

Épissignatures sur données de séquençage long read pour le diagnostic des troubles du neurodéveloppement

Résumé du projet

Contexte. Les troubles du neurodéveloppement (TND), incluant la déficience intellectuelle et les troubles du spectre de l'autisme, concernent environ 5 % de la population française, avec près de 35 000 nouveaux cas par an. Leurs tableaux cliniques, très hétérogènes et parfois chevauchants, rendent le diagnostic complexe. Un diagnostic moléculaire précis est pourtant essentiel pour mettre fin à l'errance diagnostique, adapter la prise en charge et améliorer le conseil génétique aux familles. Le séquençage à haut débit à lectures courtes, intégré au Plan France Médecine Génomique 2025, permet aujourd'hui un diagnostic dans 40 à 60 % des cas. Près de la moitié des patients restent toutefois sans diagnostic, et environ 10 % présentent des variants de signification incertaine (VSI), dont l'interprétation reste un verrou majeur de la génomique médicale.

Les *épissignatures*, profils spécifiques de méthylation de l'ADN détectables dans le sang, offrent un moyen puissant de reclassifier les VSI et de confirmer des diagnostics génétiques. Plus de 100 épissignatures ont été identifiées ces dix dernières années dans les TND, essentiellement à partir de puces de méthylation Illumina, et leur potentiel clinique est reconnu par la Haute Autorité de Santé. En parallèle, le séquençage à lectures longues PacBio HiFi lit de longues molécules d'ADN avec une très grande précision, détecte tous les types de variants (y compris les variants de structure et la plupart des répétitions) et donne directement accès à la méthylation du génome, en une seule analyse. Il devient ainsi envisageable de développer des épissignatures directement à partir du séquençage du génome, et d'explorer une marque encore peu étudiée, la méthylation 6mA, en regard de la méthylation 5mC classique. Le défi est avant tout statistique : détecter des différences de méthylation très fines (de l'ordre de 5 à 15 % entre patients et témoins) à partir de données de profondeur modeste (~30x). Les premiers modèles publiés, principalement sur données Oxford Nanopore, se limitent encore largement à l'interrogation de signatures existantes.

Objectifs. Cette thèse s'inscrit dans le projet LR2D (appel à projets MESSIDORE 2025), qui vise à évaluer la faisabilité et les bénéfices médicaux d'une analyse génomique et épigénomique multimodale par séquençage du génome HiFi chez des individus présentant des troubles du développement. L'objectif principal de la thèse est d'évaluer la transférabilité des épissignatures, depuis les puces de méthylation vers le séquençage du génome HiFi, sur une liste restreinte de syndromes présentant des épissignatures bien établies. Cet objectif repose sur plusieurs étapes que cette thèse se propose d'aborder :

1. **Caractériser les données de méthylation PacBio HiFi**, et en particulier la variabilité de ces mesures entre molécules d'ADN, le long du génome et entre individus, à partir de données HiFi déjà disponibles et de duplicats techniques puces Illumina–PacBio.
2. **Développer des méthodes robustes de découverte et d'analyse d'épissignatures** à partir des données HiFi, tirant parti de l'information portée par des régions entières plutôt que par des positions CpG isolées.
3. **Évaluer les performances prédictives** des épissignatures développées sur données HiFi, en comparaison des biomarqueurs établis sur puces, et identifier les propriétés, exigences et bénéfices techniques propres au séquençage à lectures longues.

4. **Approfondir la compréhension des corrélations méthylation–phénotype** et de l'hétérogénéité intrinsèque des épisignatures et des syndromes d'intérêt, dans une logique de médecine personnalisée.
5. **Explorer le profil génomique de la marque 6mA**, désormais accessible, et évaluer son intérêt comme biomarqueur indépendant dans des troubles présentant une épisignature 5mC connue et robuste.

Résultats attendus. D'un point de vue méthodologique, ce projet fournira des outils statistiques et d'apprentissage automatique adaptés aux données épigénomiques issues du séquençage long read, un champ encore largement ouvert. Le (la) doctorant(e) s'appuiera sur une base unique de plus de 1 300 profils de méthylation sur puces couvrant 50 gènes impliqués dans les TND (projet Epi2Diag), complétée par 400 génomes PacBio HiFi produits dans le cadre du projet LR2D : 100 trios enfant atteint–parents, 80 témoins positifs et 20 témoins négatifs pour l'évaluation des épisignatures.

D'un point de vue clinique, le projet prépare l'intégration des épisignatures dans un test unique de première intention combinant analyses génomique et épigénomique. Il contribuera à réduire l'errance diagnostique, à améliorer l'interprétation des VSI et à mieux comprendre la diversité phénotypique au sein d'un même syndrome.

