



HAL
open science

Etude de la régulation de l'expression des gènes pour l'interprétation de variations génomiques.

Kévin Cassinari

► **To cite this version:**

Kévin Cassinari. Etude de la régulation de l'expression des gènes pour l'interprétation de variations génomiques.. Biochimie, Biologie Moléculaire. Normandie Université, 2023. Français. NNT : 2023NORMR062 . tel-04621786

HAL Id: tel-04621786

<https://theses.hal.science/tel-04621786>

Submitted on 24 Jun 2024

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.



Normandie Université



THÈSE

Pour obtenir le diplôme de doctorat

Spécialité **ASPECTS MOLECULAIRES ET CELLULAIRES DE LA BIOLOGIE**

Préparée au sein de l'Université de Rouen Normandie

Etude de la régulation de l'expression des gènes pour l'interprétation de variations génomiques.

Présentée et soutenue par
KEVIN CASSINARI

Thèse soutenue le **23/10/2023**
devant le jury composé de

M. JEAN-MICHEL DUPONT	PROFESSEUR DES UNIV - PRATICIEN HOSP., UNIVERSITE PARIS CITE	Rapporteur du jury
MME FLORENCE PETIT	PROFESSEUR DES UNIV - PRATICIEN HOSP., UNIVERSITE DE LILLE	Rapporteur du jury
M. JEROME BOULIGAND	PROFESSEUR DES UNIV - PRATICIEN HOSP., Université Paris-Saclay	Membre du jury
MME ANDREE DELAHAYE-DURIEZ	PROFESSEUR DES UNIV - PRATICIEN HOSP., UNIVERSITE SORBONNE PARIS-NORD	Membre du jury
M. PASCALE SAUGIER-VEBER	MAITRE DE CONF UNIV. - PRATICIEN HOSP., Université de Rouen Normandie	Membre du jury
M. GAEL NICOLAS	PROFESSEUR DES UNIV - PRATICIEN HOSP., Université de Rouen Normandie	Directeur de thèse

Thèse dirigée par GAEL NICOLAS (CANCER AND BRAIN GENOMICS)



Normandie de Biologie Intégrative,
Santé, Environnement



Résumé

Titre. Étude de la régulation de l'expression des gènes pour l'interprétation de variations génomiques.

Au cours des quinze dernières années, les avancées en génomique ont rendu le séquençage complet du génome humain accessible, le positionnant comme une étape clé dans l'identification des causes génétiques des maladies rares. Avec l'expansion de ces analyses pangénomiques, nous sommes confrontés à une augmentation de la découverte de variations dont la signification reste incertaine. L'interprétation détaillée de ces variations nécessite constamment de nouvelles méthodes d'analyse de support. L'étude de l'expression génique constitue une approche clé pour l'interprétation de ces variants, et elle contribue également à améliorer le rendement diagnostique global. Ce travail de thèse explore l'apport des études sur l'ARN, qu'elles soient ciblées ou transcriptomiques, dans le diagnostic des maladies génétiques, tout en proposant des méthodes qui pourraient être intégrées en routine diagnostique. En se focalisant sur la génétique du neurodéveloppement et la neurogénétique, ces travaux ont adopté une double approche. Une première repose sur des analyses ciblées, de type une altération génomique : un transcrit. Dans le contexte de maladies telles que la maladie d'Alzheimer jeune, les calcifications cérébrales primaires ou le syndrome de Cornelia de Lange, des analyses ciblées de l'ARN ont été réalisées. Basées sur le RT-ddPCR ou le séquençage de l'ARN, ces analyses ont permis d'éclaircir des variations dont la signification était auparavant inconnue pour certaines, notamment des variations non codantes, avec un impact fort sur les transcrits de ces gènes, et permettent de proposer des stratégies simplifiées pour une utilisation diagnostique de ces techniques. La seconde partie porte sur un volet transcriptomique avec une approche de type une altération génomique : de multiples transcrits. Cette approche avait pour objectif de révéler une signature transcriptomique pour le syndrome de Cornelia de Lange, dont les variants pathogènes impactent la régulation

d'expression de multiples autres gènes. Nous avons pu identifier une signature transcriptomique des lignées de cellules souches pluripotentes induites (iPSC) modifiées par CRISPR/Cas9 pour intégrer des variations nucléotidiques délétères du gène *NIPBL*, offrant également une meilleure compréhension des bases moléculaires de ce syndrome. Ce travail offre de précieuses perspectives scientifiques et techniques, suggérant des façons d'optimiser l'intégration des tests transcriptomiques dans le paysage génomique actuel.

Mots Clés : Syndrome de Cornelia de Lange, calcifications cérébrales primaires, Maladie d'Alzheimer du sujet jeune, RNA-seq, PCR digitale, biomarqueurs, transcriptomique, ARN, variations non codantes, variations de structure.

Abstract

Title: Study of gene expression regulation for the interpretation of genomic variations

Over the past fifteen years, advancements in genomics have made whole genome sequencing accessible, establishing it as a critical diagnosis tool to identify genetic causes of rare diseases. With the growth of such pan-genomic analyses, we are facing an increasing discovery of variants of uncertain significance. The detailed interpretation of these variants constantly requires the development of new supportive analytical methods. The study of gene expression represents a central approach for interpreting these variants and also plays a critical role in enhancing the overall diagnostic yield. This PhD research delves into the contribution of targeted or transcriptomic RNA studies for genetic disease diagnostics, and proposes methods that could be seamlessly integrated into standard diagnostic routines. By focusing on neurodevelopmental and neurogenetic diseases, this research work adopted a twofold approach. The first one is centered on targeted analyses, one genetic alteration : one transcript. In the context of early-onset Alzheimer's disease, primary cerebral calcifications, or Cornelia de Lange syndrome, targeted RNA analyses were conducted. Leveraging RT-ddPCR or RNA sequencing, combined or not with cellular models developed through CRISPR/Cas9 genome editing, these analyses highlighted variants, including non-coding ones, the significance of some of which was previously uncertain, with a strong impact on the targeted transcript, and suggest streamlined strategies for the diagnostic application of these techniques. The second part of this work is based on a one genomic alteration : consequences on multiple transcripts approach. The main aim was to establish a transcriptomic signature for Cornelia de Lange syndrome. This signature was identified in induced pluripotent stem cell (iPSC) lines edited by CRISPR/Cas9 to incorporate deleterious nucleotide variations in the *NIPBL* gene, and allowed us to provide deeper insights into the molecular underpinnings of this syndrome. This

research furnishes valuable scientific and technical perspectives, suggesting ways to optimize the incorporation of transcriptomic tests within the current genomic landscape.

Keywords: Cornelia de Lange Syndrome, primary familial brain calcifications, early-onset Alzheimer's disease, RNA-seq, Digital PCR, biomarkers, transcriptomics, RNA, noncoding variations, structural variations.

À la mémoire du Professeur Thierry Frebourg,

Remerciements

Je tiens à remercier chaleureusement les membres de mon jury,

À la Professeure Florence Petit, pour avoir accepté d'évaluer mon travail en tant que rapportrice. J'espère que ce dernier pourra renforcer la collaboration déjà solide entre les équipes de génétique de Rouen et de Lille sur ces thématiques du post-exome.

Au Professeur Jean-Michel Dupont, merci d'avoir accepté d'être rapporteur de cette thèse et d'y apporter votre expertise de cytogénéticien.

À la Professeure Andrée Delahaye-Duriez, pour ta présence au sein de ce jury et ton accompagnement comme membre du comité de suivi de la thèse. Tu es l'un des premiers visages que j'ai associés à la génétique dès le début de mes études de médecine, et tu sais également le plaisir que j'ai de voir une membre de l'université Paris 13 être présente pour ce jury.

Au Professeur Jérôme Bouligand, merci d'avoir accepté d'évaluer ce travail, apportant ainsi votre expertise de généticien moléculaire et biologiste cellulaire.

À la Docteur Pascale Saugier-Weber, pour votre indéfectible soutien, de notre première rencontre, il y a 9 ans et jusqu'au suivi de cette thèse, en passant par l'aide à la rédaction de mon mémoire de médaille d'or, qui a amorcé mon activité de recherche. Merci, pour vos conseils et pour nos discussions qui m'aident grandement dans les moments importants.

Au Professeur Gaël Nicolas, pour la direction de cette thèse bien sûr, mais aussi pour tout ce que tu apportes au quotidien par ton savoir, ta gentillesse et ta disponibilité, et ce malgré toutes les missions qui sont les tiennes maintenant. Tu es une source constante de motivation et d'inspiration. Surtout, ne change pas.

Merci à celles et ceux qui m'ont aidé au quotidien pour la réalisation de cette thèse,

À Anne Rovelet-Lecrux, pour ton aide lors des nombreuses manip réalisées lors de cette thèse, toujours dans la bonne humeur !

À Camille Leclezio, pour ta capacité à résoudre les problèmes insolubles et pour anticiper les questions auxquelles je n'avais parfois même pas encore réfléchi.

À Céline Dérambure, Myriam Vezain, ainsi qu'à Nathalie Drouot, Anne-Claire Richard, et Stéphane Rousseau pour l'aide à la génération des données de cette thèse. Ainsi qu'à Olivier Quenez et Sophie Coutant pour leur traitement bioinformatique des résultats.

Un merci tout particulier à l'équipe de génétique clinique pour le recrutement des patients et pour leur expertise : Alice Goldenberg, Anne-Marie Guerrot, Gabriella Vera et Juliette Coursimault. Merci aussi à Lou Grangeon, pour cette belle collaboration et pour la mise en commun de nos compétences respectives.

Merci aussi à François Lecoquierre, pour nos échanges quotidiens, pour tes conseils, tes idées et tes remarques. Je suis très heureux de t'avoir comme collègue et comme ami.

Merci à Andrée Delahaye-Duriez, Vincent Gatinois, Myriam Bernaudin et Henri Gondé, d'avoir été membres de mon comité de suivi de thèse pendant ces quatre années.

Merci aux collaborateurs et collaboratrices d'autres centres qui ont également contribué à ce travail, ainsi qu'aux patients et à leurs familles.

(...Merci à celles et ceux que j'aurais oublié de citer, surtout, venez me voir pour me disputer !)

Merci à mes collègues au quotidien,

À toute l'équipe de génétique du CHU de Rouen, dirigée par Claude Houdayer. Même si ces dernières années ont été éprouvantes, nous avons su rester unis et avancer ensemble.

Un remerciement tout particulier pour le laboratoire de cytogénétique, qui m'a accueilli en tant qu'interne et m'a vu grandir (ou pas !). Je mesure la chance que j'ai de travailler au sein d'une équipe non seulement compétente mais aussi joyeuse et toujours bienveillante. Un merci spécial à Pascal Chambon pour ta patience et tes précieux conseils, à Géraldine Joly-Hélas qui m'a accueilli dans son bureau et continue de me faire aimer les chromosomes, et à Mathieu Castelain, dont les qualités vont bien au-delà de la qualité. Merci enfin au Pr Bertrand Macé, pour son accueil dès le début de l'internat et pour nos longues discussions.

Merci aussi à nos jeunes internes, toujours plus nombreux et motivés, qui alimentent constamment ma passion pour l'enseignement.

Merci à toute l'équipe de l'unité Inserm U1245 et particulièrement aux membres de l'équipe 3, pour votre disponibilité et votre capacité commune à pouvoir résoudre (presque) tous les problèmes.

Merci à mes amis,

Merci à ma famille, en particulier à mes parents, ma sœur et mon frère pour leur soutien inébranlable.

Merci à mes trois filles, Apolline, Héloïse et Capucine, pour la joie qu'elles m'apportent tous les jours et bien sûr merci à Naïma, qui a toujours été là... bien avant tout ça...



Table des matières

Résumé	1
Abstract	3
Remerciements	6
Table des matières	8
Liste des figures	11
Liste des tableaux	12
Liste des abréviations	13
Introduction	16
1. Introduction générale	16
2. L'expression génique et l'organisation chromatinienne	21
2.1. Gènes, transcription et traduction	21
2.2. Eléments régulateurs de la transcription (enhancers et autres éléments).....	22
2.2.1. Les enhancers	23
2.2.2. Les promoteurs.....	24
2.2.3. Les autres éléments régulateurs de la transcription	25
2.3. Organisation chromatinienne et territoires chromosomiques	26
2.3.1. Chromosome et territoires chromosomiques	26
2.3.1.1. La chromatine	26
2.3.1.2. Les territoires chromosomiques.....	28
2.3.2. Notion de TAD et de LAD.....	29
2.3.2.1. TAD (Topologically associating domains).....	29
2.3.2.2. LAD (lamina associated domains).....	33
2.4. Le complexe cohésine	34
2.4.1. Structure et éléments constitutifs du complexe cohésine.....	34
2.4.2. Fonctions du complexe cohésine	36
2.5. Eléments de régulation post-transcriptionnels	37
2.5.1. L'épissage	37
2.5.1. Autres éléments de régulation post transcriptionnelle	40
3. Le syndrome de Cornelia de Lange	42

4. Les techniques d'étude de l'expression génique	46
4.1. Techniques ciblées de l'expression génique	46
4.1.1.1. RT-qPCR.....	46
4.1.2. RT-MLPA et RT-QMPSF.....	48
4.1.2.1. Principe de la QMPSF	48
4.1.2.2. Principe de la MLPA	49
4.1.3. La RT-ddPCR.....	51
4.2. Techniques d'études transcriptomiques	54
4.2.1. Les puces d'expression	55
4.2.2. Le RNAseq.....	56
4.2.2.1. Techniques de RNAseq.....	56
4.2.2.2. L'analyse des données de RNAseq : quantification de transcrit et analyses différentielles	58
4.2.2.3. L'analyse des données de RNAseq : étude de l'épissage	60
5. Bases de données et modèles d'étude.....	62
5.1. Bases de données.....	62
5.1.1. La base GTEx.....	62
5.1.2. La base ENCODE	63
5.1.3. La base de données VISTA.....	64
5.1.4. La base de données Genehancer	66
6. Modèles d'études <i>ex vivo</i> et <i>in vitro</i>	68
6.1. Échantillons issus de patients	68
6.2. Utilisation de modèles cellulaires pour étudier la régulation transcriptionnelle : intérêt de l'édition génomique par CRISPR/Cas	69
7. Problématique.....	71
Résultats - Partie I. Etudes ciblées de la régulation de l'expression des gènes.....	72
1. Haploinsuffisance du gène <i>SLC20A2</i> médiée par la disruption d'un élément régulateur responsable de calcifications cérébrales primaires	73
1.1. Contexte et résumé des travaux.....	73
1.2. Article scientifique	77

2. Caractérisation des conséquences transcriptionnelles d'une tripllication du locus <i>APP</i> dans la maladie d'Alzheimer avec angiopathie amyloïde cérébrale.....	87
2.1. Contexte et résumé des travaux :	87
2.1.1. Article scientifique.....	91
3. Utilisation du RNAseq pour l'interprétation de WGS : caractérisation de l'impact de variations introniques profondes du gène <i>NIPBL</i> dans le syndrome de Cornelia de Lange	97
3.1. Contexte et résumé des travaux	97
3.2. Article scientifique	100
Résultats - Partie II. Apport des études transcriptomiques pour l'étude du syndrome de Cornelia de Lange	116
1. Contexte et résumé des travaux.....	116
2. Article scientifique	122
Discussion.....	156
Conclusion.....	167
Références	168

Liste des figures

Figure 1. Résumé schématique des interactions enhancers – promoteurs.	25
Figure 2. Représentation schématique de l'organisation du nucléosome	27
Figure 3. Représentation en FISH-3D des chromosomes dans un noyau	29
Figure 4. Représentation schématique de l'organisation des TADs	31
Figure 5. Principe résumé du Hi-C.....	33
Figure 6. Anneau constitutif du complexe cohésine	35
Figure 7. Impact du clivage des cohésines sur l'organisation des TADs.....	37
Figure 8. Schéma représentant les modes d'épissage alternatif de l'ARNm précurseur.....	38
Figure 9. Le syndrome de Cornelia de Lange	43
Figure 10. Rappel de la cinétique d'une réaction de qPCR.....	47
Figure 11. Exemple de résultat de QMPSF	48
Figure 12. Principe de la MLPA	50
Figure 13. Intérêt de la PCR digitale pour détecter des évènements rares.	52
Figure 14. Génération des microgouttelles en ddPCR	53
Figure 15. Exemple de Sashimi Plot	61
Figure 16. Résumé des informations fournies par la base Encode.....	64
Figure 17. Stratégie de tests rapporteurs utilisés pour les données de la base VISTA.....	65
Figure 18. Exemple de visualisation des données de Genhancer sur UCSC	67
Figure 19. Plan de réalisation des analyses transcriptomiques	118

Liste des tableaux

Tableau 1. Récapitulatif des signes cliniques évocateurs de syndrome de CdL

Tableau 2. Sujets inclus dans la première série du projet CoSign

Liste des abréviations

- 3-C** : Capture de la conformation chromosomique, *chromosome conformation capture*
- AAC** : Angiopathie amyloïde cérébrale
- ACLF** : Association des cytogénéticiens de langue française
- ANPGM** : Association nationale des praticiens en génétique moléculaire
- ACPA** : Analyse chromosomique sur puce à ADN
- ADN** : Acide Désoxyribonucléique
- ADNc** : ADN complémentaire
- ARN** : Acide ribonucléique
- ARNm** : ARN messenger
- ARNr** : ARN ribosomal
- CCP** : Calcifications cérébrales primaires, *primary familial brain calcification (PFBC)*
- CdLS** : Syndrome de Cornelia de Lange, *Cornelia de Lange Syndrome*
- CGH-array** : Hybridation génomique comparative, *array comparative genomic hybridization*
- CN** : Nombre de copies, *copy number*
- CNV** : Variation du nombre de copies, *copy number variation*
- CRE** : Élément cis régulateur, *cis regulatory element*
- CRISPR** : *Clustered Regularly Interspaced Short Palindromic Repeats*
- Ct** : Cycle seuil, *cycle threshold*
- dPCR** : PCR digitale, *digital PCR*
- ddPCR** : PCR digitale en émulsion, *digital droplet PCR*
- DI/AD** : Déficience intellectuelle et anomalies du développement
- FISH** : Hybridation *in situ* en fluorescence, *fluorescence in situ hybridization*
- Hi-C** : Capture de la conformation chromatinienne à haut débit, *high-throughput chromatin conformation capture*
- Indel** : Insertions-délétions

iPSC : Cellules souches pluripotentes induites, *induced pluripotent stem cells*

LAD : Domaines associés à la lamina, *lamina associated domains*

LNA : Acide nucléique verrouillé, *Locked Nucleotide Acid*

lncRNA : long ARN non codant, *long non coding RNA*

MAJ : Maladie d'Alzheimer du sujet jeune

miARN : micro-ARN

NGS : Séquençage de nouvelle génération, *next generation sequencing*

MLPA : Amplification multiplexe de sondes par ligation, *multiplex ligation-dependent probe amplification*

NMD : Dégradation des ARN non-sens, *nonsense-mediated decay*

PCR : Amplification en chaîne par polymérase, *polymerase chain reaction*

qPCR : PCR quantitative ou PCR en temps réel, *quantitative PCR, real time PCR*

QMPSF : PCR multiplex quantitative de courts fragments fluorescents, *Quantitative multiplex PCR of short fluorescent fragments*

RBPs : Facteurs de liaison à l'ARN, *RNA-binding proteins*

RIN : Mesure de l'intégrité de l'ARN, *RNA Integrity Number*

RNAseq : Séquençage de l'ARN, *RNA sequencing*

RPKM : Lectures par kilobase par million de lectures alignées, *reads per kilo base per million mapped reads*

RT : Retrotranscription

snRNA-Seq : Séquençage de l'ARN de noyaux uniques, *single nuclei RNA sequencing*

SNV : Variation nucléotidique, *Single nucleotide variation*

SV : Variation de la structure chromosomique, *structural variation*

TAD : Domaines topologiques associés, *topological associated domains*

TFBS : Site de liaison des facteurs de transcription, *transcription factor binding site*

TND : Troubles du neurodéveloppement

TPM : Transcrits par million, *transcripts per million*

TSA Troubles du spectre de l'autisme

TSS : Site de démarrage de la transcription, *transcription start site*

UTR : Régions non traduites, *untranslated regions*

WES : Séquençage de l'exome, *whole exome sequencing*

VSI : Variants de signification incertaine

WGS : Séquençage du génome, *whole genome sequencing*

Introduction

1. Introduction générale

Au cours des quinze dernières années, les technologies d'étude du génome humain ont connu des avancées significatives [1]. Le séquençage de nouvelle génération (*Next Generation Sequencing*, NGS) permet ainsi d'étudier un ensemble de gènes associés à une pathologie ou à une entité syndromique (panel de gènes), de séquençer l'exome, c'est-à-dire l'intégralité des parties codantes de notre génome (*Whole Exome Sequencing*, WES, ~ 34 Mb) [2], ou encore de réaliser le séquençage complet du génome humain (*Whole Genome Sequencing*, WGS, ~ 3 Gb) [3]. L'exemple des déficiences intellectuelles (DI) et des anomalies du développement (AD), ou, plus généralement, des troubles du neurodéveloppement (TND), illustre l'intérêt d'une exploration pangénomique en première intention. En effet, le diagnostic étiologique génétique de ces affections, qui affectent 2 à 3% de la population, est très complexe [4]. La complexité s'explique par l'hétérogénéité génétique majeure avec près de 1200 gènes connus à ce jour comme impliqués dans des formes monogéniques de DI [5], ainsi que par les plus de 5000 syndromes rares connus d'origine monogénique qui ont été décrits, parmi lesquels un nombre significatif inclut une DI [6]. S'ajoutent des formes non monogéniques, aujourd'hui encore très difficiles à caractériser sur le plan étiologique à l'échelle individuelle. Le NGS a largement contribué à augmenter le taux de diagnostic dans les DI/AD, avec un rendement pouvant aller jusqu'à 55% en WGS pour les DI sévères [7],[8]. Ces nouvelles technologies de séquençage ont également permis d'identifier les gènes responsables de nombreuses autres maladies rares dans tout le champ de la médecine.

Néanmoins, le séquençage du génome, à lui seul, ne permet pas de surmonter toutes les difficultés associées au diagnostic génétique de ces maladies. En effet, l'interprétation de ces variations à des fins médicales prend en compte de nombreux paramètres, incluant leur

fréquence dans des populations contrôles, leur ségrégation dans la famille, leur effet prédit sur la protéine et les données de la littérature. Pour les variations nucléotidiques (*Single Nucleotide Variation*, SNV) et petites insertions-délétions (*indel*) touchant des gènes associés à des maladies de transmission Mendélienne, ces interprétations font l'objet de recommandations nationales (Harmonisation de l'interprétation de variants de séquence générés par les analyses en NGS, ANPGM, 2017) et internationales [9], [10]. Il en va de même pour les variations de nombre de copies (*Copy Number Variation*, CNV) [11]. Cette stratégie d'interprétation classe les variations nucléotidiques en différentes catégories : bénignes (classe 1), probablement bénignes (classe 2), probablement délétères (classe 4), et délétères (classe 5). Certains variants nucléotidiques ou de structure (incluant les CNV) restent cependant inclassables dans une de ces catégories, ces variants sont nommés variants de signification incertaine (VSI ; classe 3). La proportion de VSI est encore plus importante dans les régions non codantes, incluant les régions introniques ou intergéniques, dont les effets sur la fonction des gènes restent difficiles à déterminer [12].

Ainsi, si l'avènement des techniques pangénomiques que sont le WES et le WGS a considérablement transformé notre capacité à étudier les maladies génétiques, en permettant une analyse approfondie et systématique du génome humain, l'interprétation des VSI demeure un défi majeur pour la génétique médicale. En effet, la fonction de ces variants reste souvent incertaine et leur rôle dans la pathogenèse des maladies génétiques est difficile à déterminer [13]. Pour pouvoir avancer dans la classification de ces variations dites de classe 3, il est nécessaire de combiner les résultats du WGS avec des analyses fonctionnelles, des études d'expression génique et/ou des modèles animaux. De plus, l'intégration de données multi-omiques, telles que la transcriptomique, la protéomique et la métabolomique, permet de mieux comprendre les mécanismes pathogéniques et d'identifier de nouvelles cibles thérapeutiques pour ces maladies [14].

Dans ce contexte, le travail présenté ici s'intéresse au rôle que peut avoir l'étude de la régulation de l'expression des gènes à l'échelle de l'ARN messager (ARNm), pour comprendre les mécanismes qui sous-tendent les variations génomiques et leurs impacts sur les phénotypes observés et aider à reclasser ces variations. Ces explorations, qu'elles soient réalisées par des approches ciblées (*e.g.* RT-ddPCR) ou transcriptomiques globales (*e.g.* RNAseq), contribuent à élucider l'effet et même les mécanismes associés à certaines variations [15], [16]. Au-delà des méthodes d'étude directe du produit de la transcription, l'ARN, de nombreux exemples illustrent l'importance d'étudier les interactions entre les régions cis-régulatrices et les variants génétiques pour moduler l'expression des gènes [17]. L'utilisation de techniques d'édition génomique, telles que CRISPR/Cas9, a également permis de démontrer directement les effets causaux de certains variants sur l'expression des gènes et les phénotypes [18].

L'étude de la régulation de l'expression des gènes est ainsi cruciale pour interpréter les variations génomiques et leurs impacts sur les phénotypes. Néanmoins, il subsiste des défis à relever, tels que la caractérisation des régions non codantes fonctionnelles et l'élucidation des mécanismes spatio-temporels de la régulation de l'expression des gènes [19]. La poursuite de ces recherches permettra d'améliorer notre capacité à interpréter les variations génomiques et, ultimement, à développer des stratégies thérapeutiques plus ciblées et personnalisées pour le traitement des maladies génétiques.

Dans cette perspective, ce travail de thèse se concentre sur l'étude de la régulation des gènes en se focalisant au niveau de l'ARN messager, afin d'interpréter les variations génomiques, nucléotidiques et structurelles. Après avoir introduit les différents concepts et technologies disponibles pour l'étude de la régulation de l'expression des gènes, ce travail sera divisé en deux grandes parties correspondant aux deux angles de vue que nous avons choisi de prendre.

La première partie sera consacrée à illustrer comment l'étude ciblée des niveaux d'ARNm d'un gène, en utilisant des techniques de RT-digital droplet PCR (RT-ddPCR), en combinaison ou non avec des modèles cellulaires ou le NGS, contribue à déterminer la pathogénicité et l'effet fonctionnel d'une variation individuelle. Nous adopterons le point de vue d'une variation affectant un transcrite et/ou un gène. Ainsi, cette première partie abordera (i) l'établissement de la pathogénicité d'une délétion non codante en amont du gène *SLC20A2* dans le contexte des calcifications cérébrales primaires à l'aide de mesures de quantités relatives des transcrits du gène cible dans le sang de patients et de lignées cellulaires HEK modifiées par CRISPR/Cas9, (ii) la caractérisation cytogénétique et transcriptionnelle de la première triplication du gène *APP* décrite à ce jour dans le cadre de la maladie d'Alzheimer et (iii) les conséquences, non plus transcriptionnelles, mais post-transcriptionnelles de mutations *de novo* introniques profondes dans le syndrome de Cornelia de Lange (CdLS) introduisant un néoexon non en phase et responsables d'une dégradation partielle des transcrits.

La seconde partie de ces travaux visera à étudier les conséquences transcriptionnelles multiples résultant d'une unique variation génétique. Plus précisément, nous nous intéresserons aux cas où une variation affecte plusieurs gènes et transcrits. Cette partie sera consacrée au CdLS, qui est considéré comme une transcriptomopathie, c'est-à-dire une maladie monogénique touchant un gène critique dans la régulation de l'expression d'autres gènes. Cette partie abordera le sujet à travers l'outil transcriptomique qu'est le RNAseq. Nous étudierons les conséquences sur l'expression de multiples gènes d'une altération individuelle d'un seul gène, *NIPBL*, gène majeur du CdLS, et la recherche d'une signature transcriptomique associée à *NIPBL*. Nous avons eu pour objectif de rechercher une potentielle signature transcriptomique du CdLS dans le sang de patients et dans des lignées de cellules souches pluripotentes induites (iPSC) dans le but final de pouvoir utiliser les enseignements

des conséquences des altérations de *NIPBL* pour l'interprétation future des variations de signification incertaine.

2. L'expression génique et l'organisation chromatinienne

2.1. Gènes, transcription et traduction

L'expression génique englobe l'ensemble des processus biochimiques et moléculaires par lesquels l'information héréditaire contenue dans un gène est convertie en molécules fonctionnelles qui participent à diverses activités cellulaires, telles que les protéines, synthétisées via des ARN messagers (ARNm), ou tels que les ARN non codants [20]. Ce processus d'expression génique comporte plusieurs étapes interconnectées et finement régulées, permettant une modulation précise de l'expression des gènes en réponse aux besoins et aux conditions spécifiques des cellules [21].

La première étape cruciale de l'expression génique est la transcription, au cours de laquelle l'ADN génomique est lu par l'ARN polymérase pour produire un ARN, qui sera ensuite mûri en ARN messager (ARNm) dans le cas des gènes codant pour des protéines [22]. L'ARNm est ainsi une réplique de l'information génétique codée dans le gène et servira de matrice pour la synthèse de polypeptides lors de la traduction [23]. Cette étape de transcription est régulée par une multitude de facteurs et de co-facteurs de transcription qui interagissent avec les séquences promotrices et les éléments régulateurs de l'ADN qui seront détaillés par la suite [24]. Après la transcription, l'ARN pré-messager (pré-ARNm) subit plusieurs modifications post-transcriptionnelles, incluant notamment le processus fondamental qu'est l'épissage. D'autres modifications post-transcriptionnelles ont également lieu, telles que l'ajout d'une coiffe 5' (*capping*), ou l'ajout d'une queue poly(A) à l'extrémité 3' des ARNm [25],[26]. Ces modifications sont nécessaires à la stabilité et au parcours de l'ARNm ainsi qu'au bon déroulé et à l'efficacité de sa traduction en chaîne polypeptidique, une fois arrivé dans le cytosol [25]. La traduction est l'étape finale de l'expression génique, au cours de laquelle les ARNm matures sont lus par les ribosomes et traduits en chaînes polypeptidiques. Ces chaînes sont ensuite repliées et mûries en protéines fonctionnelles [27]. La traduction

est également finement régulée par les facteurs d'initiation de la traduction ainsi que par des microARN (miARN), qui permettent de contrôler l'expression des gènes en modulant la traduction des ARNm ou en favorisant leur dégradation [28],[29].

L'expression génique est ainsi un processus complexe, dynamique et hautement régulé qui implique la transcription de l'ADN en ARN, des modifications post-transcriptionnelles, et la traduction de l'ARNm en une chaîne polypeptidique, qui, après modifications post-traductionnelles, conduira à des protéines matures.

La quantité d'ARNm peut ainsi être utilisée pour étudier les niveaux d'expression génique, d'où l'intérêt de sa quantification dans un tissu, un état ou un moment donné du développement. Ces mesures permettent d'identifier les gènes qui sont actifs dans une cellule donnée, ainsi que leur niveau d'expression relatif [30]. Elle peut également être utilisée pour comparer les niveaux d'expression génique entre différents types cellulaires, états ou moments du développement. Comme mentionné précédemment, ces niveaux dépendent notamment de l'effet d'éléments régulateurs de l'expression génique et reflètent un équilibre entre la synthèse, la maturation et les modifications post-transcriptionnelles, et la dégradation.

2.2. *Eléments régulateurs de la transcription (enhancers et autres éléments)*

Les séquences régulatrices correspondent à une partie de l'ADN non codant influant sur le niveau de transcription des gènes. Elles sont reconnues par des facteurs de transcription, également appelés facteurs-trans, qui agissent de différentes façons, en augmentant ou en diminuant l'expression du gène. Les séquences régulatrices interviennent ainsi au niveau de l'initiation de la transcription dans la régulation de l'expression des gènes.

2.2.1. *Les enhancers*

Les enhancers sont des éléments cis-régulateurs (*cis regulatory element, CRE*) activateurs, influençant la transcription des gènes voisins, indépendamment de leur orientation et de leur position par rapport au promoteur [31]. Structurellement, les enhancers sont des séquences d'ADN non codantes, généralement de 50 à 1500 paires de bases de longueur [32]. Des facteurs de transcription spécifiques se lient aux enhancers, via des modules de liaison aux facteurs de transcription (*Transcription factor binding site, TFBS*). Ils recrutent des complexes de protéines régulatrices et le complexe médiateur, qui modulent l'activité de l'ARN polymérase sur le promoteur d'un gène voisin. Ces modules permettent la liaison des facteurs de transcription et d'autres protéines régulatrices [33], qui à leur tour modulent l'activité de l'ARN polymérase au niveau du promoteur du gène voisin [31].

Les enhancers peuvent être situés à plusieurs kilobases en amont ou en aval du promoteur cible, incluant des régions intergéniques, mais aussi à l'intérieur des gènes eux-mêmes [34]. Ils jouent un rôle crucial dans la régulation fine de l'expression génique, et leur perturbation par des variations structurales ou nucléotidiques peut entraîner des altérations de l'expression des gènes cibles, et potentiellement conduire à des conditions pathologiques [34], [35], bien qu'il y ait encore relativement peu d'exemples étayés.

L'identification des enhancers demeure un défi majeur, il n'existe en effet pas de consensus sur la manière de les définir et de les caractériser, même si des approches sont proposées, associant l'utilisation de bases de données à des caractérisations fonctionnelles. Par ailleurs, les enhancers sont souvent redondants, c'est-à-dire qu'un même gène peut être régulé par plusieurs enhancers, ce qui rend leur identification et leur caractérisation encore plus complexe [31].

2.2.2. *Les promoteurs*

Les promoteurs sont des régions spécifiques de l'ADN situées immédiatement en amont du site d'initiation de la transcription (*Transcription start site, TSS*) d'un gène. Les promoteurs sont des séquences d'ADN conservées, généralement courtes, qui servent de points d'ancrage pour les facteurs de transcription et l'ARN polymérase. Les promoteurs sont nécessaires pour initier la transcription et assurent que les gènes sont exprimés au bon moment, dans les cellules appropriées et en réponse aux signaux environnementaux [36].

Il existe plusieurs types de promoteurs, parmi eux, on retrouve des promoteurs basaux ou constitutifs, qui sont impliqués dans l'expression de base des gènes, et les promoteurs spécifiques des tissus ou inductibles, qui sont responsables de l'expression génique dans des contextes cellulaires ou environnementaux particuliers [37]. Les promoteurs sont composés de différents éléments, dont la boîte TATA, une séquence consensus présente dans de nombreux promoteurs eucaryotes [38], et les séquences d'initiation de la transcription, séquences conservées entourant le site d'initiation de la transcription [39].

Les interactions entre les facteurs de transcription, les cofacteurs et les promoteurs permettent une régulation précise de l'expression génique. Les facteurs de transcription se lient à des séquences spécifiques d'ADN dans les promoteurs et recrutent l'ARN polymérase et d'autres protéines pour former le complexe de transcription [40], [Figure 1]. Par ailleurs, des mécanismes épigénétiques, tels que la méthylation de l'ADN et les modifications des histones, peuvent influencer la disponibilité des promoteurs pour la liaison des facteurs de transcription et l'initiation de la transcription [41].

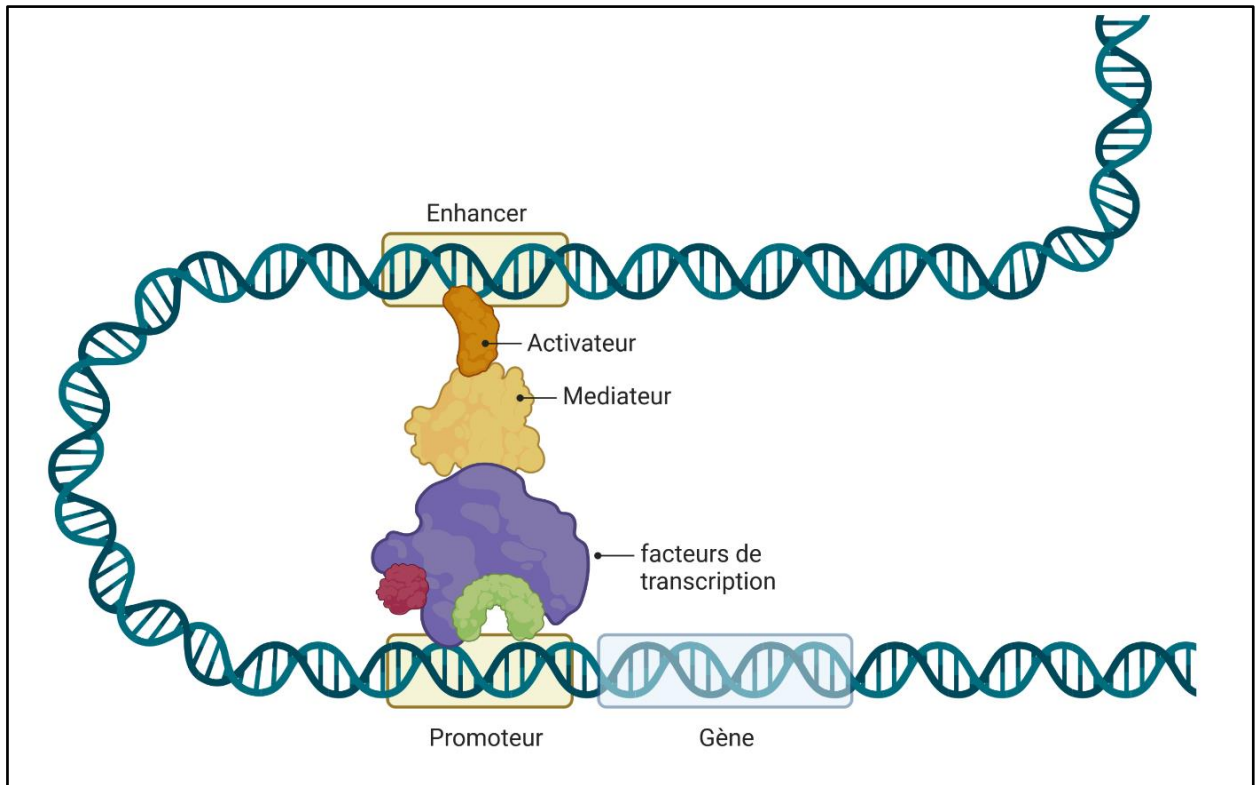


Figure 1. Résumé schématique des interactions enhancers – promoteurs.

2.2.3. Les autres éléments régulateurs de la transcription

Outre les enhancers et promoteurs, d'autres éléments régulateurs sont impliqués dans la régulation de l'expression génique.

Les éléments *silencers* (répresseurs) sont des séquences d'ADN non codantes réduisant ou inhibant l'expression des gènes. Cette modulation peut se réaliser soit via le recrutement de protéines spécifiques, appelées répresseurs qui vont se lier aux séquences d'ADN cibles [42], soit directement par le recrutement du complexe de remodelage de la chromatine qui va modifier la structure de nucléosomes et ainsi moduler la condensation de la chromatine, ce qui affecte son accessibilité et, *in fine*, l'expression du gène.

Les longs ARN non codants (lncRNA), ARN non codants d'une longueur supérieure à 200 nucléotides, jouent un rôle important dans la régulation de l'expression des gènes à différents

niveaux, notamment la régulation transcriptionnelle, post-transcriptionnelle et épigénétique [43].

Les régions de contrôle du locus sont des éléments régulateurs de l'ADN qui coordonnent l'expression de gènes spécifiques situés à proximité, en contrôlant leur activation ou leur répression. Les régions de contrôle du locus sont impliquées dans la régulation de l'expression des gènes à des niveaux élevés et appropriés dans des types cellulaires spécifiques et dans des conditions particulières [44]. Les régions de contrôle du locus sont souvent présentes dans des domaines génomiques qui contiennent des groupes de gènes fonctionnellement liés, tels que les complexes de gènes de la globine.

Enfin, les éléments *insulators* ou isolateurs également nommés « éléments frontières » agissent comme des barrières entre les éléments régulateurs et les promoteurs de gènes adjacents. Ils empêchent ainsi l'influence d'enhancers ou de silencers sur des gènes non cibles et protègent les gènes contre les effets de position. Ces rôles sont assistés par des protéines de liaison aux *insulateurs* parmi lesquelles le *CCCTC-binding factor* (CTCF) qui joue un rôle majeur dans la définition des boucles chromatinienne et la formation des domaines topologiquement associés (*topological associated domains, TAD*), autre composant majeur dans la régulation de l'expression génique.

Beaucoup de progrès ont été réalisés dans la compréhension de ces domaines, nous en décrirons les principaux enseignements dans la seconde partie de l'introduction.

2.3. Organisation chromatinienne et territoires chromosomiques

2.3.1. Chromosome et territoires chromosomiques

2.3.1.1. La chromatine

La chromatine est un complexe macromoléculaire constitué d'ADN, de protéines histones et de protéines non histones. Elle permet l'organisation de la molécule d'ADN et la régulation

de l'expression génique dans les cellules eucaryotes [45]. Fondamentalement, l'ADN s'enroule autour des histones pour former des structures appelées nucléosomes, qui sont à leur tour organisés en fibres de chromatine plus épaisses [46]. Les nucléosomes, unités de base de la structure de la chromatine, sont composés d'un octamère d'histones (comportant deux copies de chacun des histones H2A, H2B, H3 et H4) et d'environ 147 paires de bases d'ADN enroulé autour de l'octamère [46]. Les nucléosomes sont reliés entre eux par des séquences d'ADN appelées liaisons d'ADN (*DNA linkers*), auxquelles se lie l'histone H1, ou histone liaison, stabilisant la structure de la fibre de chromatine et facilitant le compactage ultérieur de l'ADN [47], [Figure 2]. La chromatine, en compactant l'ADN, permet non seulement de loger d'importantes quantités d'ADN dans le noyau des cellules eucaryotes, mais aussi de réguler finement l'expression génique.

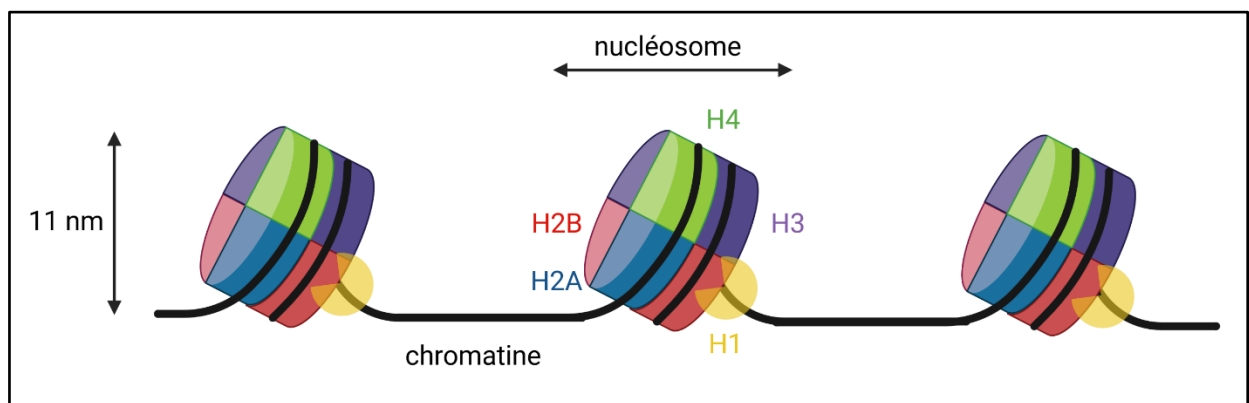


Figure 2. Représentation schématique de l'organisation du nucléosome

Les modifications post-traductionnelles des histones (acétylation, méthylation, phosphorylation, ubiquitination, ...) peuvent moduler la structure de la chromatine en changeant la charge électrique des histones et en affectant leur capacité à se lier à l'ADN, ou encore en recrutant des protéines régulatrices. Ainsi, l'acétylation des histones conduit généralement à un relâchement de la chromatine, favorisant une activité transcriptionnelle accrue. Par contre, la méthylation des histones est souvent associée à une chromatine plus condensée, entraînant une répression de l'expression génique [48]. Par ailleurs, des

modifications de la chromatine sont également possibles via l'action du complexe de remodelage de la chromatine (protéines SWI/SNF) qui va contribuer à modifier la position des nucléosomes le long de l'ADN afin de permettre ou de bloquer la fixation des facteurs de transcription ou d'autres protéines régulatrices [49].

2.3.1.2. Les territoires chromosomiques

Au cours de l'interphase, la chromatine adopte une organisation bien définie, loin d'être aléatoire : Chaque segment d'ADN, correspondant à un chromosome, occupe une région spécifique nommée territoire chromosomique [Figure 3]. Ces territoires chromosomiques sont ainsi des régions distinctes au sein du noyau cellulaire, où les chromosomes résident et s'organisent de manière spatialement spécifique, cette organisation jouant un rôle crucial dans la régulation de l'expression génétique [50]. La position des chromosomes à l'intérieur du noyau est dictée par des interactions dynamiques entre la chromatine et le cytosquelette nucléaire, ainsi que par des interactions avec d'autres éléments constitutifs du noyau [51].

C'est grâce à ces interactions spécifiques que l'organisation tridimensionnelle de la chromatine est modulée. Cette organisation a une influence directe sur l'expression génique en modulant l'accessibilité des gènes et des séquences régulatrices. Elle détermine ainsi la manière dont les facteurs de transcription et les complexes enzymatiques interagissent avec la transcription. Les mécanismes de réplication et de réparation de l'ADN sont également modulés par cette organisation [52].

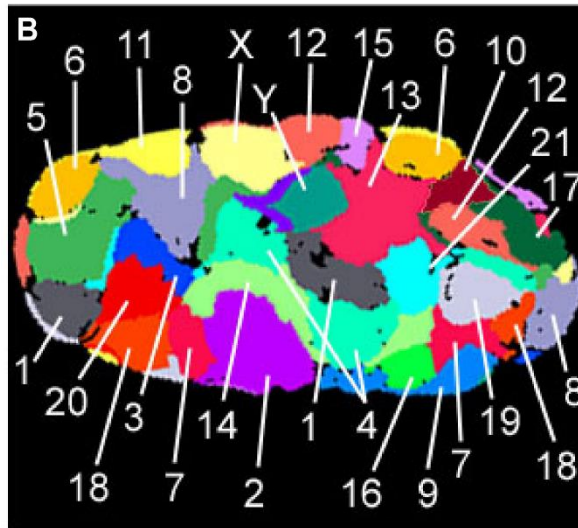


Figure 3. Représentation en FISH-3D des Chromosomes dans un noyau de fibroblaste en phase G0

d'après Bolzer et collaborateurs [53]

Par ailleurs, les interactions entre les territoires chromosomiques contribuent également à la régulation de l'expression génique en permettant une communication entre les différents chromosomes [54] ainsi que des interactions interchromosomiques. Ces interactions favorisent ainsi la formation de complexes multi-protéiques impliqués dans la régulation de l'expression génique, tels que les complexes de co-activateurs et les complexes de répresseurs [55]. La structure et l'organisation des territoires chromosomiques peuvent également être modifiées en réponse à des signaux environnementaux (lumière, stress thermique, ...) ou à des changements d'état cellulaire [51].

2.3.2. *Notion de TAD et de LAD*

2.3.2.1. *TAD (Topologically associating domains)*

Les domaines topologiquement associés (*Topological associated domain*, TAD) sont des unités structurales conservées au sein du génome, marqués par des interactions physiques fréquentes entre les séquences d'ADN qu'ils englobent [56]. Ces domaines ont une importance

capitale dans la régulation de l'expression génique : ils facilitent les contacts entre les promoteurs et les éléments régulateurs distants, comme les enhancers, tout en restreignant les interactions non désirées entre les éléments régulateurs de différents TAD [57]. La cohérence structurale des TAD est assurée par l'action concertée de protéines des familles cohésines et condensines, ainsi que par les protéines CTCF (*CCCTC-binding factor*), lesquelles forment des boucles d'ADN en se liant à des motifs spécifiques localisés dans les régions limitrophes des TAD [58], [Figure 4]. La compréhension de la dynamique et de la fonction des TAD est essentielle pour élucider les mécanismes de la régulation des gènes et les relations entre la structure du génome et les fonctions cellulaires. L'étude approfondie de la dynamique et de la fonction des TAD est une thématique récente mais féconde, visant à élucider les mécanismes qui sous-tendent la régulation de l'expression génique, ainsi que pour établir les liens entre la structure du génome et la fonction cellulaire.

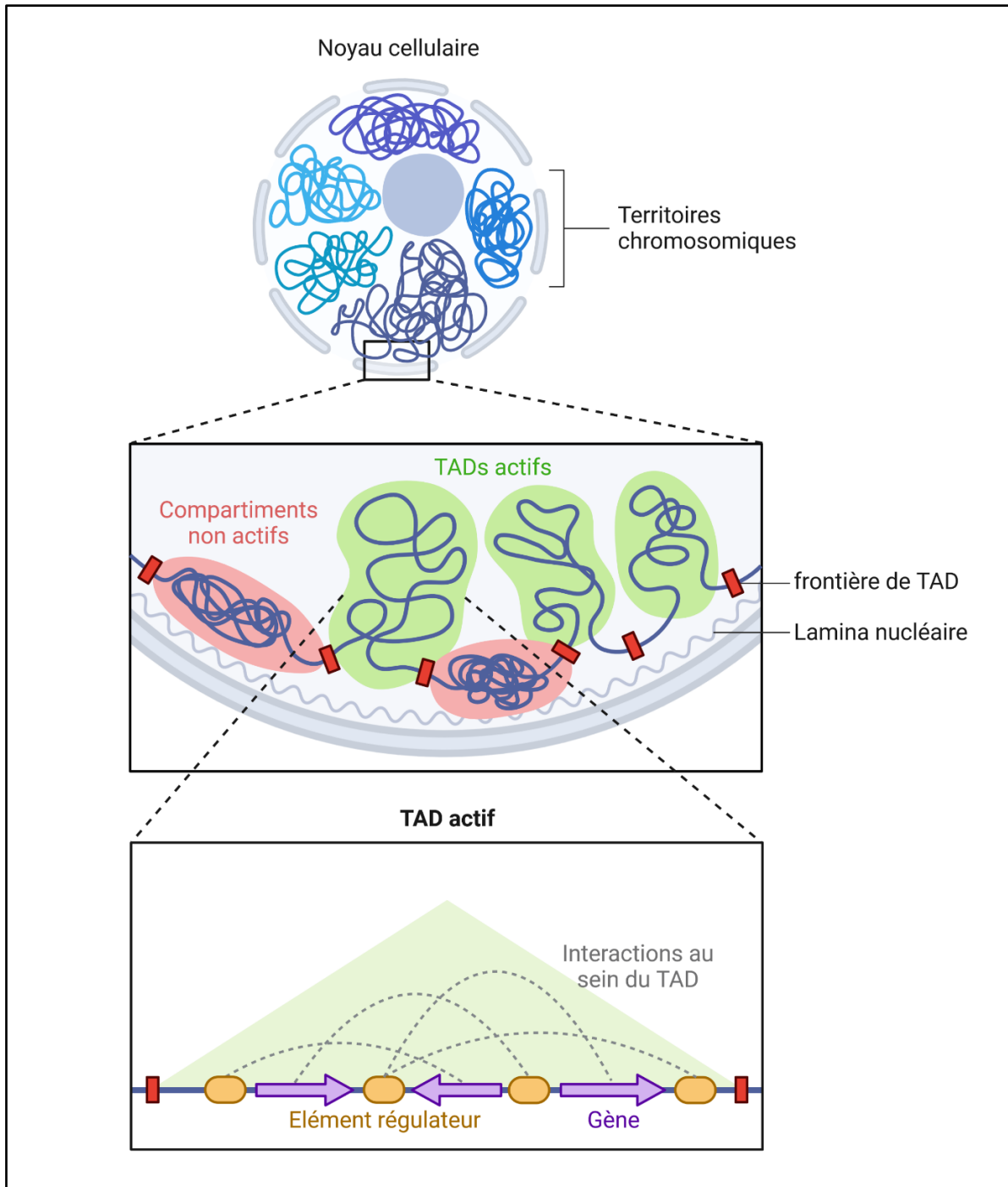


Figure 4. Représentation schématique de l'organisation des TADs

Par ailleurs, les TAD participent également à la stabilité du génome en prévenant les réarrangements chromosomiques et en restreignant les effets des variations nucléotidiques ponctuelles sur l'expression génique [59]. Les variations génomiques qui affectent les limites des TAD peuvent perturber les interactions entre les régions chromosomiques, provoquant

des anomalies dans l'expression génique et, dans certains cas, conduisant à des conditions pathologiques. Ainsi, des variations nucléotidiques ou structurales (délétions, duplications, inversions ou translocations) peuvent avoir des conséquences fonctionnelles sur les gènes situés à proximité [60], [61]. Par exemple, des réarrangements chromosomiques affectant les TAD ont été associés à des maladies du développement, telles que la polydactylie [60].

Les TAD jouent ainsi un rôle dans la régulation de l'expression génétique à plusieurs niveaux. D'une part ils contrôlent l'accessibilité de l'ADN aux facteurs de transcription [56]. Les boucles d'ADN formées par les TAD peuvent permettre ou empêcher l'accès des facteurs de transcription à leurs sites de liaison sur l'ADN, modulant ainsi l'expression des gènes en activant ou en réprimant l'interaction entre les facteurs de transcription et l'ADN. Par ailleurs, la structure des TAD a une influence dynamique sur la conformation et la structure chromatinienne [62]. Ainsi, en modifiant la structure de la chromatine, et donc l'activation de promoteurs différents, les TAD peuvent favoriser ou inhiber la production de différentes isoformes de protéines à partir d'un même gène conduisant à une diversité protéique. Enfin, les TAD peuvent être impliqués dans la régulation de l'expression génétique en contrôlant la formation de domaines d'expression et en permettant la mise en place de régions d'expression différentielle au niveau du génome [55],[62]. Les TAD peuvent ainsi contribuer à la spécificité tissulaire de l'expression des gènes et à la complexité fonctionnelle des organismes multicellulaires.

Cependant, l'association entre une perturbation d'une frontière de TAD et un phénotype lié à la dysrégulation de gènes contenus dans un TAD n'est pas constante, et des recherches plus récentes suggèrent que des interactions *enhancer*-promoteurs, même réduites par la formation d'une néo-frontière de TAD, peuvent rester suffisantes pour garantir l'expression de certains gènes du développement [64]. Ainsi, il n'est pas possible de prédire l'effet d'une variation de structure uniquement via l'impact qu'elle pourrait avoir sur une frontière de TAD.

Les méthodes de cartographie des TADs reposent principalement sur l'utilisation de techniques dérivées de la capture de la conformation chromosomique (*chromosome conformation capture*, 3-C), notamment le Hi-C, qui permet la détection des interactions spatiales entre les régions chromosomiques des interactions chromosomiques à l'échelle du génome entier [Figure 5].

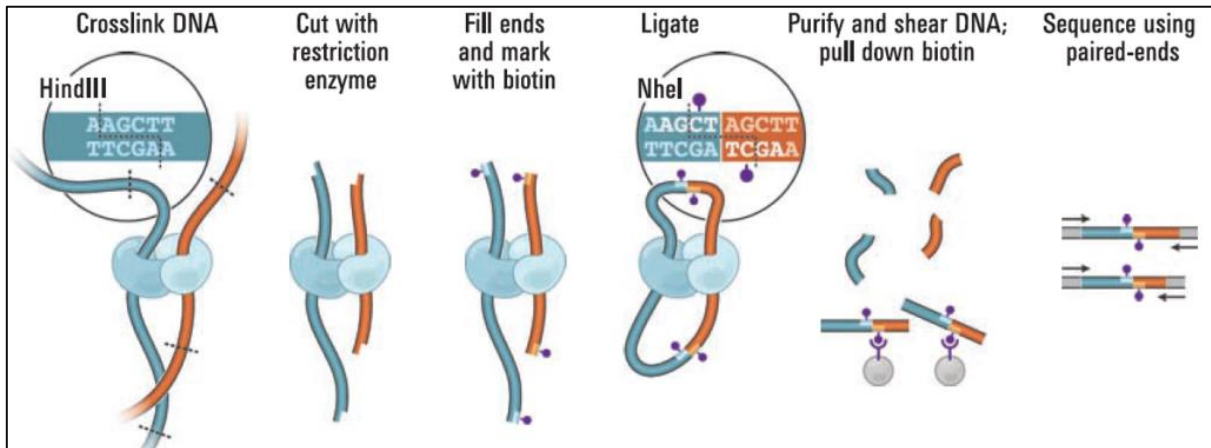


Figure 5. Principe résumé du Hi-C

La technique de Hi-C peut être découpée en 6 étapes : **1.** fixation des interactions chromosomiques : Les protéines et l'ADN sont fixés *in situ* en utilisant un agent de fixation, généralement le formaldéhyde, de manière à figer les interactions physiques entre différentes régions de la chromatine. **2.** digestion de l'ADN pendant laquelle les cellules sont lysées et l'ADN est digéré par une enzyme de restriction. **3.** marquage des extrémités des fragments avec de la biotine, facilitant l'identification des fragments. **4.** Religation des extrémités permettant que les régions d'ADN qui étaient proches dans l'espace tridimensionnel du noyau (et donc en interaction) soient ligaturées ensemble. **5.** Purification de l'ADN pour ne conserver que les fragment biotinilés et **6.** Séquençage des fragments depuis leurs deux extrémités (séquençage paired-end). D'après Lieberman-Aiden et collaborateurs [65].

2.3.2.2. LAD (lamina associated domains)

Les domaines associés à la lamina (*Lamina Associated Domains*, LAD) sont des régions chromosomiques qui interagissent directement avec la lamina nucléaire, structure fibreuse qui tapisse la face interne de l'enveloppe nucléaire et qui joue, elle aussi, un rôle crucial dans la maintenance de la structure du noyau et de la régulation de l'expression génique [66], [67]. Les LAD sont principalement constitués d'hétérochromatine, compacte, méthylée et ainsi

transcriptionnellement inactive. Les LAD ont été initialement identifiés par des techniques de capture de la chromatine en contact avec la lamina nucléaire, telles que la DamID (*DNA adenine methyltransferase identification*) [66]. Ils sont généralement de grande taille, entre 100 kb et 10 Mb. L'implication des LAD dans la régulation de l'expression génique se fait via le recrutement des gènes et des éléments régulateurs à la périphérie du noyau, où l'environnement est moins favorable à la transcription [66],[67]. Par ailleurs, la dynamique des interactions LAD-lamina est régulée au cours du développement et de la différenciation cellulaire, ce qui suggère que les LAD jouent un rôle important dans l'établissement et le maintien des programmes d'expression génique spécifiques aux différents types cellulaires [70]. Les perturbations de l'organisation des LAD et de leur interaction avec la lamina nucléaire ont été associées à plusieurs pathologies, notamment des maladies du vieillissement, des dystrophies musculaires et des cancers [65],[69].

2.4. Le complexe cohésine

2.4.1. Structure et éléments constitutifs du complexe cohésine

Le complexe cohésine est une structure protéique multi-sous-unitaire hautement conservée au cours de l'évolution. Il est composé des protéines SMC1, SMC3, RAD21 et SCC3 (SA1 ou SA2) [Figure 6]. SMC1 et SMC3 sont membres de la famille des protéines de maintenance structurale des chromosomes (*Structural Maintenance of Chromosomes*, SMC). Les protéines SMC ont deux caractéristiques structurelles principales : un domaine « tête » ayant une activité ATPase (formé par l'interaction des extrémités N-terminal et C-terminal) et un domaine « charnière » qui permet leur dimérisation. La tête et le domaine charnière sont reliés l'un à l'autre par des hélices antiparallèles. Ainsi, le socle du complexe cohésine est constitué par un hétérodimère SMC1/SMC3 relié par les domaines charnières.

