COLLEGE DOCTORAL BRETAGNE SCIENCES DE LA VIE ET DE LA SANTE

Université de Bretagne Occidentale

THESE DE DOCTORAT DE

L'UNIVERSITE
DE BRETAGNE OCCIDENTALE

ECOLE DOCTORALE N° 637
Sciences de la Vie et de la Santé
Spécialité : *Génétique, Génomique et Bioinformatique*

Par
Marie-Sophie OGLOBLINSKY

Statistical strategies leveraging population data to help with the diagnosis of rare diseases

Thèse présentée et soutenue à Brest, le 28 juin 2024
Unité de recherche : INSERM UMR1078 – Génétique, Génomique fonctionnelle et Biotechnologies

Rapporteurs avant soutenance :

Aurélie COBAT	CR – Institut Imagine UMR1163
Antonio RAUSELL	MCU-PH – Institut Imagine UMR1163

Composition du Jury :

Président :	Prénom Nom	*(à préciser après la soutenance)*
Rapporteurs :	Aurélie COBAT	CR – Institut Imagine UMR1163
	Antonio RAUSELL	MCU-PH – Institut Imagine UMR1163
Examinateurs :	Donald CONRAD	Associate Professor – Oregon National Primate Research Center
	Marie DE TAYRAC	PU-PH – IGDR UMR6290, CNRS, UR
	Gérald LE GAC	PU-PH – UMR1078, INSERM, UBO, EFS, CHRU
	Gaël NICOLAS	PU-PH – CBG UMR1245, INSERM, UNIROUEN
Dir. de thèse :	Emmanuelle GENIN	DR – UMR1078, INSERM, UBO, EFS, CHRU
Encdr. de thèse :	Gaëlle MARENNE	IR – UMR1078, INSERM, UBO, EFS, CHRU

COLLEGE SCIENCES
DOCTORAL DE LA VIE
BRETAGNE ET DE LA SANTE

Université de Bretagne Occidentale

Titre : Stratégies statistiques exploitant les données de la population générale pour aider au diagnostic des maladies rares

Mots clés : Maladies rares, hétérogénéité génétique, génome non-codant, digénisme

Résumé : La forte hétérogénéité génétique et les modes de transmission complexes des maladies rares posent le défi d'identifier le variant causal si un seul patient le porte, en utilisant des données de séquençage et des méthodes d'analyse standard. Pour aborder ce problème, la méthode PSAP utilise des distributions nulles par gène de scores de pathogénicité CADD pour évaluer la probabilité d'observer un génotype donné dans la population générale. L'objectif de ce travail était de répondre au manque de diagnostic des maladies rares grâce à des méthodes statistiques. Nous proposons PSAP-genomic-regions, une extension de la méthode PSAP au génome non codant, en utilisant comme unités de test des régions prédéfinies reflétant la contrainte fonctionnelle à l'échelle du génome entier. Nous avons implémenté PSAP-genomic-regions et sa version initiale PSAP-genes dans Easy-PSAP, un workflow Snakemake intuitif et adaptable, accessible aussi bien aux chercheurs qu'aux cliniciens. Appliqué à des familles touchées par de l'infertilité masculine, Easy-PSAP a permis la priorisation de variants candidats pertinents dans des gènes connus et nouveaux. Nous nous sommes ensuite concentrés sur le digénisme, le mode le plus simple de transmission complexe, qui implique l'altération simultanée de deux gènes pour développer une maladie. Nous avons décrit et évalué les méthodes actuelles publiées dans la littérature pour détecter le digénisme et proposé de nouvelles stratégies pour améliorer le diagnostic de ce mode de transmission complexe.

Title : Statistical strategies leveraging population data to help with the diagnosis of rare diseases

Keywords : Rare diseases, genetic heterogeneity, non-coding genome, digenism

Abstract : High genetic heterogeneity and complex modes of inheritance in rare diseases pose the challenge of identifying an n-of-one patient's causal variant using sequencing data and standard analysis methods. To tackle this issue, the PSAP method uses gene-specific null distributions of CADD pathogenicity scores to assess the probability of observing a given genotype in a healthy population. The goal of this work was to address rare disease lack of diagnosis through statistical strategies. We propose PSAP-genomic-regions an extension of the PSAP method to the non-coding genome, using as testing units predefined regions reflecting functional constraint at the scale of the whole genome. We implemented PSAP-genomic-regions and the initial PSAP-genes in Easy-PSAP a user-friendly and versatile Snakemake workflow, accessible to both researchers and clinicians. When applied to families affected by male infertility, Easy-PSAP allowed the prioritization of relevant candidate variants in known and novel genes. We then focused on digenism, the most simple mode of complex inheritance, which implicates the simultaneous alteration of two genes to develop a disease. We reviewed and benchmarked current methods in the literature to detect digenism and put forward new strategies to improve the diagnostic of this complex mode of inheritance.