Etude *In Silico* de l'effet enhancer de la fixation de TRF2 sur les régions extra-télomériques

Nori SADOUNI

Université Nice Sophia-Antipolis
Master Sciences de la vie & de la santé
Mention Génétique Immunité et Développement
Parcours Biologie Informatique et Mathématique BIM
2015-2016

Sous la direction du
Pr. Eric Gilson, chef de l'équipe « Télomeres, Sénescence et Cancer »
&
du Dr. Olivier Croce, Responsable service Bioinformatique IRCAN

Institut de recherche sur le Cancer et le Vieillissement (IRCAN)
CNRS UMR 7284 - INSERM U 1081 - UNS Faculté de Médecine 28 avenue de Valombrose 06107 Nice Cedex 02 - France Tel. : +33 (0)493377782 / +33 (0)493377703

Table des matières

Table des figures

Remerciements

Je remercie tout d'abord l'Université de Nice Sophia-Antipolis, qui m'a permis de pouvoir réaliser ce stage. Ayant côtoyé d'autres stagiaires au sein de l'institut, j'ai constaté qu'il était rare pour un Master 1 de pouvoir effectuer un stage d'une durée de 5 mois. Ainsi, je remercie l'Université de nous offrir une telle immersion dans le monde du travail, où l'on peut vraiment se faire une idée du poste que l'on occupera. Je tenais également à remercier les différents enseignants que j'ai pu avoir, et qui m'auront permis d'avoir les outils pour correctement appréhender ce stage. Je remercie tout particulièrement le Dr. Karine ROBBE-SERMESANT, le Dr. Patrick COQUILLARD qui a gentiment répondu à mon appel lorsque je l'ai contacté durant mon stage au sujet de tests statistiques, le Dr. Gilles BERNOT pour ces cours sur l'environnement linux qui m'auront été essentiels.

Je tenais également à remercier le Pr.Eric GILSON, pour m'avoir accueillis dans son équipe, il m'aura tout de suite mit à l'aise, je me suis vite senti intégrer dans l'équipe. Je tenais à le remercier également pour ses nombreuses explications sur le plan biologique, il m'aura permis de cerner le problème et m'aura apporter une méthode de raisonnement qui me servira par la suite.

Je remercie également grandement le Dr.Olivier CROCE, pour tout ses conseils, son aide sur l'aspect informatique, mais également biologique. Il m'aura été d'une aide précieuse. J'ai été au cours de ce stage très bien encadré. Nous avions des réunions hebdomadaires où l'on faisait le point. Je tenais également à le remercier pour sa patience, pour le temps qu'il a prit pour m'expliquer le fonctionnement de certains programmes, ses conseils dans certains scripts. Et pour tout le temps qu'il a du prendre sur son travail personnel pour m'épauler.

Un grand merci également au Dr. Julien CHERFILS pour son aide et ses conseils au cours de mon stage et pour toutes ses blagues et sa bonne humeur, mais également au Dr.Alexandre OTTAVIANI, et DJERBI Nadir, avec qui j'ai pu passé de très bon moments notamment à la retraite d'équipe à Breil.

Un grand merci aux autres stagiaires bio-informatique de l'IRCAN, Ludovic KOSTHOWA, Amandine AUDINO, Sidwell RIGADE, et Audrey DURAN, pour leurs conseils, les bons moments que j'ai pu passé avec eux, et leur soutiens au cours de mon stage.Ainsi que Floren TESSIER, ingénieur au pôle bioinformatique de l'IRCAN.

Et bien évidemment, ma famille pour leur soutien tout au long de ma scolarité, et également mon épouse, Lilia, qui a su être présente et m'épauler tout au long de la durée de mon stage.

Abstract

Telomeres are complex nucleo-protein localized at the extremities of chromosomes. This structure are essential for genome stability. Telomere dysfunction favors chromosomal instability involved in early carcinogenesis. Telomere's protein part are composed by complex of six proteins, named shelterins (TRF1, TRF2, POT1, RAP1, TIN2 and TPP1), which are associated to the telomeric DNA, protecting the telomere against recognition by DNA Damage Response (DDR) pathways. Telomeric repeat-binding factor 2 (TRF2) is frequently overexpressed in human tumours and has oncogenic properties. Studies has show that this protein bind TTAGGG motif in mammalian telomeres.

Experiments based on chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) identified extra-telomeric binding sites of TRF2. This extra-telomeric binding sites are composed by the repetition of the motif TTAGGG, these binding sites are named Interstitial Telomeric Sites : ITS. It was observed an activation of transcription of genes closed to ITS motifs.

Others biological experiments showed an enrichment of H3K27Ac closed to ITS confering a property of super enhancer for TRF2. Using bio-informatics tools the property of super enhancers of TRF2 was confirmed. For example the gene GPC6 has been founded to be involved in immunity system having a role in human tumor. This gene was identified as a target of TRF2 by biological experiments and find in parallel by in silico analysis. We made overlap between several files, and observed a colocalisation between H3K27Ac on humain genome and ITS. We suggest a possibly regulation of this gene by TRF2. We observed the same colocalisation of ITS with H3K27ac mark with other mammalians. This confirm the crucial role of TRF2 as a super enhancer.

Keywords : Bioinformatics, Telomere, TRF2, extra-telomerique sites,enhancer

1 Introduction

1.1 Structure et organisation des télomères

L'ADN est le support de l'information génétique. Il s'agit d'une structure condensée à certains stades cellulaires sous forme de chromosome. Il existe deux régions distinctes au niveau du chromosome : les centromères situés au centre du chromosome, qui ont un rôle lors de la division cellulaire et aux extrémités on retrouve les télomères. Les télomères sont essentiel dans la stabilité du génome. Ils sont fortement corrélés au vieillissement cellulaire, et à l'apparition de cancer.[4]

Ils sont constitués de divers composants moléculaires. Premièrement une composition nucléotidique un peu particulière, faite de répétitions en tandem d'un motif qui est TTAGGG.
On y retrouve également une reverse transcriptase essentielle, la Télomérase. Elle est constituée de deux sous unités, une catalytique à activité transcriptase inverse (Telomerase reverse transcriptase : TERT) et une sous unité ARN (Telomerase ARN component TERC) qui contient la séquence qui sert de matrice à la synthèse de la répétition télomérique. Elle va venir allonger le brin 3'. La télomérase est capitale dans le maintien de l'intégrité des cellules souches et se retrouve très souvent dans les cellules tumorales conférant à ces cellules un caractère immortel.[10] Une mutation de cette enzyme peut causer de nombreuses pathologies humaine, dûe à une érosion excessive des télomères. [21] On y trouve également un ensemble de 6 protéines nommées les Shelterins (TRF1, TRF2, POT1, RAP1, TIN2, TPP1).[8]
L'ensemble de ses composants forment le complexe télosome.

1.2 Rôle et mécanisme du télosome

Les protéines Shelterins se fixent à l'ADN télomérique et préviennent l'activation des voies de reconnaissance et de réparation de l'ADN endommagé (Domage DNA Reponse : DDR).[3] Parmi les protéines shelterins, TRF1, TRF2 et POT1 jouent un rôle central en conférant la spécificité de fixation du complexe sur cet ADN. TRF1 et TRF2 se fixent à l'ADN double brin et POT1 vient se fixer sur le brin 3' sortant.[21] Ils préviennent l'action des kinases ATM et ATR, qui auront pour conséquence d'activer la voie de réponse aux dommages de l'ADN. En effet, en l'absence de ces protéines Shelterins, la machinerie cellulaire détectera l'extrémité des chromosomes comme étant une cassure double brin d'ADN, et stoppera le cycle cellulaire. Egalement elle pourra activer des mécanismes de réparation, lesquels seront par exemple des mécanismes de jonction des extrémités non homologues qui permettront de restaurer la continuité de l'ADN endommagé. [9]

La longueur des répétitions télomériques décroit au fur et à mesure du vieillissement physiologique des tissus somatiques. Les télomères sont constitués d'une répétition du motif TTAGGG, cette séquence est non codante et permettra à la cellule d'éviter de perdre de l'information génétique au cours des divisons cellulaires. En effet, à chaque réplication de l'ADN, l'ADN polymerase est incapable de copier les derniers nucléotides présents sur les deux brins matrice d'ADN. Mais grâce à la zone «tampon» formée par les

télomères, ce défaut de réplication n'a pas d'effet délétère et la stabilité de l'information génétique est assurée.

