting that disease-causing variants can have various functional effects ranging from selective heterodimerization defects to complete lack of DNA binding or possible dominant-negative effects [24]. These authors suggested that the variety of variations could explain the phenotypic variability. Other authors suggested that seizures are more often associated to missense than truncating variants [25] but this was not confirmed afterwards [6]. It is tempting to speculate that some milder phenotype might be explained by variants having a less severe effect, but no clear correlation between the type of variation (missense, truncating) or its location and the phenotype was reported so far [6]. Actually, in the patients reported here, no correlation between the PTHS score and the type or the location of the variant was found. Some of the patients, as for instance patient APN117, had a milder PTHS score while carrying disease-causing variants previously described in classical PTHS cases (Fig. 2). Finally, the c.990 G > A variant, predicted to affect the exon 12 splice donor site, was identified in two patients poorly evocative of PTHS (patient APN149 and Patient 6 reported by Tan et al., 2014). In this specific case, the presence of normal splicing in a part of transcripts might explain the milder phenotype of these patients. Due to the large number of *TCF4* transcripts and to the tissue-variability, splicing effects are difficult to assess. Furthermore, the threshold of *TCF4* normal transcript level sufficient to avoid a pathogenic effect is not known since several cases of typical PTHS with varying levels of mosaicism have been reported [26–29]. Interruptions of the *TCF4* gene can also result in a broader phenotype than usually described, as suggested by Kalscheuer et al. in 2008 after reporting the case of a girl with mild ID, minor facial gestalt and a balanced 18;20 translocation disrupting *TCF4* in exon 4 [9]. More recently, Schluth-Bolard et al. reported a case of a girl with severe developmental delay and microcephaly who was carrier of an apparently balanced translocation between chromosomes 1 and 18, which was disrupting *TCF4* in intron 6 [30]. Similar complex chromosomal translocations have been reported in familial cases of mild ID with an autosomal dominant transmission pattern, without any feature of PTHS [10, 31]. Both breakpoints were located before exon 8. More than a dozen of transcripts isoforms are described for *TCF4*. Functional RNA studies carried on fibroblasts showed, as expected, a decrease of the long isoforms of *TCF4* (affected by the breakpoint) in the patients while the short isoforms encoding nuclear TCF4 were upregulated [31]. The authors suggested that the persistence of the expression of *TCF4* short isoforms may rescue part of PTHS phenotype. In our study, there is no correlation between the number of isoforms affected by the different disease-causing variations and the severity of the phenotype, suggesting that additional mechanisms than a rescue with short isoforms are responsible for the clinical variability. Finally, genetic background may also play a role and influence the severity of clinical manifestations caused by a disease-

Table 4 Summary of clinical information available for patients with *TCF4* mutations identified by other large-scale sequencing studies (form supplementary information of De ligt et al. [11]; Hamdan et al. [13], Tan et al. [15], Grozeva et al. [12])

Reference	Patient identification	TCF4 pathogenic variant	Clinical description	Previous genetic investigations
De Ligt et al. [11]	Trio *15*	de novo c.1727G>A, p.(Arg576Gln)	Female, 4 y-o: Moderate ID, feeding problems, recurrent otitis. Sitting: 22 months, walking 3,5 y-o; no speech. Mild dysmorphic features (epicanthic folds, a broad nasal tip and prominent ears). Hypotonia. Ataxic walking pattern. MRI: mildly enlarged ventricles without structural anomalies	CGH-array, 15q methylation, *MECP2*, *EHMT1*, *UBE3A* sequencing
Hamdan et al. [13]	Case 1045.400	de novo c.1153 C>T, p.(Arg385*)	Male, 6 y-o: Severe ID. no stand nor walk without support, no speech. Hypotonia. Hypersalivation. No breathing problem. Had seizure once. Minor dysmorphy (wide mouth, bilateral single palmar creases, bilateral clinodactyly and overlapping 2nd toes). MRI: increased T2 and FLAIR signal in the periventricular regions, thin corpus callosum, myelination delay.	CGH-array, *FMR1* CGG expansion testing, *MECP2* sequencing, 15q methylation
Tan et al. [15]	Patient 5	c.991−2 A>G, g.52927260 T>C	Female, 5 y-o: Global developmental delay, height and size: 10th percentile, minor facial dysmorphy. Myopia. MRI: tiny pineal cyst and peritrigonal white matter intensity	Karyotype, CGH-array, *ZEB2* sequencing
Tan et al. [15]	Patient 6	de novo c.990 G>A	Female, 7 y-o: Global developmental delay, facial dysmorphy (thick overfolded helix of the ear, wide mouth, coarse facial features, flat philtrum, bulbous nose). Ataxia. Strabismus. MRI: normal	CGH-array, *FMR1* CGG expansion testing, *SNRPN* methylation, 17q deletion, *MECP2* sequencing
Grozeva et al. [12]			UK10K_FINDWGA5411380	c.550−1 G>A, g.52946888 C>T
Male: NA	NA			
Grozeva et al. [12]			UK10K_FINDWGA5410771	c.505 C>T, p.(Gln169*)
Male: NA	NA			

