Développement et application d'un outil bio-informatique pour cartographier la machinerie de l'ARN polymérase I chez les mammifères

Mémoire

Marianne Sabourin-Félix

Maîtrise en biologie cellulaire et moléculaire
Maître ès sciences (M. Sc.)

Québec, Canada

Développement et application d'un outil bio-informatique pour cartographier la machinerie de l'ARN polymérase I chez les mammifères

Mémoire

Marianne Sabourin-Félix

Sous la direction de :

Thomas Moss, directeur de recherche

RÉSUMÉ

L'immunoprécipitation de la chromatine suivie du séquençage haut débit (ChIP-seq) est une technique permettant de visualiser les interactions entre l'ADN et les protéines. Toutefois, en pratique, la résolution de cette technique laisse à désirer. En étudiant les gènes de l'ARN ribosomique (ADNr), nous avons observé que le facteur majeur limitant la résolution découle du recouvrement inégal des séquences de chaque locus. Cette inégalité est superposée à la distribution réelle de la séquence d'ADN immunoprécipitée entrainant un profil de liaison protéique aberrant. Un logiciel de déconvolution a été développé afin de corriger la couverture inégale des données ChIP-seq en les normalisant par rapport aux données de l'input (*Whole Cell Extract*). Lorsqu'appliqué sur les données de l'ADNr, cet outil s'est avéré très utile en fournissant un profil de liaison détaillé de la chromatine et des facteurs de transcription le long de ce gène. D'autre part, des études de localisation des sites d'interactions protéiques d'UBF, un facteur de transcription associé à l'ADNr, à la grandeur du génome couplé à des expériences de DNase-seq et de *microarray* ont permis de mettre en lumière les rôles potentiels d'UBF dans les régions non ribosomiques. En conclusion, nous avons développé un outil permettant la normalisation par déconvolution de données de séquençage haut-débit qui permet d'augmenter la résolution du profil de liaison protéique sur l'ADNr en plus d'identifier les rôles potentiels d'UBF à l'échelle du génome.

ABSTRACT

Chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) is a technique that allows to visualize interactions between DNA and proteins. However in practice, the resolution of this technique leaves much to be desired. During our studies of the ribosomal RNA genes (rDNA), we observed that one major factor limiting resolution results from the unequal recovery of sequence data across any given locus. This inequality is superimposed on the actual distribution of immunoprecipitated DNA sequences resulting in aberrant protein binding profiles. A software was developed to correct the unequal coverage of ChIP-seq data by normalizing to the input (Whole Cell Extract) with a deconvolution protocol. When applied on the rDNA, this approach has been especially useful in providing a detailed map of chromatin and transcription factor distribution across the gene. On the other hand, genome-wide localization of protein interaction sites for UBF, a transcription factor associated to rDNA, coupled with DNase-seq and microarray experiments shed light on the potential roles of UBF in non-ribosomal regions. In conclusion, we developed a tool allowing the normalization by deconvolution of high-throughput sequencing data that allows to increase the resolution of protein binding profiles on the rDNA. In addition we identified the potential roles of UBF at genome scale.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

ABRÉVIATIONS

3C	*Chromosome conformation capture*
4HT	4-hydroxytamoxifène
Å	Ångström
ADN	Acide désoxyribonucléique
ADNr	ADN ribosomique
ARN	Acide ribonucléique
ARNm	ARN messager
ARNr	ARN ribosomique
BAM	*Binary sequence alignment map* (*Binary SAM*)
BED	*Browser extensible data*
bp	*Base pair*
Cavin-1	*Caveolae-associated protein 1*
C. elegans	*Caenorhabditis elegans* (nématode)
ChIP	*Chromatin Immunoprecipitation*
ChIP-nexus	*ChIP with nucleotide resolution using exonuclease digestion, unique barcode and single ligation*
ChIP-seq	immunoprécipitation de la chromatine suivie de séquençage haut débit
CHU	Centre Hospitalier Universitaire
CIHR	*Canadian Institute of Health Research*
CK1	Caséine kinase I
CK2	Caséine kinase II
CPE	*Core promoter element*
CTCF	*CCCTC-binding factor*
DBD	*DNA binding domain*
DNase I	Désoxyribonucléase I
DFC	*Dense fibrillar component*
D. melanogaster	*Drosophila melanogaster* (mouche à fruits)
DrUBF	UBF chez le poisson zèbre
DNMT	*DNA methyltransferases*
EDTA	Acide éthylènediaminetétraacétique
EGTA	Acide éthylènebis(oxyéthylènenitrilo)tétraacétique
ERCre	Cre recombinase fusionnée à un récepteur à l'estrogène
ERK	*Extracellular signal–regulated kinases*
ETS	*External transcribed spacer*
FC	*Fibrillar center*
FCP1	*TFIIF-associating component of CTD phosphatase*
FDR	*False Discovery Rate*
GEO	*Gene Expression Omnibus*
GC	*Granular component*
G. gallus	*Gallus gallus* (poulet)
GREAT	*Genomic Regions Enrichment of Annotations Tool*
GTF	*General transcription factors*
GWAS	*Genome-wide association study*