Ainsi quand les télomères deviennent très courts, ils signalent à la cellule un arrêt de la prolifération, en la rendant sénescente, ou en déclenchant l'apoptose. Cet arrêt se fait *via* l'activation des voies de reconnaissance de l'ADN endommagé résultant d'une diminution critique du nombre de complexes protéiques shelterins fixés aux répétitions d'ADN télomériques. Un raccourcissement anormal de la taille des télomères et l'entrée prématurée en sénescence sont à l'origine de nombreuses pathologies humaines ; que ce soit des syndromes de vieillissement accéléré comme la dyskératose congénitale, ou des pathologies liées au vieillissement comme les pathologies cardio-vasculaires, les syndromes neurodégénératifs et le diabète de type 2. Un dysfonctionnement télomérique entrainera également une instabilité chromosomique (réarrangements excessifs) et pourra entrainer une carcinogenèse précoce.

De plus, il a été observé que la structure est sensible à une large gamme de facteur endogène et à l'environnement. Ils sont sensible, au stress oxidatif, stress hormonal, stress psychologique, alcool, caféine par exemple ou encore aux chocs thermiques. Ces différents facteurs auront un impact sur les télomères [16]

Depuis peu, de nouvelles connaissances émergent quant aux rôles des télomères. Il a été démontré qu'ils peuvent réguler la transcription. Il s'agit d'un phénomène epigénétique : Telomere Position Effect (TPE). Il y aura une modification de la configuration de la chromatine des régions situées près des télomères.[11]

Ainsi, les télomères semblent avoir un rôle important dans la régulation du cycle cellulaire, mais également comme senseur de stress et prédicateur d'espérance de vie.

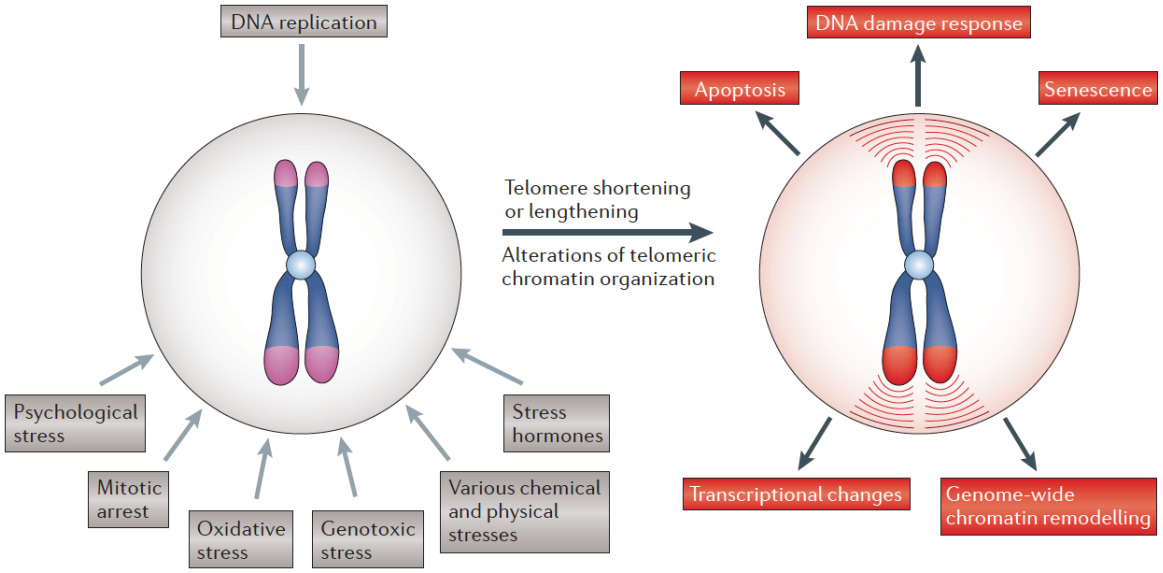

FIGURE 1 – Les télomères peuvent être vus comme des antennes émettrices ou réceptrices de signaux. Ils peuvent intégrer une large gamme de facteurs endogènes ou exogènes (Cases grises). Des modifications résulteront de ses stimulus (Cases rouges). Ye et al. 2010

1.3 Les protéines Shelterins

Les mécanismes de fixations des protéines Shelterins sur les télomères sont plutôt bien connus. Mais il y a encore peu de données quant à l'effet de la fixation de protéines Shelterins sur des zones extra-télomeriques. Il a été découvert lors d'expérience de ChIP-seq, qui est une technique permettant d'étudier les interactions ADN/protéine à l'échelle du génome, que certaines protéines Shelterins pouvaient se fixer en dehors des télomères. Ils se fixent sur le motif TTAGGG, motif répété au minimum quatre fois. On nomme cette répétion de motif Interstital Telomeric Sites : ITS. [21][17]

Par exemple, RAP1 a été observée au niveau de la chromatine dans le génome de souris ou dans des cellules cancéreuses humaines. Quant à l'effet de cette fixation, il a été suggéré que RAP1 dans les cellules de souris était impliqué dans la formation de la forme hypercondensé de la chromatine : l'hétérochromatine. RAP1 serait également impliqué dans des processus métaboliques et aurait également un effet sur des gènes intervenant dans l'immunité.[14] Les protéines Shelterins auraient donc un rôle dans la régulation de l'expression de gènes qui sont à proximité. Leurs sites de fixations étant fréquemment dans un gène ou proche de celui ci.[21]

1.4 Telomere Repeat-binding Factor 2 : TRF2

Nous allons maintenant nous intéresser aux rôles de TRF2. Il a été démontré qu'une sur-expression de TRF2 au niveau des ITS est très souvent observée dans les lignées tumorales et contribue à l'oncogènese. Par exemple il a été observé qu'un sur-dosage de TRF2 entraine une surpression de HS3ST4 qui code pour un heparane sulphate (glucosamine). Ce fort niveau de concentration de TRF2 dans les cellules tumorales va, *via* HS3ST4, réduire la faculté des cellules à recruter des cellules du système immunitaire les «Natural Killers» : NK.[2] Un effet dose dépendant reversible est observé. Ainsi une sur ou sous regulation de TRF2 dans les cellules tumorales affectent fortement la tumorigénicité sans activation de la DDR.[2]

D'autres études vont dans le même sens en démontrant l'aspect oncogène de TRF2. TRF2 est sur-exprimée dans la vascularisation de nombreux cancers. Elle est activée par le transcrit de Wilms Tumor suppressor : WT1. Elle va ensuite sur-activer la transcription du gène PDGFRß. Ainsi TRF2 va favoriser la migration, prolifération et formation des cellules endothéliales chez les tumeurs.