MRI: Magnetic resonance imaging; y-o: year-old; NA: non available. The variations are given according to hg19/GRC37 for the genomic nomenclature and the RefSeq transcript NM_001083962.1 for the cDNA nomenclature

causing variant in *TCF4*. It is interesting to note that Patient MMPN166, one the most severely affected patient, also carries an inherited 22q11.21 duplication which segregates with various neurological signs in her family. The hypothesis of a second genetic hit should be considered to account for the phenotypic difference of patients carrying a disease-causing variant in *TCF4*.

The growing number of HTS realized in routine in patients with ID may allow to provide more data about the prevalence of disease-causing variants in *TCF4* in patients with cognitive impairment and to assess its related phenotype in an unbiased manner. Our study extended the clinical spectrum associated to *TCF4* mutation from PTHS to nonspecific intellectual disability. The high prevalence (0.7%) of disease-causing variants in *TCF4* found in large cohorts of patients suffering from intellectual disability proves that the borders of PTHS are less stringent than we used to consider. This gene should therefore be included in all HTS panels used for diagnosis of unspecific ID. The use of "Pitt-Hopkins syndrome" when reporting a disease-causing variant in *TCF4* in a patient with a low PTHS clinical score should also be discussed.

Web resources

The URLs for online tools and data presented herein are:

OMIM: http://www.omim/org/

UCSC: http://genome.ucsc.edu/

dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP/

Mutation Nomenclature: http://www.hgvs.org/mutnomen/recs.html

Exome Variant Server, NHLBI Exome Sequencing Project (ESP): http://evs.gs.washington.edu/EVS/

ExAC Browser (Beta) | Exome Aggregation Consortium: http://exac.broadinstitute.org/

Integrative Genomics Viewer (IGV): http://www.broadinstitute.org/igv/

These variants were submitted to Clinvar: http://www.ncbi.nlm.nih.gov/clinvar/

Acknowledgements We thank the families for their participation to the study. We also thank the Fondation Jerome Lejeune and the Agence de la Biomédecine for their financial support. They want also to thank all the people from the Strasbourg Hospital molecular diagnostic lab, from the IGBMC and CNG/CEA sequencing platform from UMR_S 1112 (Bernard Jost, Stéphanie Le Gras, Mathieu Jung, Céline Keime, Jean Muller, Véronique Geoffroy) and from the Mondor Hospital molecular diagnostic lab (Thierry Gaillon) for their technical and bioinformatical supports.