H. influenza	*Haemophilus influenza* (Bacille de Pfeiffer)
HMG-box	*High-mobility group box*
HOMER	*Hypergeometric optimization of motif enrichment*
H. sapiens	*Homo sapiens* (homme)
hUBF	UBF chez l'humain
IBM	*International Business Machines*
IGS	*Intergenic spacer*
Indel	Insertion/délétion
IP	Immunoprécipitation
ITS	*Internal transcribed spacer*
kb	Kilo base
kDa	Kilo Dalton
KO	*Knock-out*
lncRNA	*Long non-coding RNA*
m5C	5-méthylcytosine
m6A	N6-methyladénine
MACS2	*Model-based Analysis of ChIP-Seq data*
MAPK	*Mitogen-activated protein kinase*
Mb	Méga base
MEFs	*Mouse embryonic fibroblasts*
MgCl2	Chlorure de magnésium
mM	Millimolaire
M. musculus	*Mus musculus* (souris)
mUBF	UBF chez la souris
Na	Sodium
NaCl	Chlorure de sodium (sel)
NCBI	*National Center for Biotechnology Information*
NGS	*Next-generation sequencing*
NMP	Nucléophosmine
NoLS	*Nucleolar localisation sequence*
NORs	*Nucleolar organizer regions*
NoRC	*Nucleolar remodeling complex*
NP-40	Tergitol-type NP-40
NPD	*Nucleolar protein database*
NRD	*Negative regulatory domain*
NSERC	*National Science and Engineering Council*
PAF53	*Polymerase associated factor 53*
pb	Paire de bases
PCR	*Polymerase chain reaction*
PIC	*Preinitiation complex*
Pol I	ARN polymérase I
POLR1	ARN polymérase I
PTRF	*Pol I and transcription release factor*
qPCR	*Real-time quantitative polymerase chain reaction*
RFB	*Replication fork barrier*
R. norvegicus	*Rattus norvegicus* (rat)
RPI	ARN polymérase I

RPIα	ARN polymérase I alpha
RPIβ	ARN polymérase I bêta
RPII	ARN polymérase II
RPIII	ARN polymérase III
RPL	*Ribosomal protein from the large subunit*
RPM	*Reads per million*
RPS	*Ribosomal protein from the small subunit*
Rrn3	*Transcription initiation factor IA*
rUBF	UBF chez le rat
S	Svedberg
SAM	*Sequence alignment map*
S. cerevisiae	*Saccharomyces cerevisiae* (levure de boulanger)
SDS	Dodécylsulfate de sodium
sec	Seconde
SIMD	*Single-Instruction Multiple-Data*
SL1	*Selectivity factor 1*
SMRT	*Single-molecule real time*
SnoRNA	*Small nucleolar RNA*
SnoRNP	*Small nucleolar ribonucleoprotein*
S. pombe	*Schizosaccharomyces pombe* (levure)
SpPr	*Spacer promoter*
T_0	*Proximal-promoter target site*
TAF	*TBP associated factor*
TAF1	*TBP associated factor RNA polymerase I*
TAF1B	*TBP associated factor RNA polymerase I subunit B*
TAF1C	*TBP associated factor RNA polymerase I subunit C*
TBP	*TATA-box binding protein*
TIFIA	*Transcription initiation factor IA*
TIP5	*TTF-1 interactif protein 5*
Tris	Trishydroxyméthylaminométhane
Tris-HCL	Tris hydrochloride
TrUBF	UBF chez le poisson-globe
TSS	*Transcription start site*
TTF1	*Transcription termination factor I*
UBF	*Upstream binding factor*
UBTF	*Upstream binding transcription factor*
UCE	*Upstream control element*
UPE	*Upstream promoter element*
USD	Dollar américain
WGBS	*Whole genome bisulfite sequencing*
X. laevis	*Xenoplus laevis* (grenouille)
xUBF	UBF chez la grenouille

Imagination is the power that enables us to empathize with humans whose experiences we have never shared.

J. K. Rowling

Notre vie est un livre qui s'écrit tout seul. Nous sommes des personnages de roman qui ne comprennent pas toujours bien ce que veut l'auteur.

Julien Green

REMERCIEMENTS

Je tiens premièrement à remercier tout spécialement mon directeur de recherche le Dr Tom Moss. Merci de m'avoir fait confiance afin d'assurer le soutien bio-informatique du laboratoire. Merci de m'avoir encouragée lorsque la motivation de continuer n'y était plus. Merci de m'avoir transmis la passion pour la science qui t'habite. Je t'en serai éternellement reconnaissante.

J'aimerais également remercier mes collègues de laboratoire Jean-Clément Mars, Chelsea Herdman, Michel Tremblay et Victor Stefanovsky. Merci pour toutes ces discussions, scientifiques ou non, les mots croisés, les bières et les gâteaux ! Merci à Jean-Clément de m'avoir aidé à chaque étape de ma maîtrise et de me parler de tes théories loufoques ou non. Merci à Chelsea de m'avoir montré quelques techniques de biologie moléculaire, ce fût un plaisir d'enfiler un sarrau pour venir travailler dans le laboratoire avec toi. Un merci tout spécial à Michel d'avoir su détecter quand ça n'allait pas et d'avoir averti Mariline et Maripier, tu m'as carrément sauvé la vie !

Je souhaite également remercier Mariline Béliveau et Maripier Hainse, mes acolytes de diners, de m'avoir soutenues dans les moments difficiles et de m'avoir poussé à finir ma maîtrise. Merci aussi à Vanessa Collin pour le fou rire dans la classe de Cancer ainsi que les pensées positives et les encouragements. Merci aussi à Maryann Breton de m'avoir écoutée toujours sans jugement.