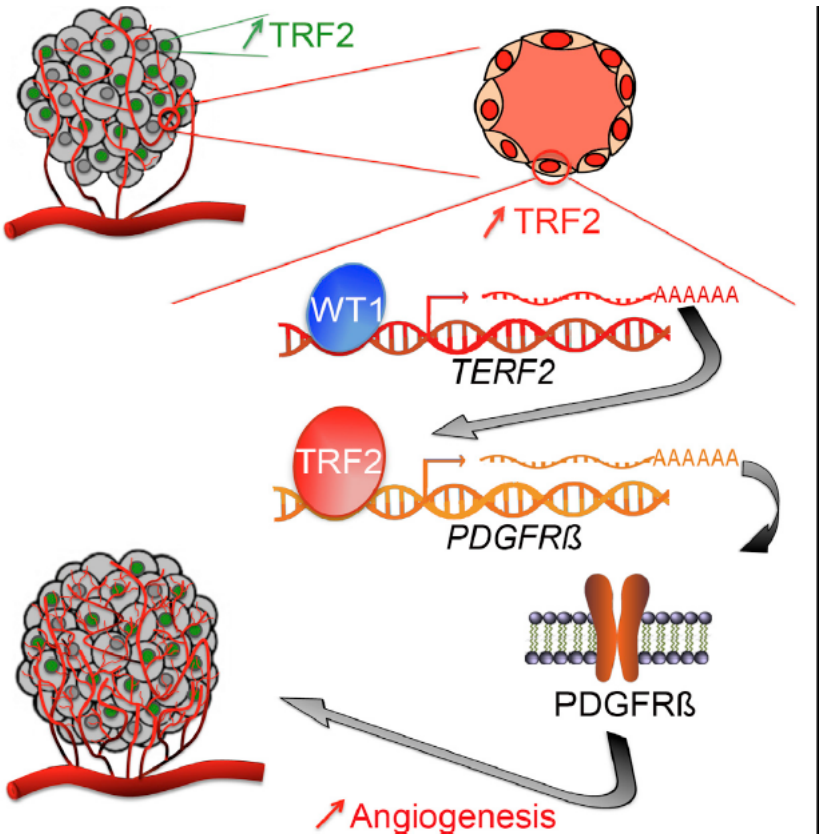

FIGURE 2 – TRF2 est retrouvé sur-exprimé dans les cellules endothéliale des tumeurs. WT1 sur-exprime TRF2 qui sur-exprimera PDGFRß qui favorisera l'angiogénèse.El Mai 2014

Mais l'effet de la fixation de TRF2 au niveau extra-télomérique restant peu comprise, il est nécessaire de connaitre les différentes cibles de cette protéine, et l'effet de cette protéine sur les gènes à proximité du site de fixation. [17]

Ainsi, l'étude des protéines Shelterins est une voie à explorer dans la recherche sur le cancer. Il s'agit là d'une partie de la recherche de l'équipe d'Eric GILSON dans laquelle j'ai réalisé mon stage.

La thématique de recherche de l'équipe d'Eric GILSON se porte sur les télomères. Les objectifs principaux étant de comprendre leur fonctionnement au niveau cellulaire, ainsi qu'identifier et caractériser les rôles des télomères, de la télomérase, et des protéines Shelterins, dans la transformation maligne des cellules humaines. Mais également leurs rôles sur la structure de la chromatine, leurs effets et l'impact du vieillissement sur ce complexe. Ce qui situe parfaitement cette équipe au sein de l'IRCAN qui est l'institut de recherche sur le cancer et le vieillissement.

Au sein du pôle bio-informatique de cette institut et encadré par le Dr.Olivier CROCE, j'ai apporté ma contribution à cette thématique de recherche avec des outils bio-informatique.

2 Objectif

Il s'agira d'étudier l'effet de la fixation de TRF2 sur les ITS et comprendre comment elle peut favoriser l'expression de ces gènes cibles. Dans un premier temps d'analyser des données issus de ChIP-seq sur TRF2. L'objectif est de comprendre et pouvoir réaliser les différentes étapes d'analyse de données issues de l'expérience. Une fois ces notions acquises, sont à déterminer les gènes à proximité des régions ciblées par la protéine et détectées lors de l'expérience. En effet ces gènes sont potentiellement régulés par la fixation de TRF2, qui agirait comme facteur de transcription.

Dans un second temps, il s'agira de croiser des données issus d'une équipe de recherche avec des données issues de l'équipe d'Eric GILSON. Il sera donc nécessaire de traiter les données, trier (=parser) des fichiers, et les analyser. Le but étant de faire ressortir des points communs entre les différentes données, la position des sites de fixation de la protéine d'intérêt : TRF2, avec des zones de transcriptions actives. Ce qui suggéra fortement un effet activateur de TRF2 au niveau de la transcription, et de plus qui permettra de faire ressortir les gènes régulés par ce dernier. L'étude sera poussée à différentes espèces de mammifères dans le but de valider l'effet activateur de transcription de TRF2.

3 Matériels et Méthodes

3.1 Analyse données de ChIP-Seq sur TRF2

Dans un premier temps, il s'agit d'analyser des données issues de ChIP-seq ciblant TRF2, la protéine Shelterin d'intérêt.
Pour ma part, j'ai du apprendre les différentes étapes d'analyse du ChIP-seq,[1] et à manipuler les divers logiciels, et mon analyse débute après les différentes phases de pré-traitement. En effet, j'ai du compléter les analyses des données issus de ChIP-seq. Afin de déterminer les gènes à proximité des sites de fixation de la protéine Shelterins TRF2.
Lors de cette analyse il s'agissait d'être moins stringent et d'utiliser des outils plus modernes pour réaliser cette recherche.
Pour cela, j'ai écrit un script en python, afin d'élargir les positions des pics. En modifiant une option du script, on peut élargir les séquences aux distances que l'on souhaite. Le script prend en compte la taille du chromosome grâce à un fichier téléchargeable sur UCSC et qui contient la taille de chaque chromosome. Ceci dans le cas où l'on élargirait trop les séquences. Ainsi on bornera les distances à la taille maximale du chromosome. Idem pour une position trop proche de zéro, si après modification la valeur est négative, elle sera bornée à zéro.

3.2 Croisement des motifs ITS avec des marques de transcription active connues chez l'Homme

Dans un deuxième temps, il s'agit de croiser des fichiers. Toutes les expériences ont été menées sur le génome humain sur l'assemblage Hg19, ce qui a nécessité la normalisation de certains fichiers.
Grâce à un script réalisé par mon tuteur le Dr.Olivier CROCE, les ITS ont pu être identifiés sur le génome humain. Ces résultats ont été croisés avec des fichiers issu d'un ChIP-seq sur H3K27Ac et H3K4me3 représentant des marques de transcription active. Ainsi j'ai à ma disposition plusieurs fichiers bed. On y trouve systématiquement le numéro du chromosome dans lequel est contenue une séquence, ensuite la position "start" et "end" de cette séquence. Il s'agit là du fichier BED type, il existe ensuite différents variants. Une liste est consultable directement sur UCSC. Ces fichiers sont croisés successivement *via* Bedtools.[15] Avec l'option de Bedtools : Intersect, les paramètres sont par défauts pour identifier les chevauchements entre les différentes positions des séquences. Les fichiers seront analysés deux à deux. Je dispose de cinq fichiers bed :

- ITS de Novo : position des ITS sur le génome.
- ChIP-seq H3K27Ac : 40.801 séquences[18]
- ChIP-seq H3K4me3 : 12.035 séquences
- ChIP-seq TRF2 : données issues de la publication de Thomas Simonet, Equipe Eric GILSON [13]
- ChIP-seq TRF2 : Yang, D et al.[20]

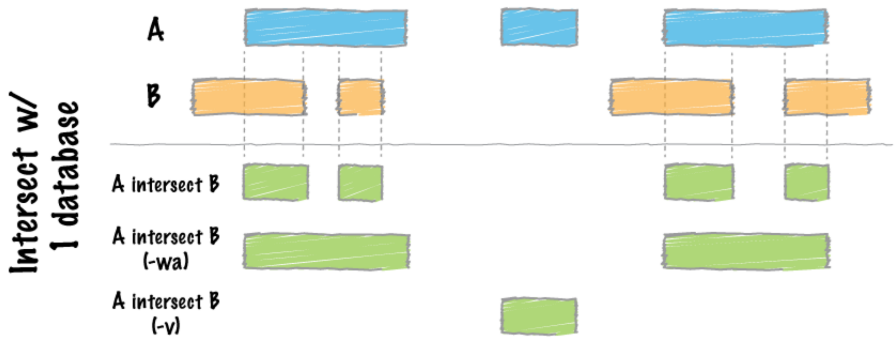

FIGURE 3 – Schéma reprenant le mode de fonctionnement de Bedtools Intersect, on récupère un fichier reprenant les chevauchements.

Les zones en communes des différents fichiers seront mis en avant. Les fichiers ont été préalablement traités (= parsés) avec des scripts personnels codés en python afin de normaliser les différents fichiers. Bedtools ne prenant qu'un type bien défini de fichier, les différentes colonnes doivent être au format prit par ce dernier. D'où la nécessité d'utiliser les scripts pour normaliser les fichiers.