Le domaine N-terminal de RAD21 contient deux hélices α qui forment un faisceau de trois hélices se reliant à SMC3 et son domaine C-terminal forme une hélice qui se reliera à SMC1, fermant ainsi le complexe lui conférant une structure annulaire. La région centrale de RAD21 est quant à elle en grande partie non structurée, mais contient plusieurs sites de liaison pour les régulateurs de la cohésine, notamment pour SA1 ou SA2. Les interfaces entre les sous-unités SMC et RAD21 peuvent s'ouvrir pour permettre à l'ADN de passer dans et hors de l'anneau de cohésine.

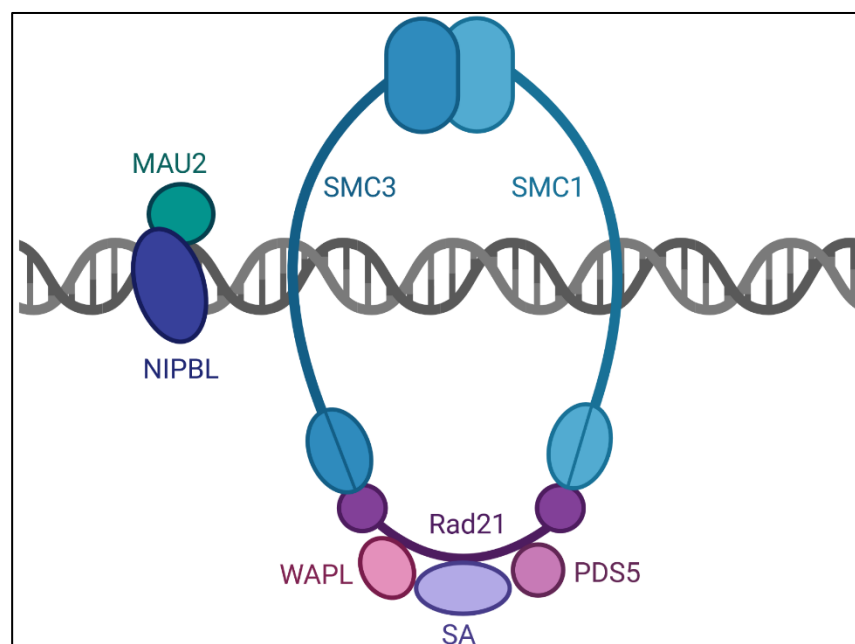


Figure 6. Anneau constitutif du complexe cohésine

En plus de ces composants constitutifs, plusieurs protéines associées au complexe ont été décrites. La première d'entre elles est NIPBL, qui joue le rôle de facteur de chargement du complexe cohésine sur les chromatides sœurs pendant la phase G1 du cycle cellulaire [72]. Pour ce faire, il est nécessaire que NIPBL se lie à la protéine MAU2, présente au niveau de l'anneau. Pendant la phase S, l'établissement de la cohésion est étroitement lié à la réplication de l'ADN et nécessite que les deux chromatides sœurs soient piégées à l'intérieur de l'anneau de cohésine. Cet anneau doit également rester fermé pour éviter la libération prématurée des

chromatides soeurs. Cette fonction est assurée par l'acétylation de SMC3 par l'établissement de la N-acétyltransférase 2 de la cohésion des chromatides soeurs (ESCO2). À la fin de la mitose, l'acétylation de SMC3 induite par ESCO2 pendant la phase S est inversée par l'Histone Déacétylase 8 (HDAC8), les SMC sont recyclées et rechargées sur la chromatine. Ainsi, le cycle de chargement, fonctionnement et déchargement de la cohésine est achevé et prêt à être répété lors du prochain cycle cellulaire.

2.4.2. *Fonctions du complexe cohésine*

Le complexe cohésine a été initialement identifié en raison de son rôle dans le maintien de la cohésion des chromatides soeurs durant la division cellulaire [73]. Depuis lors, plusieurs autres fonctions de ce complexe ont été découvertes, notamment dans la réparation de l'ADN et dans l'organisation et la régulation de l'architecture tridimensionnelle du génome [74]. La préservation de la cohésion des chromatides soeurs par le complexe cohésine est cruciale pour assurer l'intégrité du génome lors de la mitose et de la transmission de l'information génétique aux cellules filles [73]. De plus, le complexe cohésine contribue à la régulation de l'expression génétique en modulant l'accessibilité de la chromatine aux facteurs de transcription [73]. Les cohésines sont particulièrement enrichies aux frontières des TAD, qui sont constituées d'éléments protéiques variés responsables de la séparation des TAD. Cette concentration élevée de cohésines au sein des frontières des TAD suggère leur implication dans l'organisation du génome. Des études ont démontré que le clivage des cohésines entraîne une perturbation de l'architecture interne des TAD, conduisant à la perte d'interaction entre les gènes et les éléments régulateurs et, par conséquent, à une dérégulation de l'expression génique [75], [Figure 7].

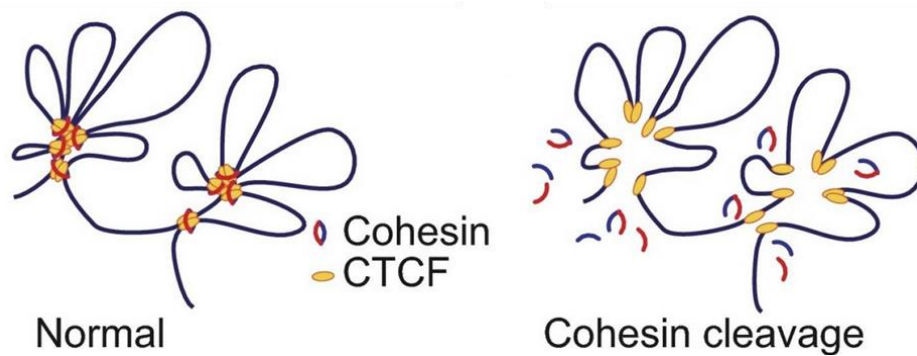


Figure 7. Impact du clivage des cohésines sur l'organisation des TADs

D'après Zuin et collaborateurs [75]

2.5. Eléments de régulation post-transcriptionnels

Avant que l'ARNm ne soit traduit en protéine, il subit plusieurs modifications. Cela inclut l'ajout d'une coiffe 5' et d'une queue poly(A) 3', ainsi que l'épissage des introns. Ces modifications peuvent affecter la stabilité de l'ARNm, son export hors du noyau, et sa traduction en protéine.

2.5.1. L'épissage

L'épissage est un mécanisme biologique essentiel ayant lieu au sein des cellules eucaryotes durant la transformation du pré-ARNm en ARNm mature [26]. Cette étape implique la suppression des introns et l'assemblage des exons en un ARNm fonctionnel. Ce processus nucléaire est catalysé par le spliceosome, complexe ribonucléoprotéique, générant une structure « en lasso » permettant l'élimination des introns [76]. Le phénomène d'épissage alternatif, consistant en un assemblage variable des exons, engendrant plusieurs isoformes d'ARNm, est quant à lui un des moteurs de la diversité protéique, en permettant d'obtenir plusieurs transcrits différents à partir d'un même gène [77], [Figure 9]. L'épissage est rendu possible du fait de la présence de sites d'épissage situés aux jonctions intron-exons, ces sites sont :

- Des sites donneurs d'épissage (ou site d'épissage 5'), situés à l'extrémité 5' de l'intron, marquant le début de la séquence intronique à éliminer
- Des sites accepteurs d'épissage (ou site d'épissage 3') : situés à l'extrémité 3' de l'intron, marquant la fin de la séquence non codante à éliminer
- Site de branchement : situé à l'intérieur de l'intron, généralement à environ 20-50 nucléotides en 5' du site accepteur, participant à la formation de la structure en lasso de l'intron facilitant l'élimination précise des séquences non codantes pour la production d'ARNm fonctionnels.

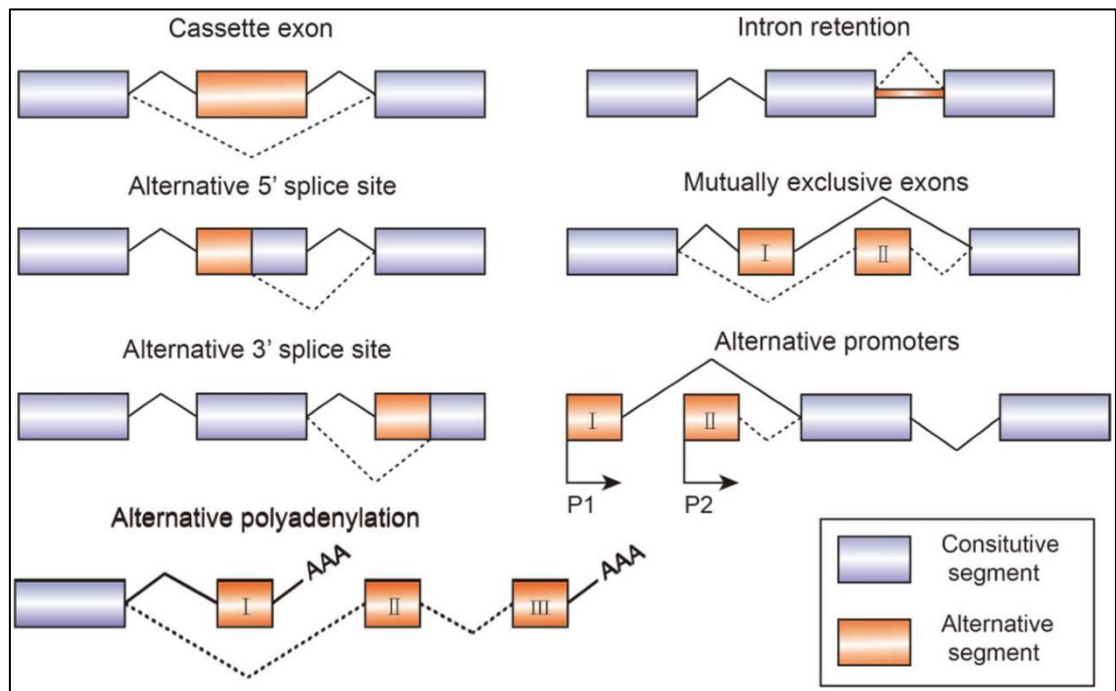


Figure 8. Schéma représentant les modes d'épissage alternatif de l'ARNm précurseur

Exon en cassette, rétention d'intron, site d'épissage 5' alternatif, exons mutuellement exclusifs, site d'épissage 3' alternatif, promoteurs alternatifs et polyadénylation alternative.

D'après Liu et collaborateurs [78]

La présence de variations nucléotidiques ou structurales impactant les sites d'épissage canoniques ou les éléments régulateurs de l'épissage ainsi que les variations conduisant à la création de sites accepteurs ou donneurs d'épissages peuvent donner lieu à des anomalies de

l'épissage. Ces anomalies d'épissage peuvent entraîner la production d'ARNm défectueux ou tronqués, et ultérieurement, de protéines non fonctionnelles ou dysfonctionnelles [79].

Plusieurs types d'anomalies d'épissage ont été identifiées, parmi lesquelles :

- Épissage alternatif aberrant : des erreurs dans la sélection des exons ou des sites d'épissage peuvent entraîner l'inclusion ou l'exclusion inappropriée d'exons dans l'ARNm mature [80].
- L'utilisation de sites d'épissage non canoniques ou cryptiques, généralement liés à des variations nucléotidiques ou structurales, pouvant causer l'insertion ou la suppression de petites séquences dans l'ARNm mature, pouvant potentiellement donner lieu à un décalage du cadre de lecture.
- Épissage en tandem : la répétition d'un exon ou d'une partie d'un exon dans l'ARNm mature peut entraîner la production de protéines anormales avec des fonctions altérées [81].

La plupart du temps, les transcrits aberrants seront alors éliminés par le système de dégradation des ARNm non-sens (*nonsense-mediated decay*, NMD), lorsqu'il existe un codon stop prématuré n'affectant pas le dernier exon ou les 50 dernières bases de l'avant dernier exon, et qui va conduire à une diminution du niveau d'ARNm du gène.

L'épissage est lui-même finement régulé. Des variations nucléotidiques peuvent altérer des sites canoniques ou cryptiques déjà connus, ou en créer de nouveaux. Elles peuvent également modifier la régulation de l'épissage en interagissant avec les facteurs activateurs ou répresseurs de l'épissage.

Des outils tels que SpliceAI (<https://github.com/broadinstitute/SpliceAI-lookup> [82] ou *Splicing Prediction Pipeline* (SPiP) [83], permettent de prédire l'impact des variations génomiques sur l'épissage d'un gène, notamment des prédictions de l'utilisation de nouveaux

sites accepteurs ou donneurs d'épissage. Ces outils sont basés sur la compilation de données issues de multiples autres outils, que nous ne décrivons pas ici.

2.5.1. *Autres éléments de régulation post transcriptionnelle*

Les microARN (miARN), petits ARN non codants d'environ 22 nucléotides de longueur, jouent un rôle crucial dans la régulation post-transcriptionnelle de l'expression des gènes chez les eucaryotes. Les miARNs agissent en se liant spécifiquement aux ARNm cibles, conduisant à leur dégradation ou à l'inhibition de leur traduction en protéines [28]. La synthèse des miRNA résulte de la transcription de gènes spécifiques pour former un pré-miRNA, une structure de plusieurs dizaines de nucléotides. Le pré-miRNA est ensuite clivé par le complexe Drosha-DGCR8 pour former un précurseur de miRNA (appelé pre-miRNA) [84]. Ce pre-miRNA est ensuite exporté du noyau vers le cytoplasme où l'endonucléase Dicer le clive pour produire un petit fragment d'ARN double brin dont l'un des brins deviendra le miRNA mature [85].

Des modifications chimiques spécifiques de l'ARN, telles que la méthylation de l'adénosine en N6-méthyladénosine (m6A), peuvent également jouer un rôle dans la régulation post-transcriptionnelle. Les modifications m6A peuvent affecter la stabilité de l'ARNm, son épissage, sa traduction et sa localisation dans la cellule. Enfin, des facteurs de liaison à l'ARN (*RNA-binding proteins*, RBPs), peuvent se lier à l'ARNm et réguler sa stabilité, son épissage, sa localisation et sa traduction.

La dégradation des ARN non-sens (*Nonsense-mediated RNA decay*, NMD) est un mécanisme de surveillance moléculaire crucial dans les cellules eucaryotes. Son rôle principal est de reconnaître et de dégrader les molécules d'ARNm contenant des codons stop prématurés, évitant ainsi la production de protéines tronquées dont l'effet pourrait être potentiellement

plus néfaste que la diminution de la quantité totale de protéine [86]. Le NMD est activé au décours de la traduction. Bien que ceci soit toujours source de débat [87], le modèle communément admis chez les mammifères est que les transcrits contenant des codons stop prématurés sont détectés en présence d'un complexe de jonction d'épissage (*exon junction complex*, EJC) exon-exon en 3' d'un codon stop, si celui-ci est localisé au moins 50-55 pb en amont (les codons stop prématurés situés dans le dernier exon ou situés moins 50-55 paires de bases en amont de la dernière jonction exon-exon ne déclenchent donc typiquement pas de NMD). Le complexe de terminaison de traduction (facteurs eRF1, eRF3 et du GTP) interagit alors avec la protéine UPF1, l'un des principaux effecteurs du NMD [88]. UPF1, en interaction avec d'autres protéines NMD, dont UPF2 et UPF3, facilite la dégradation de l'ARNm contenant un codon stop prématuré. Cette dégradation se fait par l'action d'exonucléases qui dégradent l'ARNm de l'extrémité 5' vers l'extrémité 3' ou de l'extrémité 3' vers l'extrémité 5', ou par l'endoclivage [89].

3. Le syndrome de Cornelia de Lange

Nous avons donc vu que le complexe cohésine et son chargeur *NIPBL* jouent un rôle important dans l'organisation dynamique du génome. Des variants pathogènes de plusieurs gènes de ce complexe ou de *NIPBL* sont responsables du syndrome de Cornelia de Lange (CdLS), une maladie développementale rare. Cette maladie nous a servi de modèle dans le cadre de ces travaux et nous résumons ici les principaux éléments de caractérisation phénotypique et le spectre mutationnel associé au CdLS.

Le CdLS (MIM #122470) est une maladie génétique rare qui affecte environ 1 naissance sur 50 000 [90] et constitue un syndrome malformatif cliniquement reconnaissable, associant plusieurs signes:

- Un trouble du neurodéveloppement, incluant un retard de développement et/ou un déficit intellectuel (DI) de sévérité variable, allant de sévère dans les formes classiques à une absence de DI dans les formes légères. Les patients présentent également des troubles du comportement similaires aux troubles du spectre autistique (TSA), de l'agressivité, des auto-mutilations et des troubles psychiatriques à l'âge adulte.
- Une dysmorphie faciale cliniquement reconnaissable associant une microbrachycéphalie, des sourcils arqués, un synophris, des fentes palpébrales étroites, un ptosis, de longs cils, un nez court aux narines antéversées, un philtrum long, effacé et bombant, une lèvre supérieure fine, une bouche aux coins tombants, un palais ogival, des éperons mandibulaires, un microrétrognatisme et des oreilles dysplasiques, bas implantées et en rotation postérieure.
- Une microcéphalie (90% des patients).
- Une hypertrichose avec des cheveux épais et bas implantés (80% des patients)
- Un retard de croissance pré et postnatal (95% des patients)

Les patients atteints de CdLS présentent également des anomalies des extrémités, allant de malformations réductionnelles sévères des membres supérieurs (25%) à des anomalies mineures des doigts. Parmi les anomalies des membres supérieurs, on peut citer l'agénésie ulnaire, la monodactylie, l'ectrodactylie, les oligodactylies, la brachymétopie du 1^{er} rayon, la micromélie, l'implantation proximale des pouces et la clinodactylie bilatérale des 5^{èmes} doigts. Les membres inférieurs sont plutôt caractérisés par de petits pieds et une syndactylie II-III (> 80%). D'autres organes sont fréquemment atteints, comme l'audition, les fentes palatines, les malformations oculaires, cardiaques, rénales et génitales.

Le CdLS est associé à une variabilité phénotypique, avec des formes sévères, modérées ou atypiques. Les manifestations et la sévérité de la maladie varient au cours du temps chez les patients.

La sévérité est variable tant sur le plan des malformations que sur le plan du handicap. Si la reconnaissance clinique des formes typiques est aisée, le diagnostic est parfois difficile dans les formes plus modérées [Tableau 1]. Ainsi, il est possible de distinguer le CdLS classique du CdLS modéré dont le phénotype est plus discret. Les patients présentant un CdLS modéré ont une dysmorphie faciale similaire à la forme classique mais une atteinte cognitive moins sévère ainsi que des anomalies des membres plus discrètes [91], [92], [Figure 9].



Figure 9. Le syndrome de Cornelia de Lange. (a) Anomalies des membres et (b) dysmorphie faciale.

D'après Deardorff [93]

	CdL Classique	CdL modéré	Fréquence
Malformations des membres	<u>Des membres supérieurs :</u> - de l'absence complète d'avant-bras aux formes variables d'oligodactylies (30%) - si absence de malformation sévère : micromélie, pouce implanté de façon proximale, clinodactylie du V ^{ème} doigt (100%) - Synostose radio-ulnaire <u>Des membres inférieurs :</u> - petits pieds - syndactylie II-III (> 80%)	Premier métacarpien court ou pouce implanté de façon proximale	100%
Dysmorphie	Synophris, sourcils arqués, longs cils Oreilles bas-implantées en rotation postérieure Racine du nez large ou déprimée Pointe du nez retroussée avec narines antéversées Philtrum long et peu marqué Lèvre supérieure fine, bouche aux coins tombants Palais ogival (30%), petites dents espacées Micrognathie (80%) Eperons mandibulaires (42%)		> 95%
Microcéphalie	< 2DS		100 %
Retard psychomoteur/ déficience intellectuelle	DI sévère à profonde	DI légère à modérée Troubles des apprentissages	> 95%
Retard de croissance	Prénatal et Post-natal		> 95%
Hypertrichose	Cheveux épais Cuir chevelu étendu aux régions temporales +/- visage Pilosité des oreilles, du dos et des bras		> 80 %

Tableau 1. Récapitulatif des signes cliniques évocateurs de syndrome de CdL

Sur le plan moléculaire, il existe une hétérogénéité génétique avec actuellement 6 gènes connus : *NIPBL*, *SMC1A*, *SMC3*, *HDAC8*, *RAD21* et *BRD4*. Environ 70 % des CdLS sont expliqués par des mutations d'un de ces 6 gènes. Parmi les patients atteints de CdLS avec confirmation moléculaire, 60 % portent une mutation dans le gène *NIPBL*, responsable d'un CdLS de transmission autosomique dominante [70],[73] principalement par haploinsuffisance, dans le cadre de variations nucléotidiques perte de fonction ou de variations structurales, ou bien des variations faux-sens entraînant une perte de fonction de la protéine. Deux à trois pourcent des cas sont expliqués par des mutations au sein des gènes *SMC3* [92] et *RAD21* [95], responsables d'un CdLS également transmis sur un mode autosomique dominant. Environ 10 % des cas sont expliqués par des mutations au sein des gènes *SMC1A* [92] et *HDAC8* [96], responsables d'un CdLS dont la transmission est liée au chromosome X. Enfin, le gène *BRD4* a été plus récemment décrit en 2018 [97], et également associé à un phénotype spécifique et reconnaissable [98]. Ces 6 gènes codent pour des protéines formant ou interagissant avec le complexe cohésine. Ainsi, le CdLS appartient au groupe des cohésinopathies, qui sont les pathologies liées à des anomalies de ce complexe cohésine [99]. S'il est actuellement possible d'expliquer 70 % des CdLS par des mutations de ces 6 gènes, il en résulte qu'environ 30 % des diagnostics cliniques de CdLS n'ont encore aucune base moléculaire identifiée. Le conseil génétique est alors impossible pour les familles concernées. Depuis 2011, le laboratoire de génétique moléculaire du CHU de Rouen réalise le diagnostic moléculaire du CdLS avec un recrutement national. Actuellement, la stratégie utilisée est le séquençage en parallèle par NGS des 5 premiers gènes impliqués dans le CdLS (*NIPBL*, *SMC1A*, *SMC3*, *HDAC8* et *RAD21*) et d'une liste de gènes de diagnostics différentiels appartenant au spectre des transcriptomopathies.

4. Les techniques d'étude de l'expression génique

Comme indiqué précédemment, l'évaluation du niveau d'ARNm constitue la méthode la plus simple pour explorer l'expression génique à un moment précis dans un tissu spécifique. Ces niveaux d'ARNm reflètent à la fois la production d'ARN et les modifications post-transcriptionnelles, ces dernières altérant potentiellement l'épissage ou la quantité globale de l'ARN (microARNs, NMD, etc.). De fait, il devient possible d'observer une baisse de production d'ARNm due à la perturbation de son mécanisme régulateur (suppression d'un élément régulateur, diminution des interactions entre les éléments régulateurs et leur cible, etc.), tout comme la réduction du niveau d'ARNm anormaux, potentiellement liée au NMD, mais également des augmentations d'expression, par exemple dans le contexte d'une duplication complète ou de la disruption d'une frontière de TAD. Bien que les analyses quantitatives ne permettent pas d'identifier précisément le mécanisme responsable d'une baisse ou d'une augmentation de la quantité d'ARNm dans un tissu donné, l'interprétation devra tenir compte d'autres éléments d'évaluation disponibles. Ces éléments incluent naturellement la nature de l'altération de l'ADN génomique et les éventuelles analyses fonctionnelles menées sur des modèles cellulaires. Ces analyses, souvent semi-quantitatives, sont réalisées à partir de l'ADN complémentaire (ADNc), produit par la rétrotranscription in vitro des ARN. On distingue généralement les approches ciblées, dérivées de la PCR, qui permettent d'étudier l'expression d'un gène ou d'un petit groupe de gènes, des approches globales, ou transcriptomiques, qui autorisent l'étude de l'ensemble des transcrits ou transcriptome.

4.1. Techniques ciblées de l'expression génique

4.1.1.1. RT-qPCR

La RT-qPCR (Reverse Transcription Quantitative PCR) est la méthode la plus couramment utilisée à travers le monde pour étudier de manière ciblée les niveaux d'ARNm d'un gène

spécifique. Cette technique offre la possibilité de surveiller en temps réel (à chaque cycle de PCR) l'amplification d'une région spécifique d'un ADNc par PCR. Elle permet également de déterminer un nombre de cycles seuil (*Cycle Threshold*, Ct) à partir duquel le produit de la PCR devient détectable par rapport au bruit de fond. Pour ce faire, la RT-qPCR recourt à des sondes fluorescentes spécifiques à la région à amplifier, ou à des intercalants fluorescents de l'ADN, qui sont non spécifiques, mais capables de détecter l'ADN double brin en cours d'amplification à chaque cycle [Figure 10]. Cela permet de suivre la quantité de produits amplifiés en temps réel. Le Ct est inversement proportionnel au logarithme décimal du nombre initial de copies de l'ADNc source, facilitant ainsi l'estimation de la quantité initiale d'ADNc. Cette estimation peut se faire soit par comparaison à des ADNc témoins, soit de manière relative par rapport à un gène de référence. Ces gènes de référence sont généralement choisis parmi des listes de gènes de ménage, qui ont, en théorie, une expression ne subissant pas de régulation spécifique au cours de la vie cellulaire car ils assurent des fonctions indispensables à la cellule et donc sont considérés comme stables entre les individus prélevés dans les mêmes conditions pour un tissu donné. Malgré sa popularité, cette technique a des limites. En particulier, elle peut être affectée par des réactions non spécifiques qui génèrent du bruit de fond, ce qui nécessite l'exécution de multiples répliques et de courbes d'étalonnage. De plus, la RT-qPCR peut parfois échouer à détecter des événements rares.

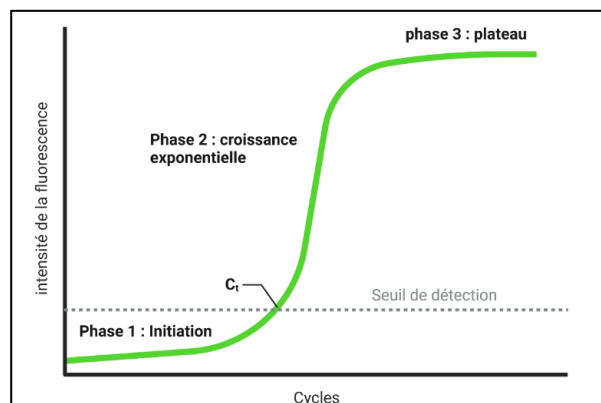


Figure 10. Rappel de la cinétique d'une réaction de qPCR

4.1.2. RT-MLPA et RT-QMPSF

4.1.2.1. Principe de la QMPSF

La PCR multiplex quantitative de fragments fluorescents courts (*Quantitative Multiplex PCR of Short fluorescent Fragments*, QMPSF) [100], développée à Rouen, repose sur l'amplification simultanée (multiplex) de plusieurs loci à l'aide d'amorces dont l'un des deux membres du couple est fluorescent. Ces amplifications sont réalisées sur des fragments de petite taille afin d'harmoniser les conditions d'amplification entre les amplicons. La réaction est arrêtée au cours de la phase exponentielle et le produit d'amplification est analysé dans un séquenceur capillaire. Les résultats seront présentés sous forme de pics dont la hauteur sera corrélée à la quantité initiale d'ADN et comparés à un ADN témoin [Figure 11].

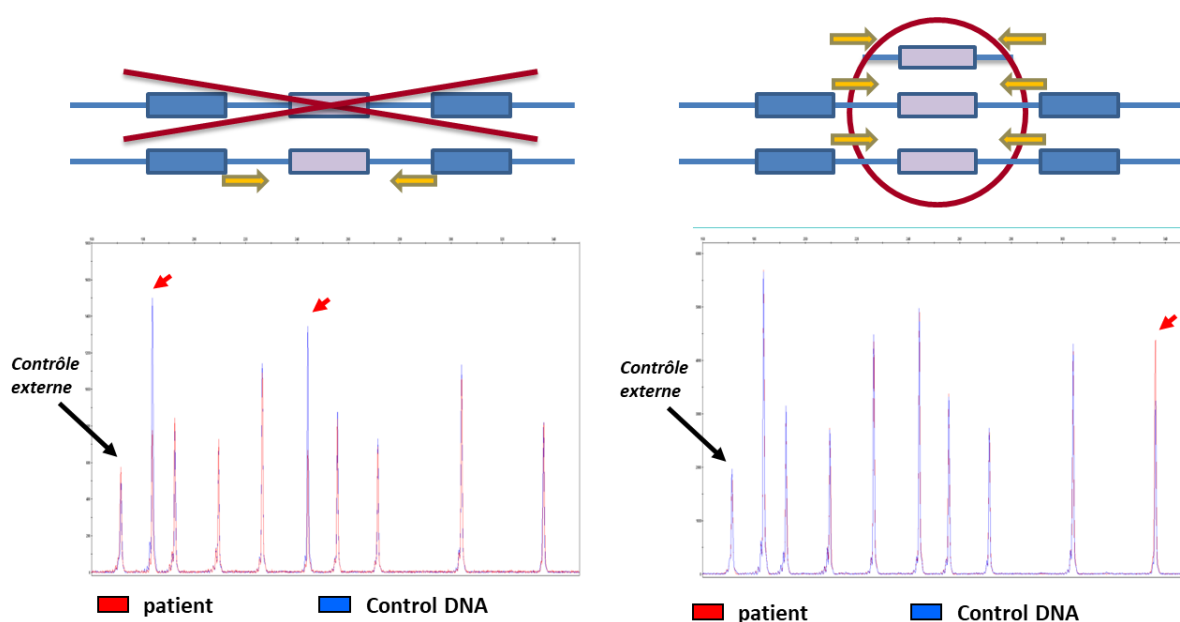


Figure 11. Exemple de résultat de QMPSF

Présentation des résultats de QMPSF : à gauche, les deux flèches rouges indiquent une délétion hétérozygote sur l'ADN du patient et, à droite, la flèche rouge indique une duplication hétérozygote sur l'ADN du patient. Sur les deux graphiques, un contrôle externe est également inclus dans l'analyse (flèche noire).

Une fois l'analyse validée, la QMPSF présente l'avantage d'une grande reproductibilité et permet également de multiplexer les réactions. Cette technique peut être utilisée à partir d'ADN génomique pour la recherche de variation de nombres de copies d'un ou plusieurs

locus chromosomiques et est particulièrement utile pour l'étude de remaniements récurrents [101]. La QMPSF est connue essentiellement pour cette application où elle concurrence la MLPA, également très répandue (cf paragraphe suivant). Néanmoins, la QMPSF peut également être utilisée pour quantifier l'expression génique de gènes cible. Dans le cadre de la RT-QMPSF. Elle est alors simplement précédée d'une réaction de reverse transcription (RT). Le choix des amplicons de référence devra porter une attention particulière au niveau d'expression attendu du ou des transcrits cibles, afin que le ou les gènes de ménage utilisés soient dans les mêmes ordres de grandeur que les transcrits d'intérêt, cette contrainte n'existant pas pour l'ADN génomique où deux copies sont attendues pour les loci de référence. Plusieurs témoins devront également être utilisés afin de contrecarrer la variabilité inter-individuelle, comme pour de nombreuses techniques de mesure d'expression ciblée. La RT-QMPSF reste néanmoins assez peu utilisée. Elle a, par exemple, été appliquée pour une première étude de l'impact des duplications ciblées du gène *APP* sur son expression [102], pour caractériser l'impact d'une variation de la région 3'UTR de ce même gène [103], ou pour tenter de caractériser l'effet d'une délétion intronique du gène *BACE2* sur l'expression de ce dernier [104]. Cette technique est également à la base d'un test fonctionnel mesurant l'activité du gène *TP53* à travers l'expression de ses gènes cible [101],[102].

4.1.2.2. Principe de la MLPA

La MLPA (*Multiplex Ligation-Dependant Probe Amplification*) permet la détection simultanée de variations de nombre de copies sur jusqu'à 60 locus différents en utilisant un seul couple d'amorces de PCR [107]. Contrairement à une PCR multiplex conventionnelle où chaque région est amplifiée avec une paire d'amorces différentes, ici tous les fragments sont amplifiés en utilisant le même couple d'amorces de PCR. La particularité de la MLPA est que ce n'est pas l'échantillon d'ADN qui est amplifié, mais des sondes MLPA qui sont ajoutées à l'échantillon, il s'agit d'une PCR universelle, réduisant les biais liés à l'efficacité de la PCR

utilisant des couples d'amorces différents. Les sondes de MLPA sont constituées de deux oligonucléotides appariés : l'un contenant la séquence d'amorce pour la PCR universelle, et l'autre une séquence complémentaire de la séquence cible d'ADN. Ces deux sondes oligonucléotidiques s'hybrident à des sites cibles immédiatement adjacents. Ce n'est que lorsque les deux sondes oligonucléotidiques sont hybridées à leur cible qu'elles peuvent être liguées en une seule sonde, contenant à la fois les séquences d'amorce sens et antisens. Ces sondes liées sont amplifiées exponentiellement au cours de la réaction de PCR, tandis que les oligonucléotides et sondes non liées ne le sont pas. Le nombre de produits de ligation de sondes dépend donc du nombre de séquences cibles dans l'échantillon, et les profils d'amplification sont comparés à un ADN témoin [Figure 12].

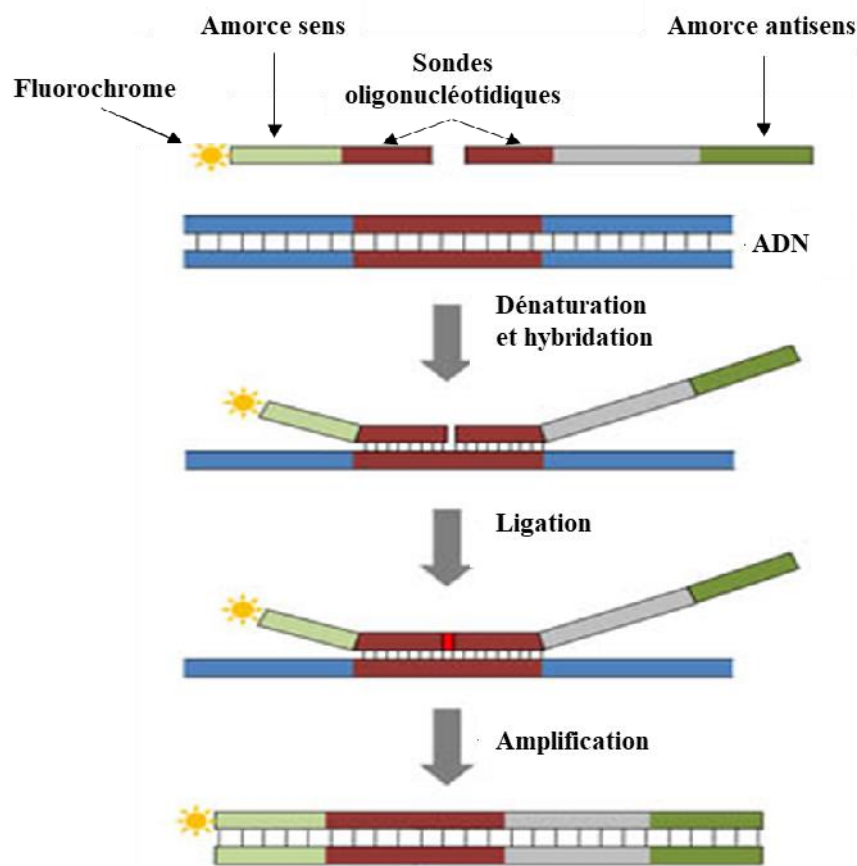


Figure 12. Principe de la MLPA

Il existe un nombre important de panels de sondes de MLPA qui sont utilisés dans des indications très variées et il est possible de créer des sondes à façon. La MLPA est plus robuste et plus reproductible que la qPCR classique et permet une lecture plus rapide des résultats avec un coût inférieur sur de grandes séries. A l'instar de la RT-QMPSF, cette technique est parfaitement utilisable pour étudier l'ARN, notamment dans le cadre de la caractérisation de transcrits, application qui fait l'objet de développement au sein de notre unité depuis plusieurs années, avec l'avantage de disposer de larges possibilités de multiplexage pour un coût relativement réduit [104],[105]. Plus récemment, un couplage avec une étape finale de NGS et la mise au point de sondes spécifiques des jonctions exon-exon a été développé dans le laboratoire, permettant une vision globale à la fois de l'expression et de l'épissage d'un gène cible [110].

4.1.3. *La RT-ddPCR*

Dans ces travaux, j'ai utilisé la reverse transcription digital PCR (RT-dPCR) pour les analyses ciblées. La RT-dPCR repose sur le principe de la PCR digitale, qui est elle-même une évolution des méthodes classiques de PCR. Cette dernière permet d'amplifier et de quantifier directement des acides nucléiques tels que l'ADN, l'ADNc, ou l'ARN.

Le développement initial de la PCR digitale et ses évolutions ultérieures sont liés au besoin de disposer d'une technique permettant d'étudier des événements génomiques rares, notamment dans le cadre de mosaïques à très faible taux, d'ADN tumoral circulant ou, plus tard, d'ADN fœtal circulant, pour lesquels les techniques de PCR classiques n'étaient pas adaptées.

Le principe, qui allait devenir la base de la PCR digitale, a été proposé pour la première fois par Sykes au début des années 1990 [111]. Il consistait à réaliser des dilutions en série d'un échantillon d'ADN jusqu'à obtenir une dilution dite limite, où chaque échantillon contient 0

ou 1 molécule d'ADN. Par la suite, une PCR en point final est réalisée dans chaque partition, permettant d'analyser séparément les différents compartiments. Cette approche permet d'appréhender des événements rares, qui ne se retrouveraient que dans certains compartiments, alors qu'ils ne seraient pas identifiables en PCR quantitative [Figure 13]. Dans ces conditions de dilution limite, on peut dénombrer le nombre total de compartiments positifs, permettant ainsi de déterminer les quantités initiales d'ADN. Cette approche est améliorée et présentée en 1999 [112] et appliquée à une nouvelle méthode d'analyse nommée PCR digitale (*digital PCR*, dPCR).

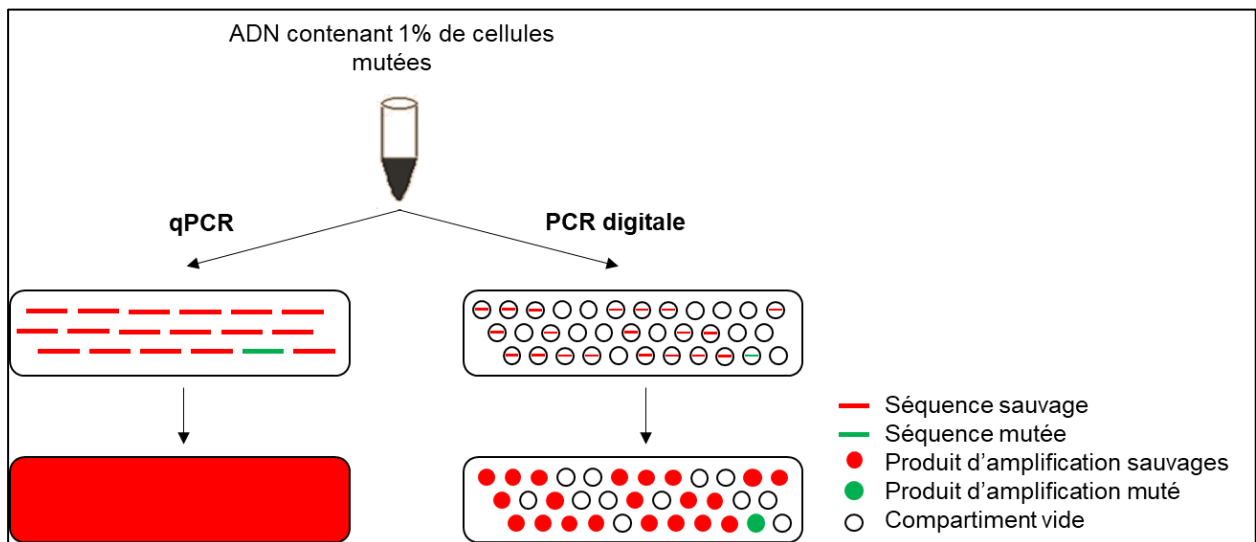


Figure 13. Intérêt de la PCR digitale pour détecter des événements rares.

La PCR digitale en émulsion (digital droplet PCR, ddPCR) consiste à fractionner le mélange de PCR et l'ADN cible en gouttelettes aqueuses qui sont émulsionnées dans de l'huile. Chaque gouttelette représente ainsi un compartiment indépendant. Cette approche permet de générer un nombre de compartiments largement supérieur à ce qui a été décrit précédemment. Théoriquement, ce nombre dépend uniquement du volume des gouttelettes et de celui de l'échantillon [113]. Cette méthode permet de réaliser des tests d'une sensibilité accrue et d'un coût réduit par rapport aux techniques de microchambres [110],[111].

Plusieurs variantes de ddPCR ont été développées. Certaines s'appuient sur la création de gouttes par agitation mécanique, ce qui génère des gouttes de tailles variables [115], pouvant ainsi conduire à des biais dans la répartition de l'ADN entre les gouttelettes. D'autres approches, notamment celles utilisant des billes magnétiques (*BEAMing*) [116], ont par la suite été proposées. Ces techniques sont très précises et sensibles, mais complexes et difficiles à adapter en routine clinique [113].

Le développement des systèmes microfluidiques a amélioré la ddPCR, en facilitant la production rapide de microgouttelettes homogènes, aisément manipulables dans des microcanaux [Figure 14].

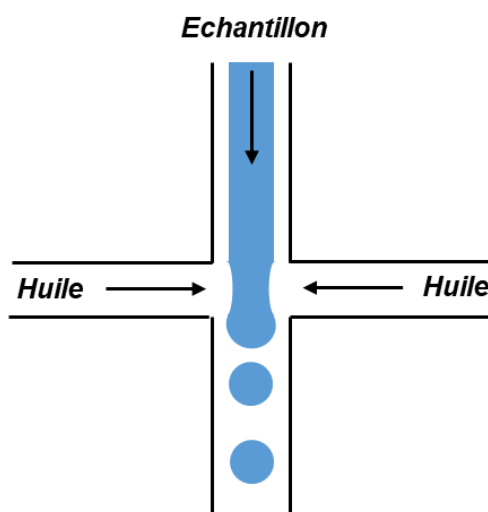


Figure 14. Génération des microgouttelettes en ddPCR

Comme pour la qPCR, les réactions positives sont détectées avec un agent intercalant (par exemple, EvaGreen®) ou avec des sondes d'hydrolyse et quantifiées de manière relative ou absolue. La ddPCR augmente la sensibilité de détection pour des anomalies à faible taux, indétectables par les techniques qPCR classiques.

Les analyses figurant dans ces travaux de thèse ont été réalisées sur une plateforme de PCR digitale en émulsion BioRad®. Cette technologie consiste en une compartimentation de

l'échantillon en 20 000 gouttelettes d'environ 5 nanolitres et permet une lecture de 2 fluorochromes différents simultanément, en analyse duplex. J'ai acquis de l'expérience avec la plateforme ddPCR et la conception de tests diagnostiques associés, notamment lors de la confirmation des CNV rares pour ma thèse d'exercice, mon activité au laboratoire de cytogénétique du CHU de Rouen, et ma recherche à l'unité Inserm 1245 avant le début de ma thèse [117], pour la détection de Sars-CoV-2 dans des prélèvements salivaires ayant une charge virale faible [118], ou, plus récemment, pour le diagnostic génétique de l'amyotrophie spinale infantile (article en cours de rédaction).

Dans ces travaux, la technique de RT-ddPCR a été utilisée via une approche de quantification relative du niveau d'ARNm d'un gène par rapport à celui d'un gène de ménage. Ces analyses sont réalisées soit directement à partir d'ARNm (RT-ddPCR one-step) ou à partir d'ADN complémentaire généré préalablement par une reverse transcription indépendante (RT-ddPCR two-step). Dans les deux cas, le principe est de réaliser une RT-ddPCR biplex, amplifiant à la fois le gène cible, marqué par un fluorochrome (généralement FAM) et le gène de ménage, marqué par un autre fluorochrome (généralement VIC ou HEX). Une fois l'analyse réalisée, il sera possible de calculer un ratio entre la quantification obtenue pour le gène cible et celle obtenue pour le gène de ménage (toutes deux déterminées à partir du nombre de gouttelettes positives). Cette méthode permet de mesurer l'ARNm dans un échantillon et, par exemple, d'évaluer comment une variation impacte l'expression génique en comparant à des ARNm de témoins.

4.2. Techniques d'études transcriptomiques

Deux méthodes principales sont classiquement utilisées pour étudier le transcriptome : les puces d'expression (micro-arrays) et le séquençage NGS de l'ARN (*RNA sequencing*, RNAseq).

4.2.1. *Les puces d'expression*

La base du fonctionnement des puces à ADN repose sur la capacité intrinsèque de l'ADN à s'hybrider à ses séquences complémentaires [119]. Ces méthodologies permettent une quantification relative des acides nucléiques et leur concept peut être extrapolé à une variété d'applications, y compris l'étude de l'expression génique. Dans un protocole standard, les ARNm sont d'abord extraits du tissu biologique d'intérêt, puis convertis en ADN complémentaire (ADNc) par le biais d'une réaction de rétrotranscription. Par la suite, cet ADNc est marqué avec un fluorochrome, typiquement la Cyanine 3 (Cy3, vert). Simultanément, un ADNc témoin peut être marqué avec un fluorochrome d'une couleur différente, généralement la Cyanine 5 (Cy5, rouge). Ces cibles marquées sont ensuite déposées sur une puce à ADN, laquelle contient des oligonucléotides complémentaires à l'ADNc de l'ensemble des gènes de l'organisme étudié. Chaque brin d'ADNc s'hybride à sa sonde complémentaire sur la puce, formant un duplex sonde/cible à double brin. Postérieurement, la puce à ADN est lavée pour éliminer les brins d'ADNc non hybridés avant d'être analysée par un scanner à haute résolution. Un logiciel analyse ensuite l'image scannée, attribuant une valeur d'intensité à chaque sonde de la puce, proportionnelle à l'expression du gène concerné. Les puces d'expression permettent ainsi l'analyse simultanée de l'expression de milliers de gènes, offrant une vue d'ensemble de l'état transcriptionnel à un moment donné. Elles peuvent rapidement générer une grande quantité de données, ce qui serait impossible à réaliser avec des méthodes ciblées tout en restant relativement flexibles, pouvant être conçues pour cibler des gènes spécifiques d'intérêt. Enfin, la possibilité de marquer différents échantillons avec des fluorochromes distincts permet une comparaison directe de l'expression génique entre différentes conditions. Cependant, les puces à ADN peuvent manquer de sensibilité pour la détection des gènes faiblement exprimés et peuvent également donner des résultats faussement positifs en raison de l'hybridation non spécifique. Par ailleurs, les puces d'expression sont généralement conçues en fonction des séquences d'ADN connues, ce qui ne

les rendent pas adaptées à la découverte de nouveaux gènes ou isoformes d'ARN. C'est pour cette raison que leur utilisation est actuellement en forte diminution, avec une tendance à leur remplacement par le RNAseq.

4.2.2. *Le RNAseq*

4.2.2.1. Techniques de RNAseq

Le séquençage de l'ARN (RNAseq) est une application de NGS permettant l'identification et la quantification des ARN d'un échantillon biologique. Dans toute expérience NGS, la première étape est de constituer une bibliothèque des fragments à séquencer. Diverses méthodes de préparation des bibliothèques sont à disposition, chacune ayant ses avantages et ses limites, et étant adaptée à un type spécifique d'étude, voici les principales :

- RNAseq à l'échelle du transcriptome entier (*total RNAseq*) : Dans cette approche, tous les types d'ARN présents dans un échantillon sont séquencés, y compris les ARNm, les ARN non codants et les ARN ribosomiaux (ARNr). L'ARN total est converti en ADNc et intégré dans la librairie. Cette méthode offre une vue globale de l'ensemble du transcriptome, notamment des ARNr, éliminés par les autres méthodes, mais est beaucoup moins sensible pour détecter les ARNm. Le RNAseq total est notamment utilisé pour des applications de métagénomique mais est peu compatible avec une utilisation pour l'étude de l'expression génique compte tenu du faible pourcentage des ARNm séquencés.
- Déplétion en ARNr : Cette technique élimine les ARNr, qui constituent une proportion significative (>95%) des ARN totaux. Une sonde spécifique des ARNr est utilisée pour les capturer et les éliminer. Après l'élimination de ces ARNr, les ARN restants sont convertis en ADNc et intégrés dans la bibliothèque. Cette technique permet de conserver tous les ARN non ribosomiques dans les bibliothèques, y compris les ARNm et les différents ARN non codants. Toutefois, l'augmentation du nombre de

cibles de séquençage nécessite une capacité de séquençage plus élevée pour assurer une couverture suffisante des ARNm.

- **Sélection Poly(A) :** Cette méthode repose sur la sélection des ARNm à l'aide d'oligonucléotides poly(T) complémentaires à la queue poly(A) présente à l'extrémité 3' des ARNm. Les ARNm capturés sont ensuite convertis en ADNc et intégrés dans la bibliothèque. Cette méthode de capture cible uniquement les ARNm, la rendant idéale pour étudier les niveaux d'expression géniques, surtout lors d'analyses d'expression différentielle. Néanmoins, elle requiert des ARN de bonne qualité, peu fragmentés, et souffre ainsi d'un biais 3', c'est-à-dire une sur-représentation des régions 3' des ARNm.
- **Séquençage spécifique des petits ARN (*small RNA sequencing*) :** Cette technique est spécialement conçue pour l'étude des petits ARN non codants, comme les microARN (miARN), les petits ARN interférents (siARN) et les piARN. Les petits ARN sont ligaturés à des adaptateurs spécifiques, convertis en ADNc, puis amplifiés par PCR pour la construction de la bibliothèque.
- **Capture de l'exome :** Cette méthode, relativement récente, vise à capturer spécifiquement les exons, après reverse transcription des ARNm par RT-PCR avec amorces poly(T). Elle s'appuie sur l'utilisation de kits de capture similaires à ceux utilisés pour la préparation des bibliothèques de séquençage de l'exome entier (WES). Des sondes complémentaires aux exons sont utilisées pour capturer les ADNc correspondants. Les ADNc capturés sont ensuite intégrés dans la bibliothèque avant le séquençage. Cette approche permet d'obtenir un nombre important de lectures par transcrit, et est donc particulièrement utile pour étudier les variations et diversité d'épissage et les variations dans les régions codantes des gènes. L'utilisation de ce type de RNAseq a récemment été mis en place au sein de notre équipe, dans le cadre de la confirmation et la caractérisation de variations nucléotidiques détectées par

séquençage de génome de long fragment (*long read sequencing*) dans le cadre d'anomalies du développement [120].

Le choix de la technique de préparation des bibliothèques dépend ainsi de l'objectif de l'étude, que ce soit l'exploration globale du transcriptome, l'étude de l'expression génique, l'identification de petits ARN non codants, ou encore l'analyse de variants exoniques et des variants d'épissage.

Une fois la bibliothèque préparée via une de ces méthodes, les échantillons sont séquencés sur séquenceur NGS, avec un nombre de reads ciblés par échantillons dépendant du mode de préparation des librairies et du type d'analyse (*e.g.* 30 à 60 millions de *reads* pour une analyse différentielle par capture polyA).

Les reads obtenus sont ensuite alignés sur un génome ou un transcriptome de référence, permettant la quantification des divers types d'ARN présents dans l'échantillon.

Le séquençage d'ARN (RNAseq) peut être scindé en deux branches d'investigation principales : la quantification des transcrits, permettant notamment les analyses différentielles, et l'analyse de l'épissage alternatif. Ces deux dimensions de l'étude du transcriptome fournissent des informations différentes mais toutes deux importantes que ce soit pour la compréhension de la complexité fonctionnelle du génome et de la régulation génétique ou pour l'interprétation et la caractérisation de variations génomiques.

4.2.2.2. L'analyse des données de RNAseq : quantification de transcrit et analyses différentielles

La quantification des transcrits constitue une approche simple en conceptualisation mais essentielle pour évaluer l'abondance relative de divers transcrits dans un échantillon biologique. Cette méthode consiste à déterminer le nombre de lectures de séquençage qui se

chevauchent à chaque position d'un gène spécifique, offrant ainsi une estimation de l'expression génique.

Pour être utilisable, cette quantification nécessite une normalisation du nombre des lectures pour permettre des comparaisons précises de l'expression des gènes entre les échantillons. Parmi les mesures de normalisation les plus couramment utilisées on retrouve les lectures par kilobase de million de lectures mappées (*Reads per kilo base per million mapped reads*, RPKM) et les transcrits par million (*Transcripts per million*, TPM). Le RPKM est une mesure de la densité de séquençage normalisée [121], il corrige le nombre brut de lectures mappées sur un gène pour la longueur du gène et pour le nombre total de lectures mappées dans l'échantillon selon la formule $RPKM = (10^9 * C) / (N * L)$ où, C est le nombre de lectures mappées sur le gène, N est le nombre total de lectures mappées dans l'échantillon, L est la longueur du cDNA du transcrit en bases (ou la longueur de l'isoforme)). Le RPKM permet ainsi une comparaison directe de la transcription des gènes entre les échantillons. Les TPM [122] sont une évolution du RPKM, normalisant également le nombre de lectures à la fois sur la longueur du transcrit et sur le nombre total de lectures, mais alors que le RPKM consiste à normaliser d'abord sur la longueur du transcrit puis pour le nombre total de lectures, le TPM fait l'inverse. Le calcul des TPM se fait *i.* en divisant, pour chaque gène, le nombre de lectures correspondant à ce transcrit par sa longueur en kilobases, ce résultat définit le nombre de lecture par kilobase (*read per kilobase*, RPK) du gène. Ensuite *ii.* En additionnant l'ensemble des RPK de l'ensemble des gènes et en le divisant par 1 000 000 pour obtenir un facteur d'échelle « par million ». Enfin, *iii.* En divisant le RPK d'un gène par ce facteur d'échelle « par million », ce résultat sera la valeur du TPM pour le gène donné. Ainsi, la somme de l'ensemble des TPM de l'ensemble des gènes exprimé dans un échantillon donné est toujours la même, ce qui autorise une meilleure représentation de la proportion de transcrits par rapport

au RPKM et facilite la comparaison entre les échantillons. Le TPM est la valeur de normalisation dont l'utilisation est recommandée par le consortium GTEx.

Plusieurs outils peuvent être utilisés pour effectuer ce décompte de lectures, tels que *high-throughput sequencing* (HTSeq) [123] ou *RNASeq by Expectation Maximization* (RSEM) [124]. Cependant, l'outil le plus couramment utilisé à ce jour, et celui que nous avons employé dans nos études, est *Salmon* [125]. Celui-ci recourt à un modèle probabiliste pour estimer l'abondance des transcrits, prenant notamment en compte les biais de contenu en GC.

Ces outils de décompte permettent de générer, pour chaque gène, une quantification des lectures, normalisées en RPKM et TPM. Ces matrices de décompte peuvent ensuite être utilisées pour visualiser de manière ciblée le niveau d'expression d'un gène dans un échantillon spécifique, en vue d'une comparaison avec des échantillons témoins. De plus, ces matrices servent de données source pour réaliser des analyses différentielles, qui seront abordées plus loin dans ce travail.

4.2.2.3. L'analyse des données de RNAseq : étude de l'épissage

Contrairement aux analyses par puce à ADN, le RNAseq offre, en plus de la quantification des transcrits, une analyse qualitative et quantitative de l'épissage. De ce fait, l'utilisation du RNAseq permet d'identifier et de quantifier les divers isoformes de transcrits issus de l'épissage des pré-ARNm. Le RNAseq autorise également une caractérisation directe de l'impact des variations nucléotidiques ou structurales influençant l'épissage, qu'il s'agisse de la création ou de la suppression de sites donneurs ou accepteurs d'épissage, et de leurs conséquences telles que les rétentions introniques ou exoniques.

Ces analyses peuvent être réalisées de manière ciblée, notamment lorsqu'il s'agit de caractériser une variation candidate sur un gène précis. Pour ce faire, une méthode simple

consiste à visualiser les données d'alignement de RNAseq (*Binary alignment files*, BAM) sur un visualisateur de génome tel que l'*Integrative Genome Viewer* (IGV) [126]. Il est notamment possible d'utiliser les *sashimi plots*, qui sont des représentations graphiques permettant de visualiser les jonctions d'épissage [Figure 15].

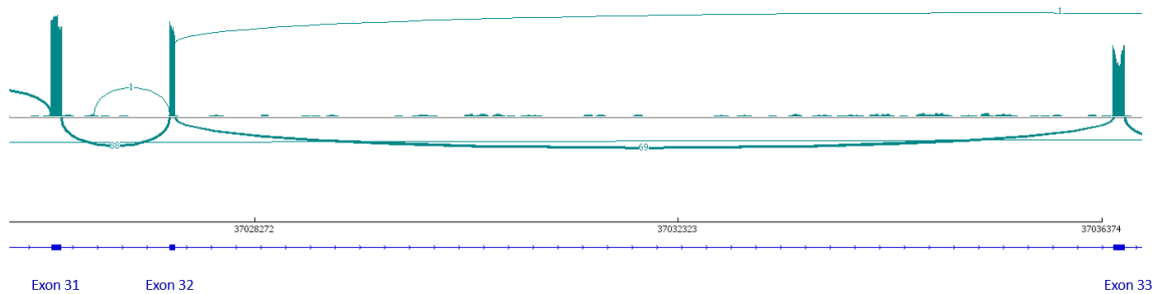


Figure 15. Exemple d'image de Sashimi Plot

Cette représentation permet de visualiser et de quantifier les jonctions d'épissage (traits fins bleus reliant les exons) et d'identifier d'éventuels épissages alternatifs. (Exemple de visualisation des exons 31 à 33 du gène NIPBL à partir d'un RNAseq issu du projet CoSign).

Par ailleurs, un certain nombre d'approches computationnelles ont à ce jour été développées pour identifier et quantifier l'épissage alternatif de gènes à partir de données RNAseq et sur l'ensemble du transcriptome, et ainsi potentiellement identifier les variations qui peuvent en être responsables. Ces outils présentent des performances diverses et les études réalisées pour les évaluer, et les comparer suggèrent qu'il peut être pertinent de les associer au sein d'un pipeline d'analyse [127] il est par exemple possible de citer, parmi les plus récents, SPLICE-q [128], SpliceTools [129], Bisbee [130] ou FRASER [131].

5. Bases de données et modèles d'étude

5.1. Bases de données

La préparation et l'analyse de données transcriptomiques ainsi que l'interprétation de conséquences des variations nucléotidiques et structurales sur l'expression génique nécessitent le recours à des bases de données fiables et largement validées. Ces bases de données sont nombreuses et en perpétuelle évolution. Nous ne détaillerons ici que les principales et/ou celles utilisées pour ces travaux de thèse.

5.1.1. La base GTEx

La principale base de données est sans doute celle du projet *Genotype-Tissue Expression* (GTEx), initié en 2013 et qui se poursuit encore aujourd'hui. L'objectif central de ce consortium est de créer une ressource publique exhaustive dédiée à l'étude de l'expression et de la régulation des gènes spécifiques à chaque type de tissu. À ce jour, cette base de données contient des échantillons provenant de 54 types de tissus sains, collectés chez près de 1000 individus. Pour chacun de ces échantillons, une large gamme d'analyses moléculaires a été réalisée, incluant le RNAseq, le séquençage de l'exome entier (WES) et le séquençage du génome entier (WGS). Le portail GTEx (<https://gtexportal.org/home/>) offre ainsi l'accès à un large éventail de données d'expression pour l'ensemble des gènes, catégorisées en fonction des tissus et même des types cellulaires. Ce portail permet notamment de visualiser rapidement le niveau d'expression d'un gène donné dans un tissu spécifique, facilitant ainsi le choix du tissu le plus adapté pour l'étude de l'expression de ce gène. De plus, il aide à sélectionner un gène de ménage approprié pour une analyse ciblée de l'ARN, à évaluer la qualité d'un résultat de RNAseq ou à diverses autres applications relatives à l'étude du niveau d'expression des gènes. Comme déjà précisé, le portail GTEx utilise, et préconise, l'utilisation des TPM pour la quantification d'expression des gènes, facilitant les comparaisons entre les gènes, entre les tissus et même entre les bases de données. À partir de ces données, le

consortium a pu enrichir son portail pour permettre la mise à disposition de données de niveau d'expression moyen de chaque gène à travers différents tissus, voire même différents types cellulaires [132] ainsi que des données de séquençage sur cellules uniques grâce à la technique de séquençage d'ARN à noyau unique (*single nuclei RNA sequencing*, snRNA-Seq, applicable à du tissu congelé, contrairement au single-cell RNAseq) [132]. La base de données de GTEx s'est également intéressée à la question de la régulation fine de l'expression des gènes, notamment à travers la caractérisation fine des variations d'expression génique via des cartographies de locus de caractères quantitatifs (*quantitative trait locus*, QTL) ou la mise en évidence de gènes d'expression tissu spécifique [133]–[136]. Toutes ces données sont complétées par des images histologiques ainsi qu'une bio-banque permettant la mise à disposition des échantillons utilisés.