Je tiens finalement à remercier ma famille, ma sœur Catherine, ma mère Francine et mon père Pierre ainsi que mon conjoint David de m'avoir soutenue pendant toutes ces années.

AVANT-PROPOS

Le travail présenté dans ce mémoire est le fruit de près de deux années de travail dans le laboratoire du Dr Tom Moss. Cela a mené à la publication de deux articles. Le premier, dont je suis le 4^{e} auteur, a été publié en 2017 dans le journal PLoS Genetics et le second, dont je suis le premier co-auteur, a été publié dans l'édition de janvier 2018 du journal Genes | Genomes | Genetics (G3).

J'ai été engagée au laboratoire afin de créer une nouvelle méthode de normalisation des données ChIP-seq appliquée spécifiquement aux gènes de l'ARN ribosomique. Cette technique s'est avérée être un simple processus de déconvolution ayant mené à une publication et qui est présentée au chapitre 2. Le script de cette procédure se retrouve à l'annexe I. L'annexe II comporte l'article intégral de PLoS Genetics où la méthode de déconvolution a été appliquée sur différents jeux de données générés au laboratoire.

J'ai d'ailleurs étudié le rôle du facteur architectural UBF à l'échelle du génome, ce qui fait l'objet du chapitre 3.

De plus, j'ai travaillé à l'analyse de plusieurs jeux de données publiques de ChIP-seq, de RNA-seq et de *Whole Genome Bisulfite Sequencing* (WGBS). Ces travaux ne sont pas présentés dans ce mémoire.

CHAPITRE 1 – INTRODUCTION

1.1 – LE NUCLÉOLE

1.1.1 – STRUCTURE DU NUCLÉOLE

Le nucléole est un sous-compartiment du noyau où s'effectue la transcription des gènes de l'ARN ribosomique (ARNr). Le nucléole a une plus grande densité électronique que le reste du noyau lorsque observé au microscope électronique (Figure 1.1).

Il existe trois sous-compartiments nucléolaires : le *Fibrillar Center* (FC), le *Dense Fibrillar Component* (DFC) et le *Granular Component* (GC) (Figure 1.1). Ces structures ont été décrites grâce au microscope électronique où elles ont été distinguées de par leur densité de fibrilles.

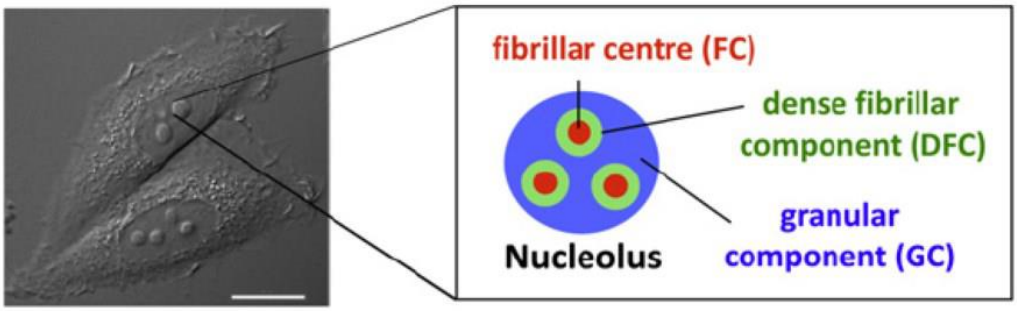

Figure 1.1 – Organisation du nucléole. Microscopie à contraste interférentiel de cellules HeLa et schéma des sous-compartiments du nucléole. Extrait de Boulon et al., 2010.

1.1.2 – NORS

À la fin de la mitose, le nucléole se reforme autour des *Nucleolar Organizer Regions* (NORs), contenant les gènes ribosomiques (Olson et al, 2002). En effet, lors de la mitose, le nucléole se désassemble et la transcription de l'ARNr s'arrête. Les NORs sont la localisation chromosomique où se retrouvent les gènes de l'ARNr. Ceux-ci sont encodés en environ 200 copies dans le génome haploïde de la souris et de l'humain. Ils sont situés sur le bras court des chromosomes 12, 15, 16, 18 et 19 murins et 13, 14, 15, 21 et 22 humains (Figure 1.2 b). Chez l'humain ils se retrouvent

uniquement sur les chromosomes acrocentriques. Dans l'humain et la souris, les NORs sont situés aux constrictions secondaires des chromosomes, la constriction primaire étant le centromère (Figure 1.2a). Dans d'autres organismes, le positionnement chromosomique des NORs est variable. De plus, la longueur de leurs NORs varient entre les individus et au sein d'un même organisme (Stults et al., 2008).

Bien que l'on dise que le nucléole se forme autour des NORs, certains NORs ne sont pas associés au nucléole. En effet, il existe une population de gènes ribosomiques hautement condensés (*silenced*) qui ne présente ni de facteur de transcription *Upstream Binding Factor* (UBF), ni de facteur reliés à l'ARN polymérase I (RPI) et qui ne peut donc pas être transcrite (Moss et al., 2007; McStay et Grummt 2008; Moss, 2011; McStay 2016).