En parallèle, j'ai utilisé Pybedtools[7] qui est similaire a Bedtools mais codé en python, qui peut donc être facilement intégrable à des scripts personnels, de plus il possède plus d'options comme par exemple Randomstats qui est très interessante.

Le module Randomstats nous permet de voir si les résultats obtenus sont significatifs. Il prend une des deux séquences que l'on envoie dans le programme, il mélange le fichier un nombre n de fois, et à chaque itération identifie les chevauchements(= overlaps) entre les deux fichiers. J'ai choisi un nombre d'itérations égal à 1000 afin d'avoir une p-value de 0.001.

Des deux programmes j'obtiens un fichier bed contenant une liste des chevauchements entre les deux fichiers. Les positions correspondent aux bornes inférieures du chevauchement. Depuis ce fichier, on peut déterminer les gènes à proximité de ces zones de chevauchements, d'après un fichier où les ITS du génome humains ont été annotés *via* la base de données UCSC.

3.3 Comparaison avec diverses espèces de mammifères

Afin d'établir les comparaisons, des analyses similaires ont été effectuées sur le génome d'autres espèces de mammifères. En effet, l'équipe ayant réalisé les ChIP-seq sur H3K27Ac et H3K4me3 sur le génome humain a également réalisé des ChIP-seq sur vingt autres espèces de mammifères.[18]
Les génomes des différentes espèces de mammifères pour lesquels on trouve une annotation sur UCSC ou sur un programme d'annotation de génome nommé PAVIS[12] sont téléchargés.
Ils sont souvent téléchargable par chromosome, ainsi ils sont concatenés *via* un script en Bash. Ceci afin d'obtenir un fichier par mammifère. Ensuite, *via* la commande makeblastdb[5], je crée une base de donnée (database) BLAST sur le serveur interne de l'IRCAN pour les différentes espèces. La commande ne prend qu'un fichier, c'est là la raison de la concaténation des chromosomes. Une fois cette BLAST database créée, *via* blastn je vais rechercher le motif TTAGGG répété au moins quatre fois. Je récupère en

sortie une liste de positions correspondant aux ITS.
Cependant, blastn renvoie en fichier de sortie, un fichier annoté des premiers caractères contenus dans le génome. S'agissant d'un fichier fasta, la première ligne est un descriptif et donc on y trouve souvent un identifiant renvoyant à un chromosome de l'espèce en question. Pour la suite de l'analyse mes scripts se basent sur la comparaison entre chromosome. J'ai donc écrit un script remplaçant dans la premiere ligne de description du fichier fasta de chaque séquence l'identifiant par le numéro du chromosome.
Il y a ensuite un filtre appliqué à ces résultats, *via* un script python du Dr.Olivier CROCE, les ITS ayant une p-value trop faible seront éliminés, ou ceux étant trop proche (distance de 400 bases, critère personnel), seront fusionnés afin d'être comptabilisé comme étant un seul et même ITS. Ces fichiers contenant les positions des ITS pour les différentes espèces seront croisés avec les ChIP-seq H3K27Ac et H3K4me3 respectivement réalisé sur chaque espèce. Il en ressortira comme pour l'Homme une liste de sites se chevauchant, et pour ces zones de chevauchement, je ressors les gènes à proximité de cette zone de fixation de la protéine TRF2 (ITS).
Les génomes sont annotés *via* UCSC, en passant par Table browser, il faut sélectionner le génome d'intérêt en tenant en compte l'assemblage. Les différents paramètres sont :

— Group : Gene and gene predictions.
— Track : Refseq Genes
— Region : on met en input le fichier contenant la liste des ITS

Get output, et on obtient en fichier de sortie un fichier annoté.
Certains génomes sont également annoté *via* PAVIS[12], qui est un programme d'annotation de génome, il est très intuitif et permet une visualisation simple. Il est possible de choisir le génome d'assemblage, et une option très interessante : Upstream/downstream, permet de choisir la zone d'élargissement souhaitée.

3.4 Création d'un réseau génétique pour identifier les processus biologiques des gènes cibles de TRF2 chez *Mus Musculus*

J'ai utilisé NetworkAnalyst[19] pour ensuite observer les processus biologiques dans lesquels étaient impliqués les gènes à proximités des ITS. L'étude a été faite sur le génome de la souris(GRCm38/mm10). L'analyse a été faite sur la souris car il s'agit du génome pour lequel le taux d'annotation est le plus fort, comparativement à Bos Taurus, Canis Lupus. Ce programme se base sur des analyses statistiques afin de donner des résultats significatifs, ainsi que sur des bases de données tel que KEGG ou Reactome. J'ai utilisé la base de donnée Reactome[6], base de donnée regroupant les processus biologiques, elle recouvre de nombreuses informations telles que les voies de signalisation, les réactions biochimiques, etc.

Ainsi avec NetworkAnalyst on peut mettre en interaction une liste de gènes dans le but de faire ressortir les processus biologiques (=pathways) dans lesquels sont impliqués ces derniers. Ainsi il faut spécifier l'espèce étudiée, et le type de données en entrée. Par exemple ID Uniprot, ou Ensembl Gene ID.
J'ai sélectionné «Official Gene Symbol», car j'entre directement les noms des gènes.

Il y a ensuite diverses options, il est possible d'étendre l'étude en sélectionnant une option permettant d'ajouter des intermédiaires pour pouvoir relier des gènes initialement non joint. Le but était de voir les interactions basiques, ainsi, il n'a pas été sélectionné d'interactant secondaire.

4 Résultats

4.1 Gènes cibles de TRF2

J'ai ré-analysé des données issues d'un ChIP-seq sur la protéine Shelterin TRF2. L'analyse précédemment réalisée avait fait ressortir une liste de gènes dont certains d'un grand intérêt pour l'équipe de Eric GILSON. Ces gènes sont par exemple VCAN, LCN2, HS3ST4. En étant moins stringent, on a pu ressortir une liste de gènes plus fournie que préalablement.

4.2 Colocalisation des ITS de novo avec des marques "enhancers"

Le but était ici de déterminer une colocalisation des ITS (rappel : Interstital Telomeric Site, correspond au motif TTAGGG répété au moins quatre fois) au sein des marques de transcription. Donc avec bedtools le croisement des régions ITS avec les positions des H3K27Ac et H3K4me3 a été effectué.

Après analyse il ressort des chevauchements entre ces différents fichiers.

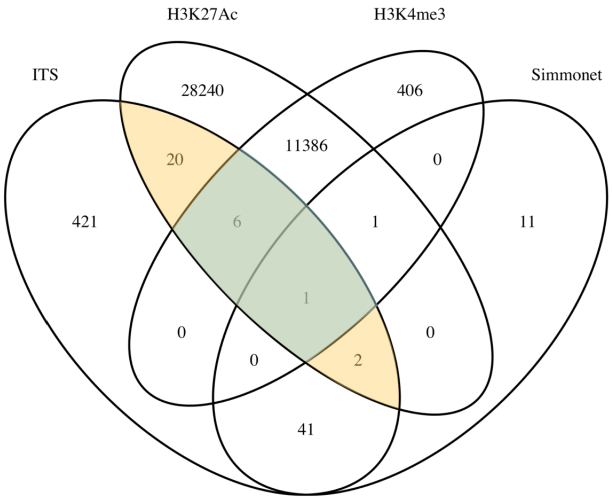

FIGURE 4 – Généré par Pybedtools. Représente le nombre de chevauchement entre les différents fichiers, en jaune les chevauchements entre les ITS et H3K27Ac (représentant les enhancers). En vert les chevauchements entre les ITS et H3K4me3(représentant les promoteurs). Un second diagramme a été généré avec le second ChIP-seq

On observe une forte colocalisation de H3K27Ac avec H3K4me3. La plupart des séquences issues de l'expérience de Simonet [17] se chevauchent avec les ITS. Vingt ITS se chevauchent strictement avec H3K27Ac et aucun strictement avec H3K4me3, cependant sept se chevauchent avec H3K4me3 quand celui ci est colocalisé avec H3K27Ac.