Livrables attendus

- Des articles scientifiques centrés sur une ou plusieurs épisignatures selon les opportunités et découvertes, ainsi qu'un article méthodologique plus large en fin de doctorat.
- La mise à disposition en libre accès des méthodes et outils via le développement d'un package R, afin de les partager avec la communauté de la génétique statistique et médicale.
- Le partage des données génomiques avec la communauté de recherche, dans le respect du RGPD (plateforme EGA, Collecteur Analyseur de Données du PFMG 2025), pour de futurs projets et méta-analyses.

Compétences requises

Ce doctorat s'adresse aux diplômés de niveau M2 en biostatistique, statistique, bioinformatique ou science des données, ou issus d'une formation ingénieure spécialisée en statistique (ENSAE, ENSAI, INSA, ...). Il nécessite de solides bases en modélisation statistique et en apprentissage automatique (*machine learning*), ainsi qu'une bonne maîtrise du logiciel R ; une pratique de Python, de l'environnement Linux et du calcul haute performance sera appréciée. Nous recherchons un(e) candidat(e) ayant une appétence pour l'analyse de données réelles de biologie et de santé, en particulier génomiques et épigénomiques, et le goût du travail en équipe pluridisciplinaire. Aucune connaissance préalable des épisignatures ou des données long read n'est exigée : une formation sera assurée par l'équipe au cours de la première année.

Laboratoire d'accueil

<https://cbg.univ-rouen.fr/>

Laboratoire Inserm UMR1245 « Génomique du Cancer et du Cerveau », direction : Pr Gaël Nicolas

Équipe 3 « Génomique pour les maladies du cerveau »

Université de Rouen Normandie - UFR Santé, 22 boulevard Gambetta, Rouen

Disposant d'un lien fort avec le CHU de Rouen mais aussi des réseaux nationaux et internationaux permettant d'accéder à des données pertinentes et de qualité, l'unité Inserm UMR1245 est une unité de recherche dynamique et fondamentalement pluridisciplinaire consacrée à des projets appliqués à l'oncogénétique et/ou la neurogénétique. Elle héberge la plateforme de génomique ASGAR, pionnière en France pour la production de données PacBio HiFi, et le CHU de Rouen est aujourd'hui un des principaux laboratoires à proposer l'analyse des épisignatures du wetlab au drylab en France (au titre de la recherche).

Au sein de l'équipe 3, aux côtés de cliniciens, biologistes moléculaires et bioinformaticiens, les statisticiens développent des méthodes statistiques permettant de mieux comprendre les prédispositions aux cancers, les maladies du cerveau et les troubles du neurodéveloppement. L'axe « épisignatures » de l'équipe bénéficie de collaborations étroites avec les pionniers du domaine à Toronto (SickKids Hospital) et avec l'Erasmus MC développant la plateforme MethaDory à Rotterdam.

Encadrement et contacts pour information

Camille Charbonnier (statisticienne, PU-PH, directrice de thèse, camille.le-clezio@chu-rouen.fr)

François Lecoquierre (généticien, MCU-PH, co-encadrant, porteur du projet LR2D, francois.lecoquierre@chu-rouen.fr)

Financement

Contrat doctoral de trois ans (Financement MESSIDORE 2025 et allocation doctorale Région Normandie). La production des données PacBio HiFi est financée par le projet LR2D (appel à projets MESSIDORE 2025).

Éléments de bibliographie

- Husson T, Lecoquierre F, Nicolas G, *et al.* Episignatures in practice: independent evaluation of published episignatures for the molecular diagnostics of ten neurodevelopmental disorders. *Eur J Hum Genet* 2024. PMID: 37872275.
- Nava C, Cogne B, Santini A, *et al.* Dominant variants in major spliceosome U4 and U5 small nuclear RNA genes cause neurodevelopmental disorders through splicing disruption. *Nature Genetics* 2025. PMID: 40379786.
- Leitão E, Santini A, Cogne B, *et al.* Systematic analysis of snRNA genes reveals frequent RNU2-2 variants in dominant and recessive developmental and epileptic encephalopathies. *Nature Genetics* 2026. PMID: 41912934.
- Santini A, Tognon A, Richard A, *et al.* Identification of an episignature for CHD3-related Snijders Blok-Campeau syndrome reveals heterogeneity in the CHARGE syndrome episignature: towards a better characterization of chromatinopathies. *Genome Medicine* 2025. PMID: 41952182.