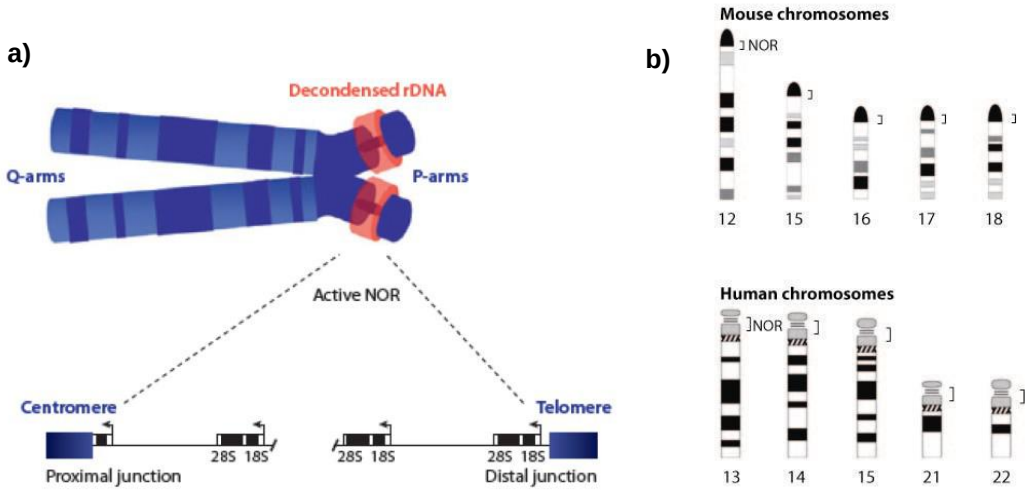

Figure 1.2 – Les NORs. a) Représentation du chromosome 15 humain. Les NORs sont indiqués en rouge. b) Positionnement des NORs sur les chromosomes de l'humain et de la souris. Extrait de McStay et Grummt, 2008.

1.1.3 – Fonctions du nucléole

Malgré que la fonction principale du nucléole soit la biogénèse des ribosomes, des études ont montré qu'il jouerait un rôle dans divers processus cellulaires. En effet, il est impliqué dans la régulation du cycle cellulaire (Visintin et Amon, 2000), dans la

capacité prolifération des cellules souches (Tsai et McKay, 2002), dans la détection du stress cellulaire (Rubbi et Milner, 2003), dans la réplication virale (Sirri et al., 2008) et dans la différentiation des cellules souches trophoblastiques (Martindill et Riley, 2008).

Des études de spectrométrie de masse ont permis de mettre en lumière plus de 700 protéines localisées au nucléole dans des cellules humaines (Sirri et al., 2008). De nos jours, la *Nucleolar Protein Database* (NPD) compte maintenant plus de 4 500 protéines nucléolaires, dont seulement une faible proportion serait impliquée directement dans la biogénèse des ribosomes (Ahmad et al., 2009).

1.2 – Le ribosome

1.2.1 – Structure du ribosome

Le ribosome est un complexe ribonucléoprotéique; il est composé au tiers de protéines ribosomiques et au deux tiers d'ARNr. Il est formé de deux sous-unités, soit le 40S et le 60S, le S indiquant leur taux de sédimentation ou *Svedberg*. La grande sous-unité (le 60S) se compose de trois ARNr : le 28S, le 5.8S et le 5S ainsi que d'environ 49 protéines ribosomiques. La petite sous-unité (le 40S), quant à elle, comporte un ARNr, le 18S, et contient environ 33 protéines ribosomiques (Hernandez-Verdun et Louvert, 2004; Moss et al., 2007). Le taux de sédimentation du ribosome en entier est de 80S.

1.2.2 – Ribosomopathies

Des mutations dans des gènes responsables de la transcription des gènes de l'ARNr, la maturation des ARNr, le transport et l'assemblage des ribosomes peuvent mener à ce qu'on appelle des ribosomopathies (Hannan et al, 2013). Ces maladies visent plusieurs processus, mais ont en commun certains symptômes tels que des dysfonctionnements immunologiques et hématologiques, un vieillissement accéléré, une déficience de la moelle osseuse ainsi qu'une prédisposition au cancer.

Dans les cas de cancer, la taille du nucléole sert de prédiction concernant le développement futur de l'état d'un patient (Derezini et al., 2000; Sirri et al., 2008). En effet, un nucléole de grande taille est signe d'un mauvais pronostic, car il indique un potentiel marqué de prolifération chez les cellules cancéreuses. La biogénèse des ribosomes est donc une cible de traitement possible contre le cancer puisqu'elle est en lien étroit avec la prolifération et la croissance cellulaire.