5.1.2. *La base ENCODE*

La base ENCODE (*ENCyclopedia Of DNA Elements*) est probablement, avec la base GTEx, la plus connue [137]. Elle fournit un catalogue des éléments fonctionnels du génome humain, et les intègre dans le but de générer des processus d'annotation et de visualisation complets. Les données sont disponibles sur le portail (<https://www.encodeproject.org/>) [138] et également intégrées à la plupart des navigateurs de génome dont celui de l'université de Californie Santa Cruz (UCSC) [139].

Les principales données d'ENCODE concernent le registre des éléments cis-régulateurs candidats (*candidate cis-Regulatory Elements*, cCREs), qui permet d'identifier des enhanceurs, promoteurs et zones riches en CTCF au moyen de la combinaison de la détection de marques épigénétiques (H3K4me3 pour les régions promotrices et H3K27ac pour les enhanceurs), l'identification des sites CTCF par ChipSeq, ainsi que celle des sites sensibles à la DNase.

Ces informations, obtenues à partir de multiples types cellulaires, sont intégrées pour générer un catalogue de cCREs [Figure 15].

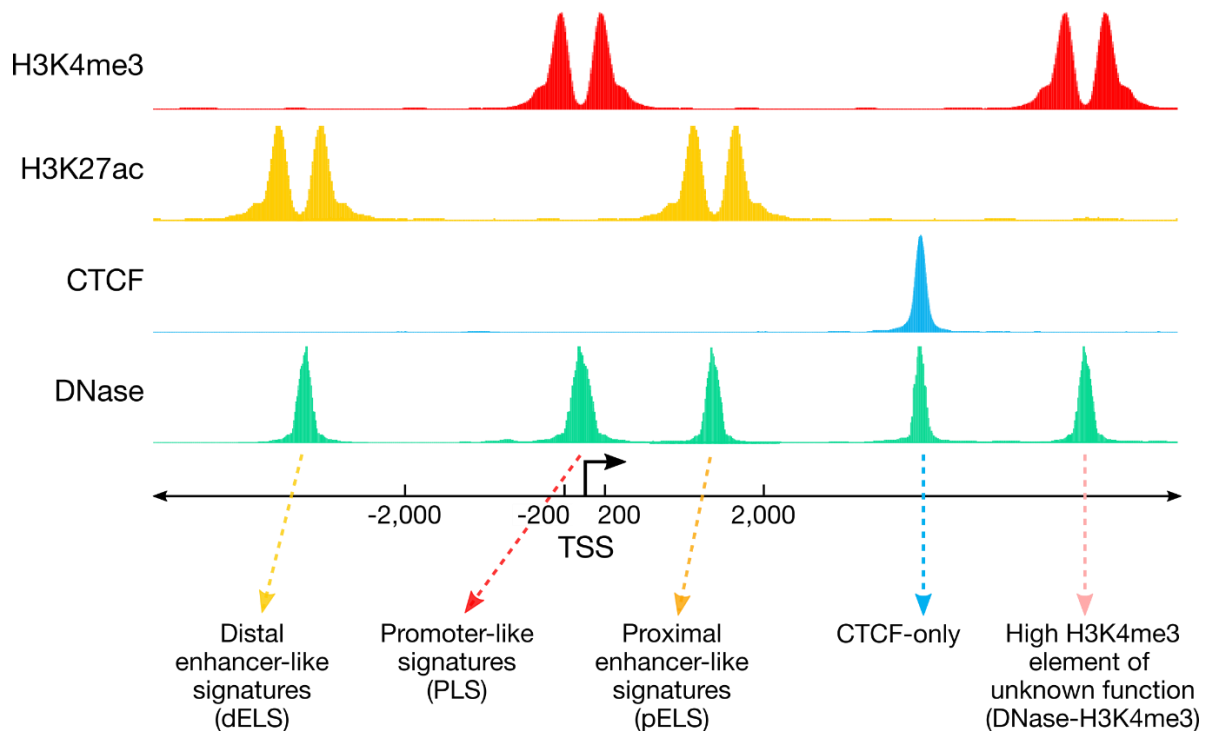


Figure 16. Résumé des informations fournies par la base Encode

d'après <https://www.encodeproject.org>

D'autres outils intégrés à la base de données permettent par ailleurs de visualiser des interactions entre les cCREs ainsi qu'entre CREs et gènes (<https://screen.encodeproject.org/>) ou de prédire l'état chromatinien d'une région génomique (<https://www.encodeproject.org/data/annotations/>).

5.1.3. La base de données VISTA

Le navigateur VISTA [140] (<https://enhancer.lbl.gov/>) est une banque de données d'enhancer humains qui ont la caractéristique d'avoir été validés expérimentalement chez des souris

transgéniques via des tests rapporteurs [Figure 17]. La plupart des éléments non codants présents dans VISTA ont été sélectionnés en raison de leur conservation extrême chez d'autres vertébrés ou de preuves épigénomiques, établies par ChIP-Seq, de marques d'*enhancer* putatives. Les résultats de cette analyse d'*enhancer in vivo* sont fournis sur le portail de VISTA et sont publiquement accessibles.

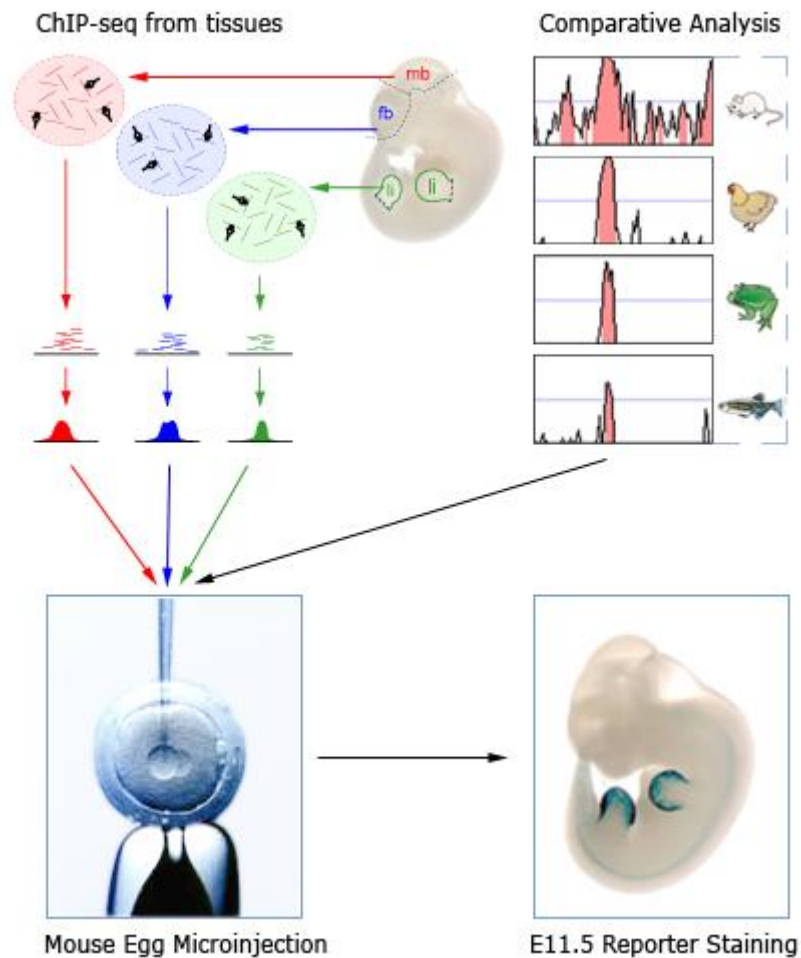


Figure 17. Stratégie de tests rapporteurs utilisés pour les données de la base VISTA

D'après <https://enhancer.lbl.gov>

5.1.4. *La base de données Genehancer*

Genehancer [141] est une base de données d'enhancers humains et de leurs gènes cibles. Cette base de données intègre un total de 434 000 enhancers issus de quatre bases de données génomiques différentes : ENCODE [137], VISTA [140], FANTOM [142] ainsi que des données complémentaires provenant de Ensembl [143]. Grâce à un algorithme d'intégration conçu pour éliminer la redondance, GeneHancer a réussi à recenser 285 000 candidats enhancers (couvrant 12,4% du génome), dont 94 000 proviennent de plus d'une source, et a attribué à chacun un score de confiance. Au-delà de la métaanalyse qui permet d'établir une liste d'enhancers humains, cette base de données propose également de lier les éléments régulateurs à leur(s) gène(s) cible(s). Pour effectuer ces associations, plusieurs processus ont été utilisés : d'une part, la corrélation de co-expression tissulaire entre les gènes et les ARN issus des enhancers (*enhancer RNA*, eRNA), ainsi qu'avec les facteurs de transcription ciblés par les enhancers ; d'autre part, les eQTL pour les variants au sein des enhancers, et enfin, des données de Hi-C, une analyse de conformation du génome spécifique du promoteur. Par la suite, les scores individuels basés sur chacune de ces quatre méthodes, ainsi que les distances génomiques gène-enhancer, ont permis d'attribuer un score de vraisemblance pour le couplage enhancer-gène. Ces scores ont servi à établir une liste de relations « élites » enhancer-gène, reflétant à la fois la définition d'enhancer de haute vraisemblance et une forte association enhancer-gène. Cette base de données est consultable directement à partir du visualisateur UCSC, ce qui permet de visualiser rapidement ces interactions [Figure 18].

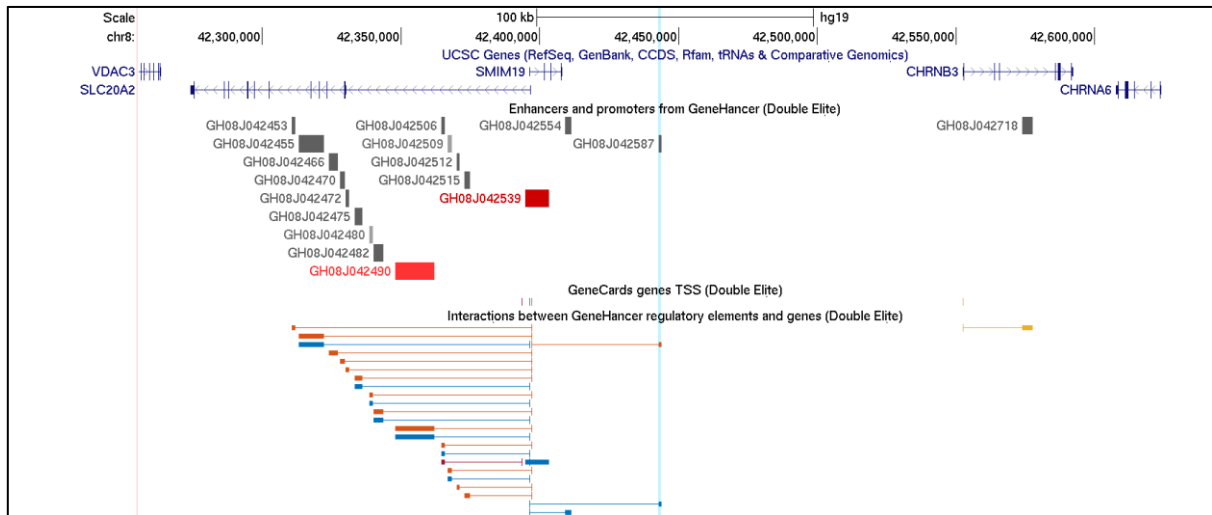


Figure 18. Exemple de visualisation des données de Genhancer sur le UCSC genome browser

Cette vue du UCSC genome browser présente, de haut en bas, *i. la représentation des gènes présents de cet intervalle, ii. La représentation des promoteurs (rouge) et des enhancers (gris), iii. la position des transcription start site (TSS) et iv. Les interactions entre les éléments régulateurs et les gènes. Figure réalisée à partir du UCSC genome browser.*

6. Modèles d'études *ex vivo* et *in vitro*

6.1. Échantillons issus de patients

Lorsqu'il s'agit d'étudier les conséquences fonctionnelles d'une variation génétique, la manière généralement la plus simple est de travailler à partir de prélèvements directement effectués chez des patients. Le sang est évidemment le tissu le plus facilement accessible. Concernant l'ARN, des méthodes largement éprouvées existent pour la collecte (par exemple, tube PAXGene) et l'extraction de l'ARN à partir d'un prélèvement sanguin. Toutefois, même lorsqu'ils sont effectués avec des tubes spécifiquement dédiés, ces prélèvements d'ARN restent fragiles et nécessitent une prise en charge technique rapide pour éviter la dégradation des ARNm, ce qui pourrait biaiser les résultats. Pour mesurer cette qualité, on utilise souvent le RNA Integrity Number (RIN) qui est une mesure standardisée utilisée pour évaluer l'intégrité de l'ARN dans un échantillon [144]. Les valeurs RIN s'étendent de 1 à 10, 10 indiquant une intégrité totale de l'ARN et 1, un ARN complètement dégradé. Un seuil de 7 est souvent considéré comme la limite minimale recommandée pour l'utilisation de l'ARNm dans des applications de microarray ou de NGS [145]. Il existe aussi d'autres méthodes d'extraction à partir d'échantillons EDTA, qu'ils soient « frais », remis en culture quelques jours, ou après des délais d'acheminement. Une autre stratégie est l'établissement de lignées lymphoblastoïdes. Bien que cette méthode soit plus longue, elle permet des comparaisons avec ou sans puromycine, facilitant la distinction de l'effet du NMD, et conduit souvent à des RIN plus élevés. Néanmoins, de nombreux gènes, en particulier ceux impliqués dans le développement, sont peu ou pas exprimés dans le sang, ce qui complique l'étude de certaines variations. Dans de tels cas, d'autres types de tissus peuvent être utilisés, comme la biopsie cutanée pour cultiver des fibroblastes. Sur le plan embryologique, la peau et le système nerveux ont une origine ectodermique. Souvent, le profil d'expression génique dans les tissus cutanés est une meilleure approximation de l'expression cérébrale que celle obtenue à partir

du sang, ce qui est notamment vrai pour de nombreux gènes impliqués dans les anomalies du neurodéveloppement [136].

6.2. Utilisation de modèles cellulaires pour étudier la régulation transcriptionnelle : intérêt de l'édition génomique par CRISPR/Cas

La technologie de *Clustered Regularly Interspaced Short Palindromic Repeats* (CRISPR/Cas), a radicalement changé le paysage de l'édition génomique. Basée sur le système de défense adaptative des bactéries, cette technologie utilise deux composants clés : une séquence d'ARN, appelée ARN guide (ARNg), et une enzyme nucléase, généralement la Cas9 [18]. L'ARNg est conçu pour être complémentaire d'une région spécifique de l'ADN cible. Cette spécificité de séquence permet à la technologie CRISPR/Cas d'être programmée pour cibler potentiellement n'importe quelle région du génome. Une fois que l'ARNg et la Cas9 se lient à l'ADN cible, la Cas9 agit comme une paire de "ciseaux moléculaires" et induit une cassure double-brin (DSB) à l'endroit précis de l'ADN [146]. La cellule répare ensuite cette cassure double-brin par l'un des deux mécanismes principaux de réparation de l'ADN : le mécanisme NHEJ (*Non-Homologous End Joining*) ou le mécanisme de recombinaison homologue (HR). Le mécanisme NHEJ, souvent inexact, peut introduire des insertions et des délétions (*indels*) à l'endroit de la cassure, ce qui peut conduire à l'inactivation du gène cible par introduction d'un décalage du cadre de lecture. En revanche, le mécanisme HR, qui nécessite un ADN donneur matrice, permet d'introduire des modifications spécifiques dans le génome [147]. CRISPR/Cas a été largement adopté et adapté pour diverses applications en laboratoire. Par exemple, il a été démontré que la technologie CRISPR/Cas est efficace pour éditer le génome des cellules souches pluripotentes induites (*induced pluripotent stem cells*, iPSC), ce qui a des implications majeures pour le développement de thérapies géniques et cellulaires [148]. En outre, la recherche continue de se développer pour améliorer la précision, l'efficacité et la sécurité de CRISPR/Cas. Par exemple, le développement de nouvelles

versions de l'enzyme Cas, comme la Cas12 et la Cas13, étend les capacités de ciblage et d'édition de la technologie [149].

Ce système permet donc d'introduire, dans un modèle cellulaire de choix, de nombreux types de variation génomiques afin de les modéliser *in vitro*. Cette approche permet de contrecarrer l'éventuelle non accessibilité d'échantillons de patients et a le grand avantage d'une meilleure comparabilité entre les conditions muté et non muté. En effet, les échantillons de patients et de témoins, même lorsqu'ils sont disponibles, ne le sont pas nécessairement en quantités (il s'agit souvent d'un seul patient à analyser), d'une part, et d'autre part, ils ne sont pas nécessairement prélevés dans les exactes mêmes conditions (appariement en âge, sexe, mais aussi possible influences environnementales sur l'expression des gènes, comme l'alimentation, la médication), ou encore l'heure de prélèvement dans la journée, le temps d'acheminement, sont des paramètres qui peuvent expliquer une certaine variabilité inter-individuelle, qui ne sera pas rencontrée dans le cadre d'un modèle cellulaire. Ces derniers nécessitent néanmoins beaucoup plus de ressources, qui ne sont pas nécessairement disponibles dans le cadre du soin.

7. Problématique

Nous avons vu dans cette introduction les différents éléments et mécanismes impliqués dans la régulation de l'expression des gènes, ainsi que les principaux outils permettant de les étudier. Ce travail de thèse s'intéressera à expliquer comment ces concepts et outils ont pu être utilisés pour explorer les conséquences transcriptomiques de certaines variations génomiques à effet fort, dans un contexte Mendélien, et solutionner des situations de variations de signification incertaine, ainsi que montrer comment des analyses transcriptomiques peuvent être un soutien aux analyses pangénomiques, comme biomarqueurs potentiels de certaines maladies monogéniques.

Résultats - Partie I. Etudes ciblées de la régulation de l'expression des gènes

Dans cette première partie, nous avons étudié les conséquences d'altérations génomiques d'un gène ou proches d'un gène sur l'expression de ce dernier (une altération génomique : un gène/transcrit). Tout d'abord, nous avons utilisé essentiellement la technique de RT-ddPCR pour caractériser l'impact transcriptionnel de variations structurales dans le cadre de deux affections neurogénétiques, les calcifications cérébrales primaires (CCP, *Primary Familial Brain Calcification, PFBC*) et la maladie d'Alzheimer du sujet jeune (MAJ, *early onset Alzheimer disease, EOAD*) avec Angiopathie Amyloïde Cérébrale (AAC). Dans l'article sur les CCP, nous avons combiné l'utilisation de la RT-ddPCR sur des prélèvements de patients et après modifications génomiques par CRISPR/Cas9 dans un modèle cellulaire HEK293, pour montrer la pathogénicité de la délétion d'un enhancer majeur du gène *SLC20A2*. Dans le travail sur la MAJ avec AAC, il s'agissait de montrer que la triplification du gène *APP*, observée pour la première fois dans une famille, est bien fonctionnelle, i.e. qu'elle s'exprime au niveau de l'ARNm et donc résulte en une multiplication par deux de la quantité de transcrits, dans le sang d'un patient porteur, et en comparaison avec des porteurs de duplication d'*APP* et des témoins, à l'aide de la technique de RT-ddPCR. Enfin, nous avons utilisé cette même technique ainsi que du RNAseq – avec une interprétation ciblée sur un locus et non pas sur tous les transcrits – pour étudier les conséquences de variations introniques du gène *NIPBL* et résoudre des variations de signification incertaine dans ce gène chez des patients avec suspicion de CdLS, cette fois-ci au niveau de l'épissage et de la dégradation par NMD des transcrits portant un exon « poison ».

1. Haploinsuffisance du gène *SLC20A2* médiée par la disruption d'un élément régulateur responsable de calcifications cérébrales primaires

1.1. Contexte et résumé des travaux

Les calcifications cérébrales primaires familiales (CCP, *Primary Familial Brain Calcification, PFBC*), anciennement appelées maladie de Fahr ou calcifications idiopathiques des noyaux gris centraux (*Idiopathic Basal Ganglia Calcification, IBGC*), est une maladie rare calcifiante du cerveau. Elle se définit par la présence d'une calcification anormale, principalement microvasculaire, des noyaux gris centraux et qui peut s'étendre à d'autres régions cérébrales [150]. Si la pénétrance radiologique est complète à 50 ans (présence de calcifications anormales sur le scanner cérébral), la pénétrance clinique est incomplète et l'expressivité de la maladie est variable. Les patients peuvent ainsi rester asymptomatiques ou présenter un large éventail de symptômes neuropsychiatriques, dont un syndrome parkinsonien, des tremblements, une dystonie, des troubles cognitifs, des symptômes psychiatriques ou une ataxie [151],[152]. Les CCP étaient jusqu'à l'identification des premiers gènes un diagnostic d'exclusion, nécessitant la recherche de nombreuses potentielles causes acquises (dont les troubles du métabolisme phospho-calciques sont les plus fréquents) ou un grand nombre de potentielles causes rares, monogéniques [153]. Le diagnostic de certitude est maintenant permis par l'identification d'une variation pathogène dans un des 6 gènes causaux connus. Les CCP sont, la plupart du temps, transmises selon un mode autosomique dominant, avec 4 gènes identifiés à ce jour : *SLC20A2* (MIM 158378) [154], *PDGFRB* (MIM 173410) [150], *PDGFBR* (MIM 190040) [155] et *XPR1* (MIM 605237) [156]. Plus récemment, des variations pathogènes bi-alléliques dans le gène *MYORG* (MIM 618255) ont été identifiées comme une cause de CCP autosomique récessive [157], avec des variations

dans la présentation clinique, puisqu'il s'agit d'un phénotype principalement moteur associé, sur le plan radiologique, à des calcifications sévères et à une atrophie cérébelleuse [152]. Enfin, des variants bi-alléliques de *JAM2* (MIM 606870) ont également été rapportées dans les CCP autosomiques récessives [158] avec un phénotype semblant similaire à celui des patients *MYORG*. Le gène *SLC20A2* (MIM 158378) est quant à lui considéré comme le principal gène autosomique dominant responsable de CCP en termes de fréquence de mutation, avec un rendement pouvant aller jusqu'à 40% chez les cas index avec antécédents familiaux identifiés [159]. Il code pour un transporteur de phosphate inorganique (Pi), l'importateur PiT2. Les variations perte de fonction de *SLC20A2* entraînent une accumulation de Pi dans le liquide périvasculaire et le liquide céphalo-rachidien, ce qui peut contribuer aux calcifications murales vasculaires dans le cerveau [160],[161]. Les variants pathogènes de *SLC20A2* introduisent généralement des codons stop prématurés (non-sens, SNV au niveau d'un site d'épissage canonique ou insertions/délétions provoquant un décalage du cadre de lecture) ou des délétions génomiques partielles ou totales, toutes ces situations conduisant à une haplo-insuffisance par la destruction des ARNm médiée par le système de dégradation des ARNm non-sens (*nonsense-mediated decay*, *NMD*) ou par insuffisance de production pour les délétions. Des variants faux-sens ont également été identifiés, avec pour conséquence une perte de la fonction d'import de Pi de PiT2 [154]. Parmi les variations pathogènes connues, seulement quelques délétions génomiques partielles ou totales de *SLC20A2* à l'origine de CCP ont été rapportées [162], [163], [164], [165]. Toutes englobaient des régions codantes et étaient interprétées comme entraînant une haplo-insuffisance ou la perte de domaines protéiques critiques. Une seule délétion rapportée était non codante : ce CNV de 578 kb a été retrouvé chez un patient finlandais et emportait le premier exon, non codant, de *SLC20A2*, son extrémité 5'UTR ainsi que la région promotrice putative du gène [166], mais sans analyse de l'expression du gène. Dans cet exemple, les conséquences négatives très probables sur l'expression du gène n'ont pas été évaluées à notre connaissance.

Dans le travail présenté ci-après, nous avons recherché, avec l'outil de détection de CNV CANOES, préalablement validé [167], des variations du nombre de copies sur les exomes de 71 cas index non apparentés présentant des CCP sans cause connue et avons identifié une délétion localisée en 8p11.21, 150 kb en amont du gène *SLC20A2*, impliquant deux gènes voisins non candidats pour les CCP et non contraints en perte de fonction dans les bases de données. Bien que la délétion n'implique aucune base codante ou du promoteur de *SLC20A2*, sa proximité relative avec le gène et le fait qu'elle en emporte un enhancer candidat d'après la base *GeneHancer* [141] nous a conduit à l'investiguer.

Le séquençage de l'exome du demi-frère atteint et l'étude ciblée de la délétion par digital droplet PCR (ddPCR) chez le cas index, son demi-frère et le fils de ce dernier, également porteurs de calcifications, ont montré que tous 3 sont porteurs de ce remaniement. L'analyse de l'ARNm des 3 patients par RT-ddPCR dans le sang a montré une diminution d'expression relative de *SLC20A2* de 45,0% par rapport à des contrôles sains ($p < 0.001$), soit dans les mêmes ordres de grandeur que 4 individus porteurs de variations nucléotidiques entraînant des codons stop prématurés de *SLC20A2* (-39,0%), responsables d'une dégradation d'ARNm par NMD. L'analyse de l'import de phosphate inorganique dans les cellules sanguines du cas index dans le cadre d'une collaboration avec l'équipe de Jean-Luc Battini (IRIM, Montpellier) a montré un défaut d'import de -39,3% par rapport aux contrôles ($p = 0.015$). Enfin, avec Anne Rovelet-Lecrux, ingénieur dans l'équipe, nous avons introduit une délétion de la région de cet enhancer supposé de *SLC20A2* dans des cellules HEK293 par CRISPR/Cas9. Après avoir confirmé la bonne introduction de la délétion par ddPCR (nombre de copie moyen dans les culots cellulaires : 1,3), nous avons mesuré une diminution d'expression de l'ARNm de *SLC20A2* de l'ordre de 35,6% dans les cellules transfectées par les ARN guides ciblant cette région, par rapport aux contrôles. Ici, nous avons fait le choix d'une introduction de la variation par CRISPR/Cas9 en « *bulk* », plutôt qu'une procédure plus longue de sélection

clonale des cellules ayant intégré la variation, comme habituellement réalisée dans le laboratoire et bien plus largement. Cette technique, plus rapide, a permis d'obtenir des résultats de manière plus directe et la comparaison avec le nombre moyen de copies permet d'assurer la fiabilité des résultats.

En conclusion, il s'agit de la première délétion non codante à l'origine d'une haploinsuffisance du gène *SLC20A2*. Cette délétion est associée à une perte d'expression du même ordre que des variations perte de fonction du gène, ainsi qu'à une perte de l'activité protéique, montrant son rôle majeur. Ce travail a permis de proposer l'utilisation d'une technique de CRISPR/Cas9 simplifiée, associée à la (RT-)ddPCR permettant une étude rapide de l'ARN et de l'ADN directement extrait d'un mélange (bulk) cellulaire transfecté, sans sélection clonale. Ces résultats, montrant l'importance que joue cet enhancer dans l'expression de ce gène, ouvrent la voie à de potentielles cibles thérapeutiques chez les patients avec CCP par haploinsuffisance de *SLC20A2*, cet enhancer pourrait devenir une cible en lui-même pour, par exemple, favoriser l'expression de l'allèle non muté.. Il s'agissait par ailleurs ici d'une stratégie simple permettant l'annotation, la mesure de l'effet sur l'ARNm, et la confirmation de l'effet *in vitro*, permettant de conclure sur l'impact de tels CNV non codants.

Outre l'article scientifique présenté ci-dessous, ce travail a fait l'objet de plusieurs présentations orales et affichées dans des congrès de génétique, à savoir :

- Communication orale lors du séminaire ARN, Assises de génétique de Tours, 2020
- Communication orale lors des journées du réseau NGS diag 2021
- Poster lors du e-congrès de l'ESHG 2020

Haploinsufficiency of the Primary Familial Brain Calcification Gene *SLC20A2* Mediated by Disruption of a Regulatory Element

Kévin Cassinari, MD,¹ Anne Rovelet-Lecrux, PhD,¹ Sandrine Tury, PhD,² Olivier Quenez, MSc,¹ Anne-Claire Richard, BSc,¹ Camille Charbonnier, PhD,¹ Robert Olasso, PhD,³ Anne Boland, PharmD, PhD,³ Jean-François Deleuze, PhD,³ Jean-François Besancenot, MD, PhD,⁴ Benoit Delpont, MD,⁴ Dorothée Pouliquen, MSc,⁵ François Lecoquierre, MD,¹ Pascal Chambon, PharmD,¹ Christel Thauvin-Robinet, MD, PhD,^{6,7,8} Dominique Champion, MD, PhD,^{1,9} Thierry Frebourg, MD, PhD,¹ Jean-Luc Battini, PhD,² and Gaël Nicolas, MD, PhD^{1*}

¹Department of Genetics and CNR-MAJ, Normandy Center for Genomic and Personalized Medicine, Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Rouen, France

²Institut de Recherche en Infectiologie de Montpellier, Université de Montpellier, CNRS, Montpellier, France

³Centre National de Recherche en Génétique Humaine, Institut de Biologie François Jacob, CEA, Université Paris-Saclay, Evry, France

⁴Department of Internal Medicine and Systemic Diseases, Dijon University Hospital, Dijon, France

⁵Department of Neurology and CNR-MAJ, Normandy Center for Genomic and Personalized Medicine, Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Rouen, France

⁶Inserm UMR 1231 GAD, Genetics of Developmental Disorders, Université de Bourgogne-Franche Comté, FHU TRANSLAD, Dijon, France

⁷CHU Dijon Bourgogne, Unité Fonctionnelle "Innovation diagnostique dans les maladies rares," laboratoire de génétique chromosomique et moléculaire, Plateau Technique de Biologie, Dijon, France

⁸Centre de Référence Maladies Rares "Déficiences Intellectuelles de causes rares," FHU-TRANSLAD, CHU Dijon Bourgogne, Dijon, France

⁹Department of Research, Rouvray Psychiatric Hospital, Sotteville-les-Rouen, France

ABSTRACT: Objective: Primary familial brain calcification (PFBC) is a rare cerebral microvascular calcifying disorder with diverse neuropsychiatric expression. Five genes were reported as PFBC causative when carrying pathogenic variants. Haploinsufficiency of *SLC20A2*, which encodes an inorganic phosphate importer, is a major cause of autosomal-dominant PFBC. However, PFBC remains genetically unexplained in a proportion of patients, suggesting the existence of additional genes or cryptic mutations. We analyzed exome sequencing data of 71 unrelated, genetically unexplained PFBC patients with the aim to detect copy number variations that may disrupt the expression of core PFBC-causing genes.

Methods: After the identification of a deletion upstream of *SLC20A2*, we assessed its consequences on gene function by reverse transcriptase droplet digital polymerase chain reaction (RT-ddPCR), an ex vivo inorganic

phosphate uptake assay, and introduced the deletion of a putative *SLC20A2* enhancer mapping to this region in human embryonic kidney 293 (HEK293) cells by clustered regularly interspaced short palindromic repeats (CRISPR) - CRISPR-associated protein 9 (Cas9).

Results: The 8p11.21 deletion, segregating with PFBC in a family, mapped 35 kb upstream of *SLC20A2*. The deletion carriers/normal controls ratio of relative *SLC20A2* mRNA levels was 60.2% ($P < 0.001$). This was comparable with that of patients carrying an *SLC20A2* premature stop codon (63.4%; $P < 0.001$). The proband exhibited a 39.3% decrease of inorganic phosphate uptake in blood ($P = 0.015$). In HEK293 cells, we observed a 39.8% decrease in relative *SLC20A2* mRNA levels after normalization on DNA copy number ($P < 0.001$).

Discussion: We identified a deletion of an enhancer of *SLC20A2* expression, with carriers showing

*Correspondence to: Dr. Gaël Nicolas, Inserm U1245, Faculté de médecine, 22, boulevard Gambetta, 76183 Rouen, France; E-mail: gaelnicolas@hotmail.com

Relevant conflicts of interests/financial disclosures: Nothing to report.

Funding agencies: This work was supported by grants from the French National Research Agency (CALCIPHOS, ANR-17-CE14-0008 to GN and J.L.B.) and from Conseil Régional de Haute Normandie—APERC 2014 no. 2014-19 in the context of Appel d'Offres Jeunes Chercheurs (CHU de

Rouen to GN). This study was cosupported by European Union and Région Normandie, more specifically in the context of the Recherche Innovation Normandie (RIN 2018 to GN). Europe is involved in Normandie with the European Regional Development Fund.

Received: 10 December 2019; **Revised:** 17 January 2020; **Accepted:** 3 April 2020

Published online 7 June 2020 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/mds.28090

haploinsufficiency in similar ranges to loss-of-function alleles, and we observed reduced mRNA levels after deleting this element in a cellular model. We propose a 3-step strategy to identify and easily assess the effect of such events. © 2020

International Parkinson and Movement Disorder Society

Key Words: CNV; enhancer; primary familial brain calcification; regulation; *SLC20A2*

For the past 10 years, the development of massive parallel sequencing has dramatically improved the knowledge on rare diseases as well as diagnostic performances. However, some proportion of patients with clinically diagnosed rare diseases remains genetically unexplained even after such a genome-wide analysis. This highlights the need to pursue research efforts on rare diseases to identify new genes. In addition, in some situations, cryptic mutations in core genes may cause disease. Noncoding variants, including variants in deep intronic regions, in untranslated regions (UTR) and variants located downstream or upstream genes, remain difficult to interpret.

Enhancers are noncoding DNA elements that *cis*-regulate the transcription of nearby genes. Their perturbation by structural or nucleotide variations has already been described in human developmental disorders such as nucleotide variations in a *PAX6* enhancer leading to aniridia^{1,2} or in an *SHH* enhancer causing polydactyly.³ This statement can apply to different fields of medical genetics, including *de novo* neurodevelopmental disorders for which deleterious events in regulatory elements have been explored on large cohorts with low diagnostic yields and a challenging interpretation.⁴ Thus, the discovery of a genomic variation close to a gene of interest should bring the attention on a putative regulation perturbation.⁵ The identification of enhancers remains challenging, although databases such as Vista,⁶ ENCODE,⁷ and Ensembl⁸ display confirmed or candidate enhancer lists. Furthermore, the use of genome-wide integration tools based on databases such as Genehancer⁹ allow facilitated mapping and visualization of these candidate regulatory elements and their links with genes. However, the identification of a variation affecting a candidate gene regulatory element is not sufficient to conclude on its putative biological effect and associated clinical consequences. Targeted mRNA assessment appears to be a mandatory step before looking for *in vitro* confirmation. CRISPR/Cas9 gene editing in cell lines combined with gene expression measures now allow the accessible assessment of such hypotheses.

Primary familial brain calcification (PFBC), previously known as Fahr's disease or idiopathic basal ganglia calcification, is a rare calcifying disorder of the brain. PFBC is defined by the presence of abnormal, mainly microvascular, calcification in the basal ganglia that may encompass other cerebral regions in the

absence of other causes.^{10,11} Patients can remain asymptomatic or present a large spectrum of neuropsychiatric symptoms, including parkinsonism, tremor, dystonia, cognitive impairment, psychiatric symptoms, and ataxia.^{12,13} PFBC is typically inherited as an autosomal-dominant trait with the following 4 genes identified so far: *SLC20A2* (MIM 158378),¹⁴ *PDGFRB* (MIM 173410),¹⁰ *PDGFB* (MIM 190040),¹⁵ and *XPR1* (MIM 605237).¹⁶ Recently, bi-allelic pathogenic variants in the *KIAA1161/MYORG* gene have been identified as a cause of autosomal-recessive PFBC¹⁷ with a predominantly motor phenotype, a severe calcification pattern, and cerebellar atrophy.¹³ More recently, bi-allelic variants in *JAM2* have been reported in autosomal-recessive PFBC.¹⁸ *SLC20A2* is considered as the major autosomal-dominant, PFBC-causing gene in terms of mutation frequency.¹⁹ It encodes an inorganic phosphate (Pi) transporter, the importer PiT2. Loss-of-function variants result in perivascular and cerebrospinal fluid accumulation of Pi that may contribute to vascular mural calcifications in the brain.^{20,21} Pathogenic *SLC20A2* variants typically introduce premature stop codons (nonsense, canonical splice site single nucleotide variants and frameshift insertions or deletions), leading to haploinsufficiency through nonsense-mediated mRNA decay. Missense variants have also been identified, with a loss of Pi import function consequences on PiT2.¹⁴ In addition, a few disease-causing partial or full *SLC20A2* deletions have been reported.²²⁻²⁶ All encompassed coding regions and were predicted to result in haploinsufficiency or the loss of critical protein domains. Interestingly, a 578 kb noncoding deletion was reported in a Finnish patient that removed the first noncoding exon of *SLC20A2* (5' UTR) as well as the putative promoter region.²⁷ However, the highly likely negative consequences on gene expression were not assessed to our knowledge.

We reanalyzed the whole exome sequencing (WES) data of 71 patients with genetically unexplained PFBC following a negative assessment of the 6 known PFBC genes by classical nucleotide and copy number variation WES analyses. We identified a novel deletion located 150 kb upstream of *SLC20A2* in a family, segregating with PFBC in all 3 affected individuals. We further mapped this deletion that encompassed a candidate enhancer and showed that carriers presented decreased *SLC20A2* expression in ranges mimicking a whole gene deletion. In addition, we observed reduced Pi uptake

in an *ex vivo* assay performed in blood cells from one carrier and assessed the effect of deleting a putative enhancer of *SLC20A2* expression by CRISPR/Cas9 *in vitro* assays.

Methods

Patients

Patients were recruited from multiple French centers as previously described.^{12,13} We included 18 previously reported patients fulfilling the diagnostic criteria of PFBC and with a screen of all 5 known genes showing no PFBC causal variant.¹³ In addition, we included 53 unrelated PFBC patients with WES showing no PFBC causal variant (total number of patients included, $n = 71$). Briefly, the diagnosis of PFBC was retained if the patients exhibited calcifications affecting at least both lenticular nuclei with a total calcification score greater than the age-specific threshold; a normal calcium–phosphorus metabolism assessment, regardless of the presence or absence of a family history; and no other known cause of brain calcification. All patients gave informed written consent for genetic analyses in a diagnostic setting and/or in the context of a study approved by the Comité de Protection des Personnes (CPP) Ile de France II ethics committee.

In family EXT-444 (Fig. 1A), the half-brother (EXT-444-002) of the proband (EXT-444-001) was also affected by PFBC. Following the first negative interpretation of the WES data of EXT-444-001, we performed WES in his half-brother as well. Then we included another affected relative for targeted analyses, EXT-444-003, the son of EXT-444-002.

An additional series of 113 unrelated probands was further assessed for targeted analyses. They all presented PFBC or unexplained brain calcifications.

Whole Exome Sequencing

Exomes were captured using the Agilent Sureselect All Exons Human V5 + UTR or V6 + UTR Kits (Agilent Technologies, Santa Clara, CA). Final libraries were sequenced on an Illumina HiSeq4000 with paired ends, 100-bp reads. The reads were mapped to the 1000 Genomes GRCh37 build using Burrows-Wheeler Aligner 0.7.5a.⁴⁷ Picard Tools 1.101 was used to flag duplicate reads. We applied Genome Analysis Toolkit 3.6 for indel realignment,⁴⁸ and single nucleotide polymorphisms and indels were called using Haplotype Caller across all samples simultaneously. The average depth of coverage was 140X. Single nucleotide variants and short insertions and deletions (indels) were annotated using SNPEff and SNPSift. Rare variants in the 5 known PFBC genes had previously been interpreted following the American College of Medical Genetics and Genomics/Association for Molecular

Pathology recommendations²⁸ and no (likely) pathogenic variant was identified. In addition, we assessed the presence of *JAM2* bi-allelic variants and, as a differential diagnosis search, the screen of variants in a list of genes known to be associated with brain calcification (Supporting Information Table e-1). We found no alternative cause following the use of the same recommendations.

Copy number variation (CNV) calling was performed using the read-depth, comparison-based CANOES tool²⁹ as previously described.^{30,31} CANOES detected no CNV encompassing the coding sequence of any of the 6 known genes with a resolution of a single exon and with good performances.³¹ With the hypothesis that regulatory elements may be disrupted by CNVs surrounding the coding regions of these genes, we focused our interpretation on rare CNVs encompassing regions mapping 500 kb around the coding sequence of the 5 core genes.

Targeted Copy Number Analyses

Targeted copy number analyses for the confirmation and characterization of the CNV in probands and for the screening of the replication series were performed using custom digital droplet polymerase chain reaction (ddPCR). This method enables the absolute quantification of the target and reference sequences and reduces the quantification of a target sequence to the enumeration of series of positive and negative end-point PCR reactions.³² This approach has been shown to have a higher analytical sensitivity for CNV detection than quantitative PCR and a better reproducibility because it does not require any calibration curve.³³ Here ddPCR analyses were performed on the QX200 ddPCR platform (Bio-Rad, Hercules, CA) using locked nucleic acid hydrolysis probes (Universal Probes Library, Roche, Basel, Switzerland) on the target (FAM tag) and *HMBS* reference gene (VIC tag), as previously described.³⁴ All primer sequences are available upon request.

Targeted mRNA Expression Analyses

The expression of *SLC20A2* in blood and cultured cells was assessed using reverse transcriptase ddPCR (RT-ddPCR), a cDNA-based version of our targeted copy number analysis. RNA was extracted from cultured cells using the Nucleospin RNA kit (Macherey-Nagel, Düren, Germany). Reverse transcription was performed using the Verso cDNA kit (Thermo Fisher Scientific, Waltham, MA) from 200 ng RNA for blood or the human embryonic kidney 293 (HEK293) cell pellets. Next, ddPCR was performed for relative quantification between *SLC20A2* and the housekeeping gene *TBP*, chosen as reference because of blood and HEK293 expression levels in the same ranges as

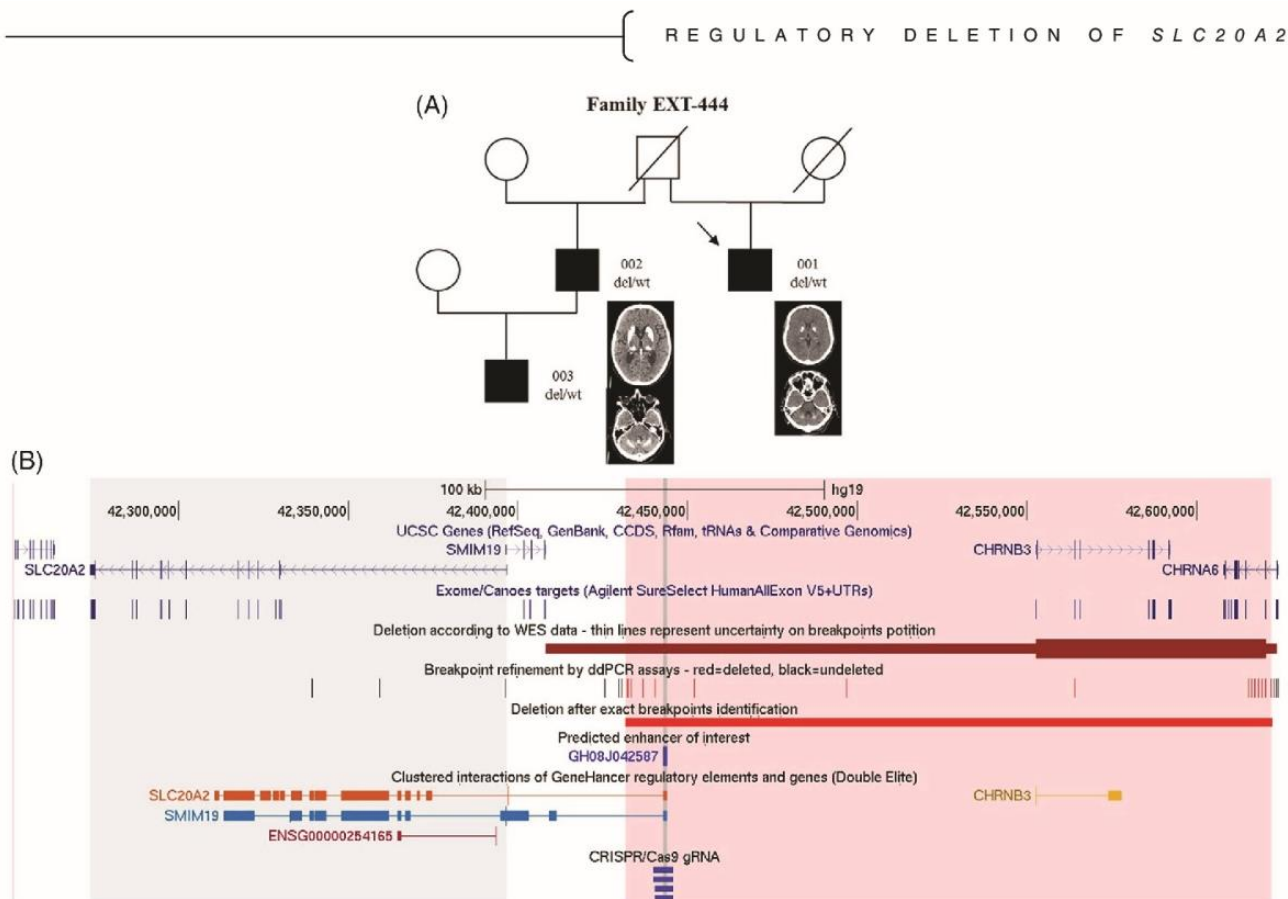


FIG. 1. Pedigree and representation of the deletion. **(A)** Reduced pedigree and brain CT scan images of the family EXT-444. Filled symbols represent affected individuals (ie, carriers of brain calcifications on CT), and white symbols represent individuals of unknown status (no CT scan). Squares indicate males; circles indicate females. The arrow points to the proband. Axial images of brain calcifications are represented below patient EXT-444-001 (CT scan performed at the age of 60 years) and EXT-444-002 (CT scan performed at the age of 67 years). CT scan images were not available for patient EXT-444-003, who is known to carry abnormal calcifications on CT scan. **(B)** Visualization on the UCSC genome browser of the region (GRCh37). From top to bottom, visualization of (1) UCSC RefSeq genes track, (2) CANOES target on WES data, (3) 8p11.21 deletion as initially detected by CANOES with thin lines representing uncertainty on breakpoints positions, (4) location of ddPCR amplicons used for characterization of the deletion (red: deleted, black: undeleted), (5) precise location of the deletion after breakpoints identification—chr8:42,431,767_42,622,181delinsT;hg19 (see Supporting Information Fig. e-2 for Sanger sequencing of the breakpoint region), the deletion did not overlap with any promoter or coding region of the *SLC20A2* gene (5) of the GH08J042587 putative enhancer of the gene (custom track from GeneHancer UCSC track)—(6) GeneHancer track of GeneHancer track “clustered interactions” highlighting the links between this regulatory element and the *SLC20A2* promoter and transcription start site, and (7), on the bottom part, the location of the deleted region in CRISPR/Cas9 assays. CRISPR/Cas9, clustered regularly interspaced short palindromic repeats/clustered regularly interspaced short palindromic repeats associated protein 9; CT, computed tomography; ddPCR, droplet digital polymerase chain reaction; del/wt, heterozygous carrier of the deletion; UCSC, University of California Santa Cruz; UTR, untranslated regions; WES, whole exome sequencing. [Color figure can be viewed at wileyonlinelibrary.com]

SLC20A2 according to genome tissue expression.³⁵ For the target, a 77 bp-amplicon was designed spanning exons 4 and 5 of *SLC20A2* (forward: 5'-AATC GGTACCAAAGGTGTGC-3', reverse: 5'CCAGAC AACAGTGGAGATATAAAC-3') in association with a FAM-labeled hydrolysis probe (5'-CTCCATCC-3') containing locked nucleic acid (Universal Probes Library, Roche). The 78-bp reference amplicon mapped to exons 1 and 2 of *TBP* (forward: 5'-CGGC TGTTAACTTCGCTTC-3', reverse: 5'-CACAGC CCAAGAAACAGTGA-3') associated with a specific VIC-labeled custom hydrolysis probe (Applied Biosystems, Thermo Fisher Scientific). Analyses were performed in the affected carriers of the deletion

(EXT-044-001, EXT-0444-002, EXT-0444-003), 4 healthy controls, and 4 positive controls carrying an *SLC20A2* premature stop codon variant predicted to result in nonsense-mediated mRNA decay: TLOF1 (ROU-5028-001) carrying the NM_006749.4: c.1158C > A p.(Tyr386*) variant, TLOF2 (ROU-5159-001) carrying the c.1017delC p.(Ser339 Argfs*116) variant, TLOF3 (EXT-1713-001) carrying the c.382delG p.(Val128Serfs*43) variant, and TLOF4 (ROU-5172) carrying the c.1152_1153delCA, p.(Asn384Lysfs*30) variant. For each sample, the analyses were performed in accordance with international guidelines for quantitative digital PCR experiments³³ in at least 3 replicates.

Phosphate Fluxes in Peripheral Blood Mononuclear Cells

Fresh blood samples were obtained from patient EXT-444-001 and 2 controls. Phosphate uptake and efflux assays in peripheral blood mononuclear cells (PBMC) were performed as previously described for XPR1 variants^{16,36,37} and more recently for *SLC20A2* variants.⁴⁰ The amount of phosphate uptake was calculated from the concentration of cold phosphate in the medium multiplied by the ratio of cellular [³³P]phosphate to total [³³P]phosphate supplemented (FF-1; Hartmann Analytic GmbH, Bern Switzerland) within a period of 30 minutes. Percentage of phosphate efflux was calculated as the ratio of released [³³P]phosphate to total cellular [³³P]phosphate.

Deletion of the *SLC20A2* Enhancer in HEK293 Cells

A genomic deletion of the putative regulatory element of *SLC20A2* was introduced in HEK293 cells using the CRISPR/Cas9 technology. The HEK293 cells were purchased from American Type Culture Collection (HEK-293; American Type Culture Collection [Manassas, VA] CRL-1573). DNA template sequences for small-guide RNAs were designed using the CRISPOR tool³⁸ (<http://crispor.tefor.net>) purchased from Eurogentec (Liège, Belgium) and cloned into the Cas9–green fluorescent protein expressing plasmid pX458 (Addgene no. 48138). Two guide RNA were designed to target a sequence located 5' of the regulatory element—named 5'-(1) and 5'-(2)—and two others on the 3' region—named 3'-(1) and 3'-(2). As control, we used a guide RNA targeting a nonrelevant region on another chromosome (*SORL1* gene, located in 11q24.1; for detailed gRNA sequences, see Supporting Information Table e-3). All selected guide RNAs (gRNAs) showed high MIT and CFD specificity scores (ranging from 80–94),³⁹ and with off-target predictions showing at least 3 nucleotide mismatches with gRNAs.

The genomic deletion of the regulatory element in the HEK293 cells was obtained by cotransfection of 2.10⁶ cells with 1 µg of a plasmid containing a guide RNA targeting the 5' region in combination with 1 µg of a plasmid targeting the 3' region using the AMAXA Nucleofactor II Device (Amaya Biosystem, USA). After 48 hours, the cells were rinsed once with phosphate-buffered saline, trypsinized, and collected in 1 mL phosphate-buffered saline. After a 5-minute centrifugation at 1000g, the cell pellet was split into 2 pellets. One was used to isolate DNA using the DNA Blood & Tissue kit (Qiagen, Hilden, Germany) and the other to isolate RNA using the Nucleospin RNA kit (Macherey-Nagel).

Statistics

Statistical analyses were performed using the Mann-Whitney test. *P* values <0.05 were considered significant.

Data Availability

Anonymized data relevant to this study and not included here will be made available upon reasonable request.

Results

Identification of a Deletion Upstream of *SLC20A2*

Following the reanalysis of WES data processed by the CANOES bioinformatics tool for CNV detection among 71 patients negatively screened for the 6 known PFBC genes, we identified a heterozygous deletion on chromosome 8 (8p11.21, chr8:42,552,547-42,620,341; GRCh37) in patient EXT-444-001. This deletion encompassed at least 66,393 bp, including 2 protein-coding genes (*CHRNA3* and *CHRNA6*) and mapped nearly 150 kb upstream of *SLC20A2*, based on exome data, which contain information restricted to exons and hence did not allow the capture of noncoding breakends. The deletion was also present in the CANOES calls obtained from WES data of the affected half-brother (EXT-444-002) of the proband. The deletion was confirmed in both patients by targeted copy number analyses using custom ddPCR (Supporting Information Fig. e-1). Finally, we obtained the blood sample of an additional affected relative, EXT-444-003, son of EXT-444-002, and found that he carried the same deletion (Fig. 1A, Supporting Information Fig. e-1).

Quantitative multiplex PCR of short fluorescent fragments with 1 amplicon in each exon of *SLC20A2*²³ confirmed that no coding region was deleted. We further mapped the deletion using additional ddPCR assays in intergenic regions. We were able to identify the breakpoints leading to the following deletion nomenclature: chr8:42,431,767_42,622,181delinsT; GRCh37 (Supporting Information Table e-2, Fig. e-2), with a proximal breakpoint located 35 kb upstream of *SLC20A2*. Although the deletion extended into the intergenic region, it did not encompass any promoter base pair of *SLC20A2* or the neighboring gene *SMIM19*. Interestingly, the deletion encompassed a strong putative *SLC20A2* enhancer GH08J042587 (chr8:42,442,818-42,443,943; GRCh37, called ENSR00000224118 in Ensembl), according to the GeneHancer database.⁹ This enhancer is predicted to closely interact with the *SLC20A2* transcription start site (Fig. 1B).

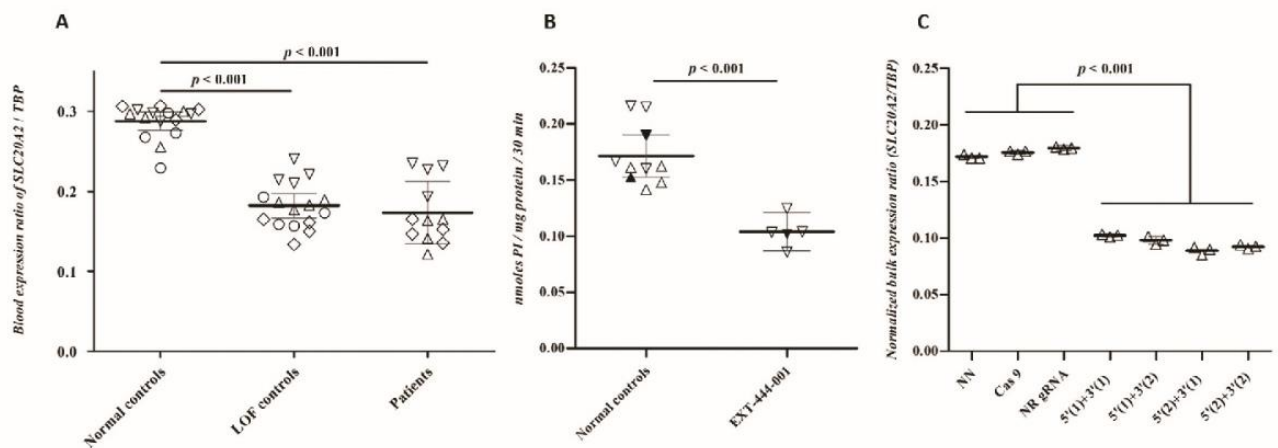


FIG. 2. Deletion of the GH08J042587 regulatory element is associated with *SLC20A2* haploinsufficiency and reduced inorganic phosphate uptake. Reduced *SLC20A2* mRNA levels and inorganic phosphate uptake in deletion carriers: complementary analyses on blood (**A,B**). (**A**) At the mRNA level, the expression assays were performed using multiplex droplet digital polymerase chain reaction with a target amplicon in the *SLC20A2* gene and a reference amplicon in the *TBP* housekeeping gene. The results are presented through an expression ratio *SLC20A2/TBP*. These data highlight a significant decrease of this ratio for the patient group (right panel, ∇ : EXT-044-001, \triangle : EXT-044-002, \diamond : EXT-044-003) and for the LOF mutation-carrier positive control group (middle panel, ∇ : TLOF1, \triangle : TLOF2, \diamond : TLOF3, \circ : TLOF4) compared with 4 normal controls (left panel, ∇ : TN1, \triangle : TN2, \diamond : TN3, \circ : TN4). (**B**) Phosphate uptake assay highlighting a significant decrease of ^{33}P uptake in EXT-044-001 peripheral blood mononuclear cells (right, ∇ : EXT-044-001) versus 2 healthy controls (left, ∇ : TN1, \triangle : TN2). (**C**) In vitro validation using CRISPR/Cas9 targeted deletion in HEK293 cells. Assays were performed after transfection of HEK293 cells by CRISPR/Cas9 using the following 4 different guide RNA combinations: 5' (1)+3' (1); 5' (1)+3' (2); 5' (2)+3' (1); 5' (2)+3' (2). Copy number analysis of the cell pellets confirmed the efficiency of the Cas9-mediated genomic deletion in all bulks of HEK293 transfected cells with a 1.36 average copy number and a normal copy number in HEK293 cells transfected in a nonrelevant genomic location (see Supporting Information Figure e-4). Then to test the *SLC20A2* mRNA level in the different populations, expression assays were performed using multiplex reverse transcriptase droplet digital polymerase chain reaction with a target amplicon in the *SLC20A2* gene and a reference amplicon in the *TBP* housekeeping gene. The results are presented through an expression ratio *SLC20A2/TBP*. The results of reverse transcriptase droplet digital polymerase chain reaction on mRNA show a significant decrease of the relative *SLC20A2* mRNA levels in all HEK293 pellets carrying the GH08J042587 deletion compared with the different control cells. Represented here are normalized *SLC20A2/TBP* mRNA levels after normalization on DNA copy number ($P < 0.001$). Vertical lines represent 95% confidence intervals. Cas9, HEK293 cells transfected with the plasmid encoding the clustered regularly interspaced short palindromic repeats associated protein 9 without gRNA; CRISPR/Cas9, clustered regularly interspaced short palindromic repeats/clustered regularly interspaced short palindromic repeats associated protein 9; HEK293, human embryonic kidney 293; LOF, loss of function; NN, non-nucleofected cells; NR, HEK293 cells transfected in a nonrelevant genomic location; Pi, inorganic phosphate.

The Deletion of the GH08J042587 Regulatory Element Is Associated With *SLC20A2* Haploinsufficiency and Reduced Pi Uptake

Because of the relative proximity between this deletion and *SLC20A2* and its content in putative regulatory elements, we assessed mRNA levels of *SLC20A2* in the blood of the carriers of the deletion by RT-ddPCR. The assay was performed on RNA isolated from PAXGene (Ozyme, Saint-Cyr-L'Ecole, France) blood tubes from all three carriers of the family EXT-444-001, EXT-444-002, EXT-444-003, from normal controls, and from positive controls carrying loss of function *SLC20A2* mutations. The deletion carriers/normal controls ratio of relative *SLC20A2* mRNA levels was 60.2% (deletion carriers vs. normal controls; $P < 0.001$; Fig. 2A). These results were comparable with that of positive controls (63.4%; $P < 0.001$), suggesting a very strong effect of this noncoding deletion.