Ces résultats ont mit en évidence une liste de sites de fixations où TRF2 peut se fixer et agir en tant qu'activateur transcriptionnel.
Bedtools Intersect renvoie pour chaque analyse un fichier comportant les régions se chevauchant. Les ITS sont annotés avec des identifiants, ainsi pour chaque ITS concerné par un chevauchement, les gènes à proximité de cette ITS sont récupérés depuis un fichier comportant l'annotation des ITS. Ils ont été préalablement annoté *via* la base de donnée UCSC.
En sortie un fichier tabulé est généré, il contient les ITS, leurs positions, et les gènes à proximité. Afin d'avoir une vision assez large, une marge de plus ou moins 100kB a été laissé à proximité des ITS. Dans le fichier, les gènes sont disposés de la façon suivante : premier gène rencontré, gènes compris dans l'intervalle ±50kB, gènes compris dans l'intervalle ±100kB.

Chro	Start	End	ID_ITS	First Gene	ITS +/- 50kb	ITS +/- 100kb
13	94944511	94944531	ITS_95	**GPC6**	GPC6	GPC6
12	121634465	121634486	ITS_119	**P2RX7**	RP11-340F14.6 CAMKK2 P2RX4 P2RX7	RP11-340F14.6 CAMKK2 P2RX4 P2RX7 SNORA70
12	125222307	125222326	ITS_121	**SCARB1**	RP11-13J12.2 SCARB1	RP11-13J12.2 SCARB1 RP11-83B20.10
11	1867061	1867086	ITS_132	**TNNI2**	SYT8 MIR7847 LSP1 AC051649.12 TNNI2 PRR33 MIR4298	SYT8 MIR7847 AC068580.5 IFITM10 AC068580.1 AC068580.6 LINC01150 LSP1 AC068580.7 CTSD TNNT3 AC051649.12 TNNI2 PRR33 RP11-295K3.1 MIR4298
11	33156759	33156777	ITS_136	**CSTF3**	Y_RNA TCP11L1 CSTF3-AS1 CSTF3	AC131263.1 Y_RNA TCP11L1 CSTF3-AS1 CSTF3 LINC00294
11	114146363	114146497	ITS_146	**NNMT**	AP002518.1 NNMT ZBTB16 RP11-64D24.2	RP11-64D24.4 AP002518.1 NNMT ZBTB16 RP11-64D24.2
10	26833346	26833364	ITS_162	**APBB1IP**	RNA5SP307 RP11-128B16.3 APBB1IP LINC00264	APBB1IP RP11-128B16.3 RNA5SP307 LINC00264 RP13-16H11.7
17	45781989	45782014	ITS_192	**TBKBP1**	TBX21 KPNB1 RP11-138C9.1 TBKBP1	TBX21 KPNB1 RP11-138C9.1 RP11-580116.2 NPEPPS TBKBP1
17	80286391	80286413	ITS_196	**SECTM1**	TEX19 SECTM1 RP13-516M14.4 CD7 RP13-20L14.2 RP13-516M14.1 UTS2R	TEX19 OGFOD3 CSNK1D Y_RNA HEXDC-IT1 SLC16A3 RP13-20L14.1 RP13-20L14.4 SECTM1 MIR6787 RP13-516M14.10 RP13-516M14.4 CD7 RP13-20L14.2 RP13-516M14.1 UTS2R HEXDC
16	790561	790582	ITS_202	**MSLN**	METRN LA16c-380A1.2 LA16c-380A1.1 HAGHL CCDC78 FAM173A MIR662 LA16c-313D11.12 MSLNL NARFL RPUSD1	LA16c-313D11.9 WDR90 RHOT2 LA16c-335H7.2 GNG13 STUB1 JMJD8 FAM173A LA16c-313D11.13 LA16c-313D11.12 WDR24 MSLNL NARFL AL022341.3 FBXL16 PRR25
16	88946666	88946668	ITS_214	**CBFA2T3**	TRAPPC2L GALNS RP11-830F9.5 CBFA2T3 PABPN1L	TRAPPC2L RP11-830F9.7 RP11-830F9.6 RP11-830F9.5 PABPN1L CDT1 PIEZO1 GALNS AC092384.1 CBFA2T3 APRT
15	42243201	42243234	ITS_220	**EHD4**	EHD4-AS1 EHD4 PLA2G4E-AS1 CTD-2382E5.6 PLA2G4E	EHD4-AS1 EHD4 PLA2G4E-AS1 RNA5SP393 CTD-2382E5.2 RP11-23P13.6 CTD-2382E5.6 SPTBN5 MIR4310 PLA2G4E
14	49573301	49573325	ITS_227	**N/A**	RP11-816J8.1	RP11-816J8.1
19	7723061	7723082	ITS_237	**STXBP2**	MCEMP1 FCER2 CAMSAP3 TRAPPC5 PCP2 RETN MIR6792 CTD-3214H19.16 CTD-3214H19.4 CTD-3214H19.6 PET100	MCEMP1 CD209 CTD-3214H19.6 FCER2 CLEC4G CAMSAP3 TRAPPC5 PCP2 RETN CTD-3214H19.4 CTD-3214H19.16 PNPLA6 PET100 STXBP2 MIR6792 XAB2
19	17666248	17666270	ITS_239	**COLGALT1**	COLGALT1 CTD-3131K8.3 CTD-3131K8.2 PGLS FAM129C UNC13A SLC27A1	COLGALT1 CTD-3131K8.3 CTD-3131K8.2 CTD-2521M24.10 CTD-3149D2.2 PGLS AC010319.1 FAM129C UNC13A NXNL1 CTD-3149D2.3 SLC27A1
19	33572734	33572756	ITS_241	**GPATCH1**	RHPN2 AC008521.1 GPATCH1	LRP3 AC008521.1 RHPN2 CTD-2540B15.10 WDR88 GPATCH1
19	58400809	58400823	ITS_245	**CTD-2583A14.**	CTD-2583A14.9 CTD-2583A14.8 ZNF587B ZNF587 ZNF418 ZNF417 CTD-2583A14.10 ZNF814	C19orf18 ZNF606 CTD-2583A14.9 CTD-2583A14.8 ZNF586 ZNF256 ZNF552 ZNF587B ZNF587 ZNF418 ZNF417 CTD-2583A14.10 ZNF814
18	61642812	61642836	ITS_258	**SERPINB8**	SERPINB8 HMSD SERPINB10	SERPINB8 SERPINB2 HMSD SERPINB10
18	74821009	74821030	ITS_261	**MBP**	RP11-4B16.3 RP11-4B16.4 MBP RP11-4B16.1	MBP RP11-751H17.1 RP11-4B16.4 RP11-4B16.3 RP11-4B16.1
22	50640291	50640309	ITS_288	**SELO**	RP3-402G11.28 SELO TRABD RP3-402G11.26 RP3-402G11.27 RP3-402G11.25 MAPK12 HDAC10 PANX2 MOV10L1 TUBGCP6	**PLXNB2** RP3-402G11.28 SELO TRABD RP3-402G11.26 RP3-402G11.27 RP3-402G11.25 AL022328.1 MAPK11 MAPK12 HDAC10 PANX2 MOV10L1 TUBGCP6
21	15915197	15915219	ITS_305	**SAMSN1**	SAMSN1 SAMSN1-AS1 AF165138.7	SAMSN1 SAMSN1-AS1 AF165138.7
6	2204329	2204348	ITS_337	**GMDS**	GMDS-AS1 GMDS	GMDS-AS1 GMDS
6	2633772	2633790	ITS_338	**C6orf195**	C6orf195 MYLK4 RP11-145H9.3	C6orf195 MYLK4 RP11-145H9.3 RP11-299J5.1
5	172208999	172209019	ITS_377	**DUSP1**	Y_RNA DUSP1 RP11-779O18.3 RP11-779O18.1 RP11-536N17.1	NEURL1B RP11-779O18.3 RP11-779O18.1 Y_RNA RP11-536N17.1 CTB-79E8.2 ERGIC1 DUSP1
4	698764	699147	ITS_381	**PCGF3**	PCGF3 RP11-1191J2.2 MFSD7 RP11-1191J2.5 PDE6B ATP5I MYL5 RP11-1191J2.4	PCGF3 RP11-1191J2.2 AC139887.4 MFSD7 RP11-1191J2.5 PDE6B RP11-440L14.4 ATP5I MYL5 RP11-440L14.1 RP11-1191J2.4 CPLX1
2	114360430	114361054	ITS_428	**RABL2A**	RABL2A FAM138B RP11-395L14.18 MIR1302-3 AL078621.1	AC017074.1 RP11-395L14.3 AC017074.2 AL078621.1 FAM138B RABL2A RNU6-744P MIR1302-3 RP11-395L14.18 PGM5P4-AS1
1	28650619	28650638	ITS_464	**MED18**	MED18 PHACTR4 SESN2	PHACTR4 DNAJC8 SESN2 Y_RNA RP5-1092A3.4 RP5-1092A3.5 MED18 ATP1F1
9	137102218	137102251	ITS_509	**WDR5**	LL09NC01-139C3.1 RNU6ATAC	LL09NC01-139C3.1 WDR5 RNU6ATAC RP11-145E17.2 RP11-145E17.1 LL09NC01-251B2.3
X	66746096	66746124	ITS_78	**AR**	AL049564.1 AR	AL049564.1 AR

