## Introduction

### Environnement de stage

L'Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGMBC) situé à Illkirch-Graffenstaten (67) est un établissement de recherche public crée en 1992. Il rassemble plusieurs organismes de recherche tels que le CNRS, l'INSERM et l'Université de Strasbourg. Divisée en 4 départements, cet Institut de 750 salariés aborde de nombreuses thématiques se rattachant aux domaines du cancer, des maladies métaboliques, des mécanismes liés aux drogues ainsi que des maladies génétiques monogéniques. La mise en commun de technologies de pointe avec un accès facilité pour les chercheurs est assurée par l’existence de plusieurs plateformes. Celles-ci sont engagées dans une démarche qualité avec pour la plupart une certification ISO 9001, acquise ou en cours d’acquisition.

La plateforme Biopuces et Séquençage au sein de laquelle mon stage s'est déroulé est subdivisée en deux activités : six personnes sont en charge de l’analyse biologique et cinq bioinformaticiens, dont mon encadrant principale, Stéphanie Le-Gras, sont en charge de l’analyse bioinformatique (obviously). La cohésion entre les deux parties est assurée par un responsable opérationnel.

Les activités développées par la plateforme se concentrent sur des thématiques de génétiques fonctionnelles, de cellules souches ou de génétique humaine. Deux technologies majeures y sont employées : celle de puces à ADN ainsi que le séquençage nouvelle génération ou *Next Generation Sequencing* (NGS) grâce à la technologie Illumina. L’activité de séquençage est équitablement répartie entre le DNA-seq, le RNA-seq et le ChIP-seq. De nombreux projets de recherche sont soumis à l’analyse bioinformatique, ils proviennent aussi bien des équipes de recherche internes qu’externes, publiques ou privées.

(L'IGBMC est engagé au sein de 14 projets européens et assure des collaborations avec diverses compagnies pharmaceutiques.)

### Contexte scientifique

Les progrès réalisés dans le domaine de la bioinformatique sont concomitants avec l’essor des technologies de séquençage à haut débit. De nombreux outils existent pour l’analyse des données issues de ces techniques.

Dans le cadre de la détection de variants, la plateforme Biopuces et Séquençage dispose d'outils de détection de polymorphismes nucléotidique ou *Single-Nucleotide Polymorphism* (SNP). Il s'agit maintenant d’implémenter un outil pour la détection de variations structurales et plus particulièrement pour la détection de *Copy Number Variation* (CNV).

Variations structurales : delections, duplications, copy-number variants, insertions, inversions, translocations.

Cause de CNV : inequal recombination. Copie inégales de segments d’ADN.

12% du génome humain sujet à CNV. Dont la moitié se situe au niveau de séquences codantes. Impact sur le niveau d’expression du gène.

Taille moyenne : entre 1 kb et 3 Mb.

Traditional methods (de cytogénétique) such as fluorescence in situ hybridization (FISH) and array comparative genomic hybridization (aCGH) suffer from low resolution of genomic regions.

Sequencing. Outils de détection de CNVs choix entre 4 méthodes. (4 algorithms différents ?).

### Objectif du travail de stage

La littérature actuelle propose déjà de nombreux outils de détection de CNVs ainsi que des comparateurs de ces outils de détection. Une des premières étapes de ce stage consiste à évaluer les outils préexistants et à sélectionner les plus pertinents en regard aux données sources à analyser. Ces dernières sont le fruit d'un re-séquençage d'ADN de patients atteints de maladies rares. Une fois les outils de détection de grands réarrangements sélectionnés, chacun sera testé sur les données acquises *via* le séquenceur Illumina. Enfin, un protocole de détection pourra être développé puis implémenté dans le pipeline d'analyses de données.