Following the identification of a decreased level of *SLC20A2* expression in blood at the mRNA level, we assessed Pi uptake in an ex vivo assay in fresh PBMCs from patient EXT-444-001 and 2 controls, as

previously described.^{16,36,37,40} Similar to what we previously reported in *SLC20A2* loss-of-function mutation carriers using the same assay,⁴⁰ we observed a 39.3% decrease of phosphate uptake with no significant effect on phosphate efflux (Fig. 2B, Supporting Information Fig. e-3). Efflux and uptake analyses were replicated with the same results.

To further assess the role of the GH08J042587 regulatory element, we deleted a genomic region containing this element in HEK293 cells using CRISPR/Cas9. We designed 2 pairs of guide RNAs surrounding the enhancer, resulting in 4 predicted deletions of 5476 to 4872 pb (Fig. 1B; detailed in Supporting Information Table e-3). Two days after transfection, cells were collected and DNA and RNA analyses were carried out. First, ddPCR analysis targeting GH08J042587 on bulk DNA revealed that transfected cells carried a 1.36 average copy number of GH08J042587 (Supporting Information Fig. e-4), confirming the efficiency of the Cas9-mediated genomic deletion while the copy number remained around 2.0 in control cells, that is, cells transfected with Cas9 plus a nonrelevant guide RNA, cells transfected with the plasmid encoding the Cas9 only, and nontransfected cells. DNA analysis of the

flanking region by ddPCR further confirmed the absence of any off-target deletion in this region, notably in the *SLC20A2* coding sequences (Supporting Information Fig. e-5). We then performed RT-ddPCR from the same samples. The average ratio of relative *SLC20A2* mRNA levels in HEK293 cells transfected with the different gRNA targeting the GH08J042587 region versus control cells was 77.4% ($P < 0.001$). Relative to the number of GH08J042587 copies lost in the cellular bulks, we observed a 39.3% decrease in *SLC20A2* mRNA levels. This can be compared with the 39.8% decreased mRNA level on average per GH08J042587 DNA copy in patients who carry the deletion in the germline (Fig. 2C, Supporting Information Fig. e-4). In addition, we screened GH08J042587 copy number across 183 genetically unsolved patients with PFBC or calcifications of unknown cause, including the 70 probands with a WES showing no PFBC causal variant. For all of these samples, a normal copy number of 2 was found in the GH08J042587 genomic localization (Supporting Information Table e-4), suggesting that such a deletion is an extremely rare cause of PFBC.

Description of the Family

The proband (EXT-444-001) presented to a neurological consultation at the age of 60 years with memory complaint, dizziness, resting tremor, muscular cramps, and cephalalgia. Neurological examination was normal with the exception of the tremor. His medical history included type 2 diabetes with peripheral neuropathy, coronary heart disease, and sleep apnea. During follow-up, he developed behavioral disturbances including irritability and aggressivity and he was treated for a depressive episode. At the age of 63, neuropsychological assessment revealed a memory impairment with dysexecutive features. The Mini Mental State Examination score was 26/30 (Norman range, N), and free and cued selective reminding test showed a total of 17/48 free recall (Pathological value, P) with inefficient cueing (sum of total recalls: 36/48, P). The Brixton test results showed 35 errors (P); Trail Making Test Part A was rated as normal (35 seconds), although he presented 1 error during part B (94 seconds); the Stroop test was rated as normal; and verbal categorical fluencies were poor (27 animals, 2 minutes, P), whereas literal fluencies were normal (24 words, letter p, 2 minutes). He showed normal gestural praxis assessment (8/8) and 39/40 correctly nominated words. Brain computed tomography scan performed at the age of 60 years detected bilateral calcifications in the lenticular (severe), caudate (faint) and dentate (faint) nuclei, and both thalami (faint) as well as faint calcifications in the subcortical white matter. Based on our visual rating scale,⁴¹ the total calcification score was 24/80. He had normal

calcium, phosphorus, and parathyroid hormone levels in blood.

His half-brother, EXT-444-002, presented an akinetic-hypertonic syndrome starting around the age of 60 years. Upon examination at the age of 61, he presented no tremor and was considered weakly dopa sensitive. The akinetic-hypertonic syndrome was considered severe. He showed additional movement disorders including dystonia and dyskinetic movements. He complained from visual nocturnal hallucinations. He also presented with cognitive impairment with memory impairment and a dysexecutive syndrome. The Mini Mental State Examination score was 20/30. The free and cued selective reminding test showed a total of 13/48 free recalls (P) with inefficient cueing (sum of total recalls: 36/48, P). The frontal assessment battery scored 12/18 (threshold), verbal fluencies were limited to 3 words (p letter, P), and the Wisconsin Card Sorting Test scale scored 2/6 (P) with 12 perseverative errors (P). The Trail Making Test Part A was rated as normal (127 seconds), although he stopped part B early. Conversely, he showed normal gestural praxis assessment (6/8) and correct denomination of 78 of 80 words (N). Upon follow-up, he presented a progressive worsening of the parkinsonian syndrome including dysarthria. He had a medical history of restless leg syndrome and high blood pressure. Brain computed tomography scan showed, at age 67, severe and confluent bilateral lenticulo-candate calcifications, severe calcifications of both thalami, severe calcifications of both cerebellar hemispheres and the vermis, faint cortical occipital calcifications, and rare punctate bilateral calcifications of the subcortical white matter. Thus, the total calcification score was 37/80. He had calcium blood levels in normal ranges.

Their father died at age 82 with no neurological or psychiatric medical history.

The son of EXT-444-002, EXT-444-003, was known to exhibit abnormal brain calcifications on computed tomography scan while asymptomatic at the age of 40 years, but the images were not available for rating, and he could not be examined.

Discussion

We report a novel 8p11.21 heterozygous deletion located 35 kb upstream of the major autosomal-dominant PFBC causative gene *SLC20A2*. Although this deletion did not encompass any coding or promoter base pair, we showed that carriers of the deletion presented decreased *SLC20A2* mRNA levels and that phosphate uptake activity of the transporter was reduced in the proband, mimicking the expected effect of heterozygous premature stop codon introducing variants. After *in vitro* assessment, we demonstrated a significant reduction of *SLC20A2* mRNA levels on cells

carrying the targeted deletion of the GH08J042587 putative enhancer, which is predicted to directly interact with the *SLC20A2* promoter.

This CNV is a nonpolymorphic deletion, with no similar event in the Database of Genomic Variants⁴² or in gnomAD data.⁴³ The deletion involves 2 protein-coding genes—*CHRNA6* and *CHRNA3*—that encode subunits of the nicotinic acetylcholine receptor. These 2 genes are not linked to human diseases and are tolerant to haploinsufficiency according to the gnomAD database⁴⁴ (probability of loss-of-function intolerance of 0 for both genes) in addition to the presence of some overlapping small deletions in the Database of Genomic Variants gold standard database. Comparatively, *SLC20A2* is strongly intolerant to loss-of-function variations (probability of loss-of-function intolerance = 0.97) with no overlapping CNV in the Database of Genomic Variants gold standard database. Hence, variant statistics in public databases are not compatible with the hypothesis that haploinsufficiency of either *CHRNA3* or *CHRNA6* can be causal of a rare Mendelian disorder such as PFBC.

Further characterization of this deletion using targeted techniques confirmed that neither the *SLC20A2* gene nor its promoter were affected. Segregation data in brain calcification carriers in this family suggested a possible impact of this deletion through a distant mechanism that we could demonstrate using *in vivo* and *in vitro* arguments. We assessed here the consequences of the deletion on *SLC20A2* expression and PiT2 function because significantly reduced expression of *SLC20A2* and/or function of PiT2 has been shown to be sufficient to cause autosomal-dominant PFBC.¹⁴ However, downstream mechanisms at the neurovascular unit remain to be better understood in patients with a loss-of-function variant of *SLC20A2*.^{45,46} We provide here compelling evidence that this deletion is the likely cause of PFBC in this family because of the measurement of haploinsufficiency in patients and the observation of reduced mRNA levels in a cellular model with a different genomic context. However, although the patients were extensively assessed by WES, we cannot rule out the existence of a nondetected mutation in another gene or a cryptic genetic alteration of *SLC20A2* affecting splicing or gene expression in other noncoding regions.

Overall, we applied here a straightforward approach both for the identification of candidate regulatory CNVs and for the assessment of their effect in 3 steps: annotation, mRNA assessment, and *in vitro* confirmation (Supporting Information Fig. e-6).

First, it is critical to perform an accurate annotation of CNV calls (including surrounding regions) from next-generation-sequencing-based bioinformatics tools using gene candidate regulatory elements information or the use of external databases including the display in web browsers, especially in regions surrounding core Mendelian genes for which dosage variation—including haploinsufficiency—is a known mechanism. To assess the presence of putative

regulatory elements, numerous tools have been made available. We chose to preferably use the integrative resource Genhancer. This tool is based on data obtained from previously published major studies (ENCODE, Ensembl, VISTA, FANTOM) and associates the results of biological assays to *in silico* predictions. From these data, combinatorial likelihood-based scores for enhancer–gene pairing were generated.⁹ Here, the interaction between GH08J042587 and the *SLC20A2* promoter was classified as “double elite,” reflecting both a high-likelihood enhancer definition and a strong enhancer–gene association.

Second, the study of the distant event effects on mRNA levels requires reliable tools for quantifying gene expression. Here, we adapted our universal locked nucleic acid–hydrolysis probe-based ddPCR protocol to cDNA to perform the relative quantification of transcripts. The same technique was successfully used as a read-out for the CRISPR/Cas9 assay in the HEK293 cells. This approach is an easy-to-use technique for measuring changes in gene expression that could be used in routine diagnostics as it takes advantage of the benefits of universal ddPCR that we have already described. It is indeed considered as a fast, robust, reliable, cost-effective method that avoids the design of new hydrolysis probes for each new assay.³⁴ These technical advantages make RT-ddPCR easier to use routinely. In addition, we provide here the results of Pi uptake in PBMCs of one carrier with the aim to assess the protein function level. Although we acknowledge the limitation of this result because of the assessment in the proband only, previous results demonstrated that it is highly reproducible.^{16,36,40}

Third, *in vitro* assessments allowed the confirmation, in a different genomic context, that the targeted deletion of the GH08J042587 was sufficient to induce an mRNA-level reduction. Here, the efficiency of CRISPR/Cas9 in HEK293 cells, the high reproducibility of RT-ddPCR, and the straightforward interpretation of the read-out allowed a bulk analysis, significantly limiting the manipulating time as compared to the generation and characterization of cellular clones. Of note, we focused here on the GH08J042587 regulatory element in our *in vitro* assays, but we cannot rule out the hypothesis that disrupting surrounding regions could also alter *SLC20A2* mRNA levels, for example, through a distinct mechanism or simply by disrupting the 3-dimensional conformation of the chromatin and hence impacting the same enhancer.

One of the limitations of CRISPR/Cas9 that needs to be controlled is the risk of off-target mutation introduction. We verified here that our assays did not introduce any copy number variant in the surrounding regions and in the coding sequence of *SLC20A2*, although some other types of variations might still be introduced. We controlled this risk by (1) selecting gRNAs carefully using the CRISPOR tool, (2) checking that putative off

targets predicted by the CRISPOR tool did not map to the *SLC20A2* region, (3) using 4 different combinations of guide RNAs, and (4) performing a bulk transfection.

Interestingly, this deletion could be detected by exome sequencing because it implicated coding sequences of 2 nearby genes. It is very likely that purely intergenic deletions of smaller size, not detectable by exome sequencing, may lead to the same effect on gene expression. The screening of our series did not reveal any other deletion using a ddPCR amplicon targeting this enhancer. However, one can expect that the growing use of whole genome sequencing as a first-tier or second-tier diagnostic tool may lead to the identification of additional genomic disruptions of gene regulatory elements in the vicinity of Mendelian genes.

In conclusion, we identified an enhancer of the expression of the *SLC20A2* gene, the deletion of which was sufficient to result in haploinsufficiency in the same ranges as the one observed in heterozygous premature stop codon variant carriers. This observation is a clear example that, when a given phenotype is strongly suggestive of a Mendelian disease, the search for cryptic and/or distant events in already known genes is as important as the search for new genes. In addition, our observation opens the way to therapeutic targets aimed at regulating gene expression. ■

Acknowledgments: We thank the family for their participation, Annick Steinmetz for her technical support, and Françoise Bille-Turc for providing clinical information. This work was performed thanks to the collaboration CEA-DRF-Jacob-CNRGH-CHU de Rouen.

References

1. Wawrocka A, Krawczynski MR. The genetics of aniridia—simple things become complicated. *J Appl Genet* 2018;59:151–159.
2. Bhatia S, Bengani H, Fish M, et al. Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved *PAX6* enhancer causes aniridia. *Am J Hum Genet* 2013;93:1126–1134.
3. Lettice LA, Heaney SJH, Purdie LA, et al. A long-range *Shh* enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* 2003;12:1725–1735.
4. Short PJ, McRae JF, Gallone G, et al. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* 2018;555:611–616.
5. Kellis M, Wold B, Snyder MP, et al. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* 2014;111:6131–6138.
6. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res* 2007;35:D88–D92.
7. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57–74.
8. Zerbino DR, Achuthan P, Akanni W, et al. Ensembl 2018. *Nucleic Acids Res* 2018;46:D754–D761.
9. Fishilevich S, Nudel R, Rappaport N, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database J Biol Databases Curation* 2017;2017. <https://pubmed.ncbi.nlm.nih.gov/28605766/> or <https://academic.oup.com/database/article/doi/10.1093/database/bax028/3737828>
10. Nicolas G, Pottier C, Maltête D, et al. Mutation of the *PDGFRB* gene as a cause of idiopathic basal ganglia calcification. *Neurology* 2013;80:181–187.
11. Ramos EM, Oliveira J, Sobrido MJ, Coppola G. Primary familial brain calcification. In: Adam MP, Ardinger HH, Pagon RA, et al., eds. *GeneReviews*®. Seattle, WA: University of Washington; 1993. <http://www.ncbi.nlm.nih.gov/books/NBK1421/>. Accessed March 7, 2019.
12. Nicolas G, Charbonnier C, de Lemos RR, et al. Brain calcification process and phenotypes according to age and sex: lessons from *SLC20A2*, *PDGFB*, and *PDGFRB* mutation carriers. *Am J Med Genet Part B Neuropsychiatr Genet* 2015;168:586–594.
13. Grangeon L, Wallon D, Charbonnier C, et al. Biallelic *MYORG* mutation carriers exhibit primary brain calcification with a distinct phenotype. *Brain J Neurol* 2019;142(6):1573–1586.
14. Wang C, Li Y, Shi L, et al. Mutations in *SLC20A2* link familial idiopathic basal ganglia calcification with phosphate homeostasis. *Nat Genet* 2012;44:254–256.
15. Keller A, Westenberger A, Sobrido MJ, et al. Mutations in the gene encoding *PDGF-B* cause brain calcifications in humans and mice. *Nat Genet* 2013;45:1077–1082.
16. Legati A, Giovannini D, Nicolas G, et al. Mutations in *XPR1* cause primary familial brain calcification associated with altered phosphate export. *Nat Genet* 2015;47:579–581.
17. Yao X-P, Cheng X, Wang C, et al. Biallelic mutations in *MYORG* cause autosomal recessive primary familial brain calcification. *Neuron* 2018;98:1116–1123.e5.
18. Cen Z, Chen Y, Chen S, et al. Biallelic loss-of-function mutations in *JAM2* cause primary familial brain calcification. *Brain J Neurol* 2020;143(2):491–502.
19. Ramos EM, Carecchio M, Lemos R, et al. Primary brain calcification: an international study reporting novel variants and associated phenotypes. *Eur J Hum Genet* 2018;26:1462–1477.
20. Hozumi I, Kurita H, Ozawa K, et al. Inorganic phosphorus (Pi) in CSF is a biomarker for *SLC20A2*-associated idiopathic basal ganglia calcification (IBGC1). *J Neurol Sci* 2018;388:150–154.
21. Jensen N, Schröder HD, Hejbøl EK, et al. Mice knocked out for the primary brain calcification-associated gene *Slc20a2* show unimpaired prenatal survival but retarded growth and nodules in the brain that grow and calcify over time. *Am J Pathol* 2018;188:1865–1881.
22. Baker M, Strongosky AJ, Sanchez-Contreras MY, et al. *SLC20A2* and *THAP1* deletion in familial basal ganglia calcification with dystonia. *Neurogenetics* 2014;15:23–30.
23. David S, Ferreira J, Quenez O, et al. Identification of partial *SLC20A2* deletions in primary brain calcification using whole-exome sequencing. *Eur J Hum Genet* 2016;24:1630–1634.
24. Grütz K, Volpato CB, Domingo A, et al. Primary familial brain calcification in the “IBGC2” kindred: all linkage roads lead to *SLC20A2*. *Mov Disord* 2016;31:1901–1904.
25. Guo X-X, Su H-Z, Zou X-H, et al. Identification of *SLC20A2* deletions in patients with primary familial brain calcification. *Clin Genet* 2019;96(1):53–60.
26. Mu W, Tochen L, Bertsch C, Singer HS, Barañano KW. Intracranial calcifications and dystonia associated with a novel deletion of chromosome 8p11.2 encompassing *SLC20A2* and *THAP1*. *BMJ Case Rep* 2019;12. <https://doi.org/10.1136/bcr-2018-228782>
27. Pasanen P, Mäkinen J, Myllykangas L, et al. Primary familial brain calcification linked to deletion of 5′ noncoding region of *SLC20A2*. *Acta Neurol Scand* 2017;136:59–63.
28. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405–424.
29. Backenroth D, Homsy J, Murillo LR, et al. CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res* 2014;42:e97.
30. Le Guennec K, Quenez O, Nicolas G, et al. 17q21.31 duplication causes prominent tau-related dementia with increased *MAPT* expression. *Mol Psychiatry* 2017;22:1119–1125.

31. Quenez O, Cassinari K, Coutant S, et al. Detection of copy number variations from NGS data using read depth information: a diagnostic performance evaluation. <https://hal-normandie-univ.archives-ouvertes.fr/hal-02317979>. Accessed November 7, 2019.
32. Vogelstein B, Kinzler KW. Digital PCR. *Proc Natl Acad Sci* 1999; 96:9236–9241.
33. Huggett JF, Foy CA, Benes V, et al. The digital MIQE guidelines: minimum information for publication of quantitative digital PCR experiments. *Clin Chem* 2013;59:892–902.
34. Cassinari K, Quenez O, Joly-Hélas G, et al. A simple, universal, and cost-efficient digital PCR method for the targeted analysis of copy number variations. *Clin Chem* 2019;65(9):1153–1160.
35. Carithers LJ, Ardlie K, Barcus M, et al. A novel approach to high-quality postmortem tissue procurement: the GTEx Project. *Bio-preserv Biobank* 2015;13:311–319.
36. Anheim M, López-Sánchez U, Giovannini D, et al. XPR1 mutations are a rare cause of primary familial brain calcification. *J Neurol* 2016;263:1559–1564.
37. López-Sánchez U, Nicolas G, Richard A-C, et al. Characterization of XPR1/SLC53A1 variants located outside of the SPX domain in patients with primary familial brain calcification. *Sci Rep* 2019;9:6776.
38. Concordet J-P, Haeussler M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res* 2018;46:W242–W245.
39. Haeussler M, Schönig K, Eckert H, et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol* 2016;17:148.
40. López-Sánchez U, Tury S, Nicolas G, et al. Interplay between PFBC-associated SLC20A2 and XPR1 phosphate transporters requires inositol polyphosphates for control of cellular phosphate homeostasis [published online ahead of print, 2020 May 11]. *J Biol Chem*. 2020; jbc.RA119.011376. <https://doi.org/10.1074/jbc.RA119.011376>
41. Nicolas G, Pottier C, Charbonnier C, et al. Phenotypic spectrum of probable and genetically-confirmed idiopathic basal ganglia calcification. *Brain J Neurol* 2013;136:3395–3407.
42. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* 2014;42: D986–D992.
43. Collins RL, Brand H, Karczewski KJ, et al. An open resource of structural variation for medical and population genetics. *bioRxiv*. <https://doi.org/10.1101/578674>
44. Karczewski KJ, Francioli LC, Tiao G, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*. <https://doi.org/10.1101/531210>
45. Zarb Y, Weber-Stadlbauer U, Kirschenbaum D, et al. Ossified blood vessels in primary familial brain calcification elicit a neurotoxic astrocyte response. *Brain* 2019;142:885–902.
46. Nahar K, Lebouvier T, Andaloussi Mäe M, et al. Astrocyte-microglial association and matrix composition are common events in the natural history of primary familial brain calcification [published online ahead of print September 27, 2019]. *Brain Pathol Zurich Switz*. <https://doi.org/10.1111/bpa.12787>
47. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*. *Bioinformatics* 2009;25(14):1754–1760.
48. McKenna A, Hanna M, Banks E. The genome analysis toolkit: a map reduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.

Supporting Data

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.

2. Caractérisation des conséquences transcriptionnelles d'une triplification du locus *APP* dans la maladie d'Alzheimer avec angiopathie amyloïde cérébrale

2.1. Contexte et résumé des travaux :

L'angiopathie amyloïde cérébrale (AAC) est une pathologie commune du sujet âgé, définie par le dépôt de substance amyloïde (peptide A β) dans la paroi des vaisseaux corticaux et leptoméningés, conduisant à des hématomes cérébraux lobaires, des microhémorragies, une hémorragie, et associée cliniquement à des troubles cognitifs, des troubles neurologiques transitoires ainsi que les conséquences cliniques des hématomes. La maladie d'Alzheimer (MA), responsable d'un trouble neurocognitif majeur de nature neurodégénérative, est également une pathologie commune chez le sujet âgé, et elle est définie sur le plan neuropathologique par l'aggrégation intraparenchymateuse du même peptide A β dans les plaques amyloïdes d'une part, et des dégénérescences neurofibrillaires intraneuronales, composées de protéine Tau hyper et anormalement phosphorylée. AAC et MA sont souvent associées chez un même individu, comme le montrent les études autopsiques [168], bien que tous les patients avec MA n'en présentent pas les symptômes ou les stigmates en IRM *in vivo*. Outre les formes communes d'AAC et de MA, de déterminisme complexe, multifactoriel, il existe de rares formes monogéniques.

La protéine précurseur de l'amyloïde- β (APP) est codée par le gène *APP*, principal gène connu pour être impliqué dans l'AAC de transmission autosomique dominante. Après la description des mutations ponctuelles, les duplications du locus *APP* ont été identifiées en 2006 comme une cause d'AAC à début précoce et/ou de la maladie d'Alzheimer à début précoce (MAJ ou *Early Onset Alzheimer Disease*, EOAD, premiers symptômes avant 65 ans) [169]. Ces variations du nombre de copies englobent au moins le gène *APP*, avec ou sans les gènes environnants sur le chromosome 21 [170]. La taille de la duplication et le contenu du gène -

au-delà du gène *APP*, critique - ne semblent pas influencer la présentation phénotypique des cas [169], [171]–[174]. Les patients atteints de trisomie 21 présentent généralement une MAJ, bien qu'ils présentent moins fréquemment des hématomes cérébraux, ce qui suggère la présence de mécanismes (relativement) protecteurs chez les patients atteints du syndrome de Down, dans le cadre d'une trisomie 21 [175]. En effet, une étude histopathologique récente a montré une AAC plus sévère mais une formation de plaques amyloïdes parenchymateuses moins importante chez les patients atteints de duplication d'*APP* que chez les patients atteints de trisomie 21 [175]. Étant donné cette diversité dans la distribution des lésions amyloïdes en fonction du contexte génétique sous-jacent, la pathogenèse de l'AAC reste à comprendre. Nous avons ici travaillé sur la première triplication du locus *APP*, en la caractérisant sur le plan génomique et de l'expression du gène *APP*.

Les duplications du gène *APP*, comme les trisomies 21, sont associées à une multiplication par 1.5 de la quantité relative d'ARNm d'*APP*, en moyenne par rapport à des témoins, dans le sang périphérique [102]. Une technique de RT-QMPSF avait été mise au point dans le laboratoire, permettant de confirmer cette hypothèse *a priori* et ainsi le mécanisme pathogénique [102]. En effet, l'hypothèse est celle que cette augmentation est la même dans le cerveau que dans le sang, et l'augmentation d'expression résulterait ainsi en une augmentation du peptide A β lui-même, issu du clivage d'*APP*. Cette technique de RT-QMPSF avait également été utilisée dans un autre travail caractérisant une délétion de deux paires de bases dans le 3'UTR d'*APP*, chez un patient avec des niveaux d'ARNm d'*APP* similaires aux patients avec duplication, dans le sang [103]. Néanmoins, cette technique, basée sur l'amplification en parallèle de plusieurs amplicons après reverse transcription, et réaction de PCR multiplexe stoppée en phase exponentielle de la PCR, est sujette à une certaine variabilité inter-individuelle inhérente à la technique. Nous avons donc souhaité développer une technique plus robuste basée sur la RT-ddPCR pour pallier ces problèmes.

Le cas index est un homme de 41 ans adressé pour l'évaluation d'un déclin cognitif progressif. L'IRM cérébrale et les biomarqueurs du liquide cérébro-spinal étaient évocateurs d'une MA avec AAC. Son père est décédé à l'âge de 48 ans d'un hématome cérébral dans un contexte d'AAC évoluant depuis ses 37 ans et de déclin cognitif progressif évoquant une MA. L'analyse par QMPSF avait révélé 4 copies du gène *APP* chez le cas index. Sa mère, 68 ans, non atteinte, est quant à elle porteuse de 2 copies du gène, suggérant un génotype 3 + 1 chez le patient et son père, génotype que nous avons pu confirmer par FISH métaphasique. En CGH-array (Agilent 180K), la triplication 21q21.3 de 506 kb (chr21:27,156,233-27,662,338;hg19) était restreinte au gène *APP*. Nous avons donc évalué les niveaux d'ARNm sanguin d'*APP* du cas index, par RT-ddPCR, et identifié un résultat conforme à l'hypothèse *a priori*, à savoir un doublement de la quantité relative par rapport aux témoins (N=10) et supérieur à la moyenne des porteurs d'une duplication d'*APP*, confirmant que les 4 copies sont fonctionnelles (N=9). Il s'agit du premier cas de triplication du locus *APP*, provoquant un doublement du niveau d'ARNm d'*APP*, dans une famille présentant une MA avec AAC autosomique dominante, parmi les plus précoces et sévères comparé aux descriptions des porteurs de duplications, bien que la différence d'âge de début avec les plus jeunes patients porteurs de duplication d'*APP* ne soit pas flagrante, et que nous ne disposions que d'une seule observation. Cette observation suggère que le mécanisme décrit dans les duplications d'*APP* (augmentation de la production d'APP et, par conséquent, du peptide A β issu de son clivage, conduisant à des dépôts amyloïdes précoces) est encore accentué par la présence d'une quatrième copie du gène, qui reste fonctionnelle. Ce type de sensibilité au dosage génique, illustré par une triplication conduisant à un phénotype proche, mais probablement plus sévère, que de celui de la duplication correspondante, constitue un exemple relativement rare en génétique médicale, à l'instar des duplications/triplications du gène *SNCA* dans les synucléinopathies (maladie de Parkinson / maladie à corps de Lewy). Ici, le recours à la RT-ddPCR a également permis une analyse fiable, bien que les analyses réalisées sur les porteurs

de duplication d'*APP* à titre de comparaison montraient une certaine variabilité inter-individuelle, possiblement en lien avec la diversité des ARN et des dates de prélèvement des porteurs, et pas uniquement pour des raisons biologiques ou techniques.

Ma contribution personnelle à ce travail a été, d'une part, la caractérisation cytogénétique de la triplication détectée chez le cas index par (i) CGH-array (Agilent 180k), qui a permis de border la triplication et d'établir son contenu génique (restreint à *APP*), (ii) caryotype (qui compte tenu de sa résolution n'a pas apporté d'information supplémentaire sur ce CNV) ainsi que la réalisation de techniques de fluorescence in situ (FISH) qui ont permis de confirmer le CNV et ont montré qu'un chromosome 21 était porteur de 3 copies de ce locus *APP* suggérant une ségrégation 3+1. Par ailleurs, j'ai réalisé les explorations cytogénétiques chez la mère du cas index, dont les explorations étaient normales. D'autre part, ma contribution principale, en lien avec ces travaux de thèse, a consisté en la caractérisation de l'impact de la triplication sur la production d'ARNm d'*APP*. J'ai mis au point une technique de RT-ddPCR one-time, basée sur la quantification relative de l'ARNm d'*APP* par rapport à un gène de ménage choisi pour ses niveaux similaires d'expression par rapport à *APP* dans le sang, et réalisé les analyses en RT-ddPCR en comparaison à des témoins positifs porteurs de duplication limitée à *APP* ou de trisomie 21 et des témoins sains.

Outre l'article scientifique présenté ci-dessous, ce travail a fait l'objet de plusieurs présentations affichées dans des congrès de génétique, à savoir :

- Poster aux Assises de Génétique en 2020, Tours
- Poster au congrès 2021 de l'ESHG, Vienne

ARTICLE OPEN ACCESS

Early-Onset Cerebral Amyloid Angiopathy and Alzheimer Disease Related to an APP Locus Triplication

Lou Grangeon, MD, Kévin Cassinari, MD, Stéphane Rousseau, BSc, Bernard Croisile, MD, Maité Formaglio, MD, Olivier Moreaud, MD, Jean Boutonnat, MD, Nathalie Le Meur, PharmD, Manuele Miné, PharmD, PhD, Thibault Coste, MD, Eva Pipiras, MD, PhD, Elisabeth Tournier-Lasserre, MD, PhD, Anne Rovelet-Lecrux, PhD, Dominique Campion, MD, PhD, David Wallon, MD, PhD, and Gael Nicolas, MD, PhD

Correspondence
Dr. Nicolas
gaelnicolas@hotmail.com

Neurol Genet 2021;7:e609. doi:10.1212/NXG.0000000000000609

Abstract

Background and Objective

To report a triplication of the amyloid- β precursor protein (*APP*) locus along with relative messenger RNA (mRNA) expression in a family with autosomal dominant early-onset cerebral amyloid angiopathy (CAA) and Alzheimer disease (AD).

Methods

Four copies of the *APP* gene were identified by quantitative multiplex PCR of short fluorescent fragments, fluorescent in situ hybridization (FISH), and array comparative genomic hybridization. *APP* mRNA levels were assessed using reverse-transcription–digital droplet PCR in the proband's whole blood and compared with 10 controls and 9 *APP* duplication carriers.


Results

Beginning at age 39 years, the proband developed severe episodic memory deficits with a CSF biomarker profile typical of AD and multiple lobar microbleeds in the posterior regions on brain MRI. His father had seizures and recurrent cerebral hemorrhage since the age of 37 years. His cerebral biopsy showed abundant perivascular amyloid deposits, leading to a diagnosis of CAA. In the proband, we identified 4 copies of a 506-kb region located on chromosome 21q21.3 and encompassing the whole *APP* gene without any other gene. FISH suggested that the genotype of the proband was 3 copies/1 copy corresponding to an *APP* locus triplication, which was consistent with the presence of 2 *APP* copies in the healthy mother and with the paternal medical history. Analysis of the *APP* mRNA level showed a 2-fold increase in the proband and a 1.8 fold increase in *APP* duplication carriers compared with controls.

Discussion

Increased copy number of *APP* is sufficient to cause AD and CAA, with likely earlier onset in case of triplication compared with duplication.

RELATED ARTICLE

 **Editorial**
The Dose Makes the
Poison
Page e610

From the Department of Neurology and CNR-MAJ (L.G., D.W.), Normandie University, UNIROUEN, Inserm U1245, CHU Rouen, CIC-CRB1404, F 76000; Department of Genetics and CNR-MAJ (K.C., S.R., N.L.M., A.R.-L., D.C., G.N.), Normandie University, UNIROUEN, Inserm U1245 and CHU Rouen, F 76000; Department of Neurology (B.C., M.F.), Lyon University Hospital; Department of Neurology (O.M.), Grenoble University Hospital; Department of Histology (J.B.), Grenoble University Hospital; AP-HP (M.M., T.C., E.T.-L.), Groupe Hospitalier Saint-Louis Lariboisière-Fernand-Widal, Service de Génétique Moléculaire Neurovasculaire, INSERM UMR 1141, NeuroDiderot, Université de Paris; Department of Histology Embryology and Cytogenetics (E.P.), Jean Verdier Hospital; Paris 13 University (E.P.), Sorbonne Paris Cité, UFR SMBH Bobigny; and PROTECT (E.P.), INSERM, Paris Diderot University, Bondy, France.

Go to Neurology.org/NG for full disclosures. Funding information is provided at the end of the article.

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND), which permits downloading and sharing the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

Glossary

aCGH = array comparative genomic hybridization; AD = Alzheimer disease; APP = amyloid- β precursor protein; CAA = cerebral amyloid angiopathy; EOAD = early-onset Alzheimer disease; MMSE = Mini Mental State Examination; mRNA = messenger RNA; QMPFS = quantitative multiplex PCR of short fluorescent fragments.

Amyloid- β precursor protein (APP) is the main gene known to be involved in autosomal dominantly inherited cerebral amyloid angiopathy (CAA). After the description of point mutations, APP locus duplications were identified in 2006 as a cause of early-onset CAA and/or early onset Alzheimer disease (EOAD, onset before 65 years).¹ Such copy number variations encompass at least the APP gene, with or without surrounding genes on chromosome 21,² and are associated with ~1.5-fold increased messenger RNA (mRNA) levels in blood compared with healthy controls, in similar ranges to patients with trisomy 21.³ The size of the duplication and the gene content—beyond the critical APP gene—do not appear to influence phenotypic presentation in cases.^{1,4-6} Patients with trisomy 21 usually show cognitive impairment similar to Alzheimer disease (AD), although they less frequently exhibit cerebral hematoma, suggesting the presence of protective mechanisms in patients with Down syndrome.⁷ Indeed, a recent histopathologic study showed more severe CAA but less parenchymal amyloid plaque formation in APP duplication than in trisomy 21 patients.⁷ Given this diversity in amyloid distribution according to the underlying genetic background, CAA pathogenesis remains to be understood. We here report an APP locus triplication, along with relative mRNA expression in the proband's blood.

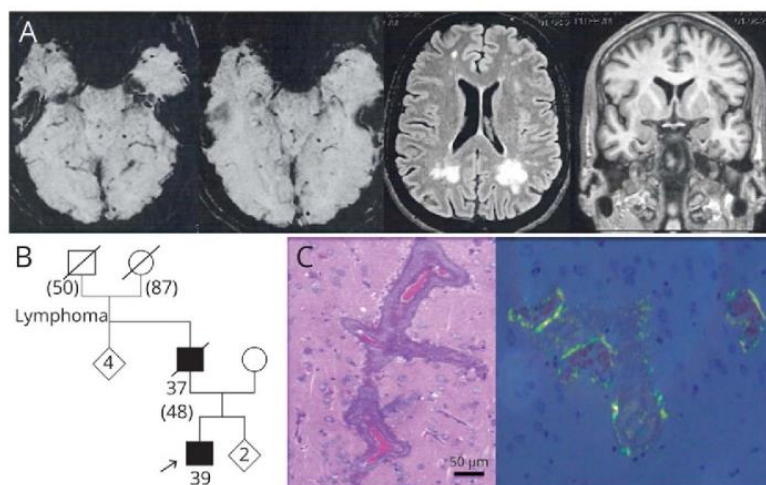
Case Presentation

A 41-year-old man without medical history was referred to a memory care center for the evaluation of progressive

cognitive decline.⁸ From age 39 years, he presented with an aggressive course of episodic memory loss making him unable to maintain his professional activity as well as attention deficit and executive dysfunction. The Mini Mental State Examination (MMSE) scored 18/30 and Frontal Assessment Battery 15/18. There was neither language impairment nor praxis or visuoconstructive dysfunction. The patient did not present any episode suggestive of stroke or seizures. Brain MRI showed multiple lobar microbleeds in the posterior fossa and occipital region and posterior periventricular leukoencephalopathy (Figure 1A) with hippocampal atrophy (bilateral Scheltens scale rating of 2). He underwent lumbar puncture for quantification of CSF AD biomarkers A β ₄₂, tau, and phosphorylated-tau. The A β ₄₂ level was decreased (404 ng/L, N > 550), with increased tau (491 ng/L, N < 400) and phospho-tau protein levels (95 ng/L, N < 60). The A β ₄₀ level was 8,546 ng/L (4,540 < N < 8,480). Overall, he fulfilled the diagnostic criteria of probable AD with evidence of the AD pathophysiologic process,⁸ in association with probable CAA following the modified Boston criteria, except the age criterion.

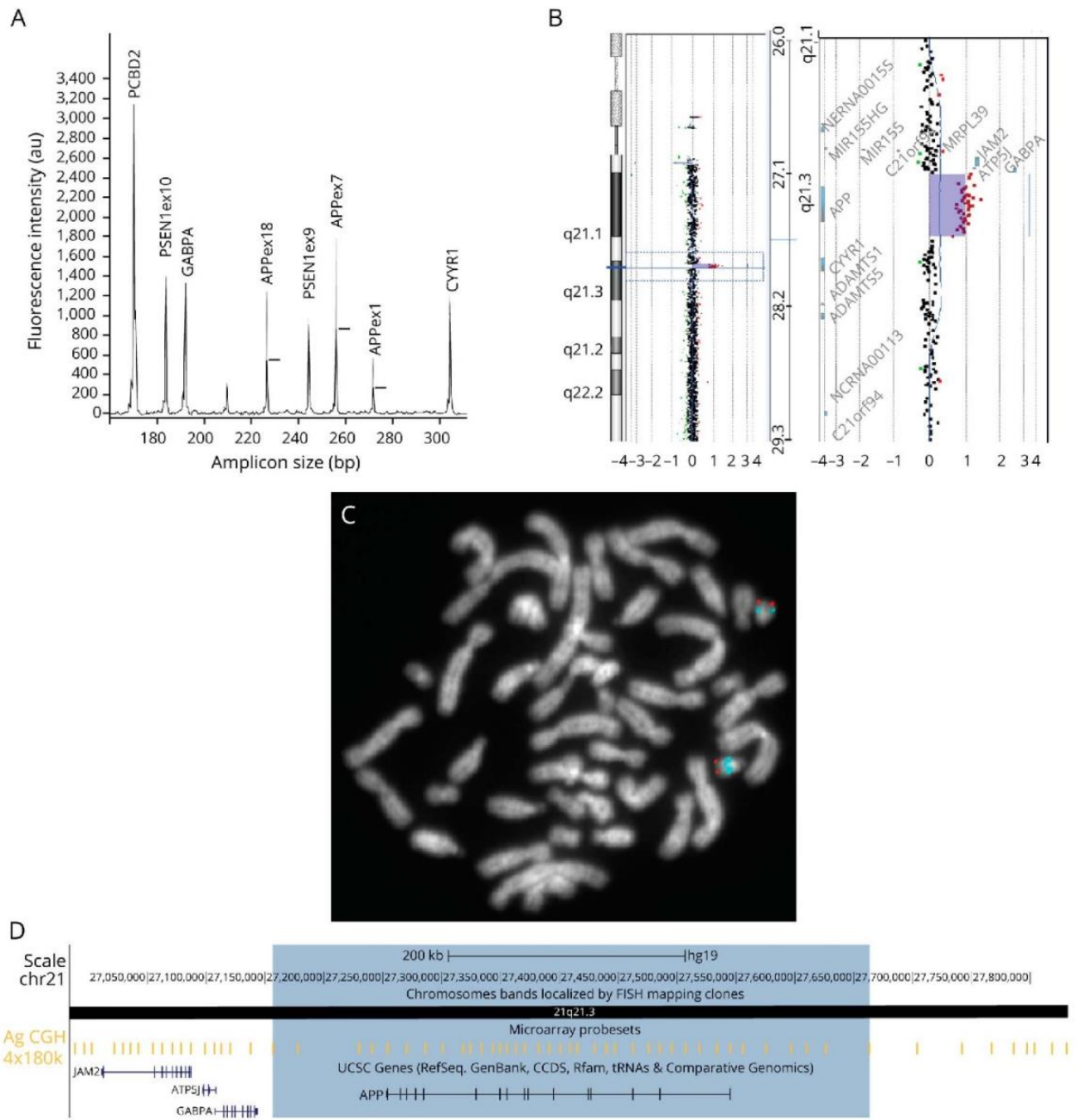
His father had had recurrent migraine with visual aura and presented at age 37 years with transient loss of consciousness (Figure 1B). Focal epilepsy was diagnosed after recurrent episodes of contact loss and electroencephalography showing slow bilateral temporal waves and then treated by carbamazepine. Subsequent cerebral MRI showed hyperintensities in

Figure 1 Partial Pedigree, Cerebral MRI of the Proband, and Histopathologic Examination of the Cerebral Biopsy Performed in His Father



(A) T2* weighted sequence showing multiple lobar microbleeds (first and second images); FLAIR-weighted sequence showing posterior leukoaraiosis (third image) and coronal view of T1-weighted sequence showing moderate bilateral hippocampal atrophy (fourth image). (B) Age at death (in parentheses) and age at onset are indicated. The proband is identified by an arrow. (C) Histopathologic examination of the cerebral biopsy of the proband's father. Bouin-fixed paraffin sections of the cerebral biopsy were stained with hematoxylin-eosin (left part) and Congo red (right part). Sections stained with Congo red were examined under crossed polarized light for analyzing vascular amyloid and revealed apple-green birefringence of amyloid material in blood vessel walls.

Figure 2 Representation of the 21q21.3 Triplication

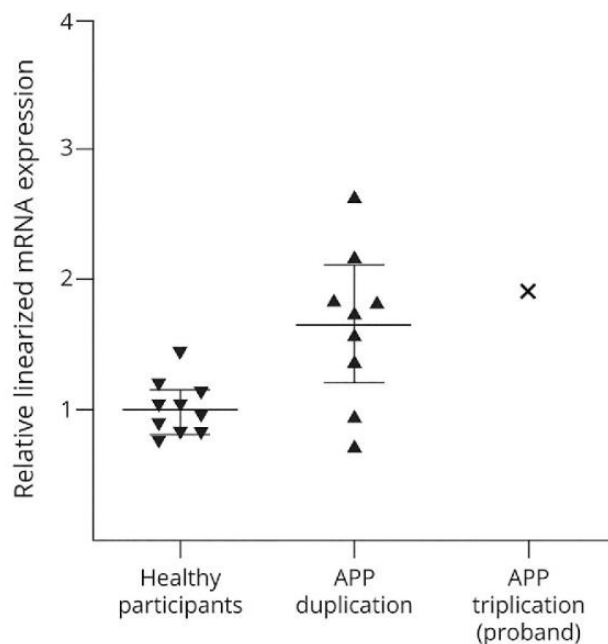


(A) Detection by QMPSF of an *APP* triplication. The electropherogram of the proband (in light gray) was superimposed on that of a normal individual (in black) by adjusting to the same level the peaks obtained from the control amplicon *PCBD2* located on chromosome 5. The vertical axis shows fluorescence in arbitrary units, and the horizontal axis indicates the size of the amplicon in base pairs. Horizontal bars indicate triplication of the amplicons, detected by a 2-fold heightening of the corresponding peaks. This QMPSF also covers 2 genes located at 21q21, *GABPA* and *CYR1* that are not duplicated. (B) Refinement of triplication breakpoints by array CGH. Representation on the Agilent Genomic Workbench 7.0 software of the 21q21.3 triplication (log ratio: 0.97) on chr21: 38,156,233–27,662,338;hg19. (C) FISH analysis of peripheral blood lymphocytes of the proband. Red: control probe located on chromosome 21 long arm subtelomeric region (VljyRM2029, Vysis; Abbott, Chicago, IL); Blue: *APP* locus specific probe (RP11-410J1; Empire Genomics, Buffalo, NY) located on 21q21.3. (D) Gene content of the triplicated region: Visualization in the UCSC genome browser of the 21q21.3 triplication chr21:27,156,233–27,662,338;hg19 (blue highlight). Yellow bars represent array CGH probes. Yellow bars at both extremes of the blue zone represent the last triplicated probes. The reference assembly used is GRCh37/hg19. Other panels represent, from top to bottom: Agilent 180K probes, cytogenetic band, UCSC genes (only RefSeq sequence). APP = amyloid- β precursor protein; FISH = fluorescent in situ hybridization.

centrum semiovale on T2-weighted sequences. CSF was acellular and showed a moderately high protein level (0.61 g/L), but AD biomarkers were not available at this time.

Progressive behavioral disorders and severe cognitive impairment occurred 4 years later and worsened over time with an MMSE of 24/30. Neuropsychological testing showed

Figure 3 Expression Analysis Using RT-ddPCR



RT-ddPCR with a target amplicon in the *APP* gene and a reference amplicon in the *DLG4* housekeeping gene for healthy controls (N = 10), patients carrying an *APP* duplication (N = 9), and our patient carrying an *APP* triplication (N = 1). Each point represents the average of 3 measures obtained for an individual (except for the proband where the point represents the average of 5 measures). Results are presented through an expression ratio *APP*/*DLG4* linearized with the average expression ratio of controls. For each group, the median relative *APP* mRNA expression is indicated with a wide horizontal line, and 95% confidence intervals are shown with short horizontal lines. *APP* = amyloid- β precursor protein; mRNA = messenger RNA; RT-ddPCR = reverse-transcription-digital droplet PCR.

episodic memory loss associated with praxis and visuoconstructive moderate dysfunction. He thus fulfilled the diagnostic criteria of probable AD⁸ and probable CAA. Tc99m-HMPAO cerebral single photon emission computerized tomography was considered normal. At age 44 years, his condition was extremely severe with repeated loss of consciousness episodes and major upper cerebral dysfunction. Cerebral MRI showed acute bilateral temporal hematomas and posterior leukoencephalopathy. Cerebral angiography did not display any sign of cerebral vasculitis. Cerebral biopsy revealed abundant perivascular amyloid deposits enabling the final diagnosis of probable CAA with supporting pathological evidence (Figure 1C). He died at age 48 years after a second hemorrhagic event with left temporal hematoma. There was no history of dementia or stroke in his family, although a censoring effect was noticed: the paternal grandfather of the proband indeed died at age 50 years from lymphoma (Figure 1B).

Genetic Assessment

We obtained informed written consent for genetic analyses by the patient and by the father's legal representative. This study was approved by the Institutional Review Board of Rouen University Hospital (CERDE #2019-55 notification).

Quantitative multiplex PCR of short fluorescent fragment (QMPSF) analyses performed from DNA isolated from fresh whole blood revealed 4 copies of the *APP* gene in the proband (Figure 2A). His *APOE* genotype was 33, and Sanger sequencing did not detect any point mutation in exons 16–17 of *APP*. The 68-year-old unaffected mother carried 2 copies of *APP*, suggesting a 3 + 1 genotype in the proband, but no DNA sample was available from his father. Fluorescent in situ hybridization using *APP* locus specific probes (RP11-410J1; Empire Genomics, Williamsville, NY) on cultured lymphocytes metaphase cells showed an asymmetric positive hybridization signal on chromosome 21 long arm without other signal hybridization anywhere else (Figure 2C), further suggesting a 3 + 1 genotype and hence autosomal dominant transmission of a triplication.

We further mapped the triplication using array comparative genomic hybridization (aCGH, Agilent SurePrint 4 × 180k; Agilent Technologies, Santa Clara, CA). The 506-kb 21q21.3 triplication (chr21:27,156,233-27,662,338;hg19) was restricted to the whole *APP* gene without any flanking gene coding sequence (Figure 2, B and D).

Finally, mRNA *APP* levels were assessed using reverse-transcription-digital droplet PCR in the proband whole blood comparatively to 10 normal controls (including the proband's mother) and 9 patients carrying *APP* duplications (see eMethods, links.lww.com/NXG/A465). The patient showed a 2-fold increase of relative *APP* mRNA levels compared with controls (Figure 3 and eFigure 1, links.lww.com/NXG/A470). Patients with *APP* duplication showed a median of 1.8 fold increase compared with controls, but no comparison could be directly made with our proband, considering this single *APP* triplication case.

Discussion

To our knowledge, this is the first report of an *APP* locus triplication, causing a two-fold upregulation of *APP* mRNA levels, in a family presenting with autosomal dominant EOAD with severe CAA. Although duplications of a given gene are now a classical cause of several autosomal dominant disorders, there are few examples of mendelian diseases caused by a gene triplication. In the field of neurodegenerative diseases, *SNCA* triplications on chromosome 4 have been reported in patients with Parkinson or Lewy body disease along with duplications in other patients.⁹ As for the *APP* gene, alpha-synuclein-encoding *SNCA* increased gene copies encompassed at least the *SNCA* gene, with or without surrounding genes (1 to 50), and displayed high diversity in clinical phenotype.¹⁰ Similarly, since its first description, some diversity of phenotypes associated with *APP* duplications has been described, mostly related to the predominance of AD or CAA-related symptoms at presentation.⁶ In this report, triplication was associated with diverse presentation, including severe cognitive disorder in the proband and recurrent ICH and seizures in his father.

The higher copy number seemed to be associated with earlier symptomatic phase in our report (37 and 39 years of age at onset) compared with *APP* duplication carriers showing ages at onset ranging from 39 to 65 years.^{1,4-6} However, we can expect some degree of diversity in ages of onset in other *APP* triplication families.

In a recent French series of EOAD, *APP* duplication carriers were more likely to present with seizures,¹¹ possibly explained by A β overproduction related to increased *APP* expression. Early seizures may be a shared clinical feature with *APP* triplication as observed in the father of the proband.

Different mechanisms can be involved in autosomal dominant CAA and AD pathogenesis. In contrast with *APP* point mutations, which can result in increased beta cleavage of *APP*, change in A β 42/38 ratio, or in A β aggregation propensity,¹² *APP* duplications lead to *APP* overproduction, with severe A β deposits in the brain parenchyma and within vessels walls.^{1,3} Here, cerebral biopsy of the proband's father confirmed severe amyloid perivascular deposits consecutive to *APP* overproduction. Indeed, we found a 2-fold upregulation of *APP* mRNA levels in blood. Unfortunately, no comparison could be made with *APP* duplication carriers given the availability of RNA in the proband only so that we cannot be sure that triplications result in significantly increased mRNA levels than in duplication.

Here, we showed by aCGH that the triplication encompassed the *APP* gene only. To our knowledge, there is a single case report of an *APP* duplication encompassing the *APP* gene solely, highlighting that increased *APP* expression is sufficient to cause EOAD and CAA.² However, the resolution of the techniques did not allow us to assess whether the breakpoints were the same in our case and in this 290–750 kb duplication.² The mechanisms underlying genomic instability of the *APP* region remain elusive, and further reports are needed to refine shared or novel breakpoints.

Overall, our report provides further evidence that increased *APP* expression is sufficient to lead to A β aggregation and subsequent EOAD and CAA. Although ages at clinical onset were among the earliest ones in our *APP* triplication carriers compared with duplication carriers, further cases would be required to conclude.

Study Funding

No targeted funding reported.

Disclosure

The authors report no disclosures relevant to the manuscript. Go to Neurology.org/NG for full disclosures.

Publication History

Received by *Neurology: Genetics* April 1, 2021. Accepted in final form June 8, 2021.

Appendix Authors

Name	Location	Contribution
Lou Grangeon, MD	Department of Neurology and CNR-MAJ, UNIROUEN, Inserm U1245, CHU Rouen, CIC-CRB1404, F-76000 Rouen, France	Collection, interpretation of data, and drafting and manuscript revision
Kévin Cassinari, MD	Department of Genetics and CNR-MAJ, Normandie University, UNIROUEN, Inserm U1245 and CHU Rouen, F-76000 Rouen, France	Collection, statistical analysis, and interpretation of data and drafting of the manuscript
Stéphane Rousseau, BSc	Department of Genetics and CNR-MAJ, Normandie University, UNIROUEN, Inserm U1245 and CHU Rouen, F-76000 Rouen, France	Collection, analysis, and interpretation of data
Bernard Croisile, MD	Department of Neurology, Lyon University Hospital, France	Collection and interpretation of data
Maité Formaglio, MD	Department of Neurology, Lyon University Hospital, France	Collection and interpretation of data
Olivier Moreaud, MD	Department of Neurology, Grenoble University Hospital, France	Collection and interpretation of data
Jean Boutonnat, MD	Department of Neurology, Grenoble University Hospital, France	Collection, analysis, and interpretation of data
Nathalie Le Meur, PharmD	Department of Genetics and CNR-MAJ, Normandie University, UNIROUEN, Inserm U1245 and CHU Rouen, F-76000 Rouen, France	Collection and interpretation of data
Manuele Miné, PharmD, PhD	AP-HP, Groupe Hospitalier Saint-Louis Lariboisière-Fernand-Widal, Service de Génétique Moléculaire Neurovasculaire, INSERM UMR 1141, NeuroDiderot, Université de Paris, France	Collection and interpretation of data
Thibault Coste, MD	AP-HP, Groupe Hospitalier Saint-Louis Lariboisière-Fernand-Widal, Service de Génétique Moléculaire Neurovasculaire, INSERM UMR 1141, NeuroDiderot, Université de Paris, France	Collection and interpretation of data
Eva Pipiras, MD, PhD	Department of Histology Embryology and Cytogenetics, Jean Verdier Hospital; Paris 13 University, Sorbonne Paris Cité, UFR SMBH Bobigny; PROTECT, INSERM, Paris Diderot University, Bondy, France	Collection and interpretation of data
Elisabeth Tournier Lasserre, MD, PhD	AP-HP, Groupe Hospitalier Saint-Louis Lariboisière-Fernand-Widal, Service de Génétique Moléculaire Neurovasculaire, INSERM UMR 1141, NeuroDiderot, Université de Paris, France	Interpretation of data and drafting and manuscript revision

Continued

Appendix (continued)

Name	Location	Contribution
Anne Rovelet-Lecrux, PhD	Department of Genetics and CNR-MAJ, Normandie University, UNIROUEN, Inserm U1245 and CHU Rouen, F-76000 Rouen, France	Collection, analysis, and interpretation of data
Dominique Campion, MD, PhD	Department of Genetics and CNR-MAJ, Normandie University, UNIROUEN, Inserm U1245 and CHU Rouen, F-76000 Rouen, France	Collection, analysis, and interpretation of data
David Wallon, MD, PhD	Department of Neurology and CNR-MAJ, Normandie University, UNIROUEN, Inserm U1245, CHU Rouen, CIC-CRB1404, F-76000 Rouen, France	Study concept, interpretation of data, and drafting and manuscript revision
Gael Nicolas, MD, PhD	Department of Genetics and CNR-MAJ, Normandie University, UNIROUEN, Inserm U1245 and CHU Rouen, F-76000 Rouen, France	Study concept, interpretation of data, and drafting and manuscript revision

References

1. Rovelet-Lecrux A, Hannequin D, Raux G, et al. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat Genet*. 2006;38(1):24-26.
2. Sleegers K, Brouwers N, Gijssels I, et al. APP duplication is sufficient to cause early onset Alzheimer's dementia with cerebral amyloid angiopathy. *Brain*. 2006;129(pt 11):2977-2983.
3. Pottier C, Wallon D, Lecrux AR, et al. Amyloid- β protein precursor gene expression in Alzheimer's disease and other conditions. *J Alzheimers Dis*. 2012;28(3):561-566.
4. Wallon D, Rousseau S, Rovelet-Lecrux A, et al. The French series of autosomal dominant early onset Alzheimer's disease cases: mutation spectrum and cerebrospinal fluid biomarkers. *J Alzheimers Dis*. 2012;30(4):847-856.
5. Lanoiselee H-M, Nicolas G, Wallon D, et al. APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: a genetic screening study of familial and sporadic cases. *PLoS Med*. 2017;14(3):e1002270.
6. Guyant-Marechal I, Berger E, Laquerriere A, et al. Intrafamilial diversity of phenotype associated with app duplication. *Neurology*. 2008;71(23):1925-1926.
7. Mann DMA, Davidson YS, Robinson AC, et al. Patterns and severity of vascular amyloid in Alzheimer's disease associated with duplications and missense mutations in APP gene, Down syndrome and sporadic Alzheimer's disease. *Acta Neuropathol (Berl)*. 2018;136(4):569-587.
8. McKhann GM, Knopman DS, Chertkow H, et al. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*. 2011;7(3):263-269.
9. Zafar F, Valappil RA, Kim S, et al. Genetic fine-mapping of the Iowan SNCA gene triplication in a patient with Parkinson's disease. *NPJ Park Dis*. 2018;4:18. Accessed October 2, 2020. [nature.com/articles/s41531-018-0054-4](https://www.nature.com/articles/s41531-018-0054-4).
10. Book A, Guella I, Candido T, et al. A meta-analysis of α -synuclein multiplication in familial Parkinsonism. *Front Neurol*. 2018;9:1021. Accessed October 2, 2020. [frontiersin.org/article/10.3389/fneur.2018.01021/full](https://www.frontiersin.org/article/10.3389/fneur.2018.01021/full).
11. Zarea A, Charbonnier C, Rovelet-Lecrux A, et al. Seizures in dominantly inherited Alzheimer disease. *Neurology*. 2016;87(9):912-919.
12. Greenberg SM, Bacskai BJ, Hernandez-Guillamon M, Pruzin J, Sperling R, van Veluw SJ. Cerebral amyloid angiopathy and Alzheimer disease—one peptide, two pathways. *Nat Rev Neurol*. 2020;16(1):30-42.

3. Utilisation du RNAseq pour l'interprétation de WGS : caractérisation de l'impact de variations introniques profondes du gène *NIPBL* dans le syndrome de Cornelia de Lange

3.1. Contexte et résumé des travaux

Ce travail est cette fois-ci dédié au syndrome de Cornelia de Lange (CdLS), dont les caractéristiques cliniques et moléculaires ont été présentés dans l'introduction de ce manuscrit de thèse. Comme cela a été précisé, des variations pathogènes dans les 5 gènes principaux de CdLS (*NIPBL*, *SMC1A*, *SMC3*, *HDAC8*, *RAD21*) expliquent environ 50% des cas, laissant les autres 50% non résolus et suggérant donc l'existence de *i.* d'autres gènes responsables de CdLS ou de syndromes cliniquement proches et *ii.* des mécanismes cryptiques responsables d'altérations dans les gènes connus. Dans ce contexte, nous avons effectué un WGS chez 5 trios (cas index atteint et parents sains) non résolus répondant aux critères suivants : (i) diagnostic clinique de CdLS classique, (ii) séquençage négatif du panel de gènes CdLS à partir de l'ADN isolé du sang et de la salive (du fait de l'existence non rare de mosaïques non détectables dans le sang), (iii) échantillons d'ADN des parents non atteints disponibles et (iv) ARN isolé du sang du cas index disponible. Une variation *de novo* pathogène a été observée dans un gène de diagnostic différentiel de CdLS chez 3/5 patients, à savoir *POU3F3*, *SPEN*, et *TAF1*. Chez les deux autres cas index, nous avons identifié deux variations introniques profondes *de novo* distinctes dans *NIPBL* avec des prédictions bioinformatiques en faveur de la création d'un nouveau site d'épissage. Pour ces deux dossiers, nous avons réalisé un RNAseq, qui a permis d'objectiver des transcrits aberrants conduisant à la création d'un nouvel exon avec décalage du cadre de lecture, conforme aux prédictions. De manière intéressante, pour un patient porteur d'un variant exonique, nous avons pu montrer que, malgré l'utilisation probablement partielle de ce nouveau site d'épissage (étant donné les comptes de jonctions aberrantes), le ratio allélique de ce marqueur exonique était proche de

50%, suggérant que le NMD ne jouait pas un rôle majeur dans ce cas précis. Ces résultats ont significativement contribué à classer ces deux variants comme probablement pathogènes. Avant ces observations, nous n'avons pu trouver qu'un seul exemple similaire dans le CdLS [176], où le patient avait été diagnostiqué avec un CdLS modéré à sévère et présentait un épissage anormal de *NIPBL* avec l'inclusion d'un nouvel exon dans la séquence de l'intron 21 introduisant un codon stop prématuré. Cependant, bien que ces évènements restent des explications très minoritaires en nombre de patients, l'accès accru au WGS associé au RNAseq révélera certainement un plus grand nombre de situations similaires, comme nous avons déjà pu le constater dans le cadre du soin à travers le plan France Médecine Génomique 2025. Il n'est actuellement pas réellement clair si l'utilisation combinée du RNAseq et du WGS est une approche plus efficace que leur utilisation séquentielle [177]–[179]. Ces résultats, bien que basés sur une petite série de patients, suggèrent qu'une utilisation séquentielle peut fournir une stratégie rentable, même si cela peut augmenter le délai de remise des résultats aux patients. Sur le plan plus fondamental, il était intéressant de noter que l'impact de l'exonisation sur le NMD n'était pas complet, suggérant qu'une perte de fonction partielle de l'allèle était suffisante pour être responsable du phénotype. Bien que la RT-PCR « simple », avec séquençage Sanger, soit souvent suffisante pour caractériser ce type d'évènements, ici le caractère quantitatif du RNAseq et plus complet au locus, a permis une vision plus globale qu'une simple RT-PCR. En effet, un des patients présentait ce marqueur exonique permettant d'aller plus loin dans la compréhension, et d'autre part, un des deux patients présentait en fait 6 variations nucléotidiques *de novo* introniques dans *NIPBL*, dont une seule avait des conséquences prédites et confirmées par RNAseq.