FIGURE 5 – liste des ITS se chevauchant avec H3K27Ac. Des tableaux identiques ont été généré pour toutes les conditions (intersection du diagramme de Venn)

Le test statistique effectué *via* Pybedtools Randomstats, confirme la significativité des

résultats. En effet, pour chaque analyse effectué, j'ai utilisé le module Randomstat de Pybedtools avec un nombre d'itération égale à 1000 pour un p-value de 0.001. Un fichier est renvoyé en sortie contenant les résultats. Il précise le nombre d'itération effectué et calcule un score reflétant le nombre moyen de chevauchement identifié à chaque itération. Pour chaque analyse de chevauchement, à une p-value de 0.001, 100% des chevauchements préalablement trouvés sont confirmés.

4.3 Confirmation des liens entre ITS et marque enhancers chez les mammifères

Dans la continuité les mêmes analyses ont été réalisées sur le génome d'autres espèces de mammifères tel que la souris (Mus Musculus), la vache (Bos Taurus), ou encore le chien (Canis Lupus). Une étude de chevauchement des séquences est réalisée dans le but de determiner une colocalisation entre marque de transcription active et sites de fixation de TRF2.
Le taux d'annotation des ITS chez les mammifères est de 90%. En comprenant l'intervalle ±100kB. Ensuite de la même façon que chez l'Homme, les données ont été croisées avec des marques de transcriptions actives. Donc les ITS avec la positions de H3K27ac et H3K4me3. Les résultats sont sous forme de fichier tabulé avec les régions se chevauchant, et les gènes à proximité.
On observe comme chez l'Homme un taux plus élevé de colocalisation des ITS avec H3K27Ac seul qu'avec H3K4me3 et même quand celui-ci est colocalisé avec H3K27Ac.

Espèces	ITS / H3K27Ac *Enhancers*	ITS / H3K4me3 *Promoteurs*
Mus Musuclus Souris	**36**	6
Bos Taurus Vache	**17**	7
Oryctolagus cuniculus Lapin	**19**	1
Rattus Norvegicus Rat	**37**	18
Canis Lupus Chien	**18**	7
Rhesus Macaque Macaque	**16**	3
Sus Scrofa Sanglier	**66**	28
Callithrix jacchus Singe - Marmoset	**26**	5

FIGURE 6 – Bilan des chevauchements chez les mammifères ; Pour les différentes espèces de mammifères, le nombre de chevauchements entre le fichier contenant les positions des ITS et les marques de transcription active

4.4 Processus biologiques des gènes associés aux ITS chez *Mus Musculus*

Nous nous sommes intéressés *via* NetworkAnalyst[19] aux processus biologiques dans lesquels sont impliqués les gènes à proximités des ITS. A été observées les principales voies dans lesquels étaient impliqués les gènes à proximité des ITS.
Pas tous les gènes listés ont pu être mit un interaction. N'ayant pas prit d'intermédiaire réactionnel, le réseau est limité au gène fourni. Ainsi sur les 950 gènes listés, le plus gros réseau génétique ayant été crée regroupe 351 gènes, et au sein de réseau génétique il ressort que le premier processus biologique dans lequel sont impliqués les gènes est le système immunitaire. Ressort fortement aussi la structure de la chromatine.

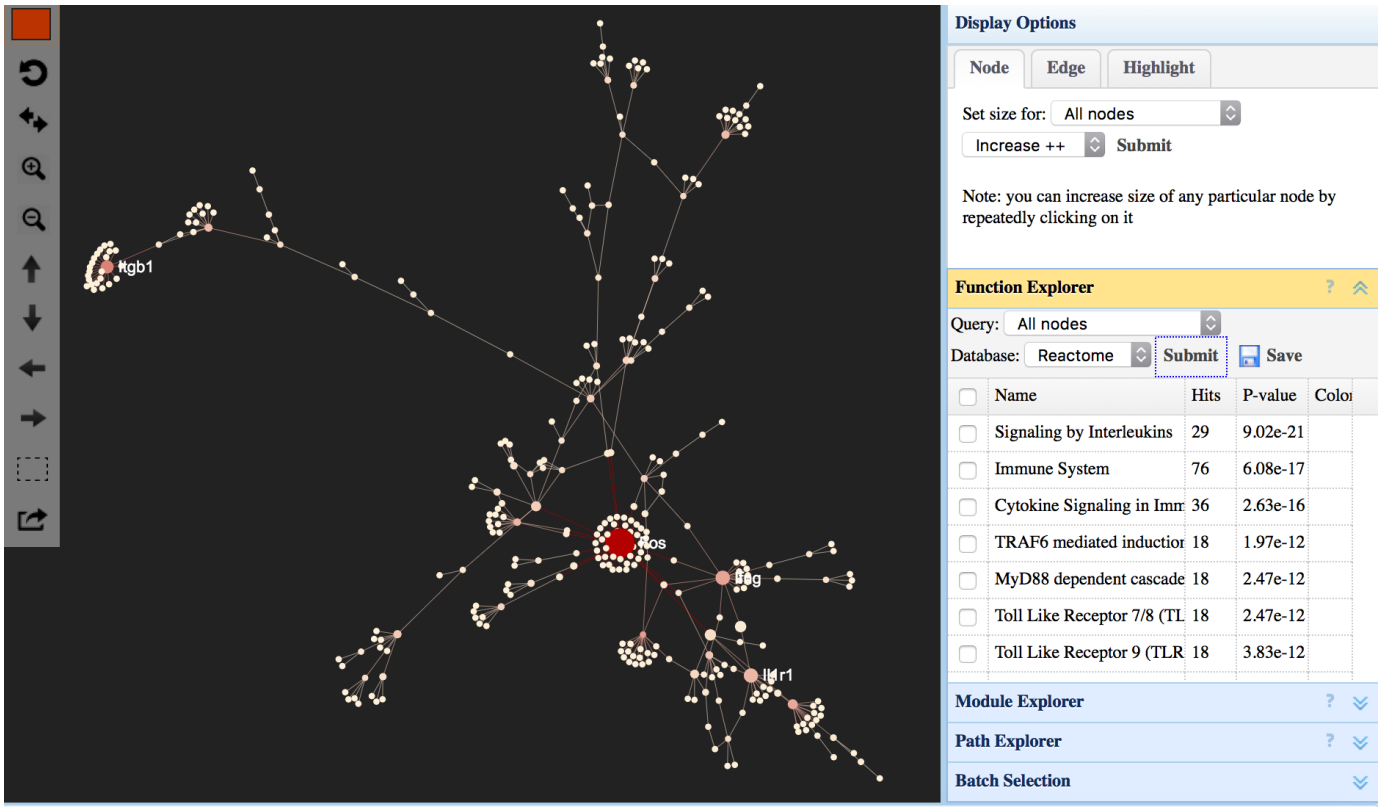

FIGURE 7 – Processus biologique des gènes à proximités des ITS.
Le graphique a été obtenu avec Networkanalyst[19], il met en relation les gènes(point jaune) fournis au logiciel et fait ressortir les processus biologiques impliqués de ces derniers. Le premier processus biologique est "système immunitaire", elle possède une p-value très significative.