Compliance with ethical standards

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

References

1. Redin C, Gerard B, Lauer J, et al. Efficient strategy for the molecular diagnosis of intellectual disability using targeted high-throughput sequencing. J Med Genet. 2014;51:724–36.
2. Amiel J, Rio M, de Pontual L, et al. Mutations in TCF4, encoding a class I basic helix-loop-helix transcription factor, are responsible for Pitt-Hopkins syndrome, a severe epileptic encephalopathy associated with autonomic dysfunction. Am J Hum Genet. 2007;80:988–93.
3. Brockschmidt A, Todt U, Ryu S, et al. Severe mental retardation with breathing abnormalities (Pitt-Hopkins syndrome) is caused by haploinsufficiency of the neuronal bHLH transcription factor TCF4. Hum Mol Genet. 2007;16:1488–94.
4. Zweier C, Peippo MM, Hoyer J, et al. Haploinsufficiency of TCF4 causes syndromal mental retardation with intermittent hyperventilation (Pitt-Hopkins syndrome). Am J Hum Genet. 2007;80:994–1001.
5. Zhuang Y, Cheng P, Weintraub H. B-lymphocyte development is regulated by the combined dosage of three basic helix-loop-helix genes, E2A, E2-2, and HEB. Mol Cell Biol. 1996;16:2898–905.
6. Whalen S, Heron D, Gaillon T, et al. Novel comprehensive diagnostic strategy in Pitt-Hopkins syndrome: clinical score and further delineation of the TCF4 mutational spectrum. Hum Mutat. 2012;33:64–72.
7. Marangi G, Ricciardi S, Orteschi D, et al. Proposal of a clinical score for the molecular test for Pitt-Hopkins syndrome. Am J Med Genet A. 2012;158A:1604–11.
8. Hamdan FF, Daoud H, Patry L, et al. Parent-child exome sequencing identifies a de novo truncating mutation in TCF4 in non-syndromic intellectual disability. Clin Genet. 2013;83:198–200.
9. Kalscheuer VM, Feenstra I, Van Ravenswaaij-Arts CM, et al. Disruption of the TCF4 gene in a girl with mental retardation but without the classical Pitt-Hopkins syndrome. Am J Med Genet A. 2008;146A:2053–9.
10. Kharbanda M, Kannike K, Lampe A, Berg J, Timmusk T, Sepp M. Partial deletion of TCF4 in three generation family with non-syndromic intellectual disability, without features of Pitt-Hopkins syndrome. Eur J Med Genet. 2016;59:310–4.
11. de Ligt J, Willemsen MH, van Bon BW, et al. Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med. 2012;367:1921–9.
12. Grozeva D, Carss K, Spasic-Boskovic O, et al. Targeted next-generation sequencing analysis of 1000 individuals with intellectual disability. Hum Mutat. 2015;36:1197–204.
13. Hamdan FF, Srour M, Capo-Chichi JM, et al. De novo mutations in moderate or severe intellectual disability. PLoS Genet. 2014;10:e1004772.
14. Rauch A, Wieczorek D, Graf E, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet. 2012;380:1674–82.
15. Tan CA, Topper S, Del Gaudio D et al. Characterization of patients referred for non-specific intellectual disability testing: the importance of autosomal genes for diagnosis. Clin Genet. 2016;89:478–483.
16. Geoffroy V, Pizot C, Redin C, et al. VaRank: a simple and powerful tool for ranking genetic variants. PeerJ. 2015;3:e796.
17. Biancalana V, Beldjord C, Taillandier A, et al. Five years of molecular diagnosis of Fragile X syndrome (1997–2001): a collaborative study reporting 95% of the activity in France. Am J Med Genet A. 2004;129A:218–24.
18. Hagerman PJ. The fragile X prevalence paradox. J Med Genet. 2008;45:498–9.