Ce projet a fait l'objet du Master 2 recherche du Dr Juliette Coursimault (DES de génétique médicale). J'ai été en charge du volet RNAseq du projet. Dans ce cadre, j'ai effectué

l'interprétation des RNAseq de ce projet et réalisé les analyses d'études de l'épissage de *NIPBL* ainsi que les analyses de quantification des transcrits à partir des données générées.

3.2. Article scientifique

Received: 11 March 2022 | Revised: 23 May 2022 | Accepted: 9 July 2022
DOI: 10.1002/humu.24438

RESEARCH ARTICLE

Human Mutation  WILEY
HUMAN GENOME VARIATION SOCIETY

Deep intronic *NIPBL* *de novo* mutations and differential diagnoses revealed by whole genome and RNA sequencing in Cornelia de Lange syndrome patients

Juliette Coursimault¹  | Kévin Cassinari¹  | François Lecoquierre¹  |
Olivier Quenez¹  | Sophie Coutant¹ | Céline Derambure¹  |
Myriam Vezain¹  | Nathalie Drouot¹ | Gabriella Vera¹ | Elise Schaefer² |
Anaïs Philippe² | Bérénice Doray³ | Laëtitia Lambert⁴ | Jamal Ghoumid⁵  |
Thomas Smol⁶  | Mélanie Rama⁷ | Marine Legendre⁸ | Didier Lacombe⁹  |
Patricia Fergelot⁹ | Robert Olaso¹⁰ | Anne Boland¹⁰ | Jean-François Deleuze¹⁰ |
Alice Goldenberg¹ | Pascale Saugier-Veber¹  | Gaël Nicolas¹ 

¹Normandie Univ, UNIROUEN, Inserm U1245 and CHU Rouen, Department of Genetics and reference center for developmental disorders, FHU-G4 Génomique, F-76000, Rouen, France

²Service de Génétique Médicale, Institut de Génétique Médicale d'Alsace (IGMA), Hôpitaux Universitaires de Strasbourg, Strasbourg, France

³Service de Génétique Médicale, Centre Hospitalier Universitaire Félix Guyon, Bellepierre Saint Denis, France

⁴Service de Génétique Clinique, CHRU NANCY, F-54000 France, UMR INSERM U 1256 N-GERE, F-54000, Nancy, France

⁵Université de Lille, ULR7364 RADEME, CHU Lille, Clinique de Génétique « Guy Fontaine », and FHU-G4 Génomique, F-59000, Lille, France

⁶Université de Lille, ULR7364 RADEME, CHU Lille, Institut de Génétique Médicale, and FHU-G4 Génomique, F-59000, Lille, France

⁷Institut de Génétique Médicale, CHU de Lille, France

⁸Service de Génétique Médicale, CHU de Bordeaux, Bordeaux, France

⁹INSERM U1211, Université de Bordeaux; Génétique Médicale, CHU de Bordeaux, Bordeaux, France

¹⁰Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine (CNRGH), 91057, Evry, France

Correspondence

Gaël Nicolas, Inserm U1245, UFR Santé, 22, boulevard Gambetta, 76183 Rouen cedex 1, Rouen, France.
Email: gaelnicolas@hotmail.com

Abstract

Cornelia de Lange syndrome (CdLS; MIM# 122470) is a rare developmental disorder. Pathogenic variants in 5 genes explain approximately 50% cases, leaving the other 50% unsolved. We performed whole genome sequencing (WGS) ± RNA sequencing (RNA-seq) in 5 unsolved trios fulfilling the following criteria: (i) clinical diagnosis of classic CdLS, (ii) negative gene panel sequencing from blood and saliva-isolated DNA, (iii) unaffected parents' DNA samples available and (iv) proband's blood-isolated RNA available. A pathogenic *de novo* mutation (DNM) was observed in a CdLS differential diagnosis gene in 3/5 patients, namely *POU3F3*, *SPEN*, and *TAF1*. In the other two, we identified two distinct deep intronic DNM in *NIPBL* predicted to create a novel splice site. RT-PCRs and RNA-Seq showed aberrant transcripts leading to the creation of a novel frameshift exon. Our findings suggest the relevance of WGS in unsolved suspected CdLS cases and that deep intronic variants may account for a proportion of them.

KEYWORDS

clustered mutations, Cornelia de Lange syndrome, kataegis, neurodevelopmental disorder, NIPBL, noncoding sequence, whole genome sequencing

1 | INTRODUCTION

Cornelia de Lange syndrome (CdLS) is a rare malformative monogenic syndrome. Five cohesin complex genes have been implicated in CdLS (*NIPBL*, *SMC1A*, *SMC3*, *RAD21*, *HDCA8*) (Deardorff et al., 2007, 2012, 2016; Kline et al., 2018; Krantz et al., 2004; Selicorni et al., 2021; Tonkin et al., 2004). *NIPBL* is the major gene as approximately 70% of suspected CdLS patients with a genetic diagnosis harbor a (likely) pathogenic variant. The other four genes together account for approximately 15% of cases, while the remaining 15% of patients exhibit pathogenic variants in differential diagnosis genes (Kline et al., 2018). *NIPBL*, *RAD21*, *SMC3* are inherited in an autosomal dominant manner whereas *HDCA8* and *SMC1A* are inherited in an X-linked manner. All types of variants have been reported, including truncating variants, missense but also splice variants and larger deletions. The diagnostic yield is about 50% overall, leaving approximately 50% patients unsolved (Gillis et al., 2004; Piché et al., 2019; Selicorni et al., 2007). Heterozygous variants in two additional genes have been reported more recently in a small number of patients with either a CdLS diagnosis or CdLS-like features, namely *MAU2* and *BRD4*, encoding cohesin complex genes (Alesi et al., 2019; Jouret et al., 2022; Olley et al., 2018; Parenti et al., 2020).

Patients with CdLS show a clinically recognizable phenotype characterized by developmental delay (DD) and/or intellectual disability (ID), growth retardation, microcephaly, limb abnormalities and dysmorphic features (Kline et al., 2018). Due to its specific phenotype, the genetic strategy in case of a suspicion of CdLS generally consists in a targeted screening first, by sequencing of a gene panel, ideally including CdLS genes along with genes associated with related disorders. Among them, so-called transcriptopathies share both pathophysiological and clinical features with CdLS, which itself is considered as a transcriptopathy and, more precisely, belongs to the group of cohesinopathies. Indeed, all 5 CdLS-causing genes encode critical cohesin complex components involved in chromatin structure maintenance and transcription regulation, in addition to *BRD4* and *MAU2*. In cases with CdLS clinical suspicion, the screening of genes associated with KBG (*ANKRD11*), Rubinstein Taybi (*EP300*), and CHOPS (*AFF4*) syndromes is also recommended, as they are classically considered as putative differential diagnoses because of phenotypic features overlapping with CdLS.

Since the advent of pangenomic sequencing techniques, it has become clear that a first or second-line access to exome sequencing (ES) or whole genome sequencing (WGS) is associated with the highest diagnostic yields for developmental disorders (Wright et al., 2015), because of a huge genetic heterogeneity. For CdLS, however, gene panel sequencing may still be considered in a

first line, because (i) the phenotype is recognizable in most cases and (ii) there is a relatively high proportion of mosaic *NIPBL* mutations (~23%), which are not detectable in blood and may be missed by classic 30–40x WGS or 60–120x ES performed from blood samples (Huisman et al., 2013; Nizon et al., 2016). Thus, it is important, before proposing ES or WGS for a classic-CdLS patient, to sequence DNA isolated from other tissues than blood (e.g., saliva or skin) with average depths allowing the identification of mosaics with 10%–20% allelic ratios (ARs).

In classic-CdLS patients negatively screened for the known genes, proposing second-line WGS appears as a promising strategy, with four main hypotheses: (i) a differential diagnosis, i.e., a pathogenic variant in a related disorder or a syndromic developmental disorder with overlapping phenotypic features (ii) a pathogenic variant in a novel gene causing CdLS, (iii) a pathogenic noncoding variant in a known CdLS-causing genes or (iv) a structural variant. Although ES may allow the assessment of the first two hypotheses, WGS offers the opportunity to identify noncoding variants as well as structural variants beyond copy number variants (CNV), and increased sensitivity for CNV detection (Hehir-Kwa et al., 2015). The contribution of deleterious variants in noncoding regions in rare diseases is of recent discovery and remains little explored. Disease-causing variants can occur in all regions outside the coding region, including promoters or enhancers (Cassinari et al., 2020; Shen et al., 2021; Zuin et al., 2017), 5'-untranslated regions (UTRs) (Borck et al., 2006; Labrousse-Colomer et al., 2020; Wright et al., 2021), 3'-UTRs (Dusl et al., 2015), intronic, near-splice regions and deep intronic regions. Some of the latter variants can create neo-exons and destabilize the protein and/or lead to a frameshift (Kim et al., 2020). In CdLS, a few examples of noncoding pathogenic variants have been reported, including four 5'-UTR variants—one altering RNA stability and three predicted to create upstream open reading frames (Borck et al., 2006; Selicorni et al., 2007; Coursimault et al., 2022). WGS also offers the opportunity to study structural variants, allowing for example the identification of complex rearrangement disrupting *NIPBL* (Plesser Duvdevani et al., 2020). Overall, it appears that WGS remains scarcely used in CdLS and could contribute to more diagnoses.

In addition to WGS, RNA-seq allows a genome-wide view of transcripts from the studied tissue that could be well complementary to WGS. Differences in messenger ribonucleic acid (mRNA) relative expression as well as splicing defects and aberrant transcripts can thus be detected. However, the input of RNA-seq in diagnostic procedures remains of rather recent assessment (Cummings et al., 2017; Lee et al., 2020; Rentas et al., 2020; Saedian et al., 2020).

By trio-based WGS, we assessed 5 patients with a clinical diagnosis of classic-CdLS, following negative gene panel sequencing. We performed RNA-seq in two of them, following the identification of strong candidate variants in deep intronic regions. Overall, we managed to identify the likely cause of the syndromic developmental disorder in all five patients.

2 | METHODS

2.1 | Patients enrollment

In this study, we considered probands referred to the Rouen University Hospital molecular genetics department for gene panel sequencing in the context of a clinical suspicion of CdLS, among patients without a (likely) pathogenic variant. The genetics laboratory of the Rouen University Hospital proposes gene panel sequencing with a national multicentric recruitment in France. Medical charts of all patients referred to our centre with a clinical suspicion of CdLS are assessed by an expert clinician and diagnoses are subsequently classified as classical or nonclassical phenotypes before the genetic molecular analysis according to Kline et al. criteria (Kline et al., 2018). Gene panel sequencing targets the coding sequence of all 5 CdLS genes and 17 differential diagnosis genes (*ESCO2*, *CREBBP*, *EP300*, *ANKRD11*, *AFF4*, *KMT2A*, *TAF6*, *SRCAP*, *ARID1A*, *ARID1B*, *SMARCB1*, *SMARCA4*, *SMARCE1*, *SMARCA2*, *SOX11*, *PHF6*, *SETD5*) following custom Agilent QXT capture and Illumina MiSeq sequencing with a approximately 700x average depth of coverage. Patients DNAs are isolated either from blood or from saliva, and sequencing is performed from either or both tissues. RNA sample is not usually required as RNA studies are not part of first-line routine diagnostics. Data are processed following standard procedures for single nucleotide variants (SNVs) and indels, and copy number variants are called using a CANOES-based workflow (Quenez et al., 2021).

In September 2020, among 184 patients referred for a CdLS diagnosis suspicion, the overall diagnostic yield was about 45%. To identify the molecular defects underlying CdLS in patients negatively screened by the above-mentioned procedures, we performed trio-based WGS in patients fulfilling the following restrictive criteria: (i) a typical (or classic) CdLS diagnosis after clinical expertise, (ii) a negative family history, (iii) absence of known molecular cause after analysis of our gene panel, performed from blood and saliva samples with a high sequencing depth (~700x) as described above, (iv) unaffected parents, with blood/DNA samples available in sufficient quantity and quality, (v) RNA of index cases available (blood sample on Paxgene tube), for possible further investigation. All patients gave informed written consent for genetic analyses. From an initial list of 102 CdLS patients without a coding pathogenic variant, 10 patients presented with a diagnosis of sporadic typical CdLS after exclusion of fetuses, and 5 patients fulfilled the above-mentioned criteria. Some other patients were lost to follow-up, while it was not possible to get DNA samples from parents or a saliva sample for the proband (when

initial screening was based on blood), or an RNA sample for the other patients.

2.2 | WGS and variant detection

Whole-genome sequencing was performed by the Centre National de Recherche en Génomique Humaine (CNRGH, Institut de Biologie François Jacob, CEA, Evry, France). After a complete quality control, genomic DNA (1 µg) was used to prepare a library for WGS, using the Illumina TruSeq DNA PCR-Free Library Preparation Kit (Illumina Inc.), according to the manufacturer's instructions. After quality control and normalization, qualified libraries were sequenced on a NovaSeq 6000 platform from Illumina (Illumina Inc.), as paired-end 150 bp reads. Samples were pooled on a NovaSeq 6000 S4 flowcell to target a minimal average sequencing depth of 30x.

Sequence quality parameters were assessed throughout the sequencing run and FASTQ file were generated for each sample. FastQ sequences were aligned on human genome hg19 using the BWA-mem program (v0.7.17). The GATK tools (v4.0.6.0) were then used for the postprocessing of the bam files (BQSR and deduplication). SNVs and short insertions and deletions (indels) were called using the GATK HaplotypeCaller tool and annotated using SNPEff and SnpSift. Structural variant analysis was performed using CANVAS (v.1.39.0) (Roller et al., 2016) and MANTA (v.1.6.0) (Chen et al., 2016) for detection of CNVs, translocations, insertions, inversions. Mobile element insertions were detected using MELT (v.2.1.5) (Gardner et al., 2017). All SVs variants detected by CANVAS, MANTA and MELT were then annotated using AnnotSV (v2.2) (Geoffroy et al., 2018). Candidate variants were confirmed by Sanger sequencing in all probands and their parents.

To accurately detect *de novo* SNVs and indels (*de novo* mutations [DNM]) in each proband, we applied complementary methods and filtration steps. We used Deepvariant with WGS model to call SNVs and indels in probands and both parents (Poplin et al., 2018). Individual GVCF were merged via glnexus with DeepVariantWGS preset to produce a multisample VCF. *De novo* candidates were obtained after filtration on GT (genotype), DP (depth), GQ (genotype quality) and VAF (variant allele fraction) fields. Filters used to call *de novo* candidates from multi sample VCF (5 trios) were: Ref genotype in parents, DP >10 and GQ >29 in proband and both parents, VAF >0.3 in proband, VAF(proband)/VAF(parents) >4, and VAF(proband)/VAF(controls) >5. For a given proband, controls corresponded to the four other probands and their parents (12 individuals). *De novo* candidate variants were manually reviewed via a custom Integrative Genomics Viewer (IGV)-based filtration interface (https://github.com/francois-lecoquierre/genomic_shortcuts/). As a quality control and to detect potential outliers regarding the count of *de novo* variants, and because the paternal age at conception is the main determinant of this biological phenotype (Jónsson et al., 2017), we plotted the number of *de novo* variants in each proband against paternal age at conception.

2.3 | Noncoding variants annotation

Variations in the 5'-UTR regions were annotated with the 5utr ['suter'] tool, allowing the search for uORF creations (<https://github.com/leklab/5utr>). Splicing was assessed using the SpliceAI tool which is a deep neural network that predicts splice junctions from a pre-mRNA transcript sequence (Jaganathan et al., 2019) (lastly assessed, June 2021). MaxEntScan, NNSPLICE, GeneSplicer and SpliceSiteFinder-like, as provided in the Alamut Visual software, were used to predict whether selected variants affected splice sites (Pertea et al., 2001; Reese et al., 1997; Yeo & Burge, 2004; Zhang, 1998).

2.4 | Detection of candidate disease-causing variants

WGS was conducted on trios composed of the affected child and both unaffected parents. Coding regions were analyzed first. Variants were selected under several filtering scenarios and were interpreted according to ACMG-AMP recommendations (Richards et al., 2015). We analyzed variants regarding the following filtration scenarios (i) *de novo* variants, (ii) variants present and/or pathogenic in patient-derived databases (Clinvar, Denovo-db) (lastly assessed, June 2021), (iii) variants that are very rare in the general population (minor allele frequency <0.001) segregating according to autosomal recessive or X-linked inheritance (Karczewski et al., 2020). Remaining gene variants underwent further prioritization and manual interpretation. Secondly, the noncoding variants were analyzed by focusing on *de novo* candidate variants in introns and 5'-UTR/3'-UTR regions and then extended the analysis to other inheritance hypotheses.

For structural variants, we filtered out those with more than one occurrence in the Database of Genomic Variant gold standard (MacDonald et al., 2014) or gnomAD-SV database (Collins et al., 2020) (both assessed in May, 2021).

2.5 | RNA-seq

Total RNAs from whole blood were extracted with PAXgene blood RNA kit according to the manufacturer's recommendations (Qiagen PreAnalytiX GmbH) and stored at -80°C until use. The quality and quantity of RNA were assessed using the 4200 TapeStation (Agilent Technologies) and the Qubit 3.0 device (Thermo Scientific). Only RNA samples with a minimal RNA integrity number of 7 were used for subsequent experiments. Libraries were prepared using the NEBNext Ultra II Directional RNA Library Kit for Illumina (New England Biolabs) kit and High-throughput sequencing of the libraries was performed on an Illumina NextSeq 500 (Illumina) using 2*75 bp sequencing to generate 60M read pairs on average per sample. Bioinformatics analysis was carried out using nf-core/RNA-seq v3.1 analysis pipeline to generate multi quality control report that uses the STAR v2.6.1d and SALMON v1.4.0 tools for alignment (Ewels

et al., 2020). Visual exploration of the BAM files was performed with the IGV tool from the Broad Institute.

2.6 | Targeted RNA analyses

Complementary DNA (cDNA) was synthesized from total RNA (collected as described above), using a high-capacity cDNA Reverse transcription kit (Applied Biosystems) with RiboLock RNase inhibitor (Thermo Fisher scientific) which was further amplified to obtain PCR products using specific primers. ThermoPrime Taq DNA polymerase from Thermo scientific was used for PCR amplification. We used the following respective primers (Figure S1). In the case of patient 4: specific primers on either side of the suspected aberrant exon were designed (PCR 1: exon 31 forward primer: 5'-CTCCAACCTCCACA CAATGACA-3' and exon 34 reverse primer: 5'- GCTGGGGTC TTATTTTGCTGA-3'). Primers were also picked within the aberrant exon (PCR 2: exon 31 forward primer: 5'-CTCCAACCTCCACA CAATGACA-3' and exon 32 reverse primer 5'-TTGGGAGGCTGA GGAAAGAG-3'); PCR 3: exon 32 forward primer 5'-TCTTTCCTCAG CCTCCAAG-3' and exon 34 reverse primer 5'-GCTGGGGTCT TATTTTGCTGA-3'). In the case of patient 5: specific primers on either side of the suspected aberrant exon were designed (PCR 1: exon 7 forward primer: 5'-AGACATGGTTCAAGTGAGGACT-3' and exon 9 reverse primer: 5'-ACATTGCCGCTTCTCACTC-3'). Primers were also picked within the aberrant exon (PCR 2: exon 7 forward primer: 5'-AGACATGGTTCAAGTGAGGACT-3' and exon 8 reverse primer 5'-TGTGGTCTTCTTTCTCCCT-3'; PCR 3: exon 8 forward primer 5'-AGGGAGAAAGAGAAGACCACA-3' and exon 9 reverse primer 5'- ACATTGCCGCTTCTCACTC-3'). PCR products were ultimately separated on a 2.5% agarose gel and validated by Sanger sequencing.

3 | RESULTS

We included five patients with a clinical diagnosis of classic CdLS assessed by a clinical expert, and negative panel sequencing performed on DNA isolated from both blood and saliva. In addition, three of them also had negative gene panel sequencing from a bulk skin biopsy sample. We performed trio-based WGS on DNA isolated from blood. Patients were aged from 5 to 18 years. Summary phenotypic data are presented in Table 1 and further described below.

The average depth of coverage was of 44x. SNVs, short insertions and deletions (indels) as well as structural variants were analyzed with a focus on known Mendelian genes (all inheritance patterns, OMIM database and home-made curated extension) and at the genome-wide level for *de novo* mutations.

We identified a likely pathogenic/pathogenic (LP/P) variant in all five patients. All were *de novo* heterozygous SNV or indels. Three patients exhibited a *de novo* LP/P coding variant in a gene causing another developmental syndrome, namely *TAF1*, *SPEN*, and *POU3F3*

TABLE 1 Clinical description of the five patients with typical CdLS analysed by WGS

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
Phenotype	Patient 1 SPEN c.4345G>T; p.(Glu1449*)	Patient 2 POU3F3 c.1084C>A; p.(Arg362Ser)	Patient 3 TAF1 c.4748A>G; p.(Tyr1583Cys)	Patient 4 NIPBL c.5862+3487C>T	Patient 5 NIPBL c.869-640G>C
Age/sex	18 years; female	10 years; female	5 years; male	15 years; female	13 years; male
Antenatal and neonatal period	Hypotrophy	Birth weight (11th P); Height (2nd P)	Severe IUGR and then hypotrophy	Hypotrophy	Severe IUGR and then hypotrophy
Psychomotor development	Global delay, mild to moderate ID (no formal diagnosis)	Global delay (no language, not walking at the age of 4 years); hypotonia	Global delay, coordination problems	Global delay (walked at 28 months, no language acquisition at the age of 15 years); anxiety, OCD	Moderate global delay; attention deficit disorder, walked at the age of 21 months, first words at the age of 3 years; has improved lexical fields with short sentences; adapted school
Organ malformations	Surgical treatment of aortic coarctation; strabismus	None	Scrotum in shawl	None	Heart defect (atrial septal defect)
Limb abnormalities	Brachymetacarpia of the 1st ray, Clinodactyly of the 5th fingers; Walking with walker; Hypertonia of the extremities; severe hip dysplasia	None	Relative brachymetacarpia of the 1st ray of the feet; wide thumbs	Clinodactyly of the 5th fingers; Brachymetacarpia of the 1st ray; III-V Brachymetatarsia	Micromelia with brachymetacarpia of the 1st ray
Head circumference	Microcephaly (-4 SD)	Microcephaly	Microcephaly (-4.5 SD)	Microcephaly (-4.2 SD)	Microcephaly (-4 SD)
Growth/eating disorders	Growth retardation	Growth retardation	Eating difficulties (especially; solid food)	Growth retardation	Early eating disorders, GERD, enteral feeding tube
Dysmorphic features	Typical; ptosis; blepharophimosis, pectus excavatum; Stretch marks and hypertrophic scars	Typical; cupped ears	Typical	Typical	Typical

Abbreviations: CdLS, Cornelia de Lange syndrome; GERD, gastroesophageal reflux disease; ID, intellectual disability; IUGR, intrauterine growth restriction; OCD, obsessive-compulsive disorder.

(Figure S2), while two patients showed a *de novo* deep intronic *NIPBL* variant predicted to create a frameshift neo-exon and thus resulting in a likely loss of function. We further confirmed the consequences of intronic *NIPBL* DNM by targeted RNA analyses and RNA-seq.

3.1 | Patient 1. Chr1(GRCh37):g.16257080G>T; NM_015001.2(SPEN):c.4345G>T, p.(Glu1449*)

3.1.1 | Clinical summary

Patient 1 is an 18-year-old girl with mild-to-moderate ID. She was born prematurely at 26 weeks of gestation in the context of maternal fever with a weight of 675 g (19.37th percentile), a length of 31 cm (12.37th percentile) and an OFC of 22.5 cm (17.91th percentile). She benefited from surgical treatment of aortic coarctation. Sucking, swallowing and digestive outcome after the neonatal period were normal. She exhibited growth retardation. At 5½ years, she weighed 12.8 kg (−2.9 SD) for a height of 99 cm (−2.5 SD). At the age of 12½ years upon last visit, height was 144 cm (−1.1 SD), weight 36 kg (−1 SD), body mass index (BMI) 17.4, even though she underwent growth hormone treatment between 3 and 9 years of age. She presented microcephaly (51 cm; −4 SD). She also had global delay. Brain magnetic resonance imaging (MRI) showed leukomalacia. She presented a pyramidal syndrome, sharp and polykinetic reflexes, and spasticity. She underwent surgery for severe hip dysplasia and

multiple tenotomies associated with botulinum toxin injections. At last visit, she could walk with walker assistance because of hypertonia of extremities. There was hypotonia of the oral sphere with difficulties of elocution, mastication and drooling. She had been treated by physiotherapy and speech therapy. She was in a regular school in a class adapted for children with special needs. She showed brachymetacarpus of the 1st ray and had typical CdLS dysmorphism with ptosis, strabismus, blepharophimosis, arched eyebrows, short nose with anteverted nostrils, thin upper lip, flat and prominent philtrum and downturned corners of the mouth (Figure 1). Hearing was normal. We also noted the presence of stretch marks with thin skin. Kline consensus clinical score was 14. CdLS gene panel sequencing was negative on blood, saliva and skin biopsy.

3.1.2 | WGS analysis

Interpretation of coding variants from the WGS data revealed three heterozygous DNM in the *SPEN* gene: a heterozygous (AR = 42.6%) truncating variant (NM_015001.2:c.4345G>T; p.(Glu1449*), ClinVar submission SUB10575763), a synonymous variant (c.3642C>T, p.(Pro1214Pro), AR = 45%), and a missense variant (c.3656C>T, p.(Thr1219Ile), AR = 41.3%). The truncating variant was prioritized. This variant is absent from the gnomAD database. The probability of loss-of-function intolerance (pLI) of this gene is 1 in the gnomAD browser. *SPEN* is reported to be enriched in *de novo* and truncating

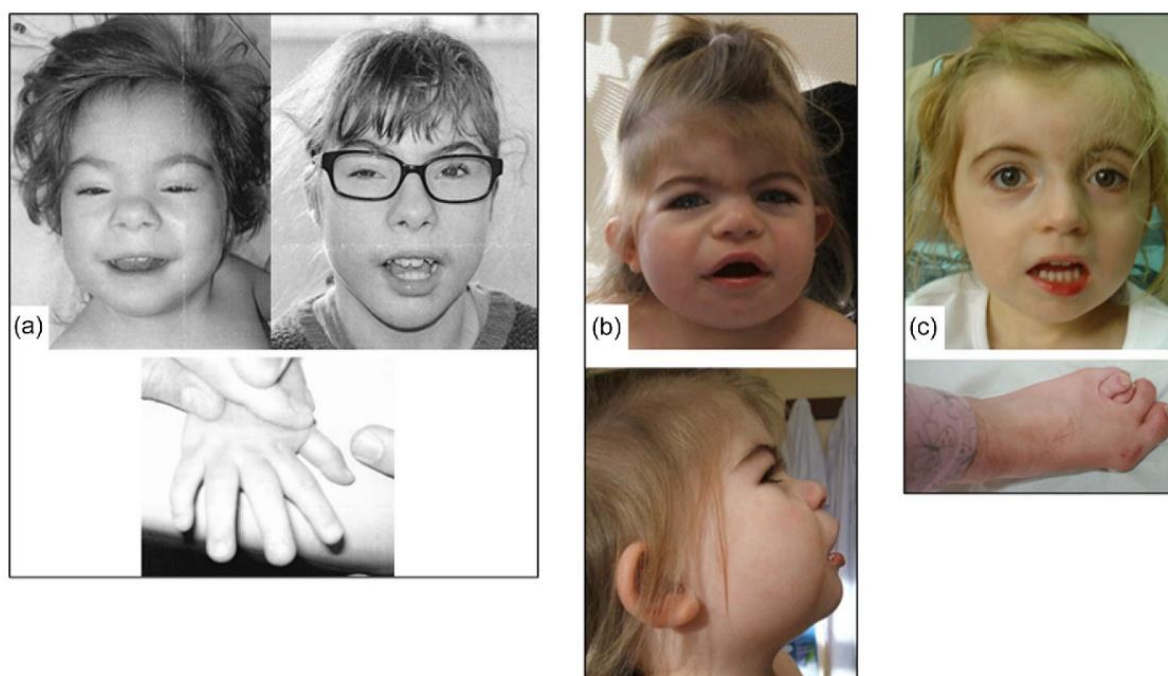


FIGURE 1 Clinical photographs of 3 individuals with clinically suspected CdLS. (a) Patient 1, front view at 1 and 11 years of age and photograph of her right hand with brachymetacarpus of the 1st ray; (b) Patient 2, front and lateral view at 2 years of age; (c) Patient 4, frontal view at the age of 5 years and photograph of the left foot

variants in patients with a developmental disorder (Radio et al., 2021; Wang et al., 2020). It encodes a transcriptional repressor with a major role in the initiation of X-chromosome inactivation (Dossin et al., 2020). Pathogenic *SPEN* variants have been very recently associated with a neurodevelopmental disease in two clinical cohorts (MIM# 619312) (Radio et al., 2021; Wang et al., 2020). The patient's phenotype appeared to be consistent with the descriptions in the literature: DD, ID, hypotonia, abnormal pyramidal signs, central nervous system (CNS) abnormalities, cardiac malformation, ophthalmologic involvement, and dysmorphism. Taken together, the c.4345G>T; p.(Glu1449*) *SPEN* variant was classified as pathogenic (Class 5) according to the ACMG-AMP guidelines (Richards et al., 2015). This allowed us to revise the diagnosis and to conclude to a neurodevelopmental disease linked to the haploinsufficiency of the *SPEN* gene in this patient. The role of the other two *SPEN* variants remains unclear. BAM files showed that they occurred on the same allele, but it was not possible to phase them with the truncating variant. The synonymous variant is not predicted to affect splicing. The missense variant was predicted deleterious by 10/18 bioinformatics software as assessed on the Varsome website in June 2021. They both remain of unknown significance. Of note, no other candidate variant was prioritized among SNV/indels and structural variants.

3.2 | Patient 2. Chr2(GRCh37): g.105473052C>A; NM_006236.3(POU3F3):c.1084C>A, p.(Arg362Ser)

3.2.1 | Clinical summary

Patient 2 is 5-year-old girl who was born at 38 weeks of gestation. Birth length (44.5 cm; 2.24th percentile) and head circumference (31.5 cm; 5th percentile) were in the lower range, and her weight was 2560 g (11.56th percentile). Her development was delayed and associated to global hypotonia. She sat at 12 months and still did not walk at 4 years of age upon last visit. She had neither gastroesophageal reflux disease (GERD) in infancy nor eating disorder. She showed growth retardation and microcephaly (87 cm, -3.5 SD). She had no language at the age of 4. The malformative assessment was normal. She had no limb abnormalities. Dysmorphic features included hirsutism, a low frontal hairline, arched eyebrows with synophris, long eyelashes, short nose with anteverted nares, flat and prominent philtrum, downturned corners of the mouth and cupped ears (Figure 1). Kline consensus clinical score was 13. CdLS gene panel sequencing was negative on blood and saliva.

3.2.2 | WGS analysis

Interpretation of coding variants from the WGS data revealed a heterozygous missense DNM in the *POU3F3* gene (NM_006236.3:c.1084C>A; p.(Arg362Ser), AR = 53%) (ClinVar Submission SUB10575785), a gene associated with Snijders

Blok-Fisher syndrome (MIM# 618604). The variant is absent from the gnomAD database and is predicted to be deleterious by 19/22 bioinformatics tools as assessed on the Varsome website in June 2021. The mutation was not previously reported in the ClinVar database but two other patients harbored distinct missense substitutions at the same codon c.1085G>T; p.(Arg362Leu) (Snijders Blok et al., 2019). This variant is located in one of the two known functional domains of *POU3F3*: the POU-specific (POU-S) domain where a clustering of missense variants has been reported. *POU3F3* encodes a transcription factor belonging to the POU family. It is involved in the regulation of many key processes in CNS development, including cortical neuron migration, specification and production of upper layers and neurogenesis. Consistent with the previous studies, Patient 2 also had hypotonia, DD, ID, morphological features with atypical cup ears, smooth philtrum, and open gendarme hat mouth. Taken together, the c.1084C>A; p.(Arg362Ser) *POU3F3* variant was classified as pathogenic (Class 5) according to the ACMG-AMP guidelines (Richards et al., 2015). The identification of this variant allowed us to make a diagnosis of Snijders Blok-Fisher syndrome in this patient, which also led to a differential diagnosis of CdLS. Of note, no other candidate variant was prioritized among SNV/indels and structural variants.

3.3 | Patient 3. ChrX(GRCh37):g.70644083A>G; NM_004606.5(TAF1):c.4748A>G, p.(Tyr1583Cys)

3.3.1 | Clinical summary

Patient 3 is a 5-year-old boy who was born at 35 weeks and 6 days of gestation with congenital torticollis. Pregnancy was marked by IUGR with hyperechogenic small intestine. Prenatal array CGH and *CFTR* gene screening revealed no abnormalities. Birth measurements were as follows: weight 1290 g (<1st percentile), length 38 cm (<1st percentile) and OFC 28 cm (<1st percentile). At the age of 20 months, his weight was 9.1 kg (1st percentile), his height was 77 cm (-2.5 SD) and his OFC was 43 cm (-4.5 SD). He showed global delay, clumsiness and hearing loss. He had a horseshoe-shaped kidney. Cardiac ultrasound was normal. He exhibited a shawl scrotum. He showed brachymetacarpia of the 1st ray of the feet and also wide thumbs. At the last examination at the age of 4 years 1 month, his weight was 12.6kg (BMI 0.4 SD) and his height 87.5 cm (-3.65 SD). OFC was 45 cm (-4.95 SD). He just started to sit unaided, tried to stand, and to walk with a walker. He could say only 3 words (mum, dad, dog) even with hearing aid. Gastroesophageal reflux remained a problem as well as constipation. He presented difficulties to eat solid food. Brain MRI at the age of 23 months showed microcephaly with suboptimal white matter myelination, a small corpus callosum and small basal ganglia as well as some degree of simplification of the cortical gyration and passive ventriculomegaly. Dysmorphic features associated plagiocephaly, arched eyebrows with synophris, long eyelashes, short nose with anteverted nares, prominent philtrum, thin upper lip with downturned corners of the mouth and large ears. He

had an achromic and a café-au-lait spot. Kline consensus clinical score was 13. CdLS gene panel sequencing was negative on blood and saliva.

3.3.2 | WGS analysis

Interpretation of coding variants from the WGS data revealed a heterozygous NM_004606.4:c.4748A>G;p.(Tyr1583Cys) (AR = 100%) DNM in the *TAF1* gene (ClinVar Submission SUB10575808). This variant is absent from the gnomAD database and predicted to be deleterious by 16/18 bioinformatics software as assessed on the Varsome website in June 2021. Many *TAF1* missense variants have been reported to cause an X-linked neurodevelopmental disease (MIM# 300966) (Cheng et al., 2020; Hurst et al., 2018; O'Rawe et al., 2015). More than 30 families, including both male and female patients, have been described. Of them, two patients had been given first a clinical diagnosis of CdLS (Cheng et al., 2020). The phenotype associated with this disease seems to be compatible with the phenotype of Patient 3 including hypotonia, DD predominantly in language, ID, autism spectrum disorders, clumsiness, IUGR, postnatal growth retardation, feeding difficulties, microcephaly, as well as dysmorphic features.

A population-scale study ranked *TAF1* 53rd among the top 1003 constrained human genes (Samocha et al., 2014) with a maximal pLI (pLI = 1 in gnomAD). *TAF1* has recently been reported as a neurodevelopmental gene enriched in *de novo* variations (Martin et al., 2021). This highly conserved gene plays a major role in the establishment of protein complexes associated with RNA Pol 2 transcription. Missense variants, distributed all along the gene, have been exclusively identified in this disease, suggesting a loss-of-function mechanism for these missense variants. Based on standards and guidelines by the ACMG-AMP, the variant was classified as likely pathogenic (Class 4). We concluded that Patient's 3 disease was likely attributable to this *de novo* missense c.4748A>G;p.(Tyr1583Cys) variant of the *TAF1* gene, again leading to a differential diagnosis of CdLS. Of note, no other candidate variant was prioritized among SNV/indels and structural variants.

3.4 | Patient 4. NC_000005.9:g.37031001C>T, c.5862+3487C>T NIPBL variant

3.4.1 | Clinical summary

Patient 4 is a 15-year-old girl with ID, anxiety and obsessive-compulsive disorder. She benefited from a prenatal karyotype because of nuchal translucency showing a normal 46,XX results. She was born at 36 weeks of gestation. Birth parameters were 2770 g (27th percentile) for weight and 48 cm for height (43rd percentile), suggesting neonatal hypotrophy. OFC was 32 cm (10th percentile). Early motor milestones were delayed: she sat at 12 months and walked first unaided at 28 months. Currently at the age of 15 years,

she has no oral language, she communicates poorly with sign language. She shows growth delay with a weight of 35 kg (−2.4 SD) and height of 143 cm (−3.2 SD) and presents severe microcephaly (49.5 cm, −4.2 SD). In addition, she has limb abnormalities, with clinodactyly of the 5th fingers, brachymetacarpia of the 1st ray and III–V brachymetatarsia (Figure 1). Cardiac and abdominal ultrasound were normal. She has no hearing loss. She has dysmorphic features including a low frontal hairline, arched eyebrows, long and prominent philtrum, downturned corners of the mouth, thin upper lip. Kline consensus clinical score was 14. CdLS gene panels sequencing was negative on blood, saliva and skin biopsy. Screening for *MECP2*, *FOXG1* and array CGH were normal.

3.4.2 | WGS analysis

Neither any candidate single nucleotide/indel variant, nor any candidate structural variant was identified following the analysis of coding regions. Analysis of the noncoding DNMs highlighted a heterozygous deep intronic DNM in the *NIPBL* gene (NM_133433.3:c.5862+3487C>T, intron 32, AR = 51.4%) (ClinVar Submission SUB10575836). This variant is absent from the gnomAD database. The variant is predicted to affect splicing by creating a novel splice donor site according to MaxEntScan, NNSPLICE, GeneSplicer and SpliceSiteFinder-like (Figure 2). The SpliceAI tool predicted a donor gain with a Δ score of 0.19 (Δ scores range from 0 to 1 and a detailed characterization is provided for 0.2, 0.5, 0.8 cutoffs). In the vicinity of this variant, several putative cryptic acceptor sites are also predicted in both wild-type and mutant contexts, of which at least one could be used, putatively leading to the creation of a novel exon between natural exons 32 and 33 (Figure 2). We performed RNA-seq from proband's whole blood collected in a PAXgene tube. Following RNA-seq, inspection of *NIPBL* alignments showed the presence of abnormal splice products mapping to intron 32 at the expected positions, showing the inclusion of the predicted 118-bp novel exon (Figure 2), with aberrant junctions to exon 32 and exon 33, respectively (chr5: 37,030,882_37,030,999). The inclusion of this neo-exon was further confirmed by RT-PCR and sequencing (Figure 2 and Figure S3.A). This insertion is out of frame and results in a premature stop codon (r.5862_5863ins118;p.Asn1954fs*50). In RNA-seq data, 20 junctions supported the existence of the neo-exon (14 between exon 32 and the neo-exon and 6 between neo-exon and exon 33), compared to 79 normal exon 32–exon 33 junctions, suggesting some degradation of neo-exon-containing transcripts by nonsense-mediated decay (NMD), and/or partial use of the newly created splicing site. In the coding sequence of *NIPBL*, one heterozygous common SNV, in exon 10 (rs3822471), was available to assess allele-specific expression. The AR was 52% at this position, suggesting that the inclusion of this neo-exon is not associated with strong degradation of transcripts containing this neo-exon by NMD. Thus, we conclude that the c.5862+3487C>T variant is associated with a significant but partial inclusion of a frameshift neo-exon, leading to a partial loss-of-function allele.

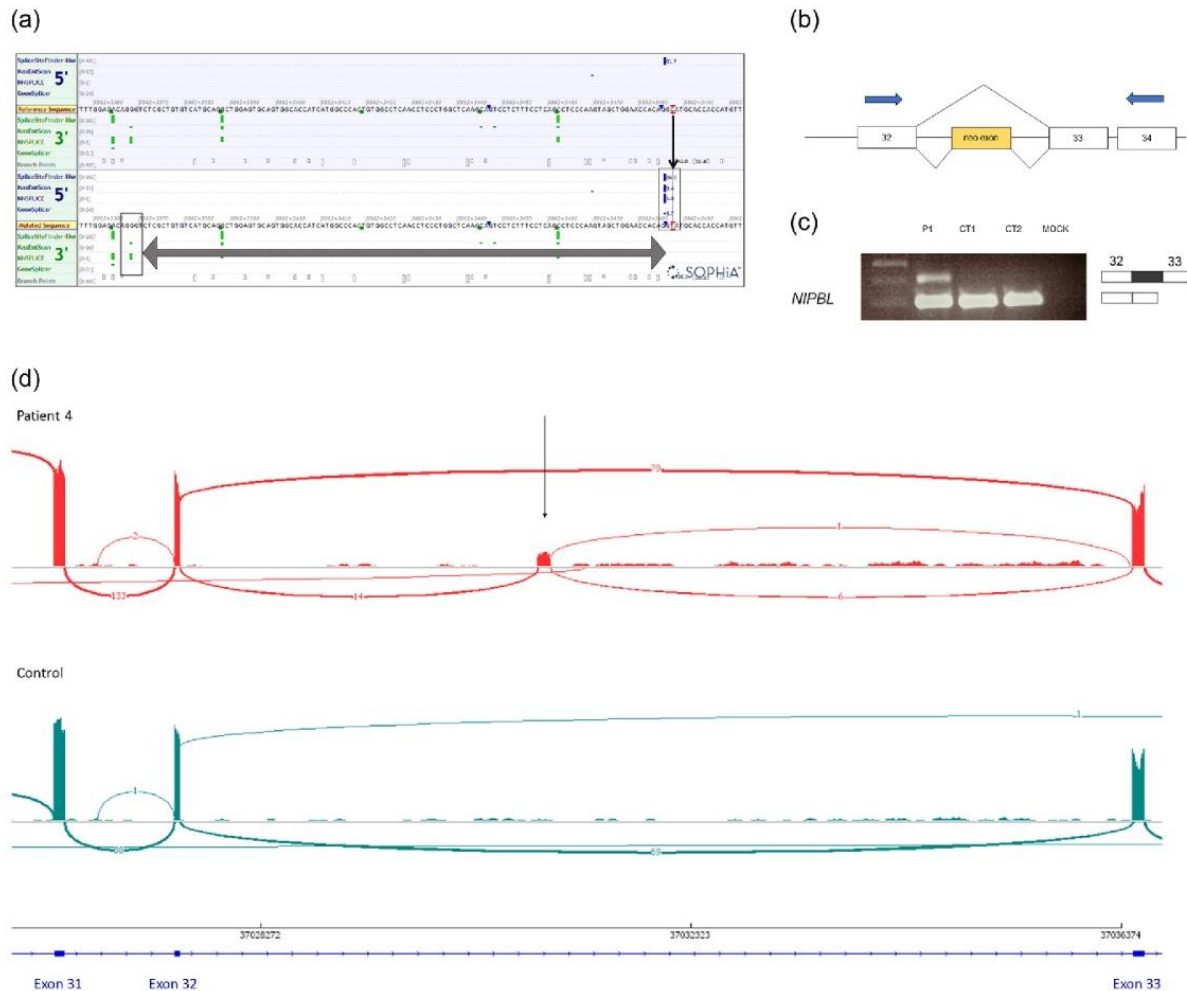


FIGURE 2 Assessment of *de novo* deep intronic *NIPBL* mutation in patient 4. (a) Splicing predictions. According to MaxEntScan, NNSPLICE, GeneSplicer and SpliceSiteFinder-like tools available in the Alamut visual software, the variant is predicted to affect splicing by creating a novel splice-donor site. The black arrow shows the mutation. The splice acceptor and donor site used are surrounded by a black square. The grey double arrow symbolizes the new exon potentially synthesized. (b) Schematic representation of the splicing alteration responsible for the creation of a new aberrant exon between exons 32 and 33. (c) Migration of RT-PCR products on agarose gel showing the presence of an extra band in the patient, compared to two controls (CT1 and CT2). This band migrates at approximately 400 bp. (d) Sashimi plots generated following the alignment of RNA-seq reads on the Integrated Genome Viewer (IGV) showing abnormal *NIPBL* splicing in patient 4 relative to a control

3.5 | Patient 5. NC_000005.9:g.36975238G>C, c.869-640G>C *NIPBL* variant

3.5.1 | Clinical summary

Patient 5 is a 13-year-old boy, born at 37 weeks of gestation after a pregnancy marked by IUGR and hydramnios. Birth measurements were abnormal with a weight of 1920 g (<1st percentile), height of 45 cm (5.44th percentile) and OFC of 29 cm (0.04th percentile). He presented in the perinatal period with eating disorder and GERD, improved by enteral feeding tube. Hearing was normal. Upon last visit at 13 years, he showed microcephaly (OFC = 50 cm, -4.5 SD), presented moderate delay and ADHD. He also had micromelia with

brachymetacarpia of the 1st ray. Cardiac ultrasound revealed an atrial septal defect. He presented facial dysmorphism including arched eyebrows with synophris, blepharophimosis, short nose with anteverted nares, flat and prominent philtrum, thin upper lip and micrognathism. Kline consensus clinical score was 13. CdLS gene panel sequencing was negative on blood, skin biopsy and saliva.

3.5.2 | WGS analysis

Neither any candidate single nucleotide/indel variant, nor any candidate structural variant was identified following the analysis of coding regions. Surprisingly, Patient 5 presented 6 distinct intronic DNM in the *NIPBL*

gene (NM_133433.3:c.610+287T>C [AR = 0.48%]; c.610+1339G>A [AR = 0.46%]; c.869-640G>C [AR = 0.57%]; c.3121+3010T>C [AR = 0.40%]; c.3305-479T>G [AR = 0.43%]; c.3856-1054G>C [AR = 0.54%]). These variants were distributed throughout the gene, in several introns (two variants in intron 6 and the others, respectively, mapping to introns

8, 10, 11, and 16) (Figure 3). Overall, Patient 5 harbored a total of 69 DNMs (61 SNVs and 8 indels), which was in the expected range given the father's age at conception (Figure S4). Therefore, these results were in favor of a mutational cluster specifically localized in the *NIPBL* gene. Among these 6 *NIPBL* variants, only one (NM_133433.3:c.869-640G>C,

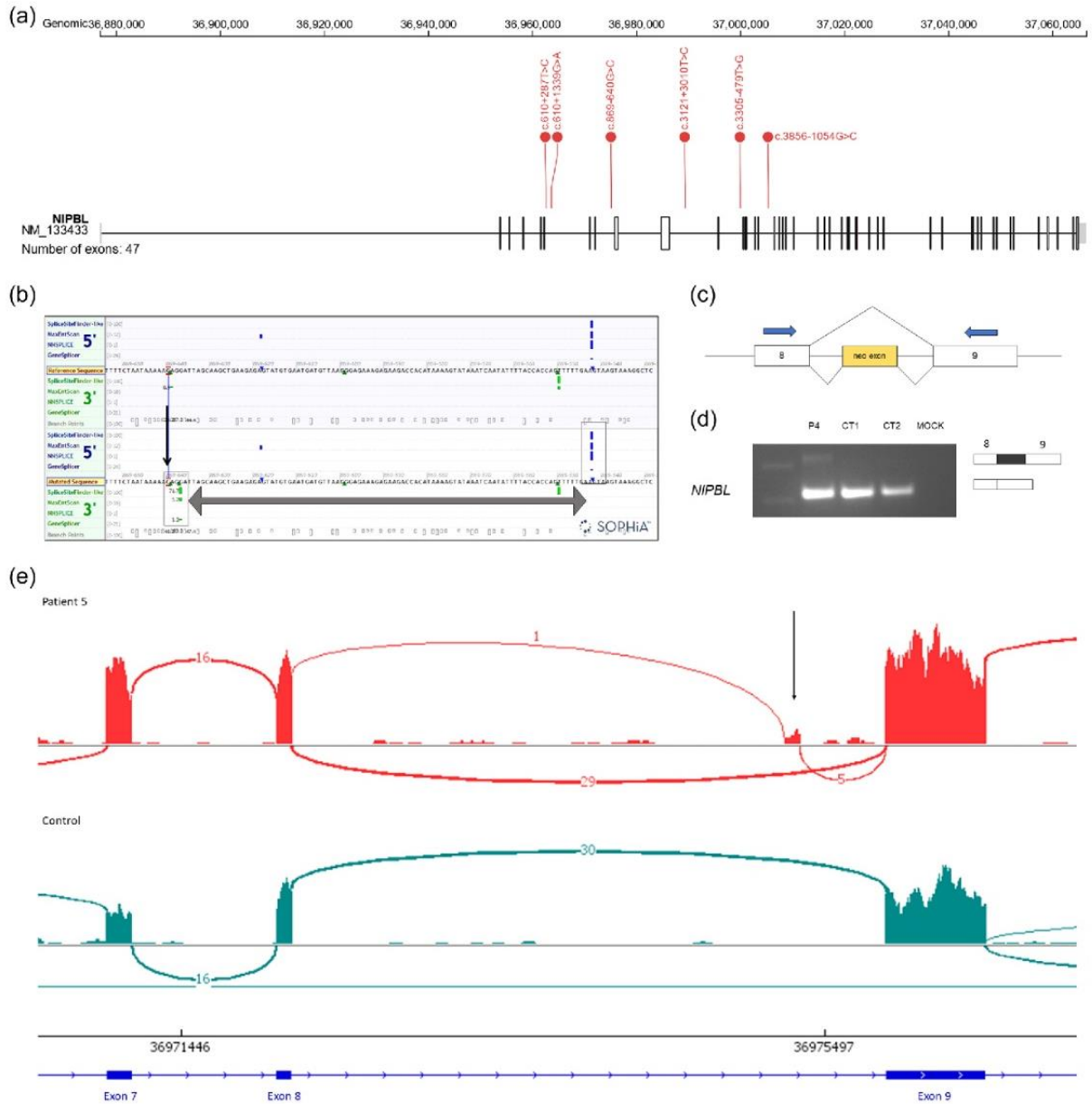


FIGURE 3 Assessment of *de novo* deep intronic *NIPBL* mutation in patient 5. (a) Distribution of the 6 intronic variants along the *NIPBL* mRNA in patient 5. GenomePaint (St. Jude Cloud; <https://genomepaint.stjude.cloud/>) has been used for this representation. (b) Splicing predictions. According to MaxEntScan, NNSPLICE, GeneSplicer and SpliceSiteFinder-like tools available in the Alamut visual software, the variant is predicted to affect splicing by creating a new splice acceptor site leading to a theoretical neo-exon. The black arrow shows the variant. The splice acceptor and donor site used are surrounded by a black square. The grey double arrow symbolizes the new exon potentially synthesized. (c) Schematic representation of the splicing alteration responsible for the creation of a new aberrant exon between exon 8 and exon 9. (d) Migration of RT-PCR products on agarose gel showing the presence of an extra band in the patient, compared to two controls (CT1 and CT2). This band migrates at approximately 320 bp. (e) Sashimi plots following the alignment of RNA-seq reads on IGV showing abnormal *NIPBL* splicing in patient 5 relative to a control. mRNA, messenger ribonucleic acid

intron 8, absent from gnomAD) showed strong predictions of effect on splicing, with a predicted creation of a new splice acceptor site (ClinVar Submission SUB10575921). The spliceAI tool predicted an acceptor gain with a Δ score of 0.83. Together with predictions of cryptic donor sites in the surrounding regions in both mutant and WT contexts, we hypothesized a possible creation of a neo-exon at position chr5:36975240_36975335 (Figure 3). We performed RNA-seq from the proband's whole blood collected in a PAXgene tube. Following RNA-seq, NIPBL analysis of aligned reads showed the presence of abnormal splice products mapping to intron 8 at the expected positions, showing the inclusion of the predicted 95-bp novel exon (Figure 3), with aberrant junctions to exon 8 and exon 9, respectively. In RNA-seq data, 6 junctions supported the existence of the neo-exon (1 between exon 8 and the neo-exon and 5 between neo-exon and exon 9), compared to 29 normal exon 8–exon 9 junctions and we could not observe any aberrant splice junction in the surrounding regions of the other *de novo* mutations identified in the other NIPBL introns. The exonized intronic sequence is out of frame and results in a premature stop codon within the exonized sequence (r.868_869ins95, p.Gly291fs*3). The neo-exon inclusion was further confirmed by RT-PCR and cDNA sequencing (Figure 3 and Figure S3.B). Unfortunately, there was no SNV in the NIPBL coding sequence, thus precluding the assessment of allele specific expression.

4 | DISCUSSION

Following the selection of 5 patients with a classic-CdLS presentation and negatively screened for the known genes, WGS analysis ended up with a probable genetic cause in all 5 patients. Nevertheless, these selected patients may not represent all patients negatively screened for CdLS genes and it is unlikely that WGS would provide such a high diagnostic rate using broader inclusion criteria. Follow-up analyses on additional patients, using less stringent inclusion criteria, may lead to a lower diagnostic rate and a higher number of variants of unknown significance. Interestingly, among the 5 (likely) pathogenic *de novo* variations, 3 were coding and are hence theoretically detectable by simplex or trio-based ES, and 2 were noncoding, highlighting one of the inputs of WGS as compared to ES.

The analysis of noncoding regions appears to be relevant, as an increasing number of reports shows a noncoding cause of neurodevelopmental diseases, sometimes after years of diagnostic odyssey (Cassinari et al., 2020; Labrousse-Colomer et al., 2020; Wright et al., 2021). In our study, we identified two different *de novo* intronic variants in two patients, both predicted to result in the creation of a novel splice site. We were able to confirm the pathogenicity of these variants by RNA-seq and RT-PCR followed by Sanger sequencing. RNA-seq was particularly useful for the patient with multiple deep intronic variants, allowing a global view, compared to targeted RT-PCR procedures.

Overall, deep intronic variants creating a frameshift neo-exon are a classic disease-causing mechanism, albeit extremely rarely reported in Mendelian disorders. It has been described for example in the DMD gene where the inclusion of a pseudo-exon was responsible of a milder Becker's muscular dystrophy phenotype (Cummings

et al., 2017). We could find only one example in CdLS (Rentas et al., 2020), where the patient was diagnosed with moderate–severe CdLS and had abnormal NIPBL splicing with inclusion of a novel exon within intron 21 sequence that was expected to introduce a premature stop codon. The increased access to WGS in the clinic will certainly unveil a larger number of similar situations.

Of note, RNA-seq was used here sequentially after WGS. It is still unclear whether combined use of RNA-seq plus WGS is a more efficient approach than sequential use (Cummings et al., 2017; Gonorazky et al., 2019; Kremer et al., 2018; Murdock et al., 2021). Integrating RNA-seq with WGS resulted in additional cases with clear diagnosis in a recent study, with an overall diagnostic rate going from 31% without RNA-seq to 38% with RNA-seq contribution (Lee et al., 2020). Moreover, 18% of all genetic diagnoses returned required RNA-seq to determine variant causality. The contribution of RNA-seq has also been highlighted recently in molecular diagnostics of rare genodermatoses (Saeidian et al., 2020) and rare muscle disorders with an overall diagnostic rate of 35% (Cummings et al., 2017). Our results, albeit in a small series of patients, suggest that a sequential use may provide a cost-effective strategy, although likely increasing the delay to patient report.

It is worth noting that Patient 5 displays a spatial aggregation of 6 intronic DNMs in the NIPBL sequence, including one that we considered as the cause of the disorder through its effect on splicing. While the mechanism associated with such an aggregation remains unclear, we hypothesize that all these variants occurred on the same parental haplotype on a single multihit event or sequence. Unfortunately, no polymorphisms could be identified nearby the variants to phase them. Long-read sequencing would be necessary to further study this hypothesis. The increase of *de novo* mutations in specific gene areas, also called clustered mutations or *kataegis*, is a recently discovered phenomenon observed in many models, both in cancers and in the germline genome, and is imperfectly elucidated at the biological level (Chan & Gordenin, 2015). It has been defined as more than five or six mutations in a range of 1000 bp of the human genome. Particular mutational patterns have been observed within these clusters, including C>G transversions and a role for CpG islands (The BRIDGES Consortium et al., 2018), however the limited number of variants in our patient precluded the identification of a specific mutational pattern. Some epigenetic marks are also associated, such as H3K36me3 nucleosome methylation or chromatin opening measured by DNase sensitivity. Different observations have incorporated a role for sex and age of the transmitting parent (Goldmann et al., 2016; Jónsson et al., 2017). Although not falling into the definition of *kataegis*, Patient 1 also harbored three DNM within the SPEN gene, including one pathogenic variant, without any enrichment in DNM overall either (Figure S4).

In addition to deep intronic DMN, our results also highlight three genes which were not usually considered as differential diagnoses until now, POU3F3, SPEN, and TAF1. They do not belong to the spectrum of cohesinopathies and are thus not included in our gene panel. Phenotypic features associated with these differential diagnoses overlapped with that of CdLS albeit with not very specific clinical signs, such as DD/ID, behavioral disorders, eating disorders

and microcephaly. Dysmorphic features associated with these conditions include synophris, arched eyebrows, short nose with anteverted nostrils, prominent philtrum, and hirsutism, hence also overlapping with CdLS. These three diagnoses highlight the difficulty of establishing a clinically solid diagnosis based solely on the patient's phenotype, even when the phenotype appears to be very specific. Previous publications of ES in patients with suspected CdLS also revealed variants in genes not involved in the cohesin complex, e.g., *ZMYND11*, *MED13L*, and *PHIP*, responsible for so-called CdLS-like phenotypes (Aoi et al., 2019). These results underline, on one hand, the interest to propose additional genetic analyses in patients without a confirmed diagnosis after gene panel sequencing, and, on the other hand, some limits to phenotype-first approaches.

To consider the description of novel differential diagnosis genes, even in so-called classic-CdLS presentations, it should be discussed either (i) to successively add new CdLS genes and differential diagnoses to gene panels, as they are identified, and thus to resequence patients without a confirmed molecular diagnosis on the first versions of the panel, or (ii) to move towards a second-line WGS or ES strategy. This second strategy seems to be the best one, from a cost-effectiveness and clinical management point of view, given the increasing accessibility of ES/WGS. This approach also allows a reasonable amount of targeted genetic tests in a context of increasing number of ID genes described. Until recently, WGS was mainly accessible for research purposes due to its cost and the amount of data generated. Very-high-throughput genomic sequencing platforms are being made accessible in a medical setting in growing number of countries, allowing an easier access to genomic medicine. However, one should remind that, in CdLS, because of the existence of mosaic *NIPBL* mutations, gene panel sequencing remains essential as a first-tier analysis. Some of the mosaics are indeed not detectable by ES or WGS methods because of too low depth of coverage. Thus, it seems important to propose CdLS gene sequencing on salivary or skin samples as a first line to patients presenting a suggestive phenotype, especially those with a typical phenotype, before proceeding to a genome-wide analysis.

