UNIVERSITE PARIS SACLAY (Paris 11)

ECOLE DOCTORALE : (Structure et Dynamique des Systèmes Vivants)

DOCTORAT

Bioinformatique

Thèse soutenue pour l‟obtention

du diplôme de doctorat par

Jia LI

Identifier les variations conduisant au cancer dans le génome et transcriptome non codant

Thèse dirigée par : Professor. Daniel GAUTHERET

Soutenue le Lundi 14 Décembre 2015

JURY

Dr. Salvatore Spicuglia,	Rapporteur
Dr. Andrei Zinovyev,	Rapporteur
Dr. Hugues Roest Crollius,	Examinateur
Pr. Olivier Lespinet,	Examinateur
Pr. Daniel Gautheret,	Directeur de thèse

2015SACLS161

Identifier les variations conduisant au cancer dans le génome et transcriptome non codant

L'annotation fonctionnelle de mutations somatiques est un point focal des études de génomique du cancer. Jusque récemment, la recherche s'est concentré sur des mutations dans la fraction codante du génome, pour lesquelles de puissants outils bioinformatiques ont été développés afin de distinguer des mutations délétères des mutations neutres. On identifie un nombre croissant de variants associés à des maladies dans le génome non-codant. L'interprétation des mutations non-codantes dans le cancer est donc devenue une tâche urgente. Des projets de grande envergure tels que ENCODE ont rendu possible l'interprétation fonctionnelle de variants dans les cancers. Plusieurs programmes ont été produits sur la base de ces informations fonctionnelles. Ces outilssont encore limités, notamment, une bas précision de la prédiction, le manque d'information de la mutation de cancer et biais de constatation importante.

Dans le chapitre 2 de cette thèse, pour interpréter fonctionnellement les mutations non-codantes dans les cancers, nous avons développé deux modèles de forêts aléatoires indépendants, appelées SNP et SOM. Compte tenu de la combinaison de caractéristiques fonctionnelles à une position donnée du génome, le modèle SNP prédit la fraction de SNP rares (une mesure de la sélection négative), et le modèle SOM prédit la densité de mutations somatiques attendue à cette position. Nous avons appliqué nos deux modèles pour évaluer des clinvariant and HGMD variants asociés à des maladies, et un ensemble de SNP-contrôle aléatoires. Les résultats ont montré que les variants associés à des maladies ont des scores plus élevés que les SNP-contrôle avec le modèle SNP et inférieures avec le modèle SOM, confortant notre hypothèse selon laquelle la sélection négative, telle que mesurée par fraction de SNP rares et de densité de mutation somatiques, nous informe sur l'impact fonctionnel des mutations tumorales dans le génome non-codant.

Jusqu'à présent, les chercheurs ont surtout considéré les gènes protéiques comme

critiques dans l'initiation et la progression des cancers. Toutefois, des preuves récentes ont montré que les ARN non-codants, en particulier les lncRNAs, sont activement impliqués dans divers processus de cancer. Un chapitre de cette thèse est consacré à cette classe de transcripts non codants. Comme pour les gènes codants, il pourrait exister un grand nombre de lncRNAs driver de cancer. Le développement d'outils bioinformatiques pour identifier et hiérarchiser les lncRNA et autres ARN non-codants est devenu un important objet de recherche en oncologie.

La dernière partie de cette thèse est consacrée à la mise en œuvre de méthodes pour découvrir des éléments non-codants potentiellement driver de cancer. Nous avons d'abord appliqué trois outils tierces, CADD, funSeq2, GWAVA, ainsi que nos modèles SNP et SOM, pour évaluer l'impact des mutations non-codantes dans tout le génome. Pour chaque locus, nous calculons la moyenne des scores de tous les variants observés à l'aide de l'un des modèles, et nous prenons au hasard le même nombre de variants et calculons leur score moyen 1 million de fois pour former une distribution nulle et obtenir une P-valeur pour ce locus. Pour valider notre hypothèse et notre modèle de permutation, nous avons testé ce système sur 452 gènes codants et 61 lncRNA liés au cancer, en utilisant des données de mutation somatique de cancer du foie, cancer du poumon, CLL et mélanome. Nous avons constaté que les lncRNAs et gènes codants associés au cancer avaient des valeurs-P significativement plus faibles que l'ensemble de lncRNAs et gènes codant. Appliquer ce test de permutation à des lncRNAs avec cinq systèmes de notation différents nous a permis de prioriser les centaines de candidats potentiellement liés au cancer.Ces candidats peuvent maintenant être soumis à validation expérimentale.