1.3 – La biogénèse des ribosomes

1.3.1 – Les gènes de l'ARNr

Les 200 copies des gènes de l'ARNr, réparties sur 5 chromosomes, sont organisées en tandem, c'est-à-dire que les unités répétées sont directement adjacentes les unes aux autres. Une unité d'ADNr contient environ 45 kilobases (kb) chez la souris et environ 43 kb chez l'humain. Elle est composée d'abord d'un segment non-transcrit d'environ 30 kb, l'*Intergenic Spacer* (IGS) qui contient les éléments régulateurs tels que le *Spacer promoter* (SpPr), l'*enhancer repeat*, le 47S promoteur ainsi que les terminateurs. Le 47S promoteur se compose de deux parties : l'*Upstream Control Element* (UCE) et le *Core Promoter Element* (CPE). Le CPE est la portion minimale requise pour une initiation de la transcription efficace; il est situé au site d'initiation de la transcription (TSS) (Moss et al., 2007). L'UCE, situé à 100 paires de bases (pb) en amont, permet, quant à lui, d'augmenter l'efficacité de la transcription (Learned et al., 1986). Le SpPr est connu pour stimuler la transcription ribosomique par RPI (Moss, 1983; de Winter et Moss, 1986; Caudy et Pikaard, 2002). Par contre, cela nécessite la présence de l'*enhancer repeat* entre le SpPr et le 47S promoteur (de Winter et Moss, 1987; Paalman et al., 1995). Le SpPr peut aussi servir de plateforme de recrutement pour les molécules de RPI qui sont ensuite livrées au 47S promoteur pour permettre une transcription rapide et efficace (Moss, 1983; Kuhn et Grummt, 1987). Il a été montré que les transcrits provenant de l'IGS sont essentiels à l'établissement et au maintien de la structure hétérochromatinienne d'un sous-ensemble d'ADNr (Mayer et al., 2006). L'unité d'ADNr est composée de plusieurs sites terminateurs : le T_0 situé à 170 pb en amont de l'UCE et le T_1 à T_{10}

situés en aval de la région codante (Grummt et al., 1986; Henderson et Sollner-Webb 1986).

L'unité d'ADNr contient aussi une région transcrite comprenant le précurseur des ARNr d'environ 14 kb (le 47S), les *External Transcribed Spacer* (ETS) aux extrémités de la région transcrite et les *Internal Transcribed Spacers* (ITS) entre le 28S et le 5.8S et entre le 5.8S et le 18S (Figure 1.3).

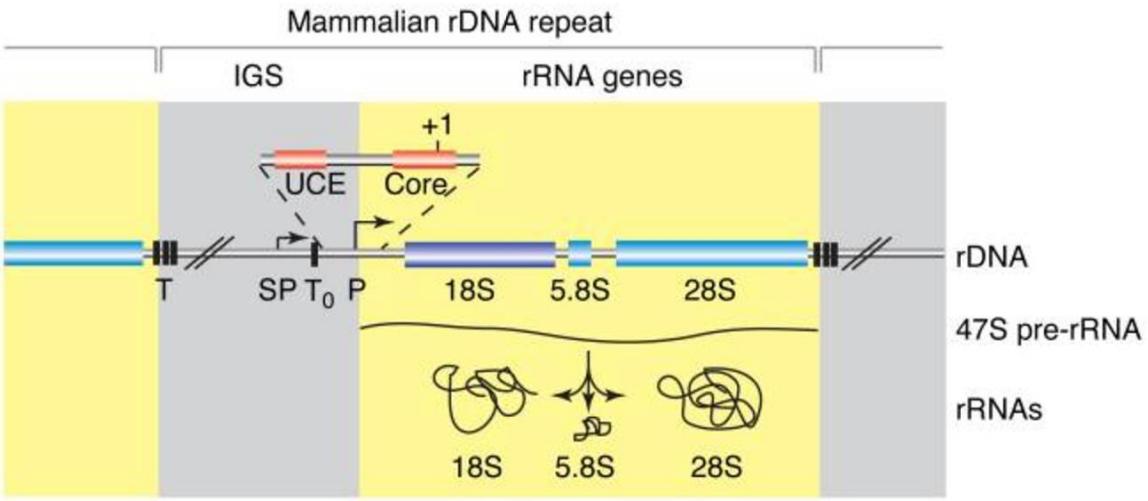

Figure 1.3 – Représentation non à l'échelle de l'ADNr. L'*intergenic spacer* (IGS) est en gris et la région transcrite est en jaune. Les promoteurs sont indiqués par SP pour le *Spacer promoter* et P pour le 47S promoteur. Les terminateurs sont indiqués par T_0 et T (T_1 à T_{10}). En rouge sont indiqués le *Upstream control element* (UCE) et le *Core promoter element* (Core). Les ARNr sont en bleu. Extrait de Russel et Zomerdijk, 2005 et reproduit avec permission.

Les ARNr sont extrêmement conservés entre les espèces. Par contre, l'IGS, les ETS et les ITS ne le sont pas (Moss et al., 2007).

Le séquençage de nouvelle génération du génome n'est pas parvenu à obtenir la séquence complète des ADNr étant donné leur nature répétitive et leur grand nombre. Par contre, une unité complète de l'ADNr a été publiée chez l'humain et la souris (Gonzalez et Sylvester, 1995; Grozdanov et al., 2003). Celles-ci sont disponibles sur *GenBank* aux numéros d'accession U13369.1 et BK000964.3 respectivement. Cela a été rendu possible grâce à la technologie de séquençage développée par Sanger dans les années 1970 (Sanger et al., 1977).

1.3.2 – TRANSCRIPTION DES GÈNES DE L'ARNR

La biogénèse des ribosomes nécessite le fonctionnement de trois machineries transcriptionnelles de la cellule (Figure 1.4). Premièrement, RPI permet la transcription du précurseur des ARNr, le 47S. L'ADNr est l'ADN qui est le plus activement transcrit dans la cellule. En effet, environ 80 % des ARN cellulaires sont d'origine ribosomique (Moss et al., 2007).