In conclusion, using trio-based WGS in highly selected patients, we have identified the genetic cause in 5/5 clinically-diagnosed CdLS patients with negative gene panel sequencing. Of them, we highlight (i) two genes, *POU3F3*, *SPEN*, the pathogenic variants of which being associated with CdLS-like phenotypic features, but also confirm that *TAF1* can be a differential diagnosis gene of CdLS (Cheng et al., 2020), and (ii) two cases of likely pathogenic deep intronic variants in *NIPBL* generating novel exons leading to a frameshift. Our data show WGS potential to diagnose unsolved patients with clinical suspicion of CdLS.

WEB RESOURCES

5utr ['suter'] tool: <https://github.com/leklab/5utr>

Custom IGV-based filtration interface: (https://github.com/francois-lecoquierre/genomic_shortcuts/)

SpliceAI: <https://spliceailookup.broadinstitute.org/>

AUTHOR CONTRIBUTIONS

Conception and design: Juliette Coursimault, Pascale Saugier-Weber, and Gaël Nicolas. *Material preparation, data collection, and analysis:* Juliette Coursimault, François Lecoquierre, Kévin Cassinari, Gabriella Vera, Nathalie Drouot, Céline Derambure, and Myriam Vezain. *Bioinformatic analysis:* Olivier Quenez, Sophie Coutant and François Lecoquierre. *First draft of the manuscript:* Juliette Coursimault. *Critical revision:* Gaël Nicolas, Kévin Cassinari, François Lecoquierre, Alice Goldenberg, Pascale Saugier-Weber. All authors contributed to data acquisition. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

Several authors of this publication are members of the European Reference Network for Developmental Anomalies and Intellectual Disability (ERN-ITHACA). The authors have no conflict of interest to declare. Collaboration CEA-DRF-Jacob-CNRGH-CHU de Rouen. This work did benefit from support of the France Génomique National infrastructure, funded as part of the «Investissements d'Avenir» program managed by the Agence Nationale pour la Recherche (contract ANR-10-INBS-09). We thank all the patients and their families as well as their referring physicians for their participation to this study. This study was co-supported by the European Union and Région Normandie in the context of Recherche Innovation Normandie (RIN2018). Europe gets involved in Normandie with the European Regional Development Fund (ERDF). This work was generated within the European Reference Network for Developmental Anomalies and Intellectual Disability. This work was performed in the framework of FHU-G4 Génomique.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

ETHICS STATEMENT

Legal representatives of patients provided informed written consents for genetic analyses in a diagnostic setting. The retrospective report on the patients' medical and genetic results was approved by the Institutional Review Board of the Rouen University Hospital (CERDE, Comité d'Ethique pour la Recherche sur Données Existantes et Hors Loi Jardé, Rouen, France) (2019/0252/OB). Informed consent was obtained from all individual participants included in the study or from legal representatives, including for photographs of patients 1, 2 and 4.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request. Deidentified clinical information tables are available upon request.

ORCID

Juliette Coursimault  <http://orcid.org/0000-0002-2668-5779>

Kévin Cassinari  <https://orcid.org/0000-0003-2683-4073>

François Lecoquierre  <https://orcid.org/0000-0002-9110-1856>

Olivier Quenez  <https://orcid.org/0000-0002-8273-8505>

Céline Derambure  <https://orcid.org/0000-0002-6722-5955>

Myriam Vezain  <https://orcid.org/0000-0002-8333-1360>

Jamal Ghomid  <http://orcid.org/0000-0002-7111-0050>

Thomas Smol  <http://orcid.org/0000-0002-0119-5896>

Didier Lacombe  <https://orcid.org/0000-0002-8956-2207>

Pascale Saugier-veber  <http://orcid.org/0000-0002-8045-6432>

Gaël Nicolas  <http://orcid.org/0000-0001-9391-7800>

REFERENCES

- Alesi, V., Dentici, M. L., Loddo, S., Genovese, S., Orlando, V., Calacci, C., Pompili, D., Dallapiccola, B., Digilio, M. C., & Novelli, A. (2019). Confirmation of *BRD4* haploinsufficiency role in Cornelia de Lange-like phenotype and delineation of a 19p13.12p13.11 gene contiguous syndrome. *Annals of Human Genetics*, 83(2), 100–109. <https://doi.org/10.1111/ahg.12289>
- Aoi, H., Mizuguchi, T., Ceroni, J. R., Kim, V. E. H., Furquim, I., Honjo, R. S., Iwaki, T., Suzuki, T., Sekiguchi, F., Uchiyama, Y., Azuma, Y., Hamanaka, K., Koshimizu, E., Miyatake, S., Mitsuhashi, S., Takata, A., Miyake, N., Takeda, S., Itakura, A., & Matsumoto, N. (2019). Comprehensive genetic analysis of 57 families with clinically suspected Cornelia de Lange syndrome. *Journal of Human Genetics*, 64(10), 967–978. <https://doi.org/10.1038/s10038-019-0643-z>
- Borck, G., Zarhrate, M., Cluzeau, C., Bal, E., Bonnefont, J.-P., Munnich, A., Cormier-Daire, V., & Colleaux, L. (2006). Father-to-daughter transmission of Cornelia de Lange syndrome caused by a mutation in the 5' untranslated region of the *NIPBL* gene. *Human Mutation*, 27(8), 731–735. <https://doi.org/10.1002/humu.20380>
- Cassinari, K., Rovelet-Lecrux, A., Tury, S., Quenez, O., Richard, A., Charbonnier, C., Olasso, R., Boland, A., Deleuze, J., Besancenot, J., Delpont, B., Pouliquen, D., Lecoquierre, F., Chambon, P., Thauvin-Robinet, C., Campion, D., Frebourg, T., Battini, J., & Nicolas, G. (2020). Haploinsufficiency of the primary familial brain calcification gene *SLC20A2* mediated by disruption of a regulatory element. *Movement Disorders*, 35(8), 1336–1345. <https://doi.org/10.1002/mds.28090>
- Chan, K., & Gordenin, D. A. (2015). Clusters of multiple mutations: incidence and molecular mechanisms. *Annual Review of Genetics*, 49(1), 243–267. <https://doi.org/10.1146/annurev-genet-112414-054714>
- Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics (Oxford, England)*, 32(8), 1220–1222. <https://doi.org/10.1093/bioinformatics/btv710>
- Cheng, H., Capponi, S., Wakeling, E., Marchi, E., Li, Q., Zhao, M., Weng, C., Piatek, S. G., Ahlfors, H., Kleynner, R., Rope, A., Lumaka, A., Lukusa, P., Devriendt, K., Vermeesch, J., Posey, J. E., Palmer, E. E., Murray, L., Leon, E., & Lyon, G. J. (2020). Missense variants in *TAF1* and developmental phenotypes: challenges of determining pathogenicity. *Human Mutation*, 41(2), 449–464. <https://doi.org/10.1002/humu.23936>
- Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., & Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature*, 581(7809), 444–451. <https://doi.org/10.1038/s41586-020-2287-8>
- Coursimault, J., Rovelet-Lecrux, A., Cassinari, K., Brischoux-Boucher, E., Saugier-veber, P., Goldenberg, A., Lecoquierre, F., Drouot, N., Richard, A., Vera, G., Coutant, S., Quenez, O., Rolain, M., Bonnet, C., Bronner, M., Lecourtois, M., & Nicolas, G. (2022). uORF-introducing variants in the 5'UTR of the *NIPBL* gene as a cause of Cornelia de Lange syndrome. *Human Mutation*. Portico. <https://doi.org/10.1002/humu.24384>
- 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., Estrella, E., Reddy, H. M., Zhao, F., Weisburd, B., Karczewski, K. J., O'Donnell-Luria, A. H., Birnbaum, D., Sarkozy, A., Hu, Y., & MacArthur, D. G. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science Translational Medicine*, 9(386), eaal5209. <https://doi.org/10.1126/scitranslmed.aal5209>
- Deardorff, M. A., Bando, M., Nakato, R., Watrin, E., Itoh, T., Minamino, M., Saitoh, K., Komata, M., Katou, Y., Clark, D., Cole, K. E., De Baere, E., Decroos, C., Di Donato, N., Ernst, S., Francey, L. J., Gyftodimou, Y., Hirashima, K., Hullings, M., & Shirahige, K. (2012). HDAC8 mutations in Cornelia de Lange syndrome affect the cohesin acetylation cycle. *Nature*, 489(7415), 313–317. <https://doi.org/10.1038/nature11316>
- Deardorff, M. A., Kaur, M., Yaeger, D., Rampuria, A., Korolev, S., Pie, J., Gil-Rodríguez, C., Arnedo, M., Loeyes, B., Kline, A. D., Wilson, M., Lillquist, K., Siu, V., Ramos, F. J., Musio, A., Jackson, L. S., Dorsett, D., & Krantz, I. D. (2007). Mutations in cohesin complex members *SMC3* and *SMC1A* cause a mild variant of Cornelia de Lange syndrome with predominant mental retardation. *The American Journal of Human Genetics*, 80(3), 485–494. <https://doi.org/10.1086/511888>
- Deardorff, M. A., Porter, N. J., & Christianson, D. W. (2016). Structural aspects of HDAC8 mechanism and dysfunction in Cornelia de Lange syndrome spectrum disorders: structural aspects of HDAC8 mechanism. *Protein Science*, 25(11), 1965–1976. <https://doi.org/10.1002/pro.3030>
- Dossin, F., Pinheiro, I., Žylic, J. J., Roensch, J., Collombet, S., Le Saux, A., Chelmicki, T., Attia, M., Kapoor, V., Zhan, Y., Dingli, F., Loew, D., Mercher, T., Dekker, J., & Heard, E. (2020). SPEN integrates transcriptional and epigenetic control of X-inactivation. *Nature*, 578(7795), 455–460. <https://doi.org/10.1038/s41586-020-1974-9>
- Dusl, M., Senderek, J., Muller, J. S., Vogel, J. G., Pertl, A., Stucka, R., Lochmuller, H., David, R., & Abicht, A. (2015). A 3'-UTR mutation creates a microRNA target site in the *GFPT1* gene of patients with congenital myasthenic syndrome. *Human Molecular Genetics*, 24(12), 3418–3426. <https://doi.org/10.1093/hmg/ddv090>
- Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, 38(3), 276–278. <https://doi.org/10.1038/s41587-020-0439-x>
- Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Pittard, W. S., Mills, R. E., 1000 Genomes Project Consortium, & Devine, S. E. (2017). The mobile element locator tool (MELT): population-scale mobile element discovery and biology. *Genome Research*, 27(11), 1916–1929. <https://doi.org/10.1101/gr.218032.116>
- Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., & Muller, J. (2018). AnnotSV: An integrated tool for structural variations annotation. *Bioinformatics (Oxford, England)*, 34(20), 3572–3574. <https://doi.org/10.1093/bioinformatics/bty304>
- Gillis, L. A., McCallum, J., Kaur, M., DeScipio, C., Yaeger, D., Mariani, A., Kline, A. D., Li, H., Devoto, M., Jackson, L. G., & Krantz, I. D. (2004). *NIPBL* mutational analysis in 120 individuals with Cornelia de Lange syndrome and evaluation of genotype-phenotype correlations. *American Journal of Human Genetics*, 75(4), 610–623. <https://doi.org/10.1086/424698>
- Goldmann, J. M., Wong, W. S. W., Pinelli, M., Farrah, T., Bodian, D., Stittrich, A. B., Glusman, G., Vissers, L. E. L. M., Hoischen, A., Roach, J. C., Vockley, J. G., Veltman, J. A., Solomon, B. D., Gillissen, C., & Niederhuber, J. E. (2016). Parent-of-origin-specific signatures of de novo mutations. *Nature Genetics*, 48(8), 935–939. <https://doi.org/10.1038/ng.3597>

- Gonorazky, H. D., Naumenko, S., Ramani, A. K., Nelakuditi, V., Mashouri, P., Wang, P., Kao, D., Ohri, K., Viththiyapaskaran, S., Tarnopolsky, M. A., Mathews, K. D., Moore, S. A., Osorio, A. N., Villanova, D., Kemaladewi, D. U., Cohn, R. D., Brudno, M., & Dowling, J. J. (2019). Expanding the boundaries of RNA sequencing as a diagnostic tool for rare Mendelian disease. *American Journal of Human Genetics*, 104(3), 466–483. <https://doi.org/10.1016/j.ajhg.2019.01.012>
- Hehir-Kwa, J. Y., Pfundt, R., & Veltman, J. A. (2015). Exome sequencing and whole genome sequencing for the detection of copy number variation. *Expert Review of Molecular Diagnostics*, 15(8), 1023–1032. <https://doi.org/10.1586/14737159.2015.1053467>
- Huisman, S. A., Redeker, E. J. W., Maas, S. M., Mannens, M. M., & Hennekam, R. C. M. (2013). High rate of mosaicism in individuals with Cornelia de Lange syndrome. *Journal of Medical Genetics*, 50(5), 339–344. <https://doi.org/10.1136/jmedgenet-2012-101477>
- Hurst, S. E., Liktov-Busa, E., Moutal, A., Parker, S., Rice, S., Szelinger, S., Senner, G., Hammer, M. F., Johnstone, L., Ramsey, K., Narayanan, V., Perez-Miller, S., Khanna, M., Dahlin, H., Lewis, K., Craig, D., Wang, E. H., Khanna, R., & Nelson, M. A. (2018). A novel variant in TAF1 affects gene expression and is associated with X-linked TAF1 intellectual disability syndrome. *Neuronal Signaling*, 2(3), NS20180141. <https://doi.org/10.1042/NS20180141>
- Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglu, S., Sanders, S. J., & Farh, K. K.-H. (2019). Predicting splicing from primary sequence with deep learning. *Cell*, 176(3), 535–548.e24. <https://doi.org/10.1016/j.cell.2018.12.015>
- Jónsson, H., Sulem, P., Kehr, B., Kristmundsdóttir, S., Zink, F., Hjartarson, E., Hardarson, M. T., Hjorleifsson, K. E., Eggertsson, H. P., Gudjonsson, S. A., Ward, L. D., Arnadottir, G. A., Helgason, E. A., Helgason, H., Gylfason, A., Jonasdóttir, A., Jonasdóttir, A., Rafnar, T., Frigge, M., & Stefansson, K. (2017). Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*, 549(7673), 519–522. <https://doi.org/10.1038/nature24018>
- Jouret, G., Heide, S., Sorlin, A., Faivre, L., Chantot-Bastaraud, S., Beneteau, C., Denis-Musquer, M., Turmpenny, P. D., Coutton, C., Vieville, G., Thevenon, J., Larson, A., Petit, F., Boudry, E., Smol, T., Delobel, B., Duban-Bedu, B., Fallerini, C., Mari, F., & Klink, B. (2022). Understanding the new BRD4-related syndrome: Clinical and genomic delineation with an international cohort study. *Clinical Genetics*, 102(2), 117–122. <https://doi.org/10.1111/cge.14141>
- Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., & MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809), 434–443. <https://doi.org/10.1038/s41586-020-2308-7>
- Kim, A., Le Douce, J., Diab, F., Ferrovova, M., Dubourg, C., Odent, S., Dupé, V., David, V., Diambra, L., Watrin, E., & de Tayrac, M. (2020). Synonymous variants in holoprosencephaly alter codon usage and impact the Sonic Hedgehog protein. *Brain*, 143(7), 2027–2038. <https://doi.org/10.1093/brain/awaa152>
- Kline, A. D., Moss, J. F., Selicorni, A., Bisgaard, A.-M., Deardorff, M. A., Gillett, P. M., Ishman, S. L., Kerr, L. M., Levin, A. V., Mulder, P. A., Ramos, F. J., Wierzba, J., Ajmone, P. F., Axtell, D., Blagowidow, N., Cereda, A., Costantino, A., Cormier-Daire, V., FitzPatrick, D., & Hennekam, R. C. (2018). Diagnosis and management of Cornelia de Lange syndrome: First international consensus statement. *Nature Reviews Genetics*, 19(10), 649–666. <https://doi.org/10.1038/s41576-018-0031-0>
- Krantz, I. D., McCallum, J., DeScipio, C., Kaur, M., Gillis, L. A., Yaeger, D., Jukofsky, L., Wasserman, N., Bottani, A., Morris, C. A., Nowaczyk, M. J. M., Toriello, H., Bamshad, M. J., Carey, J. C., Rappaport, E., Kawauchi, S., Lander, A. D., Calof, A. L., Li, H.-H., & Jackson, L. G. (2004). Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of *Drosophila melanogaster* Nipped-B. *Nature Genetics*, 36(6), 631–635. <https://doi.org/10.1038/ng1364>
- Kremer, L. S., Wortmann, S. B., & Prokisch, H. (2018). « transcriptomics »: molecular diagnosis of inborn errors of metabolism via RNA-sequencing. *Journal of Inherited Metabolic Disease*, 41(3), 525–532. <https://doi.org/10.1007/s10545-017-0133-4>
- Labrousse-Colomer, S., Soukariéh, O., Prout, C., Mouton, C., Huguenin, Y., Roux, M., Besse, C., Boland, A., Olasso, R., Constans, J., Deleuze, J.-F., Morange, P.-E., Jaspard-Vinassa, B., & Trégouët, D.-A., the GenMed consortium. (2020). A novel rare c.-39C>T mutation in the PROS1 5'UTR causing PS deficiency by creating a new upstream translation initiation codon and inhibiting the production of the natural protein [preprint]. *Clinical Science*, 134(10), 1181–1190. <https://doi.org/10.1042/CS20200403>
- Lee, H., Huang, A. Y., Wang, L.-K., Yoon, A. J., Renteria, G., Eskin, A., Signer, R. H., Dorrani, N., Nieves-Rodriguez, S., Wan, J., Douine, E. D., Woods, J. D., Dell'Angelica, E. C., Fogel, B. L., Martin, M. G., Butte, M. J., Parker, N. H., Wang, R. T., Shieh, P. B., & Nelson, S. F. (2020). Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 22(3), 490–499. <https://doi.org/10.1038/s41436-019-0672-1>
- MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2014). The database of genomic variants: A curated collection of structural variation in the human genome. *Nucleic Acids Research*, 42(Database issue), D986–D992. <https://doi.org/10.1093/nar/gkt958>
- Martin, H. C., Gardner, E. J., Samocha, K. E., Kaplanis, J., Akawi, N., Sifrim, A., Eberhardt, R. Y., Tavares, A. L. T., Neville, M. D. C., Niemi, M. E. K., Gallone, G., McRae, J., Deciphering Developmental Disorders Study, Wright, C. F., FitzPatrick, D. R., Firth, H. V., & Hurles, M. E. (2021). The contribution of X-linked coding variation to severe developmental disorders. *Nature Communications*, 12(1), 627. <https://doi.org/10.1038/s41467-020-20852-3>
- Murdock, D. R., Dai, H., Burrage, L. C., Rosenfeld, J. A., Ketkar, S., Müller, M. F., Yépez, V. A., Gagneur, J., Liu, P., Chen, S., Jain, M., Zapata, G., Bacino, C. A., Chao, H.-T., Moretti, P., Craigen, W. J., Hanchard, N. A., Undiagnosed Diseases Network, & Lee, B. (2021). Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. *The Journal of Clinical Investigation*, 131(1), 141500. <https://doi.org/10.1172/JCI141500>
- Nizon, M., Henry, M., Michot, C., Baumann, C., Bazin, A., Bessières, B., Blesson, S., Cordier-Alex, M.-P., David, A., Delahaye-Duriez, A., Delezoïde, A.-L., Dieux-Coeslier, A., Doco-Fenzy, M., Faivre, L., Goldenberg, A., Layet, V., Loget, P., Marlin, S., Martinovic, J., & Cormier-Daire, V. (2016). A series of 38 novel germline and somatic mutations of NIPBL in Cornelia de Lange syndrome. *Clinical Genetics*, 89(5), 584–589. <https://doi.org/10.1111/cge.12720>
- Olley, G., Ansari, M., Bengani, H., Grimes, G. R., Rhodes, J., von Kriegsheim, A., Blatnik, A., Stewart, F. J., Wakeling, E., Carroll, N., Ross, A., Park, S.-M., Bickmore, W. A., Pradeepa, M. M., & FitzPatrick, D. R. (2018). BRD4 interacts with NIPBL and BRD4 is mutated in a Cornelia de Lange-like syndrome. *Nature Genetics*, 50(3), 329–332. <https://doi.org/10.1038/s41588-018-0042-y>
- O'Rawe, J. A., Wu, Y., Dörfel, M. J., Rope, A. F., Au, P. Y. B., Parboosingh, J. S., Moon, S., Kousi, M., Kosma, K., Smith, C. S., Tzetzis, M., Schuette, J. L., Hufnagel, R. B., Prada, C. E., Martinez, F., Orellana, C., Crain, J., Caro-Llopis, A., Oltra, S., & Lyon, G. J. (2015). TAF1 Variants are associated with dysmorphic features, intellectual disability, and neurological manifestations. *American Journal of*

- Human Genetics*, 97(6), 922–932. <https://doi.org/10.1016/j.ajhg.2015.11.005>
- Parenti, I., Diab, F., Gil, S. R., Mulugeta, E., Casa, V., Berutti, R., Brouwer, R. W. W., Dupé, V., Eckhold, J., Graf, E., Puisac, B., Ramos, F., Schwarzmayr, T., Gines, M. M., van Staveren, T., van Ucken, W. F. J., Strom, T. M., Pié, J., Watrin, E., & Wendt, K. S. (2020). MAU2 and NIPBL variants impair the heterodimerization of the Cohesin Loader Subunits and cause Cornelia de Lange Syndrome. *Cell Reports*, 31(7), 107647. <https://doi.org/10.1016/j.celrep.2020.107647>
- Pertea, M., Lin, X., & Salzberg, S. L. (2001). GeneSplicer: A new computational method for splice site prediction. *Nucleic Acids Research*, 29(5), 1185–1190. <https://doi.org/10.1093/nar/29.5.1185>
- Piché, J., Van Vliet, P. P., Pucéat, M., & Andelfinger, G. (2019). The expanding phenotypes of cohesinopathies: One ring to rule them all. *Cell Cycle (Georgetown, Tex.)*, 18(21), 2828–2848. <https://doi.org/10.1080/15384101.2019.1658476>
- Plessner Duvdevani, M., Pettersson, M., Eisfeldt, J., Avraham, O., Dagan, J., Frumkin, A., Lupski, J. R., Lindstrand, A., & Harel, T. (2020). Whole-genome sequencing reveals complex chromosome rearrangement disrupting NIPBL in infant with Cornelia de Lange syndrome. *American Journal of Medical Genetics, Part A*, 182(5), 1143–1151. <https://doi.org/10.1002/ajmg.a.61539>
- Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10), 983–987. <https://doi.org/10.1038/nbt.4235>
- Quenez, O., Cassinari, K., Coutant, S., Lecoquierre, F., Le Guennec, K., Rousseau, S., Richard, A.-C., Vasseur, S., Bouvignies, E., Bou, J., Lienard, G., Manase, S., Fourneaux, S., Drouot, N., Nguyen-Viet, V., Vezain, M., Chambon, P., Joly-Helas, G., Le Meur, N., & Nicolas, G. (2021). Detection of copy-number variations from NGS data using read depth information: A diagnostic performance evaluation. *European Journal of Human Genetics: EJHG*, 29(1), 99–109. <https://doi.org/10.1038/s41431-020-0672-2>
- Radio, F. C., Pang, K., Ciolfi, A., Levy, M. A., Hernández-García, A., Pedace, L., Pantaleoni, F., Liu, Z., de Boer, E., Jackson, A., Bruselles, A., McConkey, H., Stellacci, E., Lo Cicero, S., Motta, M., Carrozzo, R., Dentici, M. L., McWalter, K., Desai, M., & Tartaglia, M. (2021). SPEN haploinsufficiency causes a neurodevelopmental disorder overlapping proximal 1p36 deletion syndrome with an epismutation of X chromosomes in females. *The American Journal of Human Genetics*, 108(3), 502–516. <https://doi.org/10.1016/j.ajhg.2021.01.015>
- Reese, M. G., Eckman, F. H., Kulp, D., & Haussler, D. (1997). Improved splice site detection in genie. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 4(3), 311–323. <https://doi.org/10.1089/cmb.1997.4.311>
- Rentas, S., Rathi, K. S., Kaur, M., Raman, P., Krantz, I. D., Sarmady, M., & Tayoun, A. A. (2020). Diagnosing Cornelia de Lange syndrome and related neurodevelopmental disorders using RNA sequencing. *Genetics in Medicine*, 22(5), 927–936. <https://doi.org/10.1038/s41436-019-0741-5>
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5), 405–423. <https://doi.org/10.1038/gim.2015.30>
- Roller, E., Ivakhno, S., Lee, S., Royce, T., & Tanner, S. (2016). Canvas: Versatile and scalable detection of copy number variants. *Bioinformatics*, 32(15), 2375–2377. <https://doi.org/10.1093/bioinformatics/btw163>
- Saeidian, A. H., Youssefian, L., Vahidnezhad, H., & Uitto, J. (2020). Research techniques made simple: Whole-transcriptome sequencing by RNA-Seq for diagnosis of monogenic disorders. *The Journal of Investigative Dermatology*, 140(6), 1117–1126.e1. <https://doi.org/10.1016/j.jid.2020.02.032>
- Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., Wall, D. P., MacArthur, D. G., Gabriel, S. B., DePristo, M., Purcell, S. M., Palotie, A., Boerwinkle, E., Buxbaum, J. D., Cook, E. H., & Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, 46(9), 944–950. <https://doi.org/10.1038/ng.3050>
- Selicorni, A., Mariani, M., Lettieri, A., & Massa, V. (2021). Cornelia de Lange syndrome: From a disease to a broader spectrum. *Genes*, 12(7), 1075. <https://doi.org/10.3390/genes12071075>
- Selicorni, A., Russo, S., Gervasini, C., Castronovo, P., Milani, D., Cavalleri, F., Bentivegna, A., Masciadri, M., Domi, A., Divizia, M., Sforzini, C., Tarantino, E., Memo, L., Scarano, G., & Larizza, L. (2007). Clinical score of 62 Italian patients with Cornelia de Lange syndrome and correlations with the presence and type of NIPBL mutation. *Clinical Genetics*, 72(2), 98–108. <https://doi.org/10.1111/j.1399-0004.2007.00832.x>
- Shen, Y., Shu, S., Ren, Y., Xia, W., Chen, J., Dong, L., Ge, H., Fan, S., Shi, L., Peng, B., & Zhang, X. (2021). Case report: Two novel frameshift mutations in SLC20A2 and one novel splice donor mutation in PDGFB associated with primary familial brain calcification. *Frontiers in Genetics*, 12, 643452. <https://doi.org/10.3389/fgene.2021.643452>
- Snijders Blok, L., Kleefstra, T., Venselaar, H., Maas, S., Kroes, H. Y., Lachmeijer, A. M. A., van Gassen, K. L. I., Firth, H. V., Tomkins, S., Bodek, S., Öunap, K., Wojcik, M. H., Cunniff, C., Bergstrom, K., Powis, Z., Tang, S., Shinde, D. N., Au, C., Iglesias, A. D., & Fisher, S. E. (2019). De novo variants disturbing the transactivation capacity of POU3F3 cause a characteristic neurodevelopmental disorder. *The American Journal of Human Genetics*, 105(2), 403–412. <https://doi.org/10.1016/j.ajhg.2019.06.007>
- Carlson, J., Locke, A. E., Flickinger, M., Zawistowski, M., Levy, S., Myers, R. M., Boehnke, M., Kang, H. M., Scott, L. J., Li, J. Z., & Zöllner, S., The BRIDGES Consortium. (2018). Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nature Communications*, 9(1), 3753. <https://doi.org/10.1038/s41467-018-05936-5>
- Tonkin, E. T., Smith, M., Eichhorn, P., Jones, S., Imamwerdi, B., Lindsay, S., Jackson, M., Wang, T.-J., Ireland, M., Burn, J., Krantz, I. D., Carr, P., & Strachan, T. (2004). A giant novel gene undergoing extensive alternative splicing is severed by a Cornelia de Lange-associated translocation breakpoint at 3q26.3. *Human Genetics*, 115(2), 139–148. <https://doi.org/10.1007/s00439-004-1134-6>
- Wang, T., Hoekzema, K., Vecchio, D., Wu, H., Sulovari, A., Coe, B. P., Gillentine, M. A., Wilfert, A. B., Perez-Jurado, L. A., Kvarnung, M., Slep, Y., Earl, R. K., Rosenfeld, J. A., Geisheker, M. R., Han, L., Du, B., Barnett, C., Thompson, E., Shaw, M., & Eichler, E. E. (2020). Large-scale targeted sequencing identifies risk genes for neurodevelopmental disorders. *Nature Communications*, 11(1), 4932. <https://doi.org/10.1038/s41467-020-18723-y>
- Wright, C. F., Fitzgerald, T. W., Jones, W. D., Clayton, S., McRae, J. F., van Kogelenberg, M., King, D. A., Ambridge, K., Barrett, D. M., Bayzietinova, T., Bevan, A. P., Bragin, E., Chatzimichali, E. A., Gribble, S., Jones, P., Krishnappa, N., Mason, L. E., Miller, R., Morley, K. I., & Firth, H. V. (2015). Genetic diagnosis of developmental disorders in the DDD study: A scalable analysis of genome-wide research data. *The Lancet*, 385(9975), 1305–1314. [https://doi.org/10.1016/S0140-6736\(14\)61705-0](https://doi.org/10.1016/S0140-6736(14)61705-0)
- Wright, C. F., Quaife, N. M., Ramos-Hernández, L., Danecek, P., Ferla, M. P., Samocha, K. E., Kaplanis, J., Gardner, E. J., Eberhardt, R. Y., Chao, K. R., Karczewski, K. J., Morales, J.,

- Gallone, G., Balasubramanian, M., Banka, S., Gompertz, L., Kerr, B., Kirby, A., Lynch, S. A., & Whiffin, N. (2021). Non-coding region variants upstream of MEF2C cause severe developmental disorder through three distinct loss-of-function mechanisms. *The American Journal of Human Genetics*, 108(6), 1083–1094. <https://doi.org/10.1016/j.ajhg.2021.04.025>
- Yeo, G., & Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 11(2-3), 377–394. <https://doi.org/10.1089/1066527041410418>
- Zhang, M. Q. (1998). Statistical features of human exons and their flanking regions. *Human Molecular Genetics*, 7(5), 919–932. <https://doi.org/10.1093/hmg/7.5.919>
- Zuin, J., Casa, V., Pozojevic, J., Kolovos, P., van den Hout, M. C. G. N., van Ijcken, W. F. J., Parenti, I., Braunholz, D., Baron, Y., Watrin, E., Kaiser, F. J., & Wendt, K. S. (2017). Regulation of the cohesin-loading factor NIPBL: Role of the lncRNA NIPBL-AS1 and identification of a distal enhancer element. *PLoS Genetics*, 13(12), e1007137. <https://doi.org/10.1371/journal.pgen.1007137>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Coursimault, J., Cassinari, K., Lecoquierre, F., Quenez, O., Coutant, S., Derambure, C., Vezain, M., Drouot, N., Vera, G., Schaefer, E., Philippe, A., Doray, B., Lambert, L., Ghomid, J., Smol, T., Rama, M., Legendre, M., Lacombe, D., Fergelot, P., ... Nicolas, G. (2022). Deep intronic *NIPBL* *de novo* mutations and differential diagnoses revealed by whole genome and RNA sequencing in Cornelia de Lange syndrome patients. *Human Mutation*, 43, 1882–1897. <https://doi.org/10.1002/humu.24438>

Résultats - Partie II. Apport des études transcriptomiques pour l'étude du syndrome de Cornelia de Lange

1. Contexte et résumé des travaux

Cette dernière section de la thèse se centre également sur le syndrome de Cornelia de Lange, mais cette fois-ci dans une perspective transcriptomique. Étant donné le rôle du complexe cohésine, et en particulier de NIPBL, dans la régulation génique, nous avons cherché à identifier une signature transcriptomique propre à *NIPBL* pouvant servir de biomarqueur pour ce syndrome. En effet, face à une analyse de panel ou d'exome négative, la question de la pertinence du diagnostic de CdLS peut se poser, du fait de l'existence de diagnostics différentiels. Si un biomarqueur pouvait confirmer ce diagnostic, cela ouvrirait la voie à d'autres investigations (séquençage de génome, analyse de longues molécules par séquençage ou cartographie optique, etc.). S'il existe un biomarqueur permettant de confirmer le diagnostic clinique, il sera dès lors possible de faire d'autres hypothèses. Il pourrait ainsi s'agir i. d'une altération dans un gène connu mais avec un mécanisme plus rare : inversions chromosomiques, variation dans des régions non séquencées du panel (introns, régions transcrites non traduites) ou ii. d'une altération dans un nouveau gène. La confirmation du diagnostic de CdLS par le biomarqueur permettra de justifier la poursuite des explorations génomiques, notamment par un séquençage de génome, ou de réorienter vers la recherche d'un diagnostic différentiel. Par ailleurs, dans le cas d'un VSI détecté par séquençage de panel ou de génome, ce test permettrait de confirmer le CdLS chez le patient, ce qui donnerait un argument fort en faveur de la pathogénicité du variant détecté.

C'est dans ce contexte que le projet COSIGN (Cohesine Signature) a vu le jour. Son principal objectif était donc de corréler les variations pathogènes du gène *NIPBL* à un profil

transcriptionnel spécifique du CdLS et de développer un test fonctionnel simple, rapide et économique basé sur cette signature transcriptomique.

Pour ce faire, nous avons utilisé l'ARN, prélevé sur tubes PAXgene de 12 patients, présentant un syndrome de Cornelia de Lange clinique et confirmé sur des bases moléculaires par une analyse en panel de gène, exome ou génome. En parallèle, des prélèvements PAXgene de 20 sujets atteints de pathologies non neurodéveloppementales (surdités isolées, affections dermato-génétiques pures, troubles du développement sexuel, ...) appariés en âge et en sexe ont été utilisés et inclus dans l'étude comme contrôles **[Figure 19]**. L'analyse du transcriptome des deux groupes a été réalisée par RNAseq avec 30 millions de lectures après sélection polyA, visant à identifier une liste de gènes exprimés différemment.

Parmi ces gènes, nous avons pour objectif d'en sélectionner environ 20 pour faire l'objet d'une confirmation RT-ddPCR, permettant de générer une signature transcriptionnelle du CdLS détectable par RNAseq ou analyse ciblée. Par ailleurs, pour déterminer la spécificité de la signature, il était prévu d'explorer, par RT-ddPCR, d'autres sujets dont l'ARN n'avait pas été utilisé dans l'étape de constitution de la signature. Parmi ces 12 sujets contrôles, 4 étaient des patients CdLS aux caractéristiques identiques de ceux inclus pour la définition de la signature, 4 étaient des sujets contrôles aux caractéristiques identiques des contrôles utilisés dans la définition de la signature et 4 étaient des patients présentant des anomalies du développement et/ou des déficiences intellectuelles non liées aux transcriptomopathies. Ce dernier groupe de patients avait pour vocation d'asseoir la spécificité de la signature.

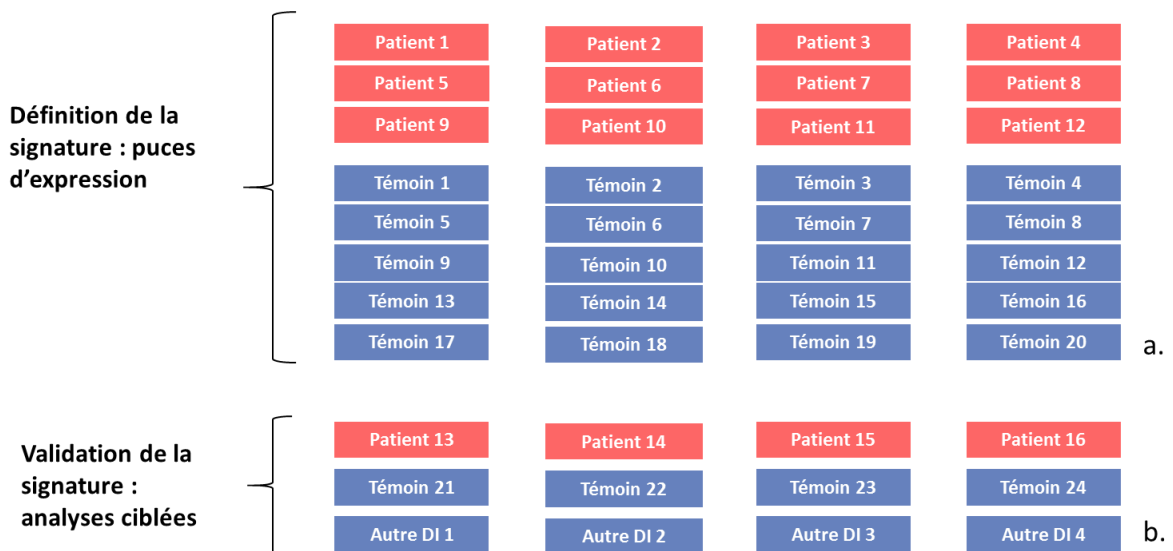


Figure 19. Plan de réalisation des analyses a. transcriptomiques pour définition de la signature et de sa robustesse et b. de la validation de la signature

Le projet CoSign a bénéficié d'un financement accordé par le CHU de Rouen dans le cadre de l'appel d'offre destiné aux jeunes chercheurs en 2019. Les échantillons ont commencé à être collectés dès 2020, avec un total de 12 cas [Tableau 2] et 20 contrôles qui ont été inclus pour la première étape des analyses transcriptomiques, visant à définir la signature. L'analyse s'est déroulée en quatre runs distincts sur Illumina NextSeq 500, utilisant l'outil *Spliced Transcripts Alignment to a Reference* (STAR) [180] et une analyse différentielle par DESeq2 [181].

RUN	NOM	PVT	RIN	Age (ans)	Sexe	Type	Transmission	Variation (NM_133433.3)
RUN 1 à 3	AYA	19-06210	5.7	0.5	F	Faux sens	De novo	c.763A>G; p.(Ser255Gly)
	SAN	19-10695	8.6	2.5	H	Faux sens	De novo	c.6893G>A; p.(Arg2298His)
	JAM	19-07704	5.4	4	H	Epissage	De novo	c.7263+4A>G; p.(?)
	SPE	19-07413	7	4,5	F	LOF	De novo	Délétion E7 à 47
	SHA	19-07679	7.5	8	F	Faux sens	De novo	c.6892C>T; (p.Arg2298Cys)
	LEP	19-04181	6.2	10,5	H	Faux sens	De novo	c.7523A>G; (p.Asp2508Gly)
	LAN	19-12808	7.4	15	H	Missense	De novo	c6470A>G; p.(Asp2157Gly)
	LAK	19-14494	7.5	8	H	LOF	De novo	Délétion E11 à 12
	YRI	20-05745	8.1	0.5	F	LOF	De novo	c.2479_2480del, p.(Arg827Glyfs*2)
RUN 4	MOH	21-08554	7.1	1	F	LOF	De novo	c.2500C>T, p.(Arg834*)
	COC	21-01063	7.8	17	F	Intronique	De novo	c.5862+3487C>T;p.(?)
	MIS	20-08371	8.3	13	H	Intronique	De novo	multiples variations introniques

Tableau 2. Sujets inclus dans la première série du projet CoSign

Malheureusement, en dépit de plusieurs tentatives, il s'est avéré impossible d'obtenir une signature transcriptionnelle fiable à partir de ces prélèvements sanguins. Deux principales causes ont été identifiées à cette difficulté : premièrement, la présence d'ARN présentant un *RNA integrity number* (RIN) faible (inférieur à 7), indiquant une dégradation significative des ARNm, ce qui a impacté la qualité de la sélection polyA et de la capture. De plus, le nombre de lectures générées par échantillon ne paraissait pas suffisant pour contrebalancer les variabilités inter-individuelles dans l'expression génique, résultant en l'impossibilité d'identifier une signature stable.

Pour pouvoir nous affranchir de ces deux points bloquant que sont la qualité des ARN et les variations inter-individuelles, nous avons décidé de poursuivre le projet en utilisant un modèle cellulaire pour la définition de la signature transcriptomique de *NIPBL*. Nous avons donc fait évoluer le projet avec un objectif double : (i) identifier des gènes dérégulés de manière solide, afin de rechercher si leur dérégulation est conservée en post-natal dans les échantillons de sang de patients et (ii) d'étudier la nature de ces gènes, profitant du modèle de cellules souches pluripotentes induites (iPSC) humaines, pouvant mimer les dérégulations précoces d'expression des gènes au cours du développement et ainsi potentiellement expliquer les symptômes observés. Ainsi, par CRISPR-Cas9, nous avons généré un total de 15 lignées d'iPSC, éditées avec différentes variations faux-sens et non-sens de *NIPBL* à l'état hétérozygote ou homozygote (p.Arg45*, p.Arg834*, p.Ser1466Lysfs*13, p.Asp2157Gly et p.Arg2298Cys), et des indels frameshift générées au décours du processus. L'analyse du transcriptome a été réalisée, avec 30M de lectures par échantillon et en réplicat, après extraction de l'ARN pour l'ensemble de ces lignées (RIN>9).

A partir de ces résultats, et d'analyses ciblées complémentaires par RT-ddPCR et Western-blot nous avons pu utiliser ces données dans le cadre des deux axes annoncés ci-dessus :

- Un premier axe correspondant à l'objectif initial du projet, à savoir générer une liste claire de gènes différentiellement exprimés entre les conditions sauvages et mutantes. Cette liste serait utilisable pour définir une signature transcriptomique du CdLS liée à l'altération de *NIPBL*. Sur ce point, la RT-ddPCR et le RNAseq ont montré une diminution des niveaux d'ARN de *NIPBL* pour toutes les cellules portant des variants tronquants, sauf pour le variant p.Arg45*, probablement à cause d'un site alternatif d'initiation de la traduction précédemment décrit [182]. Malheureusement, la majorité des gènes dérégulés dans les iPSC n'étaient pas exprimés dans le sang et le pattern d'augmentation ou de diminution de l'expression de ces gènes semblait différent entre ces deux tissus. De plus, nous nous sommes intéressés aux taux d'expression de la protéine MAU2, qui forme avec *NIPBL* un hétérodimère chargeant le complexe sur la chromatine et qui avait été montrée comme ayant une expression réduite en présence d'un variant tronquant, dans un modèle cellulaire (lignée HEK293) [182]. Pour toutes les conditions, y compris les variations faux-sens, une diminution d'environ 50% des niveaux de protéine de MAU2 a été observée, sans modification de l'ARNm de MAU2, faisant de cette protéine un biomarqueur candidat intéressant.
- Un second axe s'intéresse à améliorer la connaissance des impacts des variations de *NIPBL*. L'analyse de l'expression différentielle a mis en évidence 76 gènes dérégulés dont 45 gènes sous-exprimés (FC >0.125, FDR <5%). Parmi ces derniers, 8 gènes sont haploinsuffisants (pLi >0.9) et associés à un phénotype dans OMIM Morbid. De manière intéressante, nous avons observé que les caractéristiques cliniques associées aux variations pathogènes de ces gènes chevauchent celles du CdLS.

Les résultats de ces travaux sont détaillés dans le manuscrit ci-joint, qui sera prochainement soumis à un journal à comité de lecture. Ce travail a également été présenté sous forme de

poster lors du congrès de l'European Society of Human Genetics à Vienne en 2022. De plus, il fera l'objet d'une communication orale lors du congrès de l'Association des Cytogénéticiens de Langue Française en septembre 2023 au Havre.

2. Article scientifique

----- DEBUT DU MANUSCRIT -----

Title

Assessment of the transcriptomic consequences and MAU2 protein levels in edited induced pluripotent stem cells with *NIPBL* pathogenic variants

Authors and affiliations

Kévin Cassinari^{1,*}, Anne Rovelet-Lecrux¹, Céline Derambure¹, Myriam Vezain¹, Sophie Coutant¹, Anne-Claire Richard¹, Nathalie Drouot¹, [*Cosign collaborators*], Juliette Coursimault¹, Gabriella Vera¹, Alice Goldenberg¹, Pascale Saugier-Weber¹, Camille Charbonnier², Gaël Nicolas^{1,*}

¹. Univ Rouen Normandie, Normandie Univ, Inserm U1245 and CHU Rouen, Department of Genetics and reference Center for Developmental Disorders, F-76000 Rouen, France.

². Univ Rouen Normandie, Normandie Univ, Inserm U1245 and CHU Rouen, Department of Biostatistics, F-76000 Rouen, France.

*Corresponding authors

Kévin Cassinari, kevin.cassinari@gmail.com, and Gaël Nicolas, gaelnicolas@hotmail.com, cytogenetics laboratory, department of Genetics, Rouen University Hospital, 37 Boulevard Gambetta, Rouen, France. Tel.: +33668704102

Author contributions

Study concept and design: Kévin Cassinari, Pascale Saugier Veber, Anne Rovelet-Lecrux, Camille Charbonnier and Gaël Nicolas.

Material preparation, data collection and analysis: Kévin Cassinari, Anne Rovelet-Lecrux, Céline Derambure, Myriam Vezain, Sophie Coutant, Anne-Claire Richard, Nathalie Drouot, Juliette Coursimault, Gabriella Vera, Alice Goldenberg, Pascale Saugier-Weber, Camille Charbonnier and Gaël Nicolas.

Drafting the manuscript: Kévin Cassinari.

Critical revision: Anne Rovelet-Lecrux, Camille Charbonnier and Gaël Nicolas.

Abstract

Introduction: The cohesin complex has a critical role in chromatin structure and gene expression regulation. Cornelia de Lange syndrome (CdLS) is a transcriptomopathy linked to alterations in this complex. *NIPBL*, the main gene associated with CdLS, encodes the cohesin loading factor NIPBL, in association with its partner MAU2. Studying the transcriptomic effects of pathogenic nucleotide or structural variations in the *NIPBL* gene could lead to a better understanding of the syndrome and potentially unveil novel biomarkers.

Methods: Using CRISPR/Cas9, we introduced the following pathogenic variants of *NIPBL* into an iPSC line, at heterozygous or homozygous states: p.Arg45*, p.Arg834*, p.Ser1466Lysfs*13, p.Asp2157Gly, and p.Arg2298Cys, as well as frameshift indels in corresponding exons. We assessed mRNA levels of *NIPBL* and *MAU2* using RT-ddPCR and RNAseq, and measured protein levels of NIPBL and MAU2 protein using western blot. Subsequently, we employed RNAseq data to establish a transcriptomic signature resulting from loss of function *NIPBL* alterations.

Results: RT-ddPCR and RNAseq revealed decreased *NIPBL* mRNA levels for all clones carrying truncating variants, except for the p.Arg45* variant, likely due to a previously described alternative translation initiation site. Across all conditions, including missense variants, an approximate 50% decrease in MAU2 protein levels was observed, without alterations in *MAU2* mRNA. Differential expression analysis identified 60 dysregulated genes, including 46 downregulated genes (fold change >1.25, FDR <5%), among which 8 genes are haploinsufficient (pLi >0.9) and associated with a Mendelian disease in OMIM Morbid, with phenotypic characteristics overlapping that of CdLS.

Conclusion: Our results confirm previous findings suggesting that variations at the 5' end of *NIPBL* coding sequence escape nonsense-mediated decay. We propose MAU2 protein levels as a potential biomarker for CdLS. Lastly, we demonstrate that *NIPBL* alteration significantly leads to the decreased expression of genes, some of which recapitulate most of the CdLS phenotypic features.

Introduction

Introduction

Cornelia de Lange Syndrome (CdLS, OMIM 122470) is a clinically distinguishable intellectual disability syndrome with monogenic inheritance [1]. Distinctive clinical features include synophrys or highly arched eyebrows, long eyelashes, a small upturned nose, thin downturned lips as well as growth retardation, limb abnormalities including small hands and feet, oligodactyly or missing fingers, and intellectual disability with or without autism spectrum disorder. Additional symptoms may be encountered, such as gastroesophageal reflux disease, congenital diaphragmatic hernias, hearing loss, and vision alterations. Six genes have been associated with CdLS: *NIPBL* [2], *SMC1A* [3], *SMC3* [3], *RAD21* [4], *HDAC8* [5], and *BRD4* [6]. Nonetheless, molecular testing in patients with typical or atypical CdLS may occasionally fail in identifying a causal variant in one of these genes, suggesting the contribution of unidentified genes or cryptic variants, or uncovers a variant of uncertain significance, none of each situation allowing genetic counseling.

All six known genes encode components or proteins closely associated with the cohesin protein complex. Cohesins are conserved from yeast to humans and have been found to play a pivotal role in the regulation of gene expression, in addition to their established function in sister chromatid cohesion during mitosis. This group of proteins affects gene regulation by facilitating the organization of chromatin into topologically associating domains (TADs), thus governing long-range interactions between promoters and enhancers [7]. Cohesins also regulate transcription by modulating the recruitment and pause-release of RNA Polymerase II at promoters, thereby accentuating their primary role in gene expression.

The implication of cohesin complexes in gene expression regulation provides grounds to classify CdLS as a transcriptomopathy. Transcriptomopathies comprise a category of

monogenic disorders primarily driven by dysregulated transcription of multiple genes, resulting from DNA variants in genes encoding critical regulators of gene expression. Pathogenic variants in genes encoding the cohesin complex elements and regulators disrupt the transcription process, resulting in global changes in gene expression patterns. This is associated with a wide spectrum of developmental abnormalities observed in transcriptomopathies.

Among the six CdLS-associated genes, *NIPBL* (OMIM 608667) is the main one in frequency, accounting for approximately 60% of all cases. *NIPBL* encodes the Nipped-B-like protein, which functions as a loading factor for cohesin, assisting in attaching the cohesin complex onto the DNA in collaboration with MAU2 (OMIM 614560). Interactions between MAU2 and NIPBL are critical for cohesin functionality and in CdLS pathogenesis [8]. Both NIPBL and MAU2 play specific roles in the process of loading cohesin onto DNA, a prerequisite for the complex to exert its functions. They form a heterodimeric complex, which serves as the primary loader of the cohesin ring onto DNA.

The emergent field of transcriptomics offers a valuable path for comprehending and potentially identifying potential biomarkers and drug targets for disease modification. Previous works have already explored transcriptomic consequences of *NIPBL* alterations. Liu and collaborators have used microarrays on 16 lymphoblastic cell lines of *NIPBL* loss-of-function variants carriers, and generated a list of dysregulated genes [9]. Mills and collaborator performed RNA sequencing on induced cardiomyocytes derived from induced pluripotent stem cells (iPSCs) obtained from skin fibroblasts of four patients and generated a list of 329 up and down regulated genes in both tissues mainly linked to nucleosome [10]. Others have also performed transcriptomic approaches in various contexts [8], [11], [12], but none included both missense and protein truncating variations of *NIPBL*, introduced in iPSCs in an isogenic context.

In this study, we aimed at assessing the mRNA and protein levels of NIPBL and MAU2 as well as the transcriptomic consequences of multiple truncating and missense *NIPBL* pathogenic variants in isogenic iPSCs. We chose to work from iPSC in order to model the early defects in gene regulation during development. We confirm that early-truncating variants do not result in nonsense-mediated mRNA decay (NMD), that MAU2 protein levels are decreased in every pathogenic condition, including early-truncating, truncating and missense variants, and we highlight deregulated genes that may account for a large part of the CdLS phenotype.

Methods

Induced Pluripotent Stem Cells

The iPSc line was previously generated by our group [13]. This cell line derives from an unaffected male donor. It was analyzed for pluripotency markers expression by quantitative PCR and for chromosomal abnormalities by karyotyping, and was able to differentiate into the three germ layers. Whole exome sequencing of the parental iPSC line confirmed the absence of rare coding variants in the *NIPBL* and *MAU2* genes. iPSCs were cultured on feeder-free conditions in mTeSR Plus medium (STEMCELL Technologies, Vancouver, Canada) on Matrigel-coated culture dishes (Corning, Corning, NY, USA) diluted in DMEM-F12 according to manufacturer's instructions in a 37 °C/5% CO₂ incubator. Cells were split when they reached 80% confluency using StemPro Accutase (Thermo Fisher Scientific) and plated in 10 μM ROCK inhibitor (StemGent, Cambridge, MA, USA) supplemented medium. Medium was refreshed the next day to remove ROCK inhibitor. Cell lines were confirmed to be free of mycoplasma.

Genome editing of iPSCs

NIPBL variants were introduced in the genome of iPSC by CRISPR/Cas9. The crRNA (crRNAs) were designed using the CRISPOR.org web tool (<http://crispor.tefor.net/>). Two crRNAs were designed for each variant. (see additional file). The wild-type Cas9, tracrRNA, crRNA, and ssODN were purchased from IDT (<https://eu.idtdna.com>). The crRNA and tracrRNA were annealed to form the guide RNA, and then combined to the Cas9 to obtain RNPs. Each RNP and its corresponding single-stranded oligodeoxynucleotide (ssODN) were then nucleofected in the iPSC line, using an AMAXA nucleofector II device. Two days later, cells were diluted plated into 96-well plates as 1 cell/well. When clones reached 80% confluency, genomic DNA was isolated and the presence of the variant was assessed by

Sanger sequencing (PCR primers are available upon request). For each variant, 3 clones carrying the variant at the heterozygous state were selected, as well as 3 clones without nucleotide change, corresponding to WT controls. For the R2298C variant, 3 clones carrying the variant at the homozygous state could also be isolated.

RT-ddPCR of *NIPBL*

Total RNAs were extracted using the Nucleospin® RNA isolation kit (Macherey-Nagel), according to the manufacturer's instructions. RNA was quantified by spectrophotometry (Nanodrop; Thermo scientific). Reverse transcription was performed on 100 ng RNA, using the Verso cDNA kit with oligoDT primers (Thermo Scientific). Relative *NIPBL* gene expression in iPSCs was then assessed by digital droplet PCR (ddPCR) on a QX200 platform (Bio-Rad Laboratories). The RT-ddPCR were performed by relative quantification with TPB, used as reference gene as previously described [14]. *NIPBL* was PCR-amplified using the following primers: Fw: 5'-GCCCCATGTCCCCATTAC-3', Rv: 5'-GCAGGTAAAGGAGATGGAAGAG-3', associated with the FAM-labeled hydrolysis probe. The reference amplicon, located in the TBP gene, was PCR-amplified using the following primers: (Fw: 5'-CGGCTGTTTAACTTCGCTTC-3', Rv: 5'-CACACGCCAAGAAACAGTGA-3') associated with the HEX-labeled hydrolysis probe (IDT DNA). For each cell line, analyses were performed in two technical replicates.

Protein extraction and western blotting

Soluble proteins were extracted from each iPSC clone using RIPA buffer (Pierce, Thermofisher Scientific) and quantified using the DC protein assay kit (Bio-Rad Laboratories). To analyse the NIPBL protein, 30µg proteins were resolved on Tris-acetate NOVEX NuPAGE 3-8% gels (Invitrogen, Thermofisher Scientific). To analyse the MAU2 protein, 20µg proteins were migrated on 10% TGX Stain Free gels (Bio-Rad Laboratories).

Proteins were transferred onto a nitrocellulose membrane, blocked in 5% non-fat milk and immunoblotted with the appropriate primary antibody: anti-NIPBL (1:3,000; A301-779A, Bethyl) or anti-MAU2 antibody (1:2,000; Ab183033; Abcam). Membranes were then incubated with secondary peroxidase-labelled anti-rabbit antibody (1:10,000, Jackson ImmunoResearch Laboratories). Signals were detected with chemiluminescence reagents (ECL Clarity, Bio-Rad Laboratories) with a GBOX monitored by the Gene Snap software (Syngene). The signal intensity was quantified using the Genetools software (Syngene) and normalized to the total amount of proteins using the Stain-Free signal (ImageLab™ software, Bio-Rad Laboratories).

Patients

Patients selected for the RNAseq analysis from blood samples comprised 8 individuals diagnosed with typical CdLS, carrying heterozygous *de novo* deleterious variants of *NIPBL* identified in their blood by NGS panel sequencing [**Supplementary Table 1**]. The control group for this analysis included 20 subjects, age- and sex-matched with the 8 patients, and without any neurodevelopmental pathologies. Both patients and controls were sampled using PAXgene tubes, and informed consents were obtained from their legal representatives.

RNAseq

Total RNAs were isolated from iPSC with using the RNeasy kit from Qiagen, and with PAXgene blood RNA kit (Qiagen PreAnalytiX GmbH) for blood samples, according to the manufacturer's recommendations. RNAs were then stored at -80°C until use. The quality and quantity of RNA were assessed using the 4200 TapeStation (Agilent Technologies) and the Qubit 3.0 device (Thermo Scientific). Only RNA samples with a minimal RNA integrity number of 7 were used for subsequent experiments. Libraries were prepared using the NEBNext Ultra II Directional RNA Library Kit for Illumina (New England Biolabs) kit and

High-throughput sequencing of the libraries was performed on an Illumina NextSeq 500 (Illumina) using 2*75 bp sequencing to generate 30M read pairs on average per sample. Bioinformatics analysis was carried out using nf-core/RNA-seq v3.1 analysis pipeline to generate multi quality control report that uses the STAR v2.6.1d and SALMON v1.4.0 tools for alignment [15]-[17]. Differential analyses were performed using DESeq2 package [18] and visual exploration of the BAM files was performed with the IGV tool from the Broad Institute [19]. For the secondary analysis, the HPO term list was extracted from The Human Phenotype Ontology website (<https://hpo.jax.org/app/>) [20]. The list of haploinsufficient genes (pli >0.9) was exported from gnomAD v3.1 (<https://gnomad.broadinstitute.org>) [21]. The circular plot was generated using R package Circlize (chordDiagram function).

Results

Generation of iPSC lines carrying pathogenic NIPBL variants

By CRISPR-Cas9, we succeeded in generating a total of 15 edited induced pluripotent stem cell (iPSC) lines with five different variants. We selected 5 variants carried by patients with a typical CdLS phenotype based on the recruitment of our molecular genetics lab in Rouen, France, and successfully introduced all five in iPSC by CRISPR/Cas9 editing. We also kept for further analyses several iPSC lines with frameshift indels introduced at the targeted positions during genome editing. Overall, edited iPSC lines consisted in 2 lines with heterozygous protein-truncating variants (PTVs) in exon 3 (E3) of the *NIPBL* gene (early PTV), 9 iPSC lines carrying heterozygous PTVs in other targeted *NIPBL* locations (E10, E20, E37, or E40), 4 iPSC lines that carried heterozygous missense mutations in either E37 (c.6470A>G, p.(Asp2157Gly)) or E40 (c.6892C>T (p.Arg2298Cys)) of the *NIPBL* gene, and 2 iPSC lines with the c.6470A>G, p.(Asp2157Gly) in E37 at the homozygous state. In addition, 9 wild type (WT) iPSC lines that underwent the same selection process, without any *NIPBL* variant, were kept as controls. Overall, PTVs consisted in either patient-specific variant (c.133C>T; p.(Arg45*) in E3, c.2500C>T, p.(Arg834*) in E10 and c.4396dup, p.(Ser1466Lysfs*13 in E20)) or short insertions or deletions resulting from aberrant DNA repair at the targeted positions [Figure 1]. The PTVs in exon 3 were classified as early PTVs, because they map in 5' of a suspected alternative Translation Initiation Site [8]. Interestingly, we could not get any clone carrying a homozygous truncating variant in *NIPBL*, suggesting that a full KO of this gene is not compatible with iPSC survival and mitosis.

