Pr N. MEYER
GMRC, Service de Santé Publique du CHU
& Laboratoire de biostatistique
Faculté de Médecine de Strasbourg

Dr B. GERARD
Laboratoires de diagnostic génétique, UF de génétique moléculaire
Hôpitaux Universitaires de Strasbourg

Strasbourg, le 27/11/2016.

Proposition de sujet de stage pour un étudiant du Master M2 Master « Mathématiques et applications - Statistique : Biostatistique et statistiques industrielles »

2017 - 2018

L'analyse par séquençage à haut débit est une méthode d'analyse de l'ADN par séquençage massif en parallèle de très nombreux fragments (reads) de séquence. L'analyse par séquençage haut débit permet de détecter des mutations génétiques présentes à l'état hétérozygote (50 % de reads mutés) ou homozygote (100 % de reads mutés). Cette technologie permet également de détecter des fractions alléliques plus faibles, jusqu'à 1 % dans certaines conditions techniques. Cette technologie a révolutionné les approches diagnostiques dans de nombreuses pathologies : son application en génétique constitutionnelle et en cancérologie ne cesse de croitre et de nouveaux champs diagnostiques, jusque-là inaccessibles apparaissent comme par exemple le dépistage prénatal non invasif (ou DPNI). Le DPNI est un test de dépistage d'anomalies génétiques présentes chez un fœtus à partir d'un prélèvement de sang maternel, ce qui permet d'éviter la réalisation d'une ponction fœtale (soit ponction de liquide amniotique soit biopsie de trophoblaste). Ce test se base sur l'analyse de l'ADN (acide desoxyribonucleique) contenu dans le sang maternel : l'ADN plasmatique est en effet composé à 90-95 % d'ADN de la mère et de 5-10 % d'ADN fœtal. Les mutations génétiques présentes chez le fœtus se retrouvent donc à des fractions alléliques faibles dans le sang circulant de la mère. Une analyse par séquençage haut débit peut donc permettre de rechercher dans le sang maternel la présence ou l'absence de variations génétiques présentes spécifiquement chez le fœtus sans nécessité de prélèvement fœtal.

Nous souhaiterions développer dans le cadre de ce master un modèle statistique permettant d'évaluer la valeur prédictive de la détection d'une variation ou de l'absence de détection d'une variation génétique dans le plasma de sa mère. Pour développer ce modèle, nous avons des données issues de pool de 14 sujets dont on connaît le génotype : les ADN des 14 sujets sont mélangés de façon à obtenir une fraction allélique à hauteur d'environ 3 %. Ces sets de données seront utilisés dans un premier temps pour mettre en place le modèle statistique puis dans un second temps, le modèle statistique sera testé sur des sets de données issus de l'analyse de l'ADN plasmatique de femmes enceintes dans lesquels nous connaissons le génotype du fœtus (30 prélèvements, 100 points par prélèvement = 3000 points). La modélisation se fera préférentiellement avec des méthodes bayésiennes, dans le cadre d'un modèle hiérarchique, permettant de tenir compte des multiples niveaux de variabilités impliqué dans la modélisation : nombre total de reads sur la position génomique, nombre total de reads mutés sur la position génomique, nombre de sites à considérer pour estimer le taux moyen de la fraction allélique, etc, avec des lois de probabilités différentes à chaque niveau de la hiérarchie. Les paramètres devront être estimés à l'aide de méthode de type « bayésien empirique ». Enfin, le modèle devra pouvoir être capable d'apprentissage par inclusion de nouveau cas au fil de l'eau. La programmation se fera en R.

Le stage se fera probablement en trinôme avec un étudiant de master en bio-informatique et un étudiant en M2 de biologie moléculaire. Le stage sera encadré pour la partie biologique par le Dr Bénédicte GERARD et par le Dr Jean Muller pour la partie Bioinformatique du laboratoire de diagnostic génétique-UF de génétique moléculaire et pour la partie statistique par le Pr Nicolas MEYER, du Groupe Méthode en Recherche Clinique du Service de santé publique du CHU de Strasbourg.