5 Discussion

5.1 Critères moins strictes de l'analyse de ChIP-seq sur TRF2

Certains gènes retrouvés sont connus comme étant oncogène par exemple HS3ST4 dont on connait l'effet sur le système immunitaire. [2] Cependant, un gène intéressait tout particulièrement l'équipe, le gène de la mucine : MUC1, il aurait également un rôle oncogène, cependant il n'a pu être retrouvé dans cette séquence. Même en étant moins stringent.
En ayant été moins stringent, nous n'avons pas prit le risque d'obtenir des résultats non significatif. Ainsi, cette piste n'aura pas eu les résultats escomptés. Mais les résultats étant très dépendant des conditions (lignée cellulaire par exemple), il faudrait peut être utiliser une autre lignée afin d'identifier les autres gènes cibles potentiel de TRF2.

5.2 Correspondance entre sites ITS et H3K27Ac

On observe une forte colocalisation de H3K4me3 avec H3K27Ac. La colocalisation de ses derniers représente les promoteurs. Bio-informatiquement on retrouve les memes données que Yang et al.[18]
On observe également une bonne colocalisation des ITS avec les données issues des expériences de Simonet.[17] Cependant, comme on le rappelle, les résultats sont très dépendant de la lignée cellulaire utilisée. Ainsi, certaines conditions permettront la fixation de TRF2, d'autre non. On observe que 77% des positions TRF2 se colocalisent bien avec les sites ITS.
H3K27Ac représente les enhancers[18]. La quasi totalité des ITS se chevauchant avec ces marques de transcription active sont colocalisés avec des enhancers. TRF2 semble avoir une préférence marqué pour les sites enrichies en H3K27Ac.
Parmi les sites se colocalisant avec les enhancers se trouve un gène d'un grand intérêt pour l'équipe : GPC6, il s'agit d'un protéoglycan, il est exprimé à la surface des cellules et permet une interaction avec le système immunitaire.
Parallèlement à ces analyses bio-informatique, des ChIP-seq ont été menées sur TRF2 lors de la durée de mon stage par l'équipe. Il a été observé que lorsque la zone de fixation de TRF2 (ITS) est enrichie en H3K27Ac (marque enhancer), on note une forte augmentation de l'expression de GPC6. Ces résultats convergent et vont dans le même sens, et renforcent l'idée que TRF2 agit en enhancer.

Les ITS se chevauchant avec H3K4me3 quand ce dernier est colocalisé avec H3K27Ac représentent les promoteurs. La préférence de colocalisation semble moindre, mais elle est existante.

Cependant, les fichiers H3K27Ac et H3K4me3 ont été générés à partir de cellules hépatiques. Il aurait été très intéressant de croiser les données des ITS avec d'autres lignées cellulaires humaine. En effet, le besoin en régulation transcriptionnelle diffère selon les lignées cellulaires et donc les zones enrichies en H3K27Ac ou H3K4me3 seront différentes. Ceci pourrait nous permettre lorsque l'on croise ces fichiers avec les sites ITS, d'analyser d'autres régions et ainsi, d'observer la présence d'autres gènes.

5.3 Similarités des marques enhancers chez différentes espèces de mammfères

Les fichiers reprenant les positions des ITS ont été croisés avec les fichiers issus de ChIP-seq sur H3K27Ac et H3K4me3, et ceci sur divers espèces de mammifères.
Il en ressort que pour toutes les espèces, les sites ITS sont plus colocalisés avec les H3K27Ac qu'avec H3K4me3. Ce résultat avait déjà été obtenu chez l'homme, et l'obtenir chez plusieurs espèces de mammifère renforce l'idée que TRF2 se fixant sur les ITS agirait comme facteur enhancer. Sachant qu'environ 40% des positions H3K27Ac sont colocalisés avec H3K4me3, il est intéressant de voir qu'au sein de plusieurs espèces de mammifères, les ITS se localisent préférentiellement au niveau des marques enhancers.
Il aurait été intéressant de voir les gènes régulés par ces régions. Cependant quelques soucis d'annotations des séquences ont été rencontrées. Seuls certains génomes ont pu être annotés. Sur certains génomes on avait peu d'informations, l'assemblage n'était pas optimal et beaucoup de séquences étaient non localisées.
Mais le fait de localiser les ITS avec les marques enhancers est déjà en soit un résultat, et qui va de plus dans le même sens que ce que l'on a obtenu chez l'Homme.

5.4 Perspectives d'analyses des gènes associés aux ITS

Quelques génomes ont pu être annotés, et il était intéressant de connaitre les processus biologiques dans lesquels étaient impliqués les gènes à proximités des ITS. Ainsi les données des sites ITS annotés sur le génome de la souris ont été analysées . Il s'agit du mieux annoté.
Le premier processus biologique qui ressort est «Système Immunitaire». La bibliographie (HS3ST4) et les résultats récents (GPC6), montrent que TRF2 interagit avec des gènes intervenant dans le système immunitaire. Il serait donc intéressant de pousser ses analyses et par exemple croiser ses données avec d'autres bases de données.

5.5 Limites

Regardant les différentes analyses que j'ai pu faire notamment, les chevauchements entre plusieurs fichiers, j'ai parfois été limité. J'avais besoin d'un outil plus personnalisé répondant à mes attentes.
En effet je disposais de cinq fichiers que je devais analyser et comparer entre eux. Ces fichiers devaient être normalisés entre eux pour pouvoir être analysés. De plus les comparaisons se faisaient de deux à deux, et le diagramme de Venn généré est limité à quatre cercles.
Ainsi, au court de mon stage, j'ai développé un outil permettant d'analyser les chevauchements entre différents fichiers. Cet outil n'est pas aussi robuste et optimal que Bedtools, mais il permet d'identifier les chevauchements entre plusieurs séquences. L'avantage de cette outil est qu'il possède une option pour trouver les chevauchements non strict. J'entend par là qu'en modifiant une option, on élargit les bornes des séquences et ainsi on peut identifier des chevauchements qui initialement n'était pas observable.
On peut également y placer plus de quatre fichiers pour y obtenir une visualisation plus complète si souhaité. Cependant, ce programme reste à coupler avec le module Randomstats de Pybedtools pour trier les valeurs non significatives. Comme stipulé

précédemment, il est tout de même moins optimal, l'idée n'était pas de faire mieux que ce qui existait déjà, l'outil Bedtools étant déjà très performant, mais plus de crée un outil personnalisé.

6 Conclusion et Perspectives

L'étude la protéine TRF2 semble être prometteuse pour la lutte contre le cancer. Ainsi, il s'agissait lors de ce stage d'étudier l'effet de la fixation de TRF2 sur les gènes à proximité. Par des analyses *in silico* menées à l'aveugle, il a été déterminé une preference de colocalisation des sites de fixations de TRF2 avec des marques H3K27Ac, soit des marques enhancers. De plus, ces analyses ont permis de faire ressortir une liste de gènes potentiellement régulés par la protéine TRF2.
Des expériences biologiques ont été réalisées récemment, où TRF2 a été immunoprécipité, et il ressort que lorsque son site est enrichie en H3K27Ac, certains gènes se voient surexprimés. Un gène en particulier est ressorti dans les deux études, GPC6, qui intervient dans le système immunitaire et jouerait un rôle dans l'apparition de cancer.
Il est intéressant de voir de quelle façon les deux résultats convergent.

Bio-informatiquement, l'analyse a été poussée à d'autres génomes de mammifères, où comme chez l'Homme, les sites de fixation de TRF2 se colocalisent préférentiellement avec les marques enhancers. Ce résultat renforce l'idée que TRF2 agit en tant que enhancer.