NIPBL and MAU2 mRNA assessment

We then assessed *NIPBL* mRNA levels in iPSC lines. According to RNAseq count of linearized *NIPBL* reads (expressed in TPM), we observed a notable decrease in *NIPBL* mRNA levels in iPSC lines carrying PTVs compared to WT, after exclusion of the early PTV, showing a 41.6% decrease ($p < 0.001$). Interestingly, a non-significant increase of 22.4% ($p = 0.14$) in mRNA levels was detected in the iPSC line carrying an early PTV as compared to controls, suggesting that these transcripts are not degraded by nonsense-mediated decay (NMD), as previously reported [8]. As expected, no significant difference in *NIPBL* mRNA levels was observed for both heterozygous (-10.0%, $p = 0.41$) and homozygous (-8.6%, $p = 0.43$) missense variants [Figure 2.a]. These findings were subsequently confirmed for both PTV and early PTV variants through an independent two-step relative RT-ddPCR assay [Sup. Figure 1]. The same pattern of mRNA level variation was observed, including a significant decrease in iPSC lines carrying PTV variants (-61%, $p < 0.05$), and a surge in mRNA levels for lines with early PTV variants (+39%, $p = 0.15$).

Concerning *MAU2*, according to RNAseq data, there was no significant variation between iPSC lines carrying PTVs (after exclusion of the early PTV) and WT iPSC lines, with an observed change of only +0.01% (NS). No significant variation in *MAU2* mRNA levels was detected in iPSC lines carrying either heterozygous or homozygous missense *NIPBL* variants. Interestingly, a non-significant increase in *MAU2* mRNA levels was observed in cells carrying early PTV [Figure 2.b].

NIPBL and MAU2 protein assessment

Western blotting experiments were performed to assess NIPBL and MAU2 protein level for each clone. **[Figure 2.C,D]**. NIPBL protein levels were significantly decreased in cells line carrying PTV compared to WT (-35,8%, $p < 0.05$), with the exception of the early PTV.

Strikingly, all iPSC mutant cell lines presented a significant decrease of MAU2 protein levels as compared to controls, including PTVs (-58,4%, $p < 0.0001$), early PTVs (-49,3%, $p < 0.005$), heterozygous missense variants (-40.3%, $p < 0.0001$) and the homozygous missense variant (-64,8%, $p < 0.0001$). These results confirm previous findings for early PTVs and PTVs and extend those to missense variants **[8]**. This highlights the impact of *NIPBL* alteration in MAU2 stability.

Transcriptomic differences between mutant and WT conditions highlight deregulated genes potentially involved in CdLS phenotype

We then assessed the consequences of *NIPBL* pathogenic variants at the transcriptome level. From RNAseq data, we found that 60 genes were differentially expressed between PTV-iPSCs and WT controls, with a fold change > 1.25 or < 0.8 and a false discovery rate (FDR) below 5% **[Figure 3.a,b, Supplementary tables 3-4]**. Of them, 46 genes were downregulated, after exclusion of *NIPBL* itself, and 18 of them (39%) are haploinsufficient (gnomad pLI > 0.9), meaning that a single copy of these genes is not expected to lead to a normal phenotype. Here, fold changes of these 18 genes ranged from 0.64 to 0.43, which is in the same ranges as what is expected in case of a single copy of a given gene (x0.5 on average). Strikingly, 12 of the 46 downregulated genes are associated with an OMIM Morbid phenotype and 8 are both haploinsufficient and associated with an OMIM Morbid phenotype **[Table 1]**. These 8 genes are also referenced in the Human Phenotype Ontology (HPO). We then looked at this list of genes in other mutant conditions in iPSCs. All of them were also

significantly deregulated in the early-PTVs lines as well as in heterozygous and homozygous missense mutant lines, all in the same direction upregulation or downregulation. Coming back to the 8 genes downregulated and considered as haploinsufficient and OMIM-morbid, we compared HPO features associated with these genes to that of NIPBL-CdLS. Strikingly, the main phenotypic features of NIPBL-CdLS overlapped with that of the deregulated genes, suggesting that these genes may have a role in CdLS pathophysiology **[Figure 3.c]**.

Finally, we generated RNAseq data from fresh blood of 8 CdLS patients carrying one of the pathogenic *NIPBL* variants selected for the iPSC edition. Unfortunately, only 22 of the 60 deregulated genes in PTV-mutant iPSCs were expressed in blood (TPM cutoff set at 1) and none of the haploinsufficient plus OMIM Morbid genes belonged to this list.

Discussion

In this study, we inserted different variations of the *NIPBL* gene using CRISPR/Cas9 in order to obtain modified iPSC lines, either heterozygous or homozygous. These lines were finely characterized at the transcriptomic and protein levels. We highlighted a decrease in the protein levels of MAU2 for all lines carrying deleterious variations of *NIPBL*. Furthermore, the transcriptomic analysis by RNAseq of these lines revealed a restricted list of genes that were significantly upregulated or downregulated. We suggest that some of these genes actually contribute to the phenotype of Cornelia de Lange syndrome.

Despite an important role as the main NIPBL partner, only a few genomic alterations of *MAU2* have been described in patients so far. The only monogenic description is a *de novo* in-frame deletion of 21 bp in *MAU2*, resulting in the loss of seven amino acids in a patient with a severe phenotype and typical CdLS facial dysmorphism [8]. This variant impairs the interaction between MAU2 and the NIPBL N-terminus. Recently, a series of three patients with neurodevelopmental delay and carrying a *de novo* 19p13.11p12 deletion encompassing the entirety of *MAU2* along other genes was reported [22]. The CdLS clinical spectrum overlapped with the phenotype of these patients, with the syndrome having even been clinically suspected for one of them prior to the identification of the deletion. Altogether, these observations suggest that *MAU2* loss of function variants could be considered as causative for CdLS.

At the molecular level, it has already been shown that the fundamental interactions between MAU2 and NIPBL are crucial for their respective stability, which is mediated by heterodimerization via the N-terminal end of NIPBL [8], [23], [24]. In this study, we expand on these findings by demonstrating that distinct deleterious variations introduced into NIPBL, including missense variations located in the 3' end of the gene (E37-E40), are associated with

a decrease in MAU2 protein levels. This point appears particularly relevant since these missense variants are located within a distinct domain from the one known to interact with MAU2. It therefore seems that reductions in MAU2 protein levels could be considered as a potential universal biomarker for pathogenic *NIPBL* variations, whatever the type of the pathogenic variant.

Here, we also confirm the findings of Parenti and colleagues [8] about the impact of early truncating variations on the NIPBL protein. Indeed, Western Blot analyses using an antibody located in the central part of the protein showed no decrease in NIPBL protein levels in cells modified with truncating variations in exon 3 of *NIPBL*. However, the protein levels of MAU2 are also reduced in these cases. This remains consistent with previous findings, considering the probable absence of translation of the N-terminal end of NIPBL, which interacts with MAU2.

In establishing the transcriptomic signature of CdLS, we opted to introduce variations using CRISPR/Cas9 [25] to obtain an isogenic model for interpreting transcriptomic analyses. One of the main limitations of this type of model is, of course, that this signature unique to iPSCs may not be applicable to more differentiated tissues, especially during the early stages of differentiation, considering the changing patterns of gene expression [26]. However, given that CdLS is an early developmental disorder, it seems that iPSC cells can serve as a suitable model to capture the initial changes of this pathology.

The relatively low number of genes sufficiently expressed in the blood, among our list of genes identified as significantly dysregulated, illustrates these tissue divergences. It is well-established that the blood expression of most genes is lower than in fibroblasts, an effect amplified by inter-individual variations [27]. Together, these factors complicate the establishment of a singular transcriptional signature that can be uniformly observed across

individual blood samples. The results obtained here from poly A capture suggest that alternative RNAseq approaches, allowing better gene capture, might be more informative. This could be the case for exome captures [28]; another solution would be to use a sample derived from a fibroblast culture, offering the advantage of a higher number of expressed genes [27], but at the same time constituting a longer and more invasive approach.

A novel aspect of this work is the identification of haploinsufficient genes that are downregulated in iPSCs and overlap with phenotypic features of CdLS. Among these, several genes are linked to intellectual disability phenotypes and developmental abnormalities, which are central in CdLS. This is the case for *GRIN2A* [29], *ARFGEF1* [30], *GABRA5* [31], [32], and *WNT5A* [33]. For most of these genes, seizures are also common. Furthermore, patients carrying pathogenic variants in *GABRA5* and *GRIKA* present microcephaly, which is one of the common clinical features of CdLS. Additionally, the *WNT5* gene is associated with Robinow syndrome [34], which includes limb abnormalities, as observed in CdLS including clinodactyly of the fifth finger [35]. On the other hand, *NR3C1* is related to glucocorticoid resistance [36], leading to hirsutism in affected individuals. Hirsutism is also present in CdLS, although it does not seem to be related to a hormonal cause; the term hypertrichosis is more commonly used for CdLS. Hypertrichosis is defined as excessive hair growth anywhere on the body in either males or females. It is important to distinguish hypertrichosis from hirsutism, which is a term reserved for females with an excessive amount of terminal hairs in androgen-dependent sites [37]. Overall, this confluence of signs appears relevant to the CdLS spectrum. However, many of these features are not specific, and the fine mechanisms leading to the above-mentioned syndromes are not purely haploinsufficiency for all, so that we cannot claim that CdLS is recapitulated by the intersection of these syndromes. Further investigation is thus needed to establish causality and better understand how these (and other) deregulated genes may contribute to CdLS.

In conclusion, by establishing novel models of CdLS in isogenic iPSCs carrying distinct pathogenic *NIPBL* variants, our results strengthen the role of MAU2 protein destabilization as a shared molecular feature across variants types, thus placing the relative amount of this intracellular protein as a potential CdLS biomarker. In addition, we established a list of deregulated genes in iPSCs with pathogenic *NIPBL* variants, some of these genes being haploinsufficient, i.e. their lower expression in such orders of magnitude is not expected to provide a normal phenotype. Strikingly, 8 of them are indeed already associated with Mendelian disorders in humans when mutated, and the phenotypic features of these disorder are commonly found in CdLS.

Gène	OMIM morbid	Syndrome
ARFGEF1	619964	Developmental delay, impaired speech, and behavioral abnormalities, with or without seizures; DEDISB
CDH11	619736	Teebi hypertelorism syndrome 2; TBHS2
GABRA5	618559	Developmental and epileptic encephalopathy 79; DEE79
GRIA1	619927	Intellectual developmental disorder, autosomal dominant 67; MRD67
GRIK2	619580	Neurodevelopmental disorder with impaired language and ataxia and with or without seizures; nedlas
GRIN2A	245570	Epilepsy, focal, with speech disorder and with or without impaired intellectual development; FESD
NR3C1	615962	Glucocorticoid resistance, generalized; GCCR
WNT5A	180700	Robinow syndrome

Table 1. List of the eight downregulated genes which are haploinsufficient ($pli > 0.9$) and associated with an OMIM Morbid referenced phenotype.

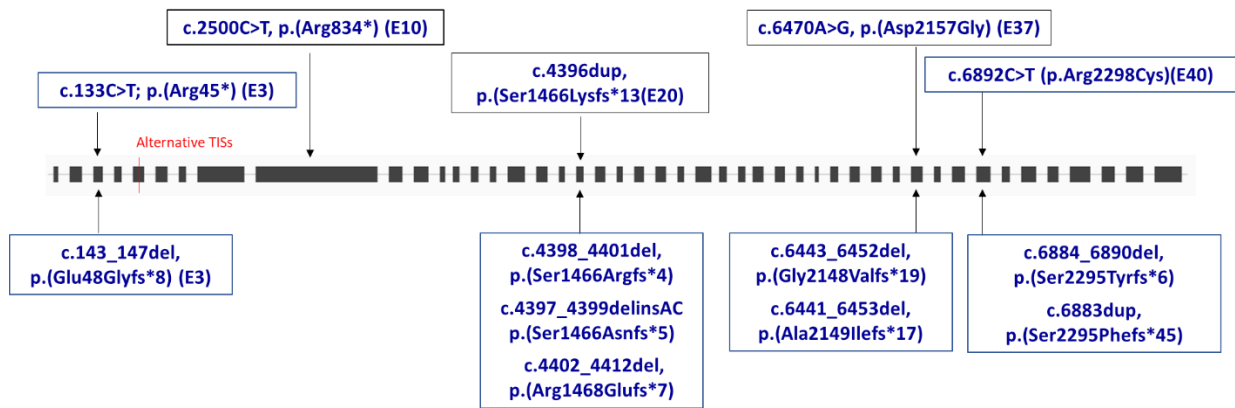


Figure 1. Variants included in iPSC lines across the *NIPBL* gene. Representation of the *NIPBL* gene sequence, where black squares represent exons and thin lines represent introns. Patient-specific variants are represented with their nomenclature at the cDNA and protein levels at the top of the figure. At the bottom of the figures are represented short insertions or deletions resulting from aberrant DNA repair at the targeted positions. TIS: Translation Initiation Site.

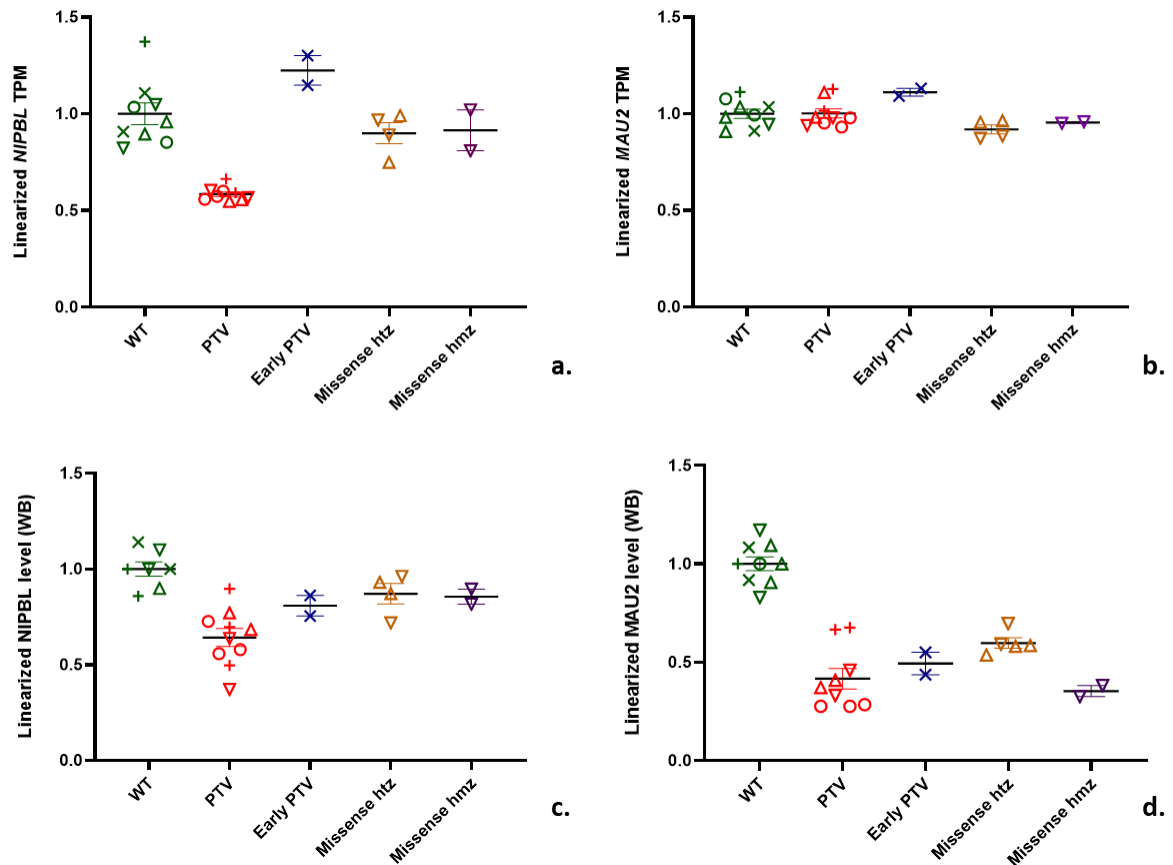
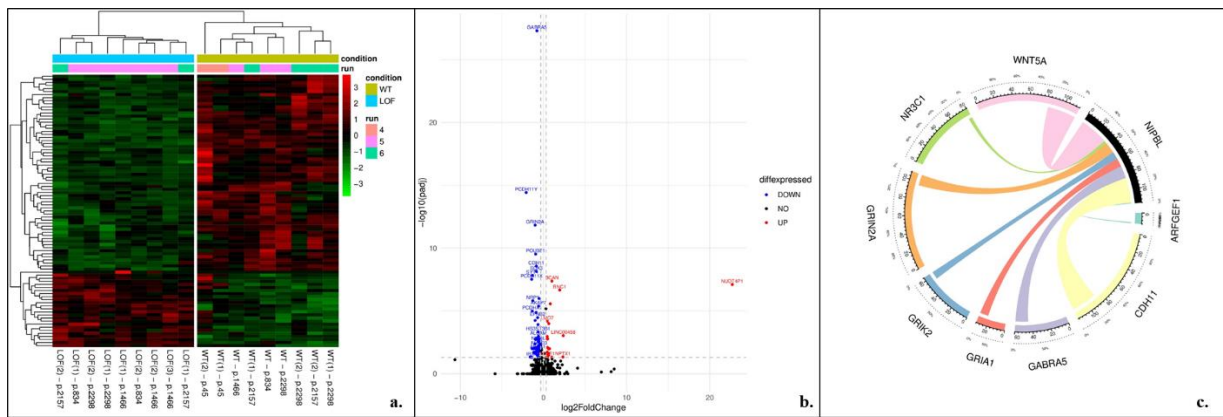


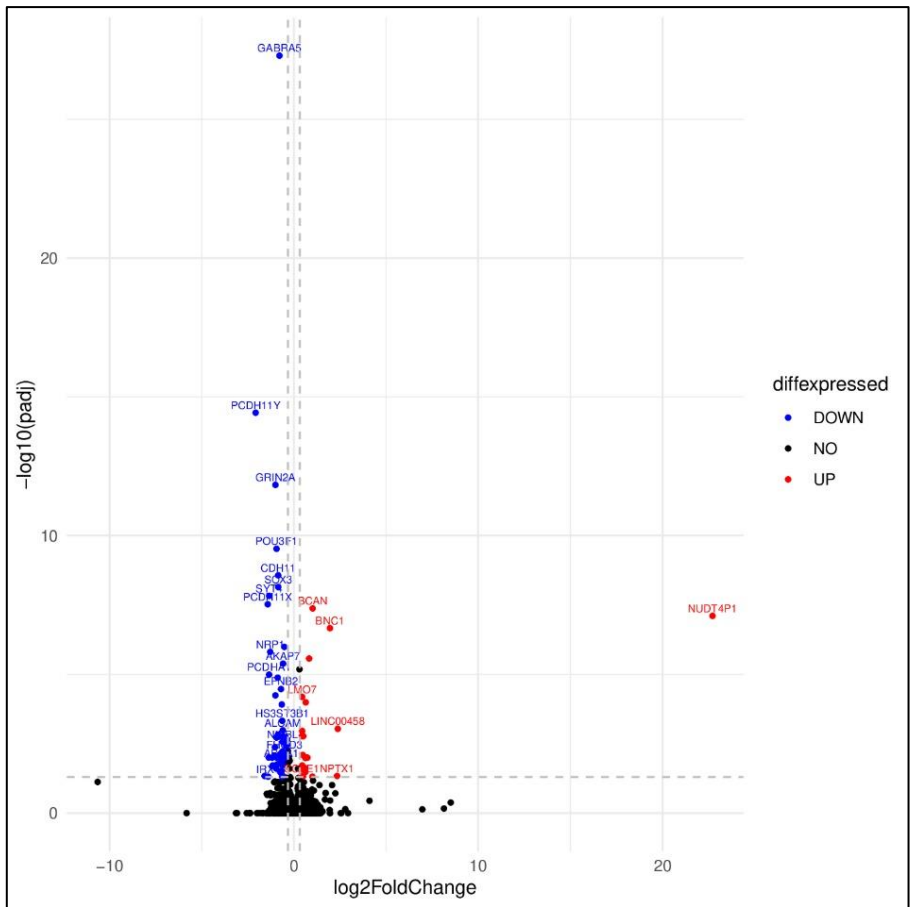
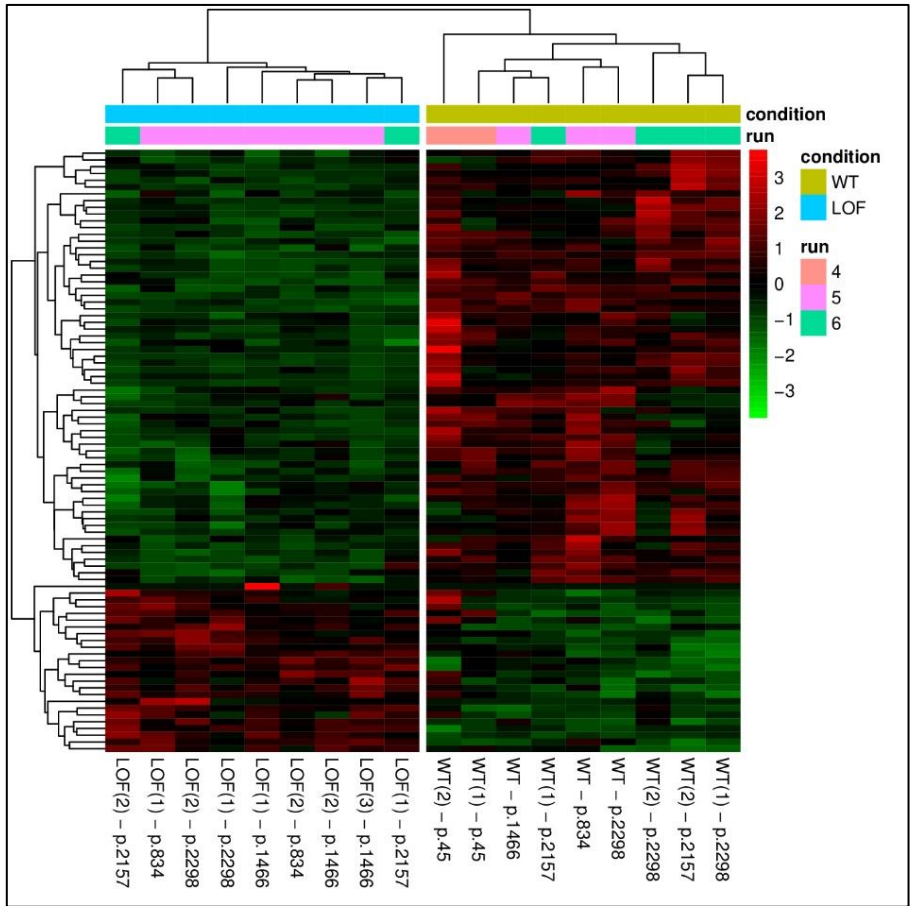
Figure 2. Assessment of mRNA and protein levels of MAU2 and NIPBL in the iPSC lines.

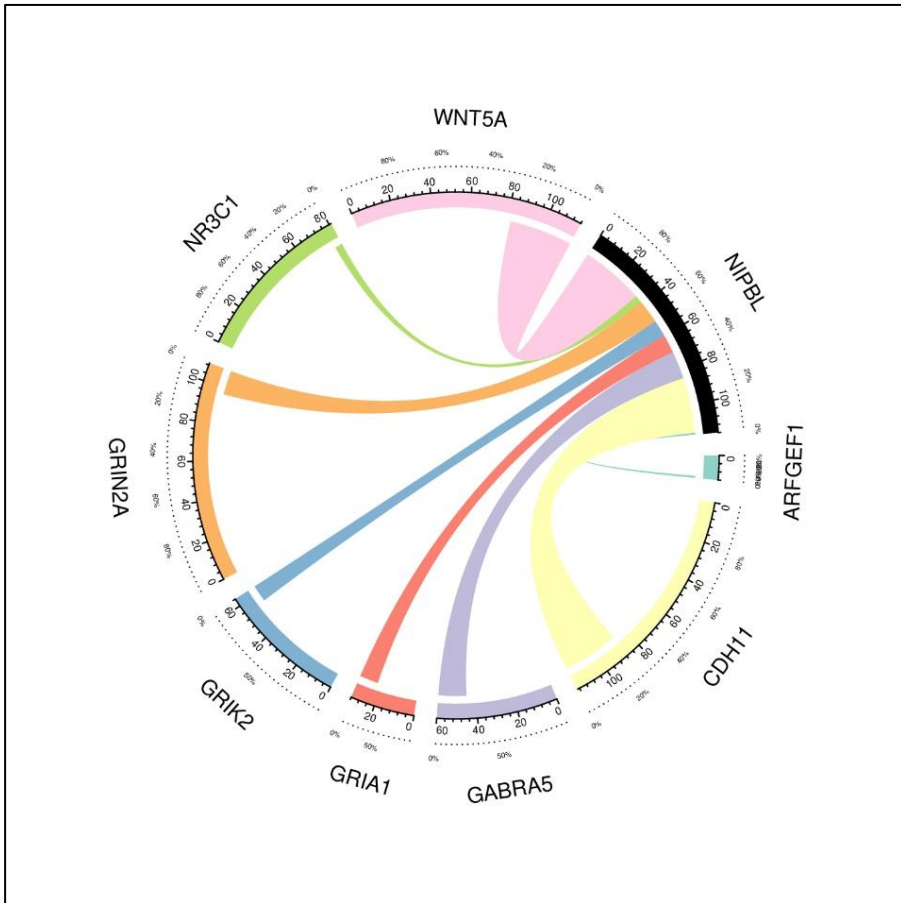
a, b. mRNA levels linearized with WT iPSC line, expressed in TPM, from RNAseq data for NIPBL and MAU2, respectively. **c, d.** Protein level linearized with WT iPSC line for NIPBL and MAU2, respectively. For each graphs, X represents variants in E3 (p.45), + represents variants in E10 (p.834), O represents variants in E20 (p.1466), ∇ represents variants in E37 (p.2157) and Δ represents variant in E40 (p.2298). Horizontal bars represent average and standard deviation.



(Pour plus de visibilité, les éléments de la figure sont présentés agrandis aux pages suivantes.)

Figure 3. Results of differential transcriptomic analyses, conducted with an FC set at 1.25 and $FDR < 0.05$. **a.** The heatmap presents an overview of the transcriptomic signature obtained by comparing iPSCs carrying LOF variations to WT iPSCs. **b.** Volcano plot showcasing downregulated genes (in blue) and upregulated genes (in red). **c.** Circular plot illustrating HPO terms shared between *NIPBL* and each of the 8 haplosensitive downregulated genes referenced in OMIM and highlighted by differential analysis. For all genes except *NIPBL*, the sector size is proportional to the number of attached HPOs. However, for *NIPBL*, the circo plot does not permit overlaps, and the sector width represents the total number of HPOs shared with other genes, with each shared HPO counted x times if shared x times.





ID	Age (years)	Sex	Variation (NM_133433.3)	Type	Transmission	RIN
19-10695	2.5	M	c.6893G>A; p.(Arg2298His)	Missense	<i>De novo</i>	8.6
19-07413	4,5	F	Deletion E7 to 47	LOF	<i>De novo</i>	7
19-07679	8	F	c.6892C>T; (p.Arg2298Cys)	Missense	<i>De novo</i>	7.5
19-12808	15	H	c6470A>G; p.(Asp2157Gly)	Missense	<i>De novo</i>	7.4
19-14494	8	H	Deletion E7 to 47	LOF	<i>De novo</i>	7.5
20-05745	0.5	F	c.2479_2480del, p.(Arg827Glyfs*2)	LOF	<i>De novo</i>	8.1
21-08554	1	F	c.2500C>T, p.(Arg834*)	LOF	<i>De novo</i>	7.1
21-01063	17	F	c.5862+3487C>T;p.(?)	Intronic	<i>De novo</i>	7.8

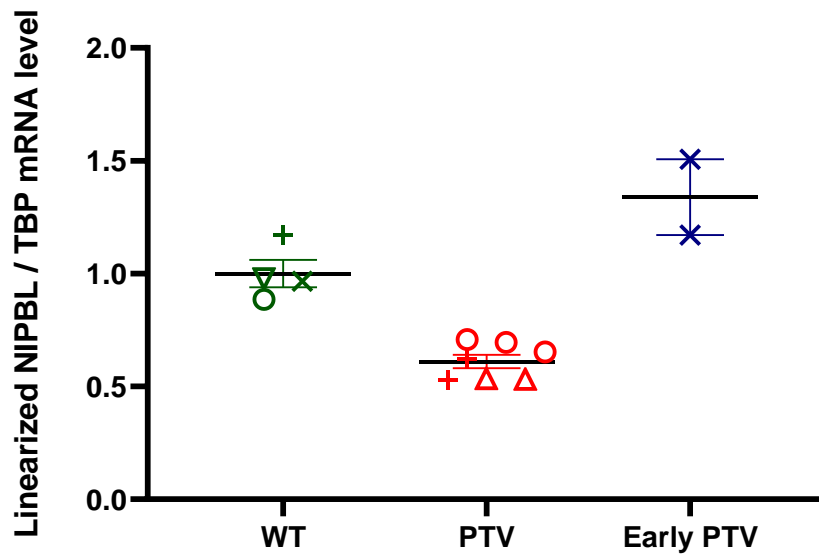
Supplementary table 1. Patients included for RNAseq blood analyses.

LOF : loss-of-function, RIN : RNA integrity number

ID	Age (years)	Sex	RIN
19-14804	1,44	F	8,7
19-15444	5,35	F	7,6
19-15548	5,44	F	7,7
19-13458	5,58	M	7,9
19-12770	6,34	M	8,4
19-13154	7,3	F	7,5
19-14591	8,02	F	8,2
19-12558	10,14	F	7
19-12634	12,21	M	8,5
19-12578	12,59	F	7,4
19-13160	13	F	7,4
19-13003	14,4	M	7,4
19-13123	14,6	F	7,8
19-15202	14,66	F	7,8
19-12635	15,19	M	7,4
19-13041	15,3	F	7,6
19-13221	15,82	F	7,8
19-12869	20,5	M	7,4
19-13607	24,53	F	7,3
19-14503	25,7	F	8

Supplementary table 2. Controls included for RNAseq blood analyses

RIN : RNA integrity number



Supplementary figure 1. NIPBL RNA level independent evaluation in the iPSC lines using RT-ddPCR.

× represent variants in E3 (p.45), + represent variants in E10 (p.834), ○ represent variants in E20 (p.1466), ▽ represent variants in E37 (p.2157), △ represent variants in E40 (p.2298). Each point represents an average of 3 technical replicates.

gene	log2FoldChange	Fold change	pvalue	padj	PLI	OMIM morbid	disease OMIM and mechanism
ABCA1	-0.671116419	0.62802051	8.42E-05	0.028455781	0	Yes	604091
AIM1	-0.381747731	0.767507242	0.000149851	0.044556925	0	No	
AJAP1	-0.786628059	0.579697408	3.53E-06	0.002014984	0.99	No	
AKAP7	-0.566333184	0.67533106	2.80E-07	0.00029037	0	No	
ALCAM	-0.602173634	0.658760687	1.56E-06	0.001175663	0.73	No	
ARFGEF1	-0.306891354	0.808381744	2.82E-05	0.011136924	1	Yes	619964
CDH11	-0.833349592	0.561224699	1.21E-10	3.35E-07	1	Yes	619736
EFNB2	-0.670760647	0.628175401	1.42E-06	0.001117353	0.99	No	
GABRA5	-0.749603341	0.594767062	4.01E-21	6.64E-17	0.91	Yes	618559
GRIA1	-0.695488904	0.617500023	2.17E-06	0.001380644	1	Yes	619927
GRIK2	-1.206259512	0.433390816	0.000173906	0.048009514	1	Yes	619580
GRIN2A	-0.975101062	0.50870421	1.81E-13	1.50E-09	1	Yes	245570
HS3ST3B1	-0.636468228	0.643285811	3.52E-07	0.00032432	0.29	No	
ID2	-0.906843838	0.533350619	2.62E-05	0.010570546	0.67	No	
IRX1	-1.440777301	0.36836878	7.78E-05	0.026833057	0	No	
KCNK12	-0.649552351	0.637478084	0.000137072	0.042045715	0.92	No	
MAF	-0.754939311	0.592571312	4.60E-05	0.016930506	0.74	Yes	601088
MAML2	-0.616083065	0.652439908	0.000150994	0.044556925	1	No	
MARCH3	-0.463034655	0.725458677	2.75E-06	0.001625069	0.17	No	
NCAM1	-0.552088728	0.68203197	3.37E-05	0.012972585	1	No	
NIPBL	-0.630664746	0.645878747	1.73E-06	0.001223604	1	Yes	122470
NR3C1	-0.631403097	0.64554828	1.46E-05	0.006733033	0.97	Yes	615962
NRP1	-1.267226608	0.415457669	4.15E-09	8.58E-06	0.99	No	
PCDH11X	-1.332236419	0.397152112	2.93E-09	6.93E-06	0.02	No	
PCDH11Y	-1.931527772	0.262151413	5.96E-13	3.29E-09	0	No	
PCDHA1	-1.236530866	0.424391932	1.25E-08	1.88E-05	0	No	
PCDHB13	-0.911558241	0.531610594	1.21E-05	0.006083638	0	No	
PCDHB14	-0.82138873	0.565896951	3.35E-07	0.00032432	0	No	
PCDHB15	-0.728538869	0.603514831	1.77E-05	0.007331885	0	No	
PCDHGB1	-0.949983241	0.517638475	9.97E-05	0.032674474	0	No	
PCDHGB2	-0.879180141	0.543676305	4.89E-05	0.017616248	0	No	
PIP5K1B	-0.374729075	0.771250233	0.00015876	0.044839225	0	No	
PLEKHA5	-0.280539531	0.823283073	0.000103905	0.033097713	1	No	
PLXDC2	-0.444704292	0.734734899	1.93E-07	0.000213079	0.13	No	
POU3F1	-0.902824351	0.534838657	9.64E-12	3.99E-08	0.56	No	
PPAP2A	-0.280301688	0.823418811	5.86E-05	0.020642286	0.1	No	
RHOU	-0.860199427	0.550876404	0.000100604	0.032674474	0.08	No	
RND3	-0.583378033	0.667399249	7.22E-08	9.20E-05	0.97	No	
SATB2	-0.818999711	0.56683482	1.43E-05	0.006733033	1	No	
SFMBT2	-0.439459046	0.737411057	1.74E-05	0.007331885	1	No	
SOX3	-0.773428919	0.585025364	5.22E-09	9.60E-06	0.45	Yes	300123
SPP1	-0.361325957	0.778448791	9.86E-06	0.005101806	0	No	

SULF2	-0.363783681	0.777123783	0.000153329	0.044556925	0.01	No	
SYT4	-1.2297827	0.426381663	1.39E-08	1.91E-05	0.44	No	
TACR3	-0.745120913	0.596617866	0.000159715	0.044839225	0	Yes	614840
WBSCR17	-0.334536116	0.793039089	1.76E-05	0.007331885	0.12	No	
WNT5A	-0.638051203	0.642580363	2.25E-06	0.001380644	0.99	Yes	180700

Supplementary table 3. List of downregulated genes in iPSC lines carrying truncating variants compared to WT controls. Haplosensitive genes are indicated in bold.

Gene	log2FoldChange	Fold change	pvalue	padj
BCAN	0.891430112	1.855014047	8.59E-08	0.000101648
BNC1	1.89664337	3.723458737	6.19E-09	1.03E-05
IFITM2	0.280774875	1.214847206	5.60E-06	0.003090545
LINC00458	2.085663796	4.244703539	3.76E-05	0.014166901
LMO7	0.41165634	1.330212137	1.77E-06	0.001223604
LOXL2	0.227624998	1.170905787	0.000309094	0.070432984
NTS	0.759272838	1.692637269	6.43E-07	0.000560608
NUDT4P1	23.37416713	10872400.23	1.24E-11	4.10E-08
PPP2R2C	0.396927289	1.31670056	1.56E-05	0.006991874
PTPN3	0.489625891	1.404080733	9.48E-06	0.005063163
RAB17	0.604572716	1.520528351	0.00012038	0.037622239
SKAP2	0.603008192	1.518880314	1.38E-06	0.001117353
SOCS3	0.446317514	1.362557879	2.00E-06	0.001326402
SPR	0.444246041	1.360602874	1.46E-05	0.006733033

Supplementary table 4. List of upregulated genes in iPSC lines carrying truncating variants compared to WT controls.

References

- [1] A. D. Kline et al., « Diagnosis and management of Cornelia de Lange syndrome: first international consensus statement », *Nat. Rev. Genet.*, vol. 19, no 10, p. 649-666, oct. 2018, doi: 10.1038/s41576-018-0031-0.
- [2] I. D. Krantz et al., « Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of *Drosophila melanogaster* Nipped-B », *Nat. Genet.*, vol. 36, no 6, Art. no 6, juin 2004, doi: 10.1038/ng1364.
- [3] M. A. Deardorff et al., « Mutations in Cohesin Complex Members SMC3 and SMC1A Cause a Mild Variant of Cornelia de Lange Syndrome with Predominant Mental Retardation », *Am. J. Hum. Genet.*, vol. 80, no 3, p. 485-494, mars 2007, doi: 10.1086/511888.
- [4] M. A. Deardorff et al., « RAD21 Mutations Cause a Human Cohesinopathy », *Am. J. Hum. Genet.*, vol. 90, no 6, p. 1014-1027, juin 2012, doi: 10.1016/j.ajhg.2012.04.019.
- [5] M. A. Deardorff et al., « HDAC8 mutations in Cornelia de Lange syndrome affect the cohesin acetylation cycle », *Nature*, vol. 489, no 7415, p. 313-317, sept. 2012, doi: 10.1038/nature11316.
- [6] G. Olley et al., « BRD4 interacts with NIPBL and BRD4 is mutated in a Cornelia de Lange-like syndrome », *Nat. Genet.*, vol. 50, no 3, p. 329-332, mars 2018, doi: 10.1038/s41588-018-0042-y.
- [7] E. Watrin, F. J. Kaiser, et K. S. Wendt, « Gene regulation and chromatin organization: relevance of cohesin mutations to human disease », *Curr. Opin. Genet. Dev.*, vol. 37, p. 59-66, avr. 2016, doi: 10.1016/j.gde.2015.12.004.

- [8] I. Parenti et al., « MAU2 and NIPBL Variants Impair the Heterodimerization of the Cohesin Loader Subunits and Cause Cornelia de Lange Syndrome », *Cell Rep.*, vol. 31, no 7, p. 107647, mai 2020, doi: 10.1016/j.celrep.2020.107647.
- [9] J. Liu et al., « Transcriptional dysregulation in NIPBL and cohesin mutant human cells », *PLoS Biol.*, vol. 7, no 5, p. e1000119, mai 2009, doi: 10.1371/journal.pbio.1000119.
- [10] J. A. Mills et al., « NIPBL^{+/-} haploinsufficiency reveals a constellation of transcriptome disruptions in the pluripotent and cardiac states », *Sci. Rep.*, vol. 8, p. 1056, janv. 2018, doi: 10.1038/s41598-018-19173-9.
- [11] F. D. Weiss et al., « Neuronal genes deregulated in Cornelia de Lange Syndrome respond to removal and re-expression of cohesin », *Nat. Commun.*, vol. 12, no 1, p. 2919, mai 2021, doi: 10.1038/s41467-021-23141-9.
- [12] P. Garcia et al., « Disruption of NIPBL/Scc2 in Cornelia de Lange Syndrome provokes cohesin genome-wide redistribution with an impact in the transcriptome », *Nat. Commun.*, vol. 12, no 1, p. 4551, juill. 2021, doi: 10.1038/s41467-021-24808-z.
- [13] L. Miguel, J. Gervais, G. Nicolas, et M. Lecourtois, « SorLA Protective Function Is Restored by Improving SorLA Protein Maturation in a Subset of Alzheimer's Disease-Associated SORL1 Missense Variants », *J. Alzheimers Dis. JAD*, juill. 2023, doi: 10.3233/JAD-230211.
- [14] K. Cassinari et al., « Haploinsufficiency of the Primary Familial Brain Calcification Gene SLC20A2 Mediated by Disruption of a Regulatory Element », *Mov. Disord. Off. J. Mov. Disord. Soc.*, vol. 35, no 8, p. 1336-1345, août 2020, doi: 10.1002/mds.28090.

- [15] P. A. Ewels et al., « The nf-core framework for community-curated bioinformatics pipelines », *Nat. Biotechnol.*, vol. 38, no 3, p. 276-278, mars 2020, doi: 10.1038/s41587-020-0439-x.
- [16] A. Dobin et al., « STAR: ultrafast universal RNA-seq aligner », *Bioinformatics*, vol. 29, no 1, p. 15-21, janv. 2013, doi: 10.1093/bioinformatics/bts635.
- [17] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, et C. Kingsford, « Salmon provides fast and bias-aware quantification of transcript expression », *Nat. Methods*, vol. 14, no 4, p. 417-419, avr. 2017, doi: 10.1038/nmeth.4197.
- [18] M. I. Love, W. Huber, et S. Anders, « Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 », *Genome Biol.*, vol. 15, no 12, p. 550, 2014, doi: 10.1186/s13059-014-0550-8.
- [19] J. T. Robinson et al., « Integrative Genomics Viewer », *Nat. Biotechnol.*, vol. 29, no 1, p. 24-26, janv. 2011, doi: 10.1038/nbt.1754.
- [20] S. Köhler et al., « The Human Phenotype Ontology in 2021 », *Nucleic Acids Res.*, vol. 49, no D1, p. D1207-D1217, janv. 2021, doi: 10.1093/nar/gkaa1043.
- [21] Q. Wang et al., « Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes », *Nat. Commun.*, vol. 11, no 1, Art. no 1, mai 2020, doi: 10.1038/s41467-019-12438-5.
- [22] M. Rieger et al., « Microdeletions at 19p13.11p12 in five individuals with neurodevelopmental delay », *Eur. J. Med. Genet.*, vol. 66, no 1, p. 104669, janv. 2023, doi: 10.1016/j.ejmg.2022.104669.
- [23] E. Watrin, A. Schleiffer, K. Tanaka, F. Eisenhaber, K. Nasmyth, et J.-M. Peters, « Human Scc4 is required for cohesin binding to chromatin, sister-chromatid cohesion, and

mitotic progression », *Curr. Biol. CB*, vol. 16, no 9, p. 863-874, mai 2006, doi: 10.1016/j.cub.2006.03.049.

[24] D. Braunholz et al., « Isolated NIBPL missense mutations that cause Cornelia de Lange syndrome alter MAU2 interaction », *Eur. J. Hum. Genet. EJHG*, vol. 20, no 3, p. 271-276, mars 2012, doi: 10.1038/ejhg.2011.175.

[25] J. A. Doudna et E. Charpentier, « Genome editing. The new frontier of genome engineering with CRISPR-Cas9 », *Science*, vol. 346, no 6213, p. 1258096, nov. 2014, doi: 10.1126/science.1258096.

[26] A. S. E. Cuomo et al., « Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression », *Nat. Commun.*, vol. 11, no 1, p. 810, févr. 2020, doi: 10.1038/s41467-020-14457-z.

[27] B. B. Cummings et al., « Improving genetic diagnosis in Mendelian disease with transcriptome sequencing », *Sci. Transl. Med.*, vol. 9, no 386, p. eaal5209, avr. 2017, doi: 10.1126/scitranslmed.aal5209.

[28] F. Lecoquierre et al., « High diagnostic potential of short and long read genome sequencing with transcriptome analysis in exome-negative developmental disorders », *Hum. Genet.*, vol. 142, no 6, p. 773-783, juin 2023, doi: 10.1007/s00439-023-02553-1.

[29] S. Endeley et al., « Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes », *Nat. Genet.*, vol. 42, no 11, p. 1021-1026, nov. 2010, doi: 10.1038/ng.677.

[30] Q. Thomas et al., « Haploinsufficiency of ARFGEF1 is associated with developmental delay, intellectual disability, and epilepsy with variable expressivity », *Genet. Med. Off. J.*

Am. Coll. Med. Genet., vol. 23, no 10, p. 1901-1911, oct. 2021, doi: 10.1038/s41436-021-01218-6.

[31] C. C. Hernandez et al., « Altered inhibitory synapses in de novo GABRA5 and GABRA1 mutations associated with early onset epileptic encephalopathies », *Brain J. Neurol.*, vol. 142, no 7, p. 1938-1954, juill. 2019, doi: 10.1093/brain/awz123.

[32] K. M. Butler et al., « De novo variants in GABRA2 and GABRA5 alter receptor function and contribute to early-onset epilepsy », *Brain J. Neurol.*, vol. 141, no 8, p. 2392-2405, août 2018, doi: 10.1093/brain/awy171.

[33] M. Roifman et al., « De novo WNT5A-associated autosomal dominant Robinow syndrome suggests specificity of genotype and phenotype », *Clin. Genet.*, vol. 87, no 1, p. 34-41, 2015, doi: 10.1111/cge.12401.

[34] M. A. Patton et A. R. Afzal, « Robinow syndrome », *J. Med. Genet.*, vol. 39, no 5, p. 305-310, mai 2002, doi: 10.1136/jmg.39.5.305.

[35] A. Abu-Ghname et al., « Extremity anomalies associated with Robinow syndrome », *Am. J. Med. Genet. A.*, vol. 185, no 12, p. 3584-3592, déc. 2021, doi: 10.1002/ajmg.a.61884.

[36] P. J. Bray et R. G. H. Cotton, « Variations of the human glucocorticoid receptor gene (NR3C1): pathological and in vitro mutations and polymorphisms », *Hum. Mutat.*, vol. 21, no 6, p. 557-568, juin 2003, doi: 10.1002/humu.10213.

[37] D. Saleh, S. N. S. Yarrarapu, et C. Cook, « Hypertrichosis », in *StatPearls, Treasure Island (FL): StatPearls Publishing, 2023. Consulté le: 20 août 2023. [En ligne]. Disponible sur: <http://www.ncbi.nlm.nih.gov/books/NBK534854/>*

----- FIN DU MANUSCRIT -----

Discussion

Ce travail de thèse a été consacré à évaluer des contextes où l'utilisation de techniques basées sur l'ARN a pu renforcer les conclusions issues des analyses pangénomiques. Dans la première partie, nous avons employé des méthodes, qu'elles soient globales ou ciblées, en association ou non avec des modèles cellulaires, afin de caractériser une variation génomique, nucléotidique ou structurale affectant un gène préalablement identifié. La seconde partie devait, à l'origine, être étroitement liée à une application clinique, visant à développer un test diagnostique ciblé, réalisable à partir d'échantillons sanguins des patients, pour confirmer le diagnostic du syndrome de Cornelia de Lange. Toutefois, cet objectif s'est avéré plus complexe que prévu, rendant nécessaire l'utilisation de modèles cellulaires pour obtenir les résultats escomptés à partir des prélèvements sanguins et ne permettant pas une translation vers le diagnostic à ce jour. Malgré ces défis, cette recherche a enrichi notre compréhension fondamentale du syndrome.

L'un des principaux défis dans l'étude de la régulation de l'expression génique et de son application en génétique médicale est la nécessité de parfaitement connaître les mécanismes de pathogénicité pour les différents gènes. L'exemple des calcifications cérébrales primaires est à cet égard très éloquent. Le gène *SLC20A2*, qui fait l'objet de l'un des travaux de cette thèse, présente une pathogénicité due à l'haploinsuffisance, c'est-à-dire une situation où le produit d'un seul allèle, bien qu'actif, ne suffit pas à assurer son activité normale pour résulter en un phénotype normal. Une fois ce mécanisme identifié, il a suffi de démontrer une perte d'expression en accord avec l'inactivité de l'un des deux allèles pour avancer un argument solide en faveur de la pathogénicité du remaniement structural. Tout comme *SLC20A2*, *XPR1* code pour un transporteur de phosphate, mais cette fois-ci en tant qu'exportateur. Toutes les variations pathogènes du gène *XPR1* sont de type faux-sens [156], et toutes les variations faux-sens identifiées pour ce gène résultent en une perte de fonction, compromettant la

fonction d'export du phosphate, que ce soit par la dégradation de la protéine, un défaut d'adressage à la membrane, ou bien un défaut de la fonction d'export. Néanmoins, pour ce gène, aucune variation non-sens, d'épissage, de délétion intragénique ou de remaniement de taille supérieure n'est rapportée. On peut ainsi faire l'hypothèse que le mécanisme de pathogénicité soit plus complexe que l'haploinsuffisance seule et que des mécanismes régulateurs pourraient y pallier. Une observation récente d'une délétion partielle de *XPRI*, détectée fortuitement par CGH-array chez un patient atteint du syndrome de Kleefstra [183], nous renforce dans cette hypothèse. En effet, cet individu ne présentait aucun signe de calcifications cérébrales au scanner (cas clinique CHU de Rouen, non publié), alors que la pénétrance radiologique de cette affection est attendue comme complète. De plus, il est particulièrement difficile voire impossible de générer des modèles de surexpression stable d'*XPRI*. Si la transfection transitoire de plasmides contenant la séquence codante de *XPRI* permet en effet d'obtenir une expression augmentée transitoirement, les tentatives de nos collègues spécialistes de ce gène à l'IGMM de Montpellier (équipe du Docteur Jean Luc Battini) ont observé des niveaux d'expression tendant vers celui des cellules n'ayant pas intégré le plasmide. Ces observations posent la question de la régulation de l'expression de certains gènes par leurs propres produits. Un exemple bien connu dans le domaine des maladies neurodégénératives est la régulation de la protéine TDP-43. TDP-43 est un facteur essentiel de liaison à l'ARN associé au métabolisme de l'ARN. Dans l'état physiologique, le maintien des niveaux normaux de protéine TDP-43 est crucial pour les fonctions physiologiques appropriées des cellules. Ainsi, l'expression de TDP-43 est étroitement régulée par une boucle de rétroaction négative autoregulatrice. Les caractéristiques clés de ces processus autoregulateurs de TDP-43 impliquent des événements d'épissage alternatifs, l'utilisation différentielle des sites de polyadénylation, la rétention nucléaire de l'ARN messager et une diminution des niveaux d'ARN messager à l'état stable [184],[185]. Des mécanismes similaires mériteraient d'être recherchés pour *XPRI*.

De manière similaire à *XPR1*, les variations pathogènes de *PDGFRB* entraînent des calcifications cérébrales lorsqu'elles sont responsable d'une perte de fonction de la protéine [150] alors que les mutations activatrices, somatiques ou constitutionnelles, sont responsables de plusieurs conditions pathologiques, comme la myofibromatose infantile ou le syndrome de Penttinen, par exemple. Néanmoins, seules des variants faux-sens ont été rapportées dans les calcifications cérébrales primaires, dont l'effet de perte de fonction a été démontré, à travers une absence de capacité de transduction du signal du récepteur transmembranaire PDGFR β , une mauvaise liaison ligand-récepteur, ou une instabilité de la protéine [186]. L'absence de variation tronquante ou de délétion pose également la question sur le mécanisme d'haploinsuffisance. A ce jour, il n'y a aucune preuve qu'un variant de type haploinsuffisance de *PDGFRB* puisse causer des calcifications cérébrales, alors que la pLI de ce gène est de 0,9. Ici, la question du mécanisme reste ouverte : régulation de l'expression génique par compensation d'expression de l'allèle sauvage, effet limite lié à la dimérisation (auto ou hétérodimérisation) ? Dans ces deux cas (*XPR1* et *PDGFRB*), il semble que le mécanisme pathogène nécessite la présence d'une traduction normale et que l'effet perte de fonction s'exerce après la traduction. Il sera certainement intéressant d'étudier les niveaux de transcrits de *PDGFRB* et d'*XPR1* chez des patients porteurs de variations tronquantes, et nous avons un projet dans le laboratoire visant à introduire des variants pathogènes et tronquants de ces deux gènes (et de *SLC20A2*) afin de mieux en comprendre les mécanismes. Ces observations doivent également inciter à la prudence et ne pas faire un amalgame entre perte de fonction et haploinsuffisance, qui sont parfois considérés comme équivalents, à tort. Les conséquences pour l'étude des conséquences des variants identifiés chez les patients sont importantes, car elles détermineront si l'ARN est une bonne cible d'étude ou non, que ce soit dans des prélèvements de patients ou des modèles cellulaires.

Deux des travaux de la première partie de cette thèse se concentrent sur l'étude de CNV : une triplification d'*APP* et une délétion non codante de *SLC20A2*. Cela nous conduit à examiner,

de manière plus approfondie, la contribution des analyses de l'ARN dans l'interprétation des variations de structure. Effectivement, même si les délétions codantes sont généralement simples à interpréter, pouvant souvent être assimilées à des variations non-sens, l'impact des duplications totales, des duplications partielles, ainsi que des SV non codants ou des SV équilibrés reste souvent plus complexe à déterminer.

Concernant les duplications géniques complètes, une simplification consisterait à dire qu'une troisième copie (apparemment) fonctionnelle d'un gène augmenterait nécessairement son expression de 1,5 fois. Toutefois, la duplication d'un gène, même complète, ne garantit pas nécessairement une augmentation de son expression, en particulier si les éléments régulateurs ne sont pas également dupliqués. La validation de cette hypothèse nécessite donc une analyse de l'ARN. Sur le plan méthodologique, et lorsque qu'elle est réalisable, une analyse allélique spécifique demeure la méthode privilégiée. C'est, par exemple, ce qui a été réalisé au sein de notre équipe pour des travaux portant sur l'étude des duplications du gène *MAPT*, qui sont associées à des démences chez les sujets jeunes se différenciant de la maladie d'Alzheimer par l'absence de dépôts β -amyloïdes. Ces duplications englobent l'entièreté du gène, au sein d'une microduplication récurrente 17q21.31 de 900 Kb, duplication en miroir de la délétion récurrente causant le syndrome de Koolen de Vries [187]. Pour démontrer la fonctionnalité de la troisième copie du gène, des analyses d'ARNm ont été effectuées, révélant une augmentation des niveaux d'ARNm du gène entre 1,6 et 1,9, attestant de la fonctionnalité des trois copies. Ces résultats, obtenus par Snapshot, soutiennent l'hypothèse d'une pathogénicité liée à une augmentation du dosage génique de *MAPT* et d'une possible augmentation des niveaux de la protéine Tau qu'il code [188] ce qui a plus tard été confirmé sur des lignées cellulaires neuronales différenciées à partir de cellules iPSC [189]. La technique de Snapshot est dérivée du séquençage Sanger, et permet de quantifier spécifiquement l'ARNm de chaque allèle d'un gène grâce à la quantification d'un SNV présent dans la séquence codante [190]. La RT-ddPCR peut tout à fait être utilisée dans ce type de situation, avec un protocole simple

consistant en une unique PCR encadrant la variation nucléotidique et deux sondes d'hydrolyse aux fluorochromes différents, l'une spécifique de la séquence sauvage et l'autre spécifique de la séquence mutante, conduisant à une quantification des deux allèles avec une haute précision [191]. Cette approche est d'autant plus facilitée que le design est hautement automatisable et le coût des réactifs, notamment des sondes d'hydrolyse, est aujourd'hui abordable. Cependant, cette méthode, si elle est à favoriser, nécessite la présence d'un SNV exonique dans le gène d'intérêt ce qui n'est pas tout le temps le cas. C'est pourquoi, dans les deux situations évoquées dans cette thèse (délétion en amont de *SLC202A2* et triplication d'*APP*), une quantification relative des niveaux totaux d'ARNm du gène par RT-ddPCR, par rapport à un gène de référence, a du être réalisée.

Bien que cette stratégie soit techniquement robuste, en témoigne la reproductibilité des réplicats techniques, elle souffre de quelques limites qui sont à considérer lors de l'interprétation des résultats. En effet, même si les gènes de référence sont largement validés [192]–[194] leur expression peut tout de même varier d'un individu à l'autre, notamment dans un échantillon aussi hétérogène sur le plan cellulaire que le sang. De même, l'expression du gène cible peut connaître également des variations physiologiques inter-individuelles. L'examen attentif des informations présentes dans les bases de données, principalement GTEX, est donc une étape cruciale pour évaluer cette variation inter-individuelle. Dans le cas d'une étude comparant deux cohortes, la comparaison de l'expression de l'ARNm entre témoins et patients permet d'éclairer les différences, à condition que les conditions de prélèvement et d'analyse soient similaires. Toutefois, comparer un cas isolé à une moyenne des témoins est plus complexe et ne permet pas une interprétation statistique robuste, d'où la prudence adoptée pour interpréter les résultats de la quantification d'ARNm d'*APP* chez le patient. À l'heure actuelle, il demeure ainsi complexe d'utiliser la RT-ddPCR en quantification relative, à partir d'un prélèvement PaxGene, pour mettre en lumière la variation des niveaux d'ARNm d'un gène en présence d'une délétion ou d'une duplication. Cela nécessite de

s'assurer de la faible variabilité de l'expression du gène chez des sujets témoins et des contrôles atteints, afin d'obtenir des valeurs de référence qui doivent être suffisamment distinctes les unes des autres pour garantir la sensibilité et la spécificité du résultat.

Le gène *APP* demeure tout de même un bon candidat pour ce type d'analyse. En effet, outre notre travail sur la triplication, une caractérisation d'un variant 3'-UTR de ce gène avait déjà été réalisée au sein de l'équipe, montrant que ce variant était à l'origine d'une augmentation des niveaux d'ARNm d'*APP* dans le sang du patient, dans un rapport de 1,47, avec une confirmation sur un modèle cellulaire [103]. Ces exemples illustrent l'intérêt de la stratégie consistant à s'assurer que le mécanisme pathologique passe par une augmentation de l'expression, puis à vérifier qu'un prélèvement sanguin est adapté à l'analyse. Dans ces situations, l'étude de l'ARNm est un outil précieux pour la caractérisation ciblée des VSI, en particulier ceux non codants.

L'adoption accrue du RNAseq pourrait permettre de surmonter certaines limitations. En recourant à une mesure standardisée, comme le TPM, et en se basant sur les progrès technologiques, notamment la capacité d'effectuer des captures d'exomes qui augmentent le nombre de lectures des gènes cibles, il est possible d'améliorer la comparaison des variations du nombre de lectures pour un gène donné. Lorsqu'elles sont combinées avec d'importantes bases de données témoins, ces méthodes faciliteront la mise en évidence d'une augmentation de l'expression due à une copie fonctionnelle supplémentaire, ou d'une diminution causée par le mécanisme de NMD ou une baisse de transcription, par exemple. Les cohortes qui ont été parmi les premières à déployer le RNAseq comme outil diagnostique pour les maladies mendéliennes [195] illustrent cette tendance. Elles permettent la détection de variations significatives du nombre de lectures pour un gène donné à l'aide d'outils tels qu'OUTRIDER [196], ainsi que l'identification de patterns d'épissage aberrants ou d'expression mono-

allélique. Nos premiers résultats de RNAseq par captures, réalisés à Rouen, ont montré une capacité accrue de mise en évidence des variations du nombre de lectures des gènes dans des situations pathologiques [120].

Cette logique peut également s'appliquer aux duplications partielles (ou intragéniques) qui demeurent à ce jour des variations structurales dont l'interprétation est complexe. Il a été montré que les gènes les moins sujets aux duplications partielles chez les patients de la base gnomAD sont aussi souvent des gènes sensibles à l'haploinsuffisance pour les variations nucléotidiques [197]. Ce résultat témoigne du fait que ces variations peuvent souvent altérer l'expression du gène et entraîner une perte de fonction. Cependant, malgré ces observations et les informations que peuvent apporter leur caractère *de novo* ou hérité, ces CNV sont fréquemment classés comme VSI à l'issue des explorations cytogénétiques traditionnelles. À cet égard, l'association du RNAseq aux techniques d'étude de longues molécules, que ce soit dans le cadre du séquençage de troisième génération (*e.g.* technologie PacBio) ou de la cartographie optique (*e.g.* Bionano Genomics), permettra une meilleure caractérisation structurale et une évaluation plus précise de leur impact sur l'expression du gène.