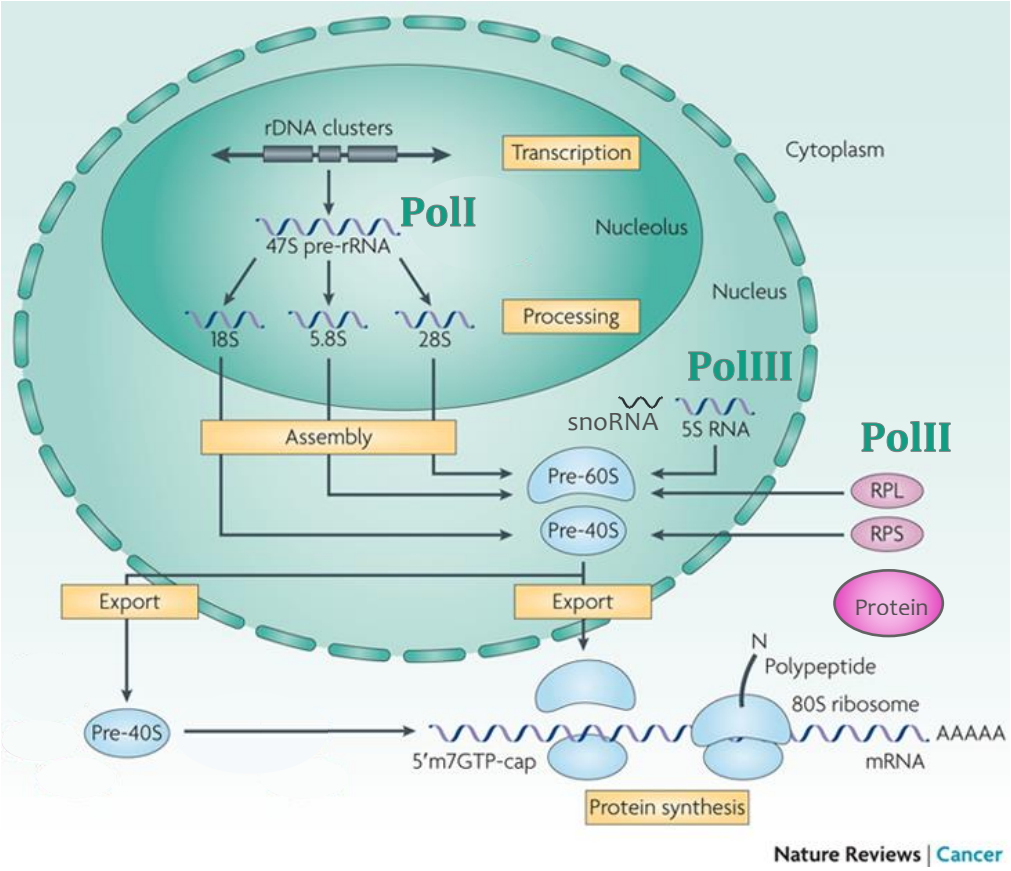

Figure 1.4 – La biogénèse des ribosomes requiert l'apport de trois polymérases. RPI transcrit le précurseur des ARNr, le 47S. RPIII synthétise le dernier ARNr, le 5S. RPII transcrit les ARNm des protéines ribosomiques (RPL = *ribosomal protein from the large subunit*, RPS = *ribosomal protein from the small subunit*) ainsi que d'autres protéines non ribosomiques. Adapté de van Riggelen et al., 2010 et reproduit avec permission.

Le processus de transcription peut même être observé au microscope électronique par la technique de *Miller Chromatin Spreading* (*Miller Spread*) (Raska, 2003). La structure résultante, rappelant un arbre de Noël, a l'ADNr et les ARN polymérase I en guise de tronc et les transcrits d'ARNr en guise de branches (Figure 1.5).

Deuxièmement, l'ARN polymérase III (RPIII) est en charge de la transcription du dernier ARNr, soit le 5S, ainsi que des *small nucleolar RNA* (snoRNAs). Ces derniers permettent les modifications nucléosidiques nécessaires aux étapes de clivage du précurseur de l'ARNr (Bachellerie et al., 2002). Troisièmement, l'ARN polymérase II effectue la transcription des ARN messagers (ARNm) des protéines ribosomiques, mais aussi des protéines non ribosomiques impliquées dans la biogénèse des ribosomes (Figure 1.4) (Chédin et al., 2007).

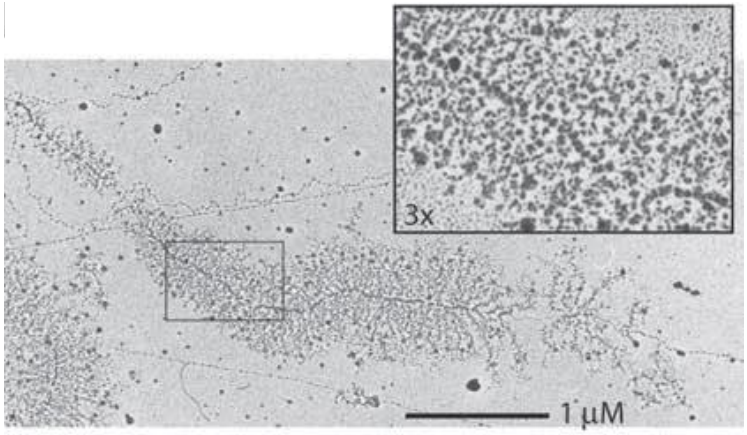

Figure 1.5 – Transcription ribosomique. Structure en arbre de Noël (*Christmas Trees*) obtenus par la technique de *Miller Spread* dans des cellules de souris Ltk$^-$ en culture. Extrait de Moss et al., 2007 et reproduit avec permission.