19. Hoyer J, Ekici AB, Endele S, et al. Haploinsufficiency of ARID1B, a member of the SWI/SNF-a chromatin-remodeling complex, is a frequent cause of intellectual disability. Am J Hum Genet. 2012;90:565–72.
20. Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542:433–8.
21. de Winter CF, Baas M, Bijlsma EK, van Heukelingen J, Routledge S, Hennekam RC. Phenotype and natural history in 101 individuals with Pitt-Hopkins syndrome through an internet questionnaire system. Orphanet J Rare Dis. 2016;11:37.
22. Zweier C, Sticht H, Bijlsma EK, et al. Further delineation of Pitt-Hopkins syndrome: phenotypic and genotypic description of 16 novel patients. J Med Genet. 2008;45:738–44.
23. Marangi G, Zollino M. Pitt-Hopkins syndrome and differential diagnosis: a molecular and clinical challenge. J Pediatr Genet. 2015;4:168–76.
24. Sepp M, Pruunsild P, Timmusk T. Pitt-Hopkins syndrome-associated mutations in TCF4 lead to variable impairment of the transcription factor function ranging from hypomorphic to dominant-negative effects. Hum Mol Genet. 2012;21:2873–88.
25. Rosenfeld JA, Leppig K, Ballif BC, et al. Genotype-phenotype analysis of TCF4 mutations causing Pitt-Hopkins syndrome shows increased seizure activity with missense mutations. Genet Med. 2009;11:797–805.
26. Giurgea I, Missirian C, Cacciagli P, et al. TCF4 deletions in Pitt-Hopkins Syndrome. Hum Mutat. 2008;29:E242–51.
27. Kousoulidou L, Tanteles G, Moutafi M, Sismani C, Patsalis PC, Anastasiadou V. 263.4 kb deletion within the TCF4 gene consistent with Pitt-Hopkins syndrome, inherited from a mosaic parent with normal phenotype. Eur J Med Genet. 2013;56:314–8.
28. Rossi M, Labalme A, Cordier MP, et al. Mosaic 18q21.2 deletions including the TCF4 gene: a clinical report. Am J Med Genet A. 2012;158A:3174–81.
29. Steinbusch CV, van Roozendaal KE, Tserpelis D, et al. Somatic mosaicism in a mother of two children with Pitt-Hopkins syndrome. Clin Genet. 2013;83:73–77.
30. Schluth-Bolard C, Labalme A, Cordier MP, et al. Breakpoint mapping by next generation sequencing reveals causative gene disruption in patients carrying apparently balanced chromosome rearrangements with intellectual deficiency and/or congenital malformations. J Med Genet. 2013;50:144–50.
31. Maduro V, Pusey BN, Cherukuri PF, et al. Complex translocation disrupting TCF4 and altering TCF4 isoform expression segregates as mild autosomal dominant intellectual disability. Orphanet J Rare Dis. 2016;11:62.
32. de Pontual L, Mathieu Y, Golzio C, et al. Mutational, functional, and expression studies of the TCF4 gene in Pitt-Hopkins syndrome. Hum Mutat. 2009;30:669–76.
33. Peippo MM, Simola KO, Valanne LK, et al. Pitt-Hopkins syndrome in two patients and further definition of the phenotype. Clin Dysmorphol. 2006;15:47–54.

Francesca MATTIOLI

Identification of Novel Genetic Cause of Monogenic Intellectual Disability

Résumé

La déficience intellectuelle (DI) est une trouble du neurodéveloppement caractérisée par une extrême hétérogénéité génétique, avec plus de 700 gènes impliqués dans des formes monogéniques de DI. Cependant un nombre important de gènes restent encore à identifier et les mécanismes physiopathologiques de ces maladies neurodéveloppementales restent encore à comprendre. Mon travail de doctorat a consisté à identifier de nouvelles causes génétiques impliquées dans la DI. En utilisant différentes techniques de séquençage de nouvelle génération, j'ai pu augmenter le taux de diagnostic chez les patients avec DI et identifié plusieurs nouvelles mutations (dans *AUTS2, THOC6*, etc) et nouveaux gènes (*BRPF1, NOVA2,* etc) impliqués dans la DI. Pour les moins caractérisés, j'ai effectué des investigations fonctionnelles pour valider leur pathogénicité, caractériser les mécanismes moléculaires qu'ils affectent et identifier leur rôle dans cette maladie.
Mes travaux de doctorat permettront d'améliorer et d'accélérer la possibilité d'obtenir un diagnostic moléculaire qui donnera accès à un meilleur suivi et à une meilleure prise en charge pour les patients. Cela permettra également de mieux comprendre les mécanismes physiopathologiques impliqués dans ces troubles neurodéveloppementaux. Ces connaissances aideront éventuellement à identifier de nouvelles cibles thérapeutiques.

Mot clés : Déficience Intellectuelle, NGS, *AUTS2, THOC6, BRPF1, NOVA2*

Résumé en anglais

Intellectual disability (ID) is a group of neurodevelopmental disorders characterized by an extreme genetic heterogeneity, with more than 700 genes currently implicated in Mendelian forms of ID but still some are not yet identified. My PhD project investigates the genetic causes of these monogenic ID by using and combining different NGS techniques. By using this strategy, I reached a relative high diagnostic yield and identified several novel mutations (in *AUTS2, THOC6,* etc) and genes (*BRPF1, NOVA2*, etc) involved in ID. For the less characterized ones, I performed functional investigations to prove their pathogenicity, delineate the molecular mechanisms altered and identify their role in this disease.
Overall, this work improved and provided new strategies to increase the molecular diagnosis in patients with ID, which is important for their healthcare and better management. Furthermore, the identification and the characterization of novel mutations and genes implicated in ID better delineate the implicated pathophysiological mechanisms, opening the way to potential therapeutic targets.

Key words: Intellectual Disability, NGS, *AUTS2*, *THOC6*, *BRPF1*, *NOVA2*