Réflexions sur l'implantation clinique des analyses transcriptomiques

Le RNAseq se présente donc comme un des outils incontournables de la stratégie diagnostique des maladies génétiques, et ce particulièrement dans le champ des DI/AD. Outre l'intérêt de visualisation des variations des niveaux d'expression des gènes, les outils pour identifier les épissages aberrants se stabilisent, de même que ceux utilisés pour mettre en évidence des expressions mono-alléliques [198]. Cette transversalité des outils semble essentielle, car permettant de couvrir différents mécanismes de pathologie, mais également d'élucider des situations plus complexes ou incomplètes. Ainsi, l'exemple des variations introniques rapportées dans l'article relatif à la cohorte de patients CdLS étudiés en WGS est

assez éloquent. Dans ces situations, les variations, *de novo*, induisent un épissage aberrant conduisant à des néo-exons, engendrant un décalage du cadre de lecture et l'apparition d'un codon stop prématuré. Or, dans ces situations, le NMD semblait très partiel, malgré la présence d'une protéine tronquée (faible diminution du nombre de TPM de *NIPBL*, et également présence d'un SNP hétérozygote visualisé sur le transcriptome de l'un des deux patients). Sur le plan biologique, l'hypothèse était ici que ces variations introniques profondes de *NIPBL* entraînent un défaut d'épissage sur une proportion des transcrits les portant, seulement, et le NMD ne concerne qu'une proportion de cette dernière. Sur l'aspect technique, cela montre à quel point la superposition des approches est nécessaire pour tirer pleinement bénéfice du RNAseq car, à elle seule, une analyse de quantité de transcrit n'aurait pas permis d'élucider les conséquences de ces variations.

La question suivante, qui fait déjà l'objet de discussions, sera de définir le meilleur positionnement du RNAseq dans la stratégie diagnostique. En effet, actuellement, la tendance est plutôt à l'utilisation de cette technologie dans un second temps, après les analyses de séquençage d'exome ou de génome, et, bien souvent, dans le but de pouvoir reclasser des VSI, avec donc une connaissance préalable du gène à visualiser. Le potentiel du RNAseq à ce sujet est réel puisque, au-delà des exemples développés ici, il permettrait, par exemple, de reclasser pathogène ou bénin, jusqu'à 31% des SNV classés VSI après première analyse, car pouvant potentiellement affecter l'épissage. Cependant, outre son utilisation comme un outil ciblé pour caractériser un VSI, c'est également de l'analyse du transcriptome entier dont il est question. Dans ce contexte, il est estimé que le RNAseq augmente le rendement diagnostique de 8 à 36% par rapport à l'exome seul [177]–[179],[195],[199]–[201]. Elle permet de déterminer des défauts d'expression (quantitative et/ou monoallélique) et d'épissage sans avoir identifié de variation candidate lors de la lecture de l'exome et ainsi de guider la (re)lecture du génome. Cette augmentation du rendement diagnostique pose donc la question suivante : dans le cas du diagnostic des maladies rares, notamment celles du développement, est-

il plus pertinent de proposer analyse concomitante avec les analyses d'exome ou de génome, ou d'opter pour l'analyse en seconde intention ? Les arguments principaux en faveur d'une analyse simultanée seraient bien sûr à la fois une possibilité d'analyse intégrée des données, avec la possibilité de pouvoir d'emblée confirmer ou infirmer l'effet d'un variant sur l'expression d'un gène et d'augmenter le rendement diagnostique. Cependant, des obstacles demeurent à cette stratégie, d'une part le surcoût engendré par la réalisation d'une analyse qui ne se révèle pas toujours nécessaire (en cas de variant classifiable d'emblée), mais aussi en termes de tissu d'étude. En effet, si les analyses d'exome et de génome en situation postnatale sont la plupart du temps réalisées à partir d'un prélèvement sanguin, la plus-value du RNAseq semble quant à elle plus grande si l'analyse est réalisée sur culture de fibroblastes cutanés, du fait d'un nombre de gènes exprimés bien plus important [195]. La nécessité de réalisation de ce type de prélèvements, plus invasifs qu'un prélèvement sanguin, et de la réalisation d'une culture cellulaire, sont un frein à une systématisation de la réalisation du RNAseq en première intention pour l'ensemble des situations d'exploration des causes génétiques de DI/AD. Cependant, cette stratégie pourrait être envisageable dans des situations particulières, associées à une certaine urgence. Par exemple, cette stratégie pourrait avoir un intérêt dans des situations prénatales, avec une analyse d'exome et un transcriptome réalisée à partir de culture de liquide amniotique. Cette stratégie permettrait d'avoir une amélioration du rendement et du temps diagnostiques, tout en permettant de caractériser d'emblée les variations identifiées dans l'exome au sein de gènes connus, et ce d'autant plus que le niveau d'expression de gènes dans le liquide amniotique cultivé est sensiblement le même que dans des fibroblastes cultivés [202].

De manière générale, cette thèse souligne l'importance croissante de disposer d'infrastructures dédiées à la culture cellulaire au sein des laboratoires diagnostiques de génétique. Comme mentionné précédemment, la disponibilité de cultures cellulaires, qu'il s'agisse de lignées lymphoblastoïdes, fibroblastiques provenant de patients ou de cellules éditées comme les

iPSC ou d'autres lignées, enrichit considérablement la qualité des analyses de l'ARN. Ces cultures offrent aussi une source potentielle d'ADN de grande taille, essentielle pour le séquençage de longs fragments ou la cartographie optique. L'exemple de la caractérisation de la délétion non codante de *SLC20A2*, associé au modèle cellulaire créé pour fournir une preuve supplémentaire de pathogénéicité du variant, est un exemple de processus qui pourrait, à l'avenir, être utilisé dans le cadre du diagnostic. En effet, la démocratisation des approches de CRISPR/Cas9, associée à des simplifications de processus, telles que l'analyse en *bulk* cellulaire proposé ici, laissent entrevoir la possibilité de disposer de ces tests de complexité intermédiaire, dans nos laboratoires.

Dans un contexte d'évolution technologique et de convergence physique, organisationnelle et scientifique entre la génétique moléculaire et la cytogénétique, il est primordial de garantir que les compétences et infrastructures en matière de culture cellulaire, acquises par les laboratoires de cytogénétique au cours des six dernières décennies, soient non seulement préservées mais également adaptées à ces nouvelles orientations.

Enfin, la question des signatures est l'axe final de ce travail. Comme cela a été rappelé, les signatures ont l'intérêt d'être utiles en tant que biomarqueurs pour aider à reclasser des variants de signification incertaine ou pour confirmer un syndrome clinique sans bases moléculaires connues. La définition d'une signature transcriptomique ou épigénétique passe généralement par une analyse différentielle visant à comparer les profils d'expression ou de méthylation d'individus présentant le syndrome d'intérêt, confirmé sur des bases moléculaires, afin de définir des paramètres variant entre les deux conditions. Ces paramètres correspondent à des gènes sous-exprimés ou sur-exprimés dans le cadre des analyses transcriptomiques telles que celles que nous avons réalisées, ou à des localisations différentiellement méthylées lorsqu'il s'agit de signatures épigénétiques. Ainsi, il reste à établir la plus-value que peut avoir le développement d'une signature transcriptomique dans un contexte où des signatures de

méthylation existent déjà. Plusieurs signatures épigénétiques ont déjà été publiées pour un certain nombre de maladies mendéliennes du développement, dont le CdLS [203], [204]. Ces signatures n'ont, pour l'instant, pas fait l'objet d'évaluations indépendantes, et un travail réalisé au sein de notre laboratoire montre que la diversité des méthodes peut entraîner des sensibilités différentes [205]. Par ailleurs, concernant un syndrome comme le CdLS, les échantillons sont analysés ensemble, et il est difficile de prédire si cette signature sera robuste d'un gène à l'autre et face à tous les mécanismes de pathogénicité, ni de savoir s'il y aura une corrélation entre signature et phénotype. Ainsi, le CdLS est un bon exemple de pathologie pour laquelle continuer à travailler à partir de l'ARNm, malgré les contraintes présentées, demeure pertinent. En outre, la disponibilité dans notre équipe des données transcriptomiques obtenues lors de ce travail, et de celles générées dans le cadre des signatures épigénétiques, y compris pour les patients porteurs des variations introduites dans les iPSC, nous permet d'envisager la comparaison de ces résultats, voire peut-être même, à l'avenir, de les intégrer dans un outil multi-omique commun.

Conclusion

La régulation de l'expression génique est un phénomène complexe et notre connaissance est encore largement incomplète. Les choix méthodologiques et thématiques de cette thèse ont été largement inspirés par les interrogations récurrentes dans la pratique quotidienne de la (cyto)génétique biologique. Cette démarche, associée au choix de s'intéresser à diverses pathologies génétiques, a permis d'appréhender différents outils d'analyse de l'ARN et de proposer plusieurs stratégies potentiellement reproductibles dans le contexte diagnostique, tout en contribuant à faire progresser notre connaissance moléculaire et cellulaire de ces maladies. Cependant, au-delà d'apporter des solutions nouvelles, ces travaux ont également permis de faire le bilan des points de blocage qui limitent notre capacité actuelle à exploiter le plein potentiel des analyses basées sur l'ARNm. Ces limites devraient être prises en compte dans les évolutions futures de notre discipline, que ce soit dans les réflexions sur son organisation ou sur les éléments de développement technologiques à fournir pour y remédier.

Références

- [1] E. R. Mardis, « The impact of next-generation sequencing technology on genetics », *Trends Genet*, vol. 24, n° 3, p. 133-141, mars 2008, doi: 10.1016/j.tig.2007.12.007.
- [2] S. B. Ng *et al.*, « Targeted capture and massively parallel sequencing of 12 human exomes », *Nature*, vol. 461, n° 7261, p. 272-276, sept. 2009, doi: 10.1038/nature08250.
- [3] M. L. Metzker, « Sequencing technologies - the next generation », *Nat Rev Genet*, vol. 11, n° 1, p. 31-46, janv. 2010, doi: 10.1038/nrg2626.
- [4] A. Rauch *et al.*, « Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study », *Lancet*, vol. 380, n° 9854, p. 1674-1682, nov. 2012, doi: 10.1016/S0140-6736(12)61480-9.
- [5] Z. Stark *et al.*, « Scaling national and international improvement in virtual gene panel curation via a collaborative approach to discordance resolution », *The American Journal of Human Genetics*, vol. 108, n° 9, p. 1551-1557, sept. 2021, doi: 10.1016/j.ajhg.2021.06.020.
- [6] K. M. Boycott *et al.*, « International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases », *Am J Hum Genet*, vol. 100, n° 5, p. 695-705, mai 2017, doi: 10.1016/j.ajhg.2017.04.003.
- [7] C. Gilissen *et al.*, « Genome sequencing identifies major causes of severe intellectual disability », *Nature*, vol. 511, n° 7509, p. 344-347, juill. 2014, doi: 10.1038/nature13394.
- [8] 100,000 Genomes Project Pilot Investigators *et al.*, « 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report », *N Engl J Med*, vol. 385, n° 20, p. 1868-1880, nov. 2021, doi: 10.1056/NEJMoa2035790.
- [9] S. Richards *et al.*, « Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology », *Genet Med*, vol. 17, n° 5, p. 405-424, mai 2015, doi: 10.1038/gim.2015.30.
- [10] D. G. MacArthur *et al.*, « Guidelines for investigating causality of sequence variants in human disease », *Nature*, vol. 508, n° 7497, p. 469-476, avr. 2014, doi: 10.1038/nature13127.
- [11] E. R. Riggs *et al.*, « Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical

- Genome Resource (ClinGen) », *Genet Med*, vol. 22, n° 2, p. 245-257, févr. 2020, doi: 10.1038/s41436-019-0686-8.
- [12] L. D. Ward et M. Kellis, « Interpreting noncoding genetic variation in complex traits and human disease », *Nat Biotechnol*, vol. 30, n° 11, p. 1095-1106, nov. 2012, doi: 10.1038/nbt.2422.
- [13] L. G. Biesecker, R. L. Nussbaum, et H. L. Rehm, « Distinguishing Variant Pathogenicity From Genetic Diagnosis: How to Know Whether a Variant Causes a Condition », *JAMA*, vol. 320, n° 18, p. 1929-1930, nov. 2018, doi: 10.1001/jama.2018.14900.
- [14] Y. Hasin, M. Seldin, et A. Lusic, « Multi-omics approaches to disease », *Genome Biology*, vol. 18, n° 1, p. 83, mai 2017, doi: 10.1186/s13059-017-1215-1.
- [15] Z. Wang, M. Gerstein, et M. Snyder, « RNA-Seq: a revolutionary tool for transcriptomics », *Nat Rev Genet*, vol. 10, n° 1, p. 57-63, janv. 2009, doi: 10.1038/nrg2484.
- [16] P. J. Park, « ChIP-seq: advantages and challenges of a maturing technology », *Nat Rev Genet*, vol. 10, n° 10, p. 669-680, oct. 2009, doi: 10.1038/nrg2641.
- [17] M. Spivakov, « Spurious transcription factor binding: non-functional or genetically redundant? », *Bioessays*, vol. 36, n° 8, p. 798-806, août 2014, doi: 10.1002/bies.201400036.
- [18] J. A. Doudna et E. Charpentier, « Genome editing. The new frontier of genome engineering with CRISPR-Cas9 », *Science*, vol. 346, n° 6213, p. 1258096, nov. 2014, doi: 10.1126/science.1258096.
- [19] M. Kellis *et al.*, « Defining functional DNA elements in the human genome », *Proc Natl Acad Sci U S A*, vol. 111, n° 17, p. 6131-6138, avr. 2014, doi: 10.1073/pnas.1318948111.
- [20] J. S. Mattick et I. V. Makunin, « Non-coding RNA », *Hum Mol Genet*, vol. 15 Spec No 1, p. R17-29, avr. 2006, doi: 10.1093/hmg/ddl046.
- [21] M. Levine et R. Tjian, « Transcription regulation and animal diversity », *Nature*, vol. 424, n° 6945, p. 147-151, juill. 2003, doi: 10.1038/nature01763.
- [22] F. Jacob et J. Monod, « Genetic regulatory mechanisms in the synthesis of proteins », *J Mol Biol*, vol. 3, p. 318-356, juin 1961, doi: 10.1016/s0022-2836(61)80072-7.
- [23] F. H. Crick, « Codon--anticodon pairing: the wobble hypothesis », *J Mol Biol*, vol. 19, n° 2, p. 548-555, août 1966, doi: 10.1016/s0022-2836(66)80022-0.

- [24] T. I. Lee et R. A. Young, « Transcriptional regulation and its misregulation in disease », *Cell*, vol. 152, n° 6, p. 1237-1251, mars 2013, doi: 10.1016/j.cell.2013.02.014.
- [25] D. L. Bentley, « Coupling mRNA processing with transcription in time and space », *Nat Rev Genet*, vol. 15, n° 3, p. 163-175, mars 2014, doi: 10.1038/nrg3662.
- [26] M. J. Moore et N. J. Proudfoot, « Pre-mRNA processing reaches back to transcription and ahead to translation », *Cell*, vol. 136, n° 4, p. 688-700, févr. 2009, doi: 10.1016/j.cell.2009.02.001.
- [27] F. Crick, « Central dogma of molecular biology », *Nature*, vol. 227, n° 5258, p. 561-563, août 1970, doi: 10.1038/227561a0.
- [28] D. P. Bartel, « MicroRNAs: target recognition and regulatory functions », *Cell*, vol. 136, n° 2, p. 215-233, janv. 2009, doi: 10.1016/j.cell.2009.01.002.
- [29] N. Sonenberg et A. G. Hinnebusch, « Regulation of translation initiation in eukaryotes: mechanisms and biological targets », *Cell*, vol. 136, n° 4, p. 731-745, févr. 2009, doi: 10.1016/j.cell.2009.01.042.
- [30] K. M. Lelli, M. Slattery, et R. S. Mann, « Disentangling the many layers of eukaryotic transcriptional regulation », *Annu Rev Genet*, vol. 46, p. 43-68, 2012, doi: 10.1146/annurev-genet-110711-155437.
- [31] F. Spitz et E. E. M. Furlong, « Transcription factors: from enhancer binding to developmental control », *Nat Rev Genet*, vol. 13, n° 9, p. 613-626, sept. 2012, doi: 10.1038/nrg3207.
- [32] N. D. Heintzman *et al.*, « Histone modifications at human enhancers reflect global cell-type-specific gene expression », *Nature*, vol. 459, n° 7243, p. 108-112, mai 2009, doi: 10.1038/nature07829.
- [33] M. Bulger et M. Groudine, « Functional and mechanistic diversity of distal transcription enhancers », *Cell*, vol. 144, n° 3, p. 327-339, févr. 2011, doi: 10.1016/j.cell.2011.01.024.
- [34] D. Shlyueva, G. Stampfel, et A. Stark, « Transcriptional enhancers: from properties to genome-wide predictions », *Nat Rev Genet*, vol. 15, n° 4, p. 272-286, avr. 2014, doi: 10.1038/nrg3682.
- [35] J. B. Zaugg *et al.*, « Current challenges in understanding the role of enhancers in disease », *Nat Struct Mol Biol*, vol. 29, n° 12, p. 1148-1158, déc. 2022, doi: 10.1038/s41594-022-00896-3.

- [36] S. T. Smale et J. T. Kadonaga, « The RNA polymerase II core promoter », *Annu Rev Biochem*, vol. 72, p. 449-479, 2003, doi: 10.1146/annurev.biochem.72.121801.161520.
- [37] D. S. Latchman, « Transcription factors: an overview », *Int J Biochem Cell Biol*, vol. 29, n° 12, p. 1305-1312, déc. 1997, doi: 10.1016/s1357-2725(97)00085-x.
- [38] C. Benoist, K. O'Hare, R. Breathnach, et P. Chambon, « The ovalbumin gene-sequence of putative control regions », *Nucleic Acids Res*, vol. 8, n° 1, p. 127-142, janv. 1980, doi: 10.1093/nar/8.1.127.
- [39] R. Javahery, A. Khachi, K. Lo, B. Zenzie-Gregory, et S. T. Smale, « DNA sequence requirements for transcriptional initiator activity in mammalian cells », *Mol Cell Biol*, vol. 14, n° 1, p. 116-127, janv. 1994, doi: 10.1128/mcb.14.1.116-127.1994.
- [40] P. J. Mitchell et R. Tjian, « Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins », *Science*, vol. 245, n° 4916, p. 371-378, juill. 1989, doi: 10.1126/science.2667136.
- [41] S. L. Berger, « The complex language of chromatin regulation during transcription », *Nature*, vol. 447, n° 7143, p. 407-412, mai 2007, doi: 10.1038/nature05915.
- [42] A. H. Swirnoff et J. Milbrandt, « DNA-binding specificity of NGFI-A and related zinc finger transcription factors », *Mol Cell Biol*, vol. 15, n° 4, p. 2275-2287, avr. 1995, doi: 10.1128/MCB.15.4.2275.
- [43] J. J. Quinn et H. Y. Chang, « Unique features of long non-coding RNA biogenesis and function », *Nat Rev Genet*, vol. 17, n° 1, Art. n° 1, janv. 2016, doi: 10.1038/nrg.2015.10.
- [44] G. A. Maston, S. K. Evans, et M. R. Green, « Transcriptional Regulatory Elements in the Human Genome », *Annual Review of Genomics and Human Genetics*, vol. 7, n° 1, p. 29-59, 2006, doi: 10.1146/annurev.genom.7.080505.115623.
- [45] R. D. Kornberg et Y. Lorch, « Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome », *Cell*, vol. 98, n° 3, p. 285-294, août 1999, doi: 10.1016/s0092-8674(00)81958-3.
- [46] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, et T. J. Richmond, « Crystal structure of the nucleosome core particle at 2.8 Å resolution », *Nature*, vol. 389, n° 6648, p. 251-260, sept. 1997, doi: 10.1038/38444.
- [47] N. Happel et D. Doenecke, « Histone H1 and its isoforms: contribution to chromatin structure and function », *Gene*, vol. 431, n° 1-2, p. 1-12, févr. 2009, doi: 10.1016/j.gene.2008.11.003.

- [48] T. Kouzarides, « Chromatin modifications and their function », *Cell*, vol. 128, n° 4, p. 693-705, févr. 2007, doi: 10.1016/j.cell.2007.02.005.
- [49] C. R. Clapier et B. R. Cairns, « The biology of chromatin remodeling complexes », *Annu Rev Biochem*, vol. 78, p. 273-304, 2009, doi: 10.1146/annurev.biochem.77.062706.153223.
- [50] T. Cremer et C. Cremer, « Chromosome territories, nuclear architecture and gene regulation in mammalian cells », *Nat Rev Genet*, vol. 2, n° 4, p. 292-301, avr. 2001, doi: 10.1038/35066075.
- [51] T. Misteli, « Beyond the sequence: cellular organization of genome function », *Cell*, vol. 128, n° 4, p. 787-800, févr. 2007, doi: 10.1016/j.cell.2007.01.028.
- [52] J. Dekker et L. Mirny, « The 3D Genome as Moderator of Chromosomal Communication », *Cell*, vol. 164, n° 6, p. 1110-1121, mars 2016, doi: 10.1016/j.cell.2016.02.007.
- [53] A. Bolzer *et al.*, « Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes », *PLoS Biol*, vol. 3, n° 5, p. e157, mai 2005, doi: 10.1371/journal.pbio.0030157.
- [54] W. A. Bickmore et B. van Steensel, « Genome architecture: domain organization of interphase chromosomes », *Cell*, vol. 152, n° 6, p. 1270-1284, mars 2013, doi: 10.1016/j.cell.2013.02.001.
- [55] S. Wang *et al.*, « Spatial organization of chromatin domains and compartments in single chromosomes », *Science*, vol. 353, n° 6299, p. 598-602, août 2016, doi: 10.1126/science.aaf8084.
- [56] J. R. Dixon *et al.*, « Topological domains in mammalian genomes identified by analysis of chromatin interactions », *Nature*, vol. 485, n° 7398, p. 376-380, avr. 2012, doi: 10.1038/nature11082.
- [57] J. H. Gibcus et J. Dekker, « The hierarchy of the 3D genome », *Mol Cell*, vol. 49, n° 5, p. 773-782, mars 2013, doi: 10.1016/j.molcel.2013.02.011.
- [58] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, et L. A. Mirny, « Formation of Chromosomal Domains by Loop Extrusion », *Cell Rep*, vol. 15, n° 9, p. 2038-2049, mai 2016, doi: 10.1016/j.celrep.2016.04.085.
- [59] A.-L. Valton et J. Dekker, « TAD disruption as oncogenic driver », *Curr Opin Genet Dev*, vol. 36, p. 34-40, févr. 2016, doi: 10.1016/j.gde.2016.03.008.

- [60] D. G. Lupiáñez *et al.*, « Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions », *Cell*, vol. 161, n° 5, p. 1012-1025, mai 2015, doi: 10.1016/j.cell.2015.04.004.
- [61] M. Franke *et al.*, « Formation of new chromatin domains determines pathogenicity of genomic duplications », *Nature*, vol. 538, n° 7624, p. 265-269, oct. 2016, doi: 10.1038/nature19800.
- [62] B. Bonev et G. Cavalli, « Organization and function of the 3D genome », *Nat Rev Genet*, vol. 17, n° 11, p. 661-678, oct. 2016, doi: 10.1038/nrg.2016.112.
- [63] E. P. Nora *et al.*, « Spatial partitioning of the regulatory landscape of the X-inactivation centre », *Nature*, vol. 485, n° 7398, Art. n° 7398, mai 2012, doi: 10.1038/nature11049.
- [64] S. Chakraborty *et al.*, « Enhancer–promoter interactions can bypass CTCF-mediated boundaries and contribute to phenotypic robustness », *Nat Genet*, vol. 55, n° 2, Art. n° 2, févr. 2023, doi: 10.1038/s41588-022-01295-6.
- [65] E. Lieberman-Aiden *et al.*, « Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome », *Science*, vol. 326, n° 5950, p. 289-293, oct. 2009, doi: 10.1126/science.1181369.
- [66] L. Guelen *et al.*, « Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions », *Nature*, vol. 453, n° 7197, p. 948-951, juin 2008, doi: 10.1038/nature06947.
- [67] B. van Steensel et A. S. Belmont, « Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression », *Cell*, vol. 169, n° 5, p. 780-791, mai 2017, doi: 10.1016/j.cell.2017.04.022.
- [68] K. L. Reddy, J. M. Zullo, E. Bertolino, et H. Singh, « Transcriptional repression mediated by repositioning of genes to the nuclear lamina », *Nature*, vol. 452, n° 7184, Art. n° 7184, mars 2008, doi: 10.1038/nature06727.
- [69] M. I. Robson *et al.*, « Tissue-Specific Gene Repositioning by Muscle Nuclear Membrane Proteins Enhances Repression of Critical Developmental Genes during Myogenesis », *Molecular Cell*, vol. 62, n° 6, p. 834-847, juin 2016, doi: 10.1016/j.molcel.2016.04.035.
- [70] D. Peric-Hupkes *et al.*, « Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation », *Molecular Cell*, vol. 38, n° 4, p. 603-613, mai 2010, doi: 10.1016/j.molcel.2010.03.016.

- [71] R. Czapiewski, M. I. Robson, et E. C. Schirmer, « Anchoring a Leviathan: How the Nuclear Membrane Tethers the Genome », *Front Genet*, vol. 7, p. 82, 2016, doi: 10.3389/fgene.2016.00082.
- [72] J. Piché, P. P. Van Vliet, M. Pucéat, et G. Andelfinger, « The expanding phenotypes of cohesinopathies: one ring to rule them all! », *Cell Cycle*, vol. 18, n° 21, p. 2828-2848, sept. 2019, doi: 10.1080/15384101.2019.1658476.
- [73] C. Michaelis, R. Ciosk, et K. Nasmyth, « Cohesins: chromosomal proteins that prevent premature separation of sister chromatids », *Cell*, vol. 91, n° 1, p. 35-45, oct. 1997, doi: 10.1016/s0092-8674(01)80007-6.
- [74] E. Watrin, F. J. Kaiser, et K. S. Wendt, « Gene regulation and chromatin organization: relevance of cohesin mutations to human disease », *Curr Opin Genet Dev*, vol. 37, p. 59-66, avr. 2016, doi: 10.1016/j.gde.2015.12.004.
- [75] J. Zuin *et al.*, « Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells », *Proc. Natl. Acad. Sci. U.S.A.*, vol. 111, n° 3, p. 996-1001, janv. 2014, doi: 10.1073/pnas.1317788111.
- [76] M. C. Wahl, C. L. Will, et R. Lührmann, « The Spliceosome: Design Principles of a Dynamic RNP Machine », *Cell*, vol. 136, n° 4, p. 701-718, févr. 2009, doi: 10.1016/j.cell.2009.02.009.
- [77] T. W. Nilsen et B. R. Graveley, « Expansion of the eukaryotic proteome by alternative splicing », *Nature*, vol. 463, n° 7280, p. 457-463, janv. 2010, doi: 10.1038/nature08909.
- [78] Q. Liu, L. Fang, et C. Wu, « Alternative Splicing and Isoforms: From Mechanisms to Diseases », *Genes*, vol. 13, n° 3, Art. n° 3, mars 2022, doi: 10.3390/genes13030401.
- [79] G.-S. Wang et T. A. Cooper, « Splicing in disease: disruption of the splicing code and the decoding machinery », *Nat Rev Genet*, vol. 8, n° 10, Art. n° 10, oct. 2007, doi: 10.1038/nrg2164.
- [80] M. M. Scotti et M. S. Swanson, « RNA mis-splicing in disease », *Nat Rev Genet*, vol. 17, n° 1, p. 19-32, janv. 2016, doi: 10.1038/nrg.2015.3.
- [81] A. Mironov, S. Denisov, A. Gress, O. V. Kalinina, et D. D. Pervouchine, « An extended catalogue of tandem alternative splice sites in human tissue transcriptomes », *PLoS Comput Biol*, vol. 17, n° 4, p. e1008329, avr. 2021, doi: 10.1371/journal.pcbi.1008329.

- [82] K. Jaganathan *et al.*, « Predicting Splicing from Primary Sequence with Deep Learning », *Cell*, vol. 176, n° 3, p. 535-548.e24, janv. 2019, doi: 10.1016/j.cell.2018.12.015.
- [83] R. Leman *et al.*, « SPiP: Splicing Prediction Pipeline, a machine learning tool for massive detection of exonic and intronic variant effects on mRNA splicing », *Hum Mutat*, vol. 43, n° 12, p. 2308-2323, déc. 2022, doi: 10.1002/humu.24491.
- [84] J. Han, Y. Lee, K.-H. Yeom, Y.-K. Kim, H. Jin, et V. N. Kim, « The Drosha-DGCR8 complex in primary microRNA processing », *Genes Dev*, vol. 18, n° 24, p. 3016-3027, déc. 2004, doi: 10.1101/gad.1262504.
- [85] D. P. Bartel, « MicroRNAs: genomics, biogenesis, mechanism, and function », *Cell*, vol. 116, n° 2, p. 281-297, janv. 2004, doi: 10.1016/s0092-8674(04)00045-5.
- [86] Y.-F. Chang, J. S. Imam, et M. F. Wilkinson, « The nonsense-mediated decay RNA surveillance pathway », *Annu Rev Biochem*, vol. 76, p. 51-74, 2007, doi: 10.1146/annurev.biochem.76.050106.093909.
- [87] S. Brogna et J. Wen, « Nonsense-mediated mRNA decay (NMD) mechanisms », *Nat Struct Mol Biol*, vol. 16, n° 2, Art. n° 2, févr. 2009, doi: 10.1038/nsmb.1550.
- [88] S. Kervestin et A. Jacobson, « NMD: a multifaceted response to premature translational termination », *Nat Rev Mol Cell Biol*, vol. 13, n° 11, p. 700-712, nov. 2012, doi: 10.1038/nrm3454.
- [89] C. J. Shoemaker et R. Green, « Translation drives mRNA quality control », *Nat Struct Mol Biol*, vol. 19, n° 6, p. 594-601, juin 2012, doi: 10.1038/nsmb.2301.
- [90] L. A. Gillis *et al.*, « NIPBL mutational analysis in 120 individuals with Cornelia de Lange syndrome and evaluation of genotype-phenotype correlations », *Am J Hum Genet*, vol. 75, n° 4, p. 610-623, oct. 2004, doi: 10.1086/424698.
- [91] S. Rohatgi *et al.*, « Facial diagnosis of mild and variant CdLS: Insights from a dysmorphologist survey », *Am J Med Genet A*, vol. 152A, n° 7, p. 1641-1653, juill. 2010, doi: 10.1002/ajmg.a.33441.
- [92] M. A. Deardorff *et al.*, « Mutations in cohesin complex members SMC3 and SMC1A cause a mild variant of cornelia de Lange syndrome with predominant mental retardation », *Am J Hum Genet*, vol. 80, n° 3, p. 485-494, mars 2007, doi: 10.1086/511888.
- [93] M. A. Deardorff, S. E. Noon, et I. D. Krantz, « Cornelia de Lange Syndrome », in *GeneReviews®*, M. P. Adam, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, L. J. Bean, K. W. Gripp, et A. Amemiya, Éd., Seattle (WA): University of Washington, Seattle, 1993.

Consulté le: 13 août 2023. [En ligne]. Disponible sur:

<http://www.ncbi.nlm.nih.gov/books/NBK1104/>

- [94] I. D. Krantz *et al.*, « Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of *Drosophila melanogaster* Nipped-B », *Nat Genet*, vol. 36, n° 6, p. 631-635, juin 2004, doi: 10.1038/ng1364.
- [95] M. A. Deardorff *et al.*, « RAD21 Mutations Cause a Human Cohesinopathy », *The American Journal of Human Genetics*, vol. 90, n° 6, p. 1014-1027, juin 2012, doi: 10.1016/j.ajhg.2012.04.019.
- [96] M. A. Deardorff *et al.*, « HDAC8 mutations in Cornelia de Lange syndrome affect the cohesin acetylation cycle », *Nature*, vol. 489, n° 7415, p. 313-317, sept. 2012, doi: 10.1038/nature11316.
- [97] G. Olley *et al.*, « BRD4 interacts with NIPBL and BRD4 is mutated in a Cornelia de Lange-like syndrome », *Nat Genet*, vol. 50, n° 3, p. 329-332, mars 2018, doi: 10.1038/s41588-018-0042-y.
- [98] G. Jouret *et al.*, « Understanding the new BRD4-related syndrome: Clinical and genomic delineation with an international cohort study », *Clin Genet*, vol. 102, n° 2, p. 117-122, août 2022, doi: 10.1111/cge.14141.
- [99] J. L. Barbero, « Genetic basis of cohesinopathies », *Appl Clin Genet*, vol. 6, p. 15-23, 2013, doi: 10.2147/TACG.S34457.
- [100] F. Charbonnier *et al.*, « Detection of exon deletions and duplications of the mismatch repair genes in hereditary nonpolyposis colorectal cancer families using multiplex polymerase chain reaction of short fluorescent fragments », *Cancer Res*, vol. 60, n° 11, p. 2760-2763, juin 2000.
- [101] P. Saugier-veber *et al.*, « Simple detection of genomic microdeletions and microduplications using QMPSF in patients with idiopathic mental retardation », *Eur J Hum Genet*, vol. 14, n° 9, p. 1009-1017, sept. 2006, doi: 10.1038/sj.ejhg.5201661.
- [102] C. Pottier *et al.*, « Amyloid- β protein precursor gene expression in alzheimer's disease and other conditions », *J Alzheimers Dis*, vol. 28, n° 3, p. 561-566, 2012, doi: 10.3233/JAD-2011-111148.
- [103] G. Nicolas *et al.*, « Mutation in the 3'untranslated region of APP as a genetic determinant of cerebral amyloid angiopathy », *Eur J Hum Genet*, vol. 24, n° 1, p. 92-98, janv. 2016, doi: 10.1038/ejhg.2015.61.

- [104] A. Rovelet-Lecrux *et al.*, « De novo deleterious genetic variations target a biological network centered on A β peptide in early-onset Alzheimer disease », *Mol Psychiatry*, vol. 20, n° 9, p. 1046-1056, sept. 2015, doi: 10.1038/mp.2015.100.
- [105] Y. Zerdoumi *et al.*, « A new genotoxicity assay based on p53 target gene induction », *Mutat Res Genet Toxicol Environ Mutagen*, vol. 789-790, p. 28-35, août 2015, doi: 10.1016/j.mrgentox.2015.05.010.
- [106] S. Raad *et al.*, « Blood functional assay for rapid clinical interpretation of germline TP53 variants », *J Med Genet*, vol. 58, n° 12, p. 796-805, déc. 2021, doi: 10.1136/jmedgenet-2020-107059.
- [107] J. P. Schouten, C. J. McElgunn, R. Waaijer, D. Zwijnenburg, F. Diepvens, et G. Pals, « Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification », *Nucleic Acids Res*, vol. 30, n° 12, p. e57, juin 2002, doi: 10.1093/nar/gnf056.
- [108] V. Bobée *et al.*, « Determination of Molecular Subtypes of Diffuse Large B-Cell Lymphoma Using a Reverse Transcriptase Multiplex Ligation-Dependent Probe Amplification Classifier: A CALYM Study », *J Mol Diagn*, vol. 19, n° 6, p. 892-904, nov. 2017, doi: 10.1016/j.jmoldx.2017.07.007.
- [109] S. Mareschal *et al.*, « Accurate Classification of Germinal Center B-Cell-Like/Activated B-Cell-Like Diffuse Large B-Cell Lymphoma Using a Simple and Rapid Reverse Transcriptase-Multiplex Ligation-Dependent Probe Amplification Assay: A CALYM Study », *J Mol Diagn*, p. S1525-1578(15)00046-X, avr. 2015, doi: 10.1016/j.jmoldx.2015.01.007.
- [110] C. Levacher *et al.*, « Disequilibrium between BRCA1 and BRCA2 Circular and Messenger RNAs Plays a Role in Breast Cancer », *Cancers (Basel)*, vol. 15, n° 7, p. 2176, avr. 2023, doi: 10.3390/cancers15072176.
- [111] P. J. Sykes, S. H. Neoh, M. J. Brisco, E. Hughes, J. Condon, et A. A. Morley, « Quantitation of targets for PCR by use of limiting dilution », *Biotechniques*, vol. 13, n° 3, p. 444-449, sept. 1992.
- [112] B. Vogelstein et K. W. Kinzler, « Digital PCR », *Proc Natl Acad Sci U S A*, vol. 96, n° 16, p. 9236-9241, août 1999, doi: 10.1073/pnas.96.16.9236.
- [113] K. Perez-Toralla *et al.*, « PCR digitale en micro-compartiments - I. Détection sensible de séquences d'acides nucléiques rares », *Med Sci (Paris)*, vol. 31, n° 1, p. 84-92, janv. 2015, doi: 10.1051/medsci/20153101017.

- [114] J. F. Huggett *et al.*, « The digital MIQE guidelines: Minimum Information for Publication of Quantitative Digital PCR Experiments », *Clin Chem*, vol. 59, n° 6, p. 892-902, juin 2013, doi: 10.1373/clinchem.2013.206375.
- [115] R. Williams, S. G. Peisajovich, O. J. Miller, S. Magdassi, D. S. Tawfik, et A. D. Griffiths, « Amplification of complex gene libraries by emulsion PCR », *Nat Methods*, vol. 3, n° 7, p. 545-550, juill. 2006, doi: 10.1038/nmeth896.
- [116] D. Dressman, H. Yan, G. Traverso, K. W. Kinzler, et B. Vogelstein, « Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations », *Proc Natl Acad Sci U S A*, vol. 100, n° 15, p. 8817-8822, juill. 2003, doi: 10.1073/pnas.1133470100.
- [117] K. Cassinari *et al.*, « A Simple, Universal, and Cost-Efficient Digital PCR Method for the Targeted Analysis of Copy Number Variations », *Clin. Chem.*, juill. 2019, doi: 10.1373/clinchem.2019.304246.
- [118] K. Cassinari *et al.*, « Assessment of Multiplex Digital Droplet RT-PCR as a Diagnostic Tool for SARS-CoV-2 Detection in Nasopharyngeal Swabs and Saliva Samples », *Clin Chem*, vol. 67, n° 5, p. 736-741, avr. 2021, doi: 10.1093/clinchem/hvaa323.
- [119] M. Schena, D. Shalon, R. W. Davis, et P. O. Brown, « Quantitative monitoring of gene expression patterns with a complementary DNA microarray », *Science*, vol. 270, n° 5235, p. 467-470, oct. 1995, doi: 10.1126/science.270.5235.467.
- [120] F. Lecoquierre *et al.*, « High diagnostic potential of short and long read genome sequencing with transcriptome analysis in exome-negative developmental disorders », *Hum Genet*, vol. 142, n° 6, p. 773-783, juin 2023, doi: 10.1007/s00439-023-02553-1.
- [121] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, et B. Wold, « Mapping and quantifying mammalian transcriptomes by RNA-Seq », *Nat Methods*, vol. 5, n° 7, p. 621-628, juill. 2008, doi: 10.1038/nmeth.1226.
- [122] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, et C. N. Dewey, « RNA-Seq gene expression estimation with read mapping uncertainty », *Bioinformatics*, vol. 26, n° 4, p. 493-500, févr. 2010, doi: 10.1093/bioinformatics/btp692.
- [123] S. Anders, P. T. Pyl, et W. Huber, « HTSeq--a Python framework to work with high-throughput sequencing data », *Bioinformatics*, vol. 31, n° 2, p. 166-169, janv. 2015, doi: 10.1093/bioinformatics/btu638.

- [124] B. Li et C. N. Dewey, « RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome », *BMC Bioinformatics*, vol. 12, p. 323, août 2011, doi: 10.1186/1471-2105-12-323.
- [125] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, et C. Kingsford, « Salmon provides fast and bias-aware quantification of transcript expression », *Nat Methods*, vol. 14, n° 4, p. 417-419, avr. 2017, doi: 10.1038/nmeth.4197.
- [126] J. T. Robinson *et al.*, « Integrative genomics viewer », *Nat Biotechnol*, vol. 29, n° 1, p. 24-26, janv. 2011, doi: 10.1038/nbt.1754.
- [127] A. Mehmood, A. Laiho, M. S. Venäläinen, A. J. McGlinchey, N. Wang, et L. L. Elo, « Systematic evaluation of differential splicing tools for RNA-seq studies », *Briefings in Bioinformatics*, vol. 21, n° 6, p. 2052-2065, déc. 2020, doi: 10.1093/bib/bbz126.
- [128] V. R. de Melo Costa, J. Pfeuffer, A. Louloui, U. A. V. Ørom, et R. M. Piro, « SPLICE-q: a Python tool for genome-wide quantification of splicing efficiency », *BMC Bioinformatics*, vol. 22, n° 1, p. 368, juill. 2021, doi: 10.1186/s12859-021-04282-6.
- [129] E. K. Flemington *et al.*, « SpliceTools, a suite of downstream RNA splicing analysis tools to investigate mechanisms and impact of alternative splicing », *Nucleic Acids Research*, vol. 51, n° 7, p. e42, avr. 2023, doi: 10.1093/nar/gkad111.
- [130] R. F. Halperin *et al.*, « Improved methods for RNAseq-based alternative splicing analysis », *Sci Rep*, vol. 11, n° 1, Art. n° 1, mai 2021, doi: 10.1038/s41598-021-89938-2.
- [131] C. Mertes *et al.*, « Detection of aberrant splicing events in RNA-seq data using FRASER », *Nat Commun*, vol. 12, n° 1, p. 529, janv. 2021, doi: 10.1038/s41467-020-20573-7.
- [132] G. Eraslan *et al.*, « Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function », *Science*, vol. 376, n° 6594, p. eabl4290, mai 2022, doi: 10.1126/science.abl4290.
- [133] THE GTEx CONSORTIUM, « The GTEx Consortium atlas of genetic regulatory effects across human tissues », *Science*, vol. 369, n° 6509, p. 1318-1330, sept. 2020, doi: 10.1126/science.aaz1776.
- [134] E. R. Gamazon *et al.*, « Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation », *Nat Genet*, vol. 50, n° 7, Art. n° 7, juill. 2018, doi: 10.1038/s41588-018-0154-4.

- [135] B. E. Stranger *et al.*, « Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease », *Nat Genet*, vol. 49, n° 12, Art. n° 12, déc. 2017, doi: 10.1038/ng.3969.
- [136] F. Aguet *et al.*, « Genetic effects on gene expression across human tissues », *Nature*, vol. 550, n° 7675, Art. n° 7675, oct. 2017, doi: 10.1038/nature24277.
- [137] ENCODE Project Consortium, « An integrated encyclopedia of DNA elements in the human genome », *Nature*, vol. 489, n° 7414, p. 57-74, sept. 2012, doi: 10.1038/nature11247.
- [138] Y. Luo *et al.*, « New developments on the Encyclopedia of DNA Elements (ENCODE) data portal », *Nucleic Acids Res*, vol. 48, n° D1, p. D882-D889, janv. 2020, doi: 10.1093/nar/gkz1062.
- [139] D. Karolchik *et al.*, « The UCSC Table Browser data retrieval tool », *Nucleic Acids Res*, vol. 32, n° Database issue, p. D493-496, janv. 2004, doi: 10.1093/nar/gkh103.
- [140] A. Visel, S. Minovitsky, I. Dubchak, et L. A. Pennacchio, « VISTA Enhancer Browser--a database of tissue-specific human enhancers », *Nucleic Acids Res*, vol. 35, n° Database issue, p. D88-92, janv. 2007, doi: 10.1093/nar/gkl822.
- [141] S. Fishilevich *et al.*, « GeneHancer: genome-wide integration of enhancers and target genes in GeneCards », *Database (Oxford)*, vol. 2017, p. bax028, janv. 2017, doi: 10.1093/database/bax028.
- [142] R. Andersson *et al.*, « An atlas of active enhancers across human cell types and tissues », *Nature*, vol. 507, n° 7493, p. 455-461, mars 2014, doi: 10.1038/nature12787.
- [143] D. R. Zerbino, S. P. Wilder, N. Johnson, T. Juettemann, et P. R. Flicek, « The ensembl regulatory build », *Genome Biol*, vol. 16, n° 1, p. 56, mars 2015, doi: 10.1186/s13059-015-0621-5.
- [144] A. Schroeder *et al.*, « The RIN: an RNA integrity number for assigning integrity values to RNA measurements », *BMC Molecular Biology*, vol. 7, n° 1, p. 3, janv. 2006, doi: 10.1186/1471-2199-7-3.
- [145] I. Gallego Romero, A. A. Pai, J. Tung, et Y. Gilad, « RNA-seq: impact of RNA degradation on transcript quantification », *BMC Biology*, vol. 12, n° 1, p. 42, mai 2014, doi: 10.1186/1741-7007-12-42.
- [146] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, et E. Charpentier, « A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity », *Science*, vol. 337, n° 6096, p. 816-821, août 2012, doi: 10.1126/science.1225829.

- [147] A. Pickar-Oliver et C. A. Gersbach, « The next generation of CRISPR-Cas technologies and applications », *Nat Rev Mol Cell Biol*, vol. 20, n° 8, p. 490-507, août 2019, doi: 10.1038/s41580-019-0131-5.
- [148] R. Ben Jehuda, Y. Shemer, et O. Binah, « Genome Editing in Induced Pluripotent Stem Cells using CRISPR/Cas9 », *Stem Cell Rev and Rep*, vol. 14, n° 3, p. 323-336, juin 2018, doi: 10.1007/s12015-018-9811-3.
- [149] O. O. Abudayyeh *et al.*, « C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector », *Science*, vol. 353, n° 6299, p. aaf5573, août 2016, doi: 10.1126/science.aaf5573.
- [150] G. Nicolas *et al.*, « Mutation of the PDGFRB gene as a cause of idiopathic basal ganglia calcification », *Neurology*, vol. 80, n° 2, p. 181-187, janv. 2013, doi: 10.1212/WNL.0b013e31827ccf34.
- [151] G. Nicolas *et al.*, « Brain calcification process and phenotypes according to age and sex: Lessons from SLC20A2, PDGFB, and PDGFRB mutation carriers », *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 168, n° 7, p. 586-594, 2015, doi: 10.1002/ajmg.b.32336.
- [152] L. Grangeon *et al.*, « Biallelic MYORG mutation carriers exhibit primary brain calcification with a distinct phenotype », *Brain*, vol. 142, n° 6, p. 1573-1586, juin 2019, doi: 10.1093/brain/awz095.
- [153] E. Masson, « Calcifications idiopathiques des noyaux gris centraux (maladie de Fahr) », *EM-Consulte*. <https://www.em-consulte.com/article/841588/article/calcifications-idiopathiques-des-noyaux-gris-centr> (consulté le 26 mai 2023).
- [154] C. Wang *et al.*, « Mutations in SLC20A2 link familial idiopathic basal ganglia calcification with phosphate homeostasis », *Nat Genet*, vol. 44, n° 3, p. 254-256, févr. 2012, doi: 10.1038/ng.1077.
- [155] A. Keller *et al.*, « Mutations in the gene encoding PDGF-B cause brain calcifications in humans and mice », *Nat Genet*, vol. 45, n° 9, p. 1077-1082, sept. 2013, doi: 10.1038/ng.2723.
- [156] A. Legati *et al.*, « Mutations in XPR1 cause primary familial brain calcification associated with altered phosphate export », *Nat Genet*, vol. 47, n° 6, p. 579-581, juin 2015, doi: 10.1038/ng.3289.

- [157] X.-P. Yao *et al.*, « Biallelic Mutations in MYORG Cause Autosomal Recessive Primary Familial Brain Calcification », *Neuron*, vol. 98, n° 6, p. 1116-1123.e5, juin 2018, doi: 10.1016/j.neuron.2018.05.037.
- [158] Z. Cen *et al.*, « Biallelic loss-of-function mutations in JAM2 cause primary familial brain calcification », *Brain*, vol. 143, n° 2, p. 491-502, févr. 2020, doi: 10.1093/brain/awz392.
- [159] E. M. Ramos *et al.*, « Primary brain calcification: an international study reporting novel variants and associated phenotypes », *Eur J Hum Genet*, vol. 26, n° 10, p. 1462-1477, oct. 2018, doi: 10.1038/s41431-018-0185-4.
- [160] N. Jensen *et al.*, « Mice Knocked Out for the Primary Brain Calcification-Associated Gene Slc20a2 Show Unimpaired Prenatal Survival but Retarded Growth and Nodules in the Brain that Grow and Calcify Over Time », *Am J Pathol*, vol. 188, n° 8, p. 1865-1881, août 2018, doi: 10.1016/j.ajpath.2018.04.010.
- [161] I. Hozumi *et al.*, « Inorganic phosphorus (Pi) in CSF is a biomarker for SLC20A2-associated idiopathic basal ganglia calcification (IBGC1) », *Journal of the Neurological Sciences*, vol. 388, p. 150-154, mai 2018, doi: 10.1016/j.jns.2018.03.014.
- [162] M. Baker *et al.*, « SLC20A2 and THAP1 deletion in familial basal ganglia calcification with dystonia », *Neurogenetics*, vol. 15, n° 1, p. 23-30, mars 2014, doi: 10.1007/s10048-013-0378-5.
- [163] S. David *et al.*, « Identification of partial SLC20A2 deletions in primary brain calcification using whole-exome sequencing », *Eur J Hum Genet*, vol. 24, n° 11, p. 1630-1634, nov. 2016, doi: 10.1038/ejhg.2016.50.
- [164] K. Grütz *et al.*, « Primary familial brain calcification in the “IBGC2” kindred: All linkage roads lead to SLC20A2 », *Mov Disord*, vol. 31, n° 12, p. 1901-1904, déc. 2016, doi: 10.1002/mds.26768.
- [165] X.-X. Guo *et al.*, « Identification of SLC20A2 deletions in patients with primary familial brain calcification », *Clin Genet*, vol. 96, n° 1, p. 53-60, juill. 2019, doi: 10.1111/cge.13540.
- [166] P. Pasanen *et al.*, « Primary familial brain calcification linked to deletion of 5' noncoding region of SLC20A2 », *Acta Neurol Scand*, vol. 136, n° 1, p. 59-63, juill. 2017, doi: 10.1111/ane.12697.
- [167] O. Quenez *et al.*, « Detection of copy-number variations from NGS data using read depth information: a diagnostic performance evaluation », *Eur J Hum Genet*, vol. 29, n° 1, p. 99-109, janv. 2021, doi: 10.1038/s41431-020-0672-2.

- [168] A. Biffi et S. M. Greenberg, « Cerebral Amyloid Angiopathy: A Systematic Review », *J Clin Neurol*, vol. 7, n° 1, p. 1-9, mars 2011, doi: 10.3988/jcn.2011.7.1.1.
- [169] A. Rovelet-Lecrux *et al.*, « APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy », *Nat Genet*, vol. 38, n° 1, p. 24-26, janv. 2006, doi: 10.1038/ng1718.
- [170] K. Sleegers *et al.*, « APP duplication is sufficient to cause early onset Alzheimer's dementia with cerebral amyloid angiopathy », *Brain*, vol. 129, n° Pt 11, p. 2977-2983, nov. 2006, doi: 10.1093/brain/awl203.
- [171] D. Wallon *et al.*, « The French series of autosomal dominant early onset Alzheimer's disease cases: mutation spectrum and cerebrospinal fluid biomarkers », *J Alzheimers Dis*, vol. 30, n° 4, p. 847-856, 2012, doi: 10.3233/JAD-2012-120172.
- [172] H.-M. Lanoiselée *et al.*, « APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: A genetic screening study of familial and sporadic cases », *PLoS Med*, vol. 14, n° 3, p. e1002270, mars 2017, doi: 10.1371/journal.pmed.1002270.
- [173] I. Guyant-Marechal *et al.*, « Intrafamilial diversity of phenotype associated with app duplication », *Neurology*, vol. 71, n° 23, p. 1925-1926, déc. 2008, doi: 10.1212/01.wnl.0000339400.64213.56.
- [174] L. Grangeon *et al.*, « Phenotype and imaging features associated with APP duplications », *Alzheimers Res Ther*, vol. 15, n° 1, p. 93, mai 2023, doi: 10.1186/s13195-023-01172-2.
- [175] D. M. A. Mann *et al.*, « Patterns and severity of vascular amyloid in Alzheimer's disease associated with duplications and missense mutations in APP gene, Down syndrome and sporadic Alzheimer's disease », *Acta Neuropathol*, vol. 136, n° 4, p. 569-587, 2018, doi: 10.1007/s00401-018-1866-3.
- [176] S. Rentas *et al.*, « Diagnosing Cornelia de Lange syndrome and related neurodevelopmental disorders using RNA sequencing », *Genet Med*, vol. 22, n° 5, p. 927-936, mai 2020, doi: 10.1038/s41436-019-0741-5.
- [177] B. B. Cummings *et al.*, « Improving genetic diagnosis in Mendelian disease with transcriptome sequencing », *Science Translational Medicine*, vol. 9, n° 386, p. eaal5209, avr. 2017, doi: 10.1126/scitranslmed.aal5209.
- [178] H. D. Gonorazky *et al.*, « Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease », *The American Journal of Human Genetics*, vol. 104, n° 3, p. 466-483, mars 2019, doi: 10.1016/j.ajhg.2019.01.012.

- [179] D. R. Murdock *et al.*, « Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing », *J Clin Invest*, vol. 131, n° 1, janv. 2021, doi: 10.1172/JCI141500.
- [180] A. Dobin *et al.*, « STAR: ultrafast universal RNA-seq aligner », *Bioinformatics*, vol. 29, n° 1, p. 15-21, janv. 2013, doi: 10.1093/bioinformatics/bts635.
- [181] M. I. Love, W. Huber, et S. Anders, « Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 », *Genome Biology*, vol. 15, n° 12, p. 550, déc. 2014, doi: 10.1186/s13059-014-0550-8.
- [182] I. Parenti *et al.*, « MAU2 and NIPBL Variants Impair the Heterodimerization of the Cohesin Loader Subunits and Cause Cornelia de Lange Syndrome », *Cell Rep*, vol. 31, n° 7, p. 107647, mai 2020, doi: 10.1016/j.celrep.2020.107647.
- [183] T. Kleefstra et N. de Leeuw, « Kleefstra Syndrome », in *GeneReviews®*, M. P. Adam, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, L. J. Bean, K. W. Gripp, et A. Amemiya, Éd., Seattle (WA): University of Washington, Seattle, 1993. Consulté le: 16 août 2023. [En ligne]. Disponible sur: <http://www.ncbi.nlm.nih.gov/books/NBK47079/>
- [184] M. Pons *et al.*, « Splicing factors act as genetic modulators of TDP-43 production in a new autoregulatory TDP-43 Drosophila model », *Hum Mol Genet*, vol. 26, n° 17, p. 3396-3408, sept. 2017, doi: 10.1093/hmg/ddx229.
- [185] M. Pons *et al.*, « Identification of TCERG1 as a new genetic modulator of TDP-43 production in Drosophila », *Acta Neuropathol Commun*, vol. 6, n° 1, p. 138, déc. 2018, doi: 10.1186/s40478-018-0639-5.
- [186] S. Lenglez *et al.*, « Distinct functional classes of PDGFRB pathogenic variants in primary familial brain calcification », *Hum Mol Genet*, vol. 31, n° 3, p. 399-409, févr. 2022, doi: 10.1093/hmg/ddab258.
- [187] D. A. Koolen, A. Morgan, et B. B. de Vries, « Koolen-de Vries Syndrome », in *GeneReviews®*, M. P. Adam, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, L. J. Bean, K. W. Gripp, et A. Amemiya, Éd., Seattle (WA): University of Washington, Seattle, 1993. Consulté le: 15 août 2023. [En ligne]. Disponible sur: <http://www.ncbi.nlm.nih.gov/books/NBK24676/>
- [188] K. Le Guennec *et al.*, « 17q21.31 duplication causes prominent tau-related dementia with increased MAPT expression », *Mol Psychiatry*, vol. 22, n° 8, Art. n° 8, août 2017, doi: 10.1038/mp.2016.226.

- [189] L. Miguel *et al.*, « Generation of 17q21.31 duplication iPSC-derived neurons as a model for primary tauopathies », *Stem Cell Research*, vol. 61, p. 102762, mai 2022, doi: 10.1016/j.scr.2022.102762.
- [190] D. Evanko, « Snapshots of gene expression », *Nat Methods*, vol. 3, n° 10, Art. n° 10, oct. 2006, doi: 10.1038/nmeth1006-774.
- [191] N. Kamitaki, C. L. Usher, et S. A. McCarroll, « Using Droplet Digital PCR to Analyze Allele-Specific RNA Expression », *Methods Mol Biol*, vol. 1768, p. 401-422, 2018, doi: 10.1007/978-1-4939-7778-9_23.
- [192] E. Eisenberg et E. Y. Levanon, « Human housekeeping genes, revisited », *Trends in Genetics*, vol. 29, n° 10, p. 569-574, oct. 2013, doi: 10.1016/j.tig.2013.05.010.
- [193] J. Xiao, X. Li, J. Liu, X. Fan, H. Lei, et C. Li, « Identification of reference genes in blood before and after entering the plateau for SYBR green RT-qPCR studies », *PeerJ*, vol. 5, p. e3726, sept. 2017, doi: 10.7717/peerj.3726.
- [194] K. Hieronymus *et al.*, « Validation of reference genes for whole blood gene expression analysis in cord blood of preterm and full-term neonates and peripheral blood of healthy adults », *BMC Genomics*, vol. 22, n° 1, p. 489, juin 2021, doi: 10.1186/s12864-021-07801-0.
- [195] V. A. Yépez *et al.*, « Clinical implementation of RNA sequencing for Mendelian disease diagnostics », *Genome Medicine*, vol. 14, n° 1, p. 38, avr. 2022, doi: 10.1186/s13073-022-01019-9.
- [196] F. Brechtmann *et al.*, « OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data », *Am J Hum Genet*, vol. 103, n° 6, p. 907-917, déc. 2018, doi: 10.1016/j.ajhg.2018.10.025.
- [197] R. L. Collins *et al.*, « A structural variation reference for medical and population genetics », *Nature*, vol. 581, n° 7809, Art. n° 7809, mai 2020, doi: 10.1038/s41586-020-2287-8.
- [198] S. E. Castel, A. Levy-Moonshine, P. Mohammadi, E. Banks, et T. Lappalainen, « Tools and best practices for data processing in allelic expression analysis », *Genome Biol*, vol. 16, n° 1, p. 195, sept. 2015, doi: 10.1186/s13059-015-0762-6.
- [199] L. Frésard *et al.*, « Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts », *Nat Med*, vol. 25, n° 6, p. 911-919, juin 2019, doi: 10.1038/s41591-019-0457-8.

- [200] H. Lee *et al.*, « Diagnostic utility of transcriptome sequencing for rare Mendelian diseases », *Genet Med*, vol. 22, n° 3, p. 490-499, mars 2020, doi: 10.1038/s41436-019-0672-1.
- [201] R. Truty *et al.*, « Spectrum of splicing variants in disease genes and the ability of RNA analysis to reduce uncertainty in clinical interpretation », *Am J Hum Genet*, vol. 108, n° 4, p. 696-708, avr. 2021, doi: 10.1016/j.ajhg.2021.03.006.
- [202] M. Lee *et al.*, « Diagnostic potential of the amniotic fluid cells transcriptome in deciphering mendelian disease: a proof-of-concept », *npj Genom. Med.*, vol. 7, n° 1, Art. n° 1, déc. 2022, doi: 10.1038/s41525-022-00347-4.
- [203] B. Sadikovic, E. Aref-Eshghi, M. A. Levy, et D. Rodenhiser, « DNA methylation signatures in mendelian developmental disorders as a diagnostic bridge between genotype and phenotype », *Epigenomics*, vol. 11, n° 5, p. 563-575, avr. 2019, doi: 10.2217/epi-2018-0192.
- [204] M. A. Levy *et al.*, « Novel diagnostic DNA methylation epesignatures expand and refine the epigenetic landscapes of Mendelian disorders », *HGG Adv*, vol. 3, n° 1, p. 100075, janv. 2022, doi: 10.1016/j.xhgg.2021.100075.
- [205] T. Husson *et al.*, « Epesignatures in practice: independent evaluation of published epesignatures for the molecular diagnostics of ten neurodevelopmental disorder ». 2023. doi: 10.21203/rs.3.rs-2924104/v1.