De plus, il en est ressorti une liste de processus biologiques dans lesquels serait impliqués les gènes sujets à la régulation de TRF2. Ce résultat offre des perspectives d'études. En effet il serait intéressant de voir quels gènes interviennent dans ce processus. Ceci permettrait grâce à une approche bio-informatique de privilégier l'étude de certains gènes cibles de TRF2. D'autant que le champ est vaste, en effet il ressort d'autres processus biologique qui semble intéressants tel que les processus impliqués dans l'organisation de la chromatine.

7 Bilan des compétences acquises

Lors de ce stage, j'aurais autant appris sur le plan professionnel que personnel.
En effet, pour pouvoir réaliser les différentes analyses que j'ai pu effectuer, j'ai du apprendre à utiliser plusieurs logiciels. Par exemple pour les analyse de ChIP-seq, j'ai du apprendre les étapes de pré-traitement de celui ci, ainsi que pour chaque étape étudier, comparer différents logiciels. Ainsi les différentes pré-analyse de données issus d'un ChIP-seq sont :

- Le controle de qualité du séquençage : FASTX-Toolkit est un logiciel permettant de réaliser parfaitement cette étape. On peut grâce à ce logiciel contrôler la qualité des séquences, et on peut également contrôler la distributions des nucléotides (une mauvaise distribution pourrait être révélateur d'un mauvais séquençage).
- L'alignement : L'alignement est une étape commune à toutes les données de séquençages, il s'agit simplement de chercher l'origine de la séquence sur le génome. J'ai utilisé Bowtie.[13] Il s'agit d'un des logiciels les couramment utilisé pour le mapping, il est très efficace pour les petites répétitions (parfait donc pour analyser la répétions du motif TTAGGG). Il faudra ensuite contrôler la qualité de la séquence après alignement. Cette étape consiste à vérifier le nombre de couvertures à une position unique sur le génome.
- La recherche de pics (Peakcalling) : C'est l'étape essentielle du ChIP-seq. Selon la taille des pics que l'on a, divers programmes sont utilisables, par exemple pour des pics larges on préférera SICER qui est plus adapté et donnera de meilleurs résultats. Pour ma part, j'ai utilisé MACS pour effectuer cette étape.

Cela aurait été formateur pour moi, j'ai du apprendre à comparer différents logiciels et choisir celui qui était le plus adapté à mes données.

De plus, j'aurais approfondi ma connaissance de l'informatique, notamment du langage python, langage dans lequel j'ai développé la plupart de mes scripts et notamment mon pipeline «overlaps». J'ai également apprit en bash (langage linux), ainsi qu'en R(statistiques).

J'ai pu réellement me rendre compte du rôle qu'occupe un bio-informaticien dans une équipe et je comprends mieux pourquoi ce poste semble important pour l'avenir. Les données générées par des expériences biologique deviennent de plus en plus importantes. Il est nécessaire d'avoir une personne compétente en informatique, mais qui comprenne également les aspects biologiques.Une double compétence est donc requise à ce poste afin de pouvoir de comprendre, assimiler, et apporter des solutions aux divers problèmes qu'une équipe de recherche peut rencontrer.

Mais plus que les différents logiciels et scripts mis en oeuvre, j'ai également appris et observé comment se déroulait la gestion d'un laboratoire, la repartition des différentes tâches communes (gestion des stocks, préparation des solutions, planning microscopie, etc.) lors des réunions de laboratoires. Car même si je faisais partie du pôle bio-informatique de l'IRCAN, j'assistais aux réunions, où, autour d'une table, les uns après les autres, nous échangions, racontant ce que l'on avait pu faire au cours de la semaine. Cela nous permettait d'être toujours à jour sur les experiences de nos collègues. J'ai vraiment apprécié l'entre-aide que j'ai pu voir, quand une personne rencontrait un problème, c'est toute l'équipe qui essayait d'apporter une solution. J'ai réellement eu la sensation de voir plus qu'une équipe, une famille, où il y avait des hauts et des bas, où il y a toujours quelqu'un pour nous remonter le moral.
Ce stage m'aura vraiment permis de m'enrichir sur le plan professionnel et personnel, et je garderai un très bon souvenir des moments que j'ai pu passer ici.

Références

[1] Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C. and Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol *9*, e1003326.

[2] Biroccio, A., Cherfils-Vicini, J., Augereau, A., Pinte, S., Bauwens, S., Ye, J., Simonet, T., Horard, B., Jamet, K., Cervera, L. et al. (2013). TRF2 inhibits a cell-extrinsic pathway through which natural killer cells eliminate cancer cells. Nature cell biology *15*, 818–828.

[3] Blackburn, E. H. (2001). Switching and signaling at the telomere. Cell *106*, 661–673.

[4] Blasco, M. A. (2005). Telomeres and human disease : ageing, cancer and beyond. Nature Reviews Genetics *6*, 611–622.

[5] Camacho, C., Madden, T., Ma, N., Tao, T., Agarwala, R. and Morgulis, A. (2013). BLAST command line applications user manual. Internet .

[6] Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B. et al. (2010). Reactome a database of reactions, pathways and biological processes. Nucleic acids research *39*, D691–D697.

[7] Dale, R. K., Pedersen, B. S. and Quinlan, A. R. (2011). Pybedtools a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics *27*, 3423–3424.

[8] De Lange, T. (2005). Shelterin : the protein complex that shapes and safeguards human telomeres. Genes & development *19*, 2100–2110.

[9] El Maï, M., Wagner, K.-D., Michiels, J.-F., Ambrosetti, D., Borderie, A., Destree, S., Renault, V., Djerbi, N., Giraud-Panis, M.-J., Gilson, E. et al. (2014). The telomeric protein TRF2 regulates angiogenesis by binding and activating the PDGFRβ promoter. Cell reports *9*, 1047–1060.

[10] Gilson, E. and Géli, V. (2007). How telomeres are replicated. Nature Reviews Molecular Cell Biology *8*, 825–838.

[11] Gottschling, D. E., Aparicio, O. M., Billington, B. L. and Zakian, V. A. (1990). Position effect at S. cerevisiae telomeres : reversible repression of Pol II transcription. Cell *63*, 751–762.

[12] Huang, W., Loganantharaj, R., Schroeder, B., Fargo, D. and Li, L. (2013). PAVIS a tool for Peak Annotation and Visualization. Bioinformatics *29*, 3097–3099.

[13] Langmead, B., Trapnell, C., Pop, M., Salzberg, S. L. et al. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biol *10*, R25.

[14] Martinez, P., Thanasoula, M., Carlos, A. R., Gómez-López, G., Tejera, A. M., Schoeftner, S., Dominguez, O., Pisano, D. G., Tarsounas, M. and Blasco, M. A. (2010). Mammalian Rap1 controls telomere function and gene expression through binding to telomeric and extratelomeric sites. Nature cell biology *12*, 768–780.

[15] Quinlan, A. R. and Hall, I. M. (2010). BEDTools a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

[16] Romano, G. H., Harari, Y., Yehuda, T., Podhorzer, A., Rubinstein, L., Shamir, R., Gottlieb, A., Silberberg, Y., Pe'er, D., Ruppin, E. et al. (2013). Environmental stresses disrupt telomere length homeostasis. PLoS Genet *9*, e1003721.

[17] Simonet, T., Zaragosi, L.-E., Philippe, C., Lebrigand, K., Schouteden, C., Augereau, A., Bauwens, S., Ye, J., Santagostino, M., Giulotto, E. et al. (2011). The human TTAGGG repeat factors 1 and 2 bind to a subset of interstitial telomeric sequences and satellite repeats. Cell research *21*, 1028–1038.

[18] Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J. et al. (2015). Enhancer evolution across 20 mammalian species. Cell *160*, 554–566.

[19] Xia, J., Gill, E. E. and Hancock, R. E. (2015). NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nature protocols *10*, 823–844.

[20] Yang, D., Xiong, Y., Kim, H., He, Q., Li, Y., Chen, R. and Songyang, Z. (2011). Human telomeric proteins occupy selective interstitial sites. Cell research *21*, 1013–1027.

[21] Ye, J., Renault, V. M., Jamet, K. and Gilson, E. (2014). Transcriptional outcome of telomere signalling. Nature Reviews Genetics *15*, 491–503.