

Episignatures from Long-Read Sequencing Data for the Diagnosis of Neurodevelopmental Disorders

Project Summary

Context. Neurodevelopmental disorders (NDD), including intellectual disability and autism spectrum disorders, affect about 5% of the French population, with nearly 35,000 new cases each year. Their highly heterogeneous and sometimes overlapping clinical presentations make diagnosis challenging, while an accurate molecular diagnosis is key to ending the diagnostic odyssey, adapting care and improving genetic counselling for families. Short-read sequencing, now part of the French Genomic Medicine Plan 2025, yields a diagnosis in 40–60% of cases. Yet about half of patients remain undiagnosed, and around 10% carry variants of uncertain significance (VUS), whose interpretation remains a major bottleneck in medical genomics.

Episignatures – disease-specific DNA methylation profiles detectable in blood – are a powerful tool to reclassify VUS and confirm genetic diagnoses. Over the past decade, more than 100 episignatures have been identified in NDD, mostly from Illumina methylation arrays, and their clinical potential is recognised by the French National Authority for Health (HAS). Meanwhile, PacBio HiFi long-read sequencing reads long DNA molecules with very high accuracy, detects all types of variants (including structural variants and most repeats) and gives direct access to genome-wide methylation in a single assay. This opens the way to deriving episignatures directly from genome sequencing, and to exploring a still poorly studied mark, 6mA methylation, alongside classical 5mC methylation. The challenge is primarily statistical: detecting very subtle methylation differences (about 5–15% between patients and controls) from moderate-depth data ($\sim 30\times$). The first published models, mostly on Oxford Nanopore data, remain largely limited to querying existing signatures.

Objectives. This PhD is part of the LR2D project (MESSIDORE 2025 call), which aims to assess the feasibility and medical benefits of a multimodal genomic and epigenomic analysis by HiFi genome sequencing in individuals with developmental disorders. The main objective of the PhD is to assess the transferability of episignatures from methylation arrays to HiFi genome sequencing, for a restricted list of syndromes with well-established episignatures. This involves several steps:

- 1- To characterise PacBio HiFi methylation data, in particular the variability of measurements across DNA molecules, along the genome and between individuals, using available HiFi data and Illumina–PacBio technical duplicates.
- 2- To develop robust methods for the discovery and analysis of episignatures from HiFi data, leveraging information from whole regions rather than isolated CpG sites.
- 3- To assess the predictive performance of HiFi-based episignatures compared with array-based biomarkers, and to identify the specific properties, requirements and technical benefits of long-read sequencing.
- 4- To deepen the understanding of methylation–phenotype correlations and of the intrinsic heterogeneity of episignatures and syndromes of interest, towards personalised medicine.
- 5- To explore the genome-wide profile of the newly accessible 6mA mark and assess its value as an independent biomarker in disorders with a known, robust 5mC episignature.

Expected results. From a methodological point of view, this project will deliver statistical and machine-learning tools tailored to epigenomic data from long-read sequencing, a largely open field. The PhD candidate will build on a unique resource of more than 1,300 array-based methylation profiles covering 50 NDD genes (Epi2Diag project), complemented by 400 PacBio HiFi genomes produced within LR2D: 100 trios (affected child and parents), 80 positive and 20 negative controls for episinature evaluation. From a clinical point of view, the project paves the way for a single first-line test combining genomic and epigenomic analyses, helping to shorten the diagnostic odyssey, improve VUS interpretation and better understand phenotypic diversity within a given syndrome.

Expected deliverables in practice:

- Scientific articles focusing on one or several episinatures depending on findings, and a broader methodological article at the end of the PhD.
- Open access to methods and tools through the development of an R package, shared with the statistical and medical genetics community.
- Sharing of genomic data with the research community in compliance with GDPR (EGA platform, French Genomic Medicine Plan data collector), for future projects and meta-analyses.

Required Skills

This PhD position is open to Master's graduates in biostatistics, statistics, bioinformatics or data science, or to graduates of engineering schools specialising in statistics (ENSAE, ENSAI, INSA, etc.). The project requires solid foundations in statistical modelling and machine learning, and good proficiency in R; experience with Python, Linux and high-performance computing is a plus. We are looking for a candidate with a strong interest in real biological and health data, especially genomic and epigenomic data, who enjoys working in a multidisciplinary team. No prior knowledge of episinatures or long-read data is required: training will be provided by the team during the first year.

Host Laboratory

Inserm UMR1245 “Genomics of Cancer and the Brain,” directed by Prof. Gaël Nicolas
University of Rouen Normandy – UFR Santé, 22 boulevard Gambetta, Rouen, France
Team 3: “Genomics of Brain Diseases”

With strong links to the Rouen University Hospital (CHU) as well as national and international networks providing access to high-quality datasets, Inserm UMR1245 is a dynamic and fundamentally multidisciplinary research unit dedicated to projects in oncogenetics and neurogenetics. It hosts the ASGARD genomics platform, a French pioneer in PacBio HiFi data production, and the Rouen University Hospital is currently among the leading laboratories in France offering episinature analysis from wetlab to drylab (on a research basis).

Within Team 3, statisticians work alongside clinicians, molecular biologists and bioinformaticians to develop statistical methods aimed at better understanding predispositions to cancer, brain diseases and neurodevelopmental disorders. The team's episinature research benefits from close collaborations with pioneers of the field in Toronto (SickKids Hospital) and with Erasmus MC in Rotterdam (developing the MethaDory platform).

Visit <https://cbg.univ-rouen.fr/> for more details.

Supervision and contact for additional information

- Camille Charbonnier (Statistician, Professor, PhD director, camille.le-clezio@chu-rouen.fr)
- François Lecoquierre (Geneticist, Associate Professor, co-supervisor, LR2D project lead, francois.lecoquierre@chu-rouen.fr)

Funding

Three-year doctoral contract (MESSIDORE 2025 / Normandy Region doctoral grant). PacBio HiFi data production is funded by the LR2D project (MESSIDORE 2025 call).

Elements of bibliography

- 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.