1.3.3 – Maturation des ARNr et assemblage du ribosome

La biogénèse des ribosomes est un processus complexe; elle comprend plusieurs étapes et fait intervenir de nombreuses protéines. Celles-ci sont, entre autres, des ATP/GTPases, des hélicases, des chaperonnes, des ribonucléases, des méthyltransférases, des endonucléases, des exonucléases et des *Small nucleolar ribonucleoprotein* (SnoRNP) comprenant des SnoRNA.

Les snoRNA permettent le clivage du précurseur de l'ARNr, et ce, même avant la fin de sa transcription, c'est-à-dire de manière co-transcriptionnelle (Granneman et Baserga, 2005). Les snoRNP permettent de guider les enzymes jusqu'aux sites de modifications.

Plus de 200 nucléotides de l'ARNr sont modifiés soit par pseudouridination ou par méthylation (Nazar, 2004). L'ARNr subit, d'autre part, plusieurs clivages successifs effectués par des exo et des endonucléases (Mullineux et Lafontaine, 2012).

Les transcrits pré-ribosomiques sont modifiés et clivés dans le DFC. Ils sont ensuite transportés au GC où ils subissent la maturation finale et sont assemblés aux protéines ribosomiques. Les deux sous-unités du ribosome, soit le 40S et le 60S, sont alors exportés au cytoplasme pour accomplir leur fonction (Figure 1.6).

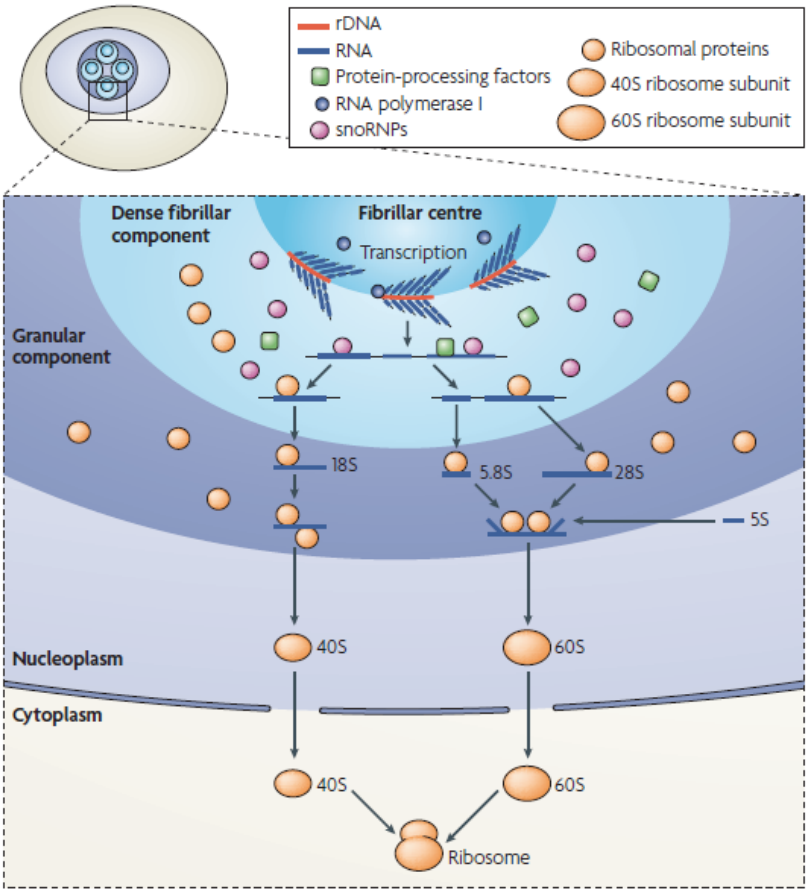

Figure 1.6 – Assemblage du ribosome. Les transcrits pré-ribosomiques sont traités par les snoRNP au sein du DFC. Dans le GC, le 5.8S et le 28S forment le 60S avec le 5S et des ribonucléoprotéines, les protéines ribosomiques. Le 18S quant à lui s'assemble avec d'autres ribonucléoprotéines pour former le 40S. Ce dernier ainsi que le 60S sont exportés au cytoplasme et se lient à l'ARNm pour former le ribosome fonctionnel. Extrait de Boisvert et al., 2007 et reproduit avec permission.

1.3.4 – La méthylation aux promoteurs des gènes de l'ARNr

La méthylation de l'ADN, c'est-à-dire la méthylation d'une cytosine en 5-méthylcytosine (m5C), est une modification épigénétique considérée comme étant caractéristique de la répression de la transcription des gènes (Bernstein et al., 2007). Chez les vertébrés, cette marque survient principalement sur des séquences C-G connues sous le nom de CpG. Chez les plantes, la méthylation de l'ADN existe dans les contextes CHG et CHH où H est égal à A, C ou T. Chez les bactéries et les eucaryotes unicellulaires, des méthylations peuvent aussi survenir sur les adénines en N6-méthyladénine (m6A) (O'Brown et Greer, 2016).

Environ 1 % des base totales sont méthylées, cela représente donc de 70 à 80 % de tous les CpG du génome (Bird, 2002). Il existe, par contre, des régions riches en CpG non méthylées aux promoteurs des gènes que l'on nomme des îlots CpG (*CpG islands*). On estime qu'il y a environ 30 000 îlots CpG dans le génome haploïde humain (Bird et al., 1987).

La méthylation des gènes est effectuée par trois *DNA methyltransferases* (DNMT) DNMT1, 3a et 3b. Il existe aussi DNMT2 dont le rôle n'est pas encore bien caractérisé et DNMT3l qui est inactif enzymatiquement parlant. DNMT3a et 3b sont responsables des méthylations aux sites CpG n'étant pas méthylés (méthylations *de novo*) tandis que DNMT1 est responsable de la maintenance, lors de la réplication, du patron de méthylation présent sur l'ADN (Klose et Bird, 2006).

La méthylation de l'ADN par DNMT3a et 3b intervient dans plusieurs fonctions biologiques telles que l'empreinte parentale (*imprinting*), l'inactivation du chromosome X, l'extinction des transposons, le vieillissement ainsi que le cancer (Law et Jacobsen, 2010; Galamb et al., 2016). De plus, le profil épigénétique est hérité durant la division cellulaire grâce à DMNT1 (Meehan et Stancheva, 2001). La méthylation de l'ADN est un processus réversible (Zhang et al., 2017).

Le promoteur des gènes de l'ARNr de l'humain comprend 26 sites de méthylation potentiels (Pietrzak et al., 2016). Il a été montré par des études *in vitro* chez la souris que la méthylation d'un seul CpG, à la position -133 par rapport au site d'initiation de la transcription de l'ADNr, inhibe la transcription (Santoro et Grummt, 2001). Il a été suggéré que cela est dû au fait qu'UBF est incapable de se lier au promoteur. Il n'y a donc pas de recrutement de *Selectivity Factor 1* (SL1) pour former le complexe de préinitiation ni d'initiation de la transcription. On estime qu'environ 50 % des gènes de l'ARNr sont méthylés, et donc éteints (*silenced*) dans les cellules différenciées (Santoro et Grummt, 2001).

1.4 – Les facteurs de régulation de la biogénèse des ribosomes

1.4.1 – Introduction des différents facteurs de régulation

Les facteurs de régulation de la biogénèse des ribosomes sont nombreux. Premièrement, il y a UBF qui agit à titre de facteur architectural et qui permet la formation d'un complexe de préinitiation stable. Deuxièmement, on retrouve SL1 qui permet l'activation de la transcription en recrutant *Transcription Initiation Factor IA* (Rrn3). Troisièmement, RPI a pour rôle la transcription de l'ADNr à proprement parler. Quatrièmement, Rrn3 fait le lien entre SL1 et RPI. Finalement, *Transcription Termination Factor I* (TTF1) est responsable de la terminaison de la transcription ribosomique. La figure suivante représente les différents facteurs de régulation de la biogénèse des ribosomes (Figure 1.7).

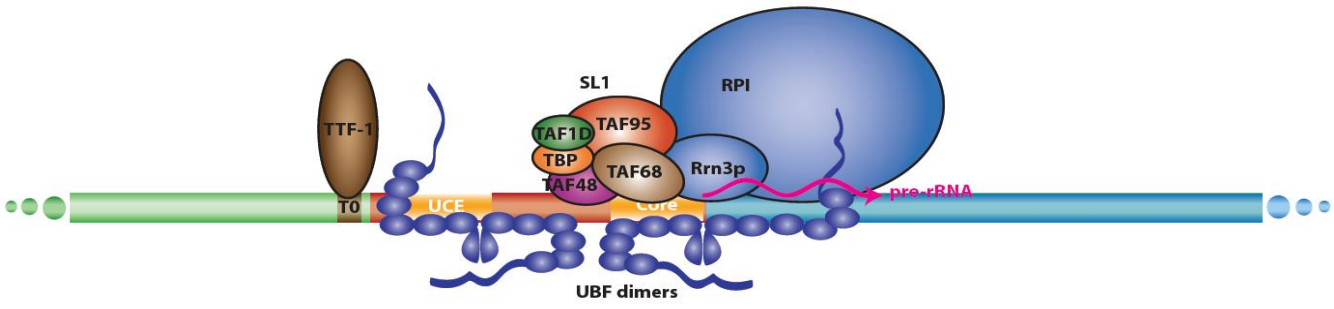

Figure 1.7 – Modèle du complexe de préinitiation au promoteur des gènes de l'ARNr. SL1 est composé de TBP et au moins 3 TAFs. UBF est lié sur l'ADN sous forme de dimères. Deux dimères d'UBF sont représentés en bleu sur cette figure. TTF1, en brun, se lie sur l'ADN à la hauteur du T_0. L'UCE et le CPE sont représentés en orange. Rrn3 et RPI sont en bleu.

1.4.2 – UBF

UBF est une protéine de 97 kilo Dalton (kDa) composée de 764 acides aminés qui fait partie de la famille des protéines *High Mobility Group box* (HMG-box) (Bell et al., 1988; Jantzen et al., 1990). Les HMG-box sont composées de 3 hélices alpha. UBF comprend 6 de ces HMG-box permettant de lier l'ADNr sous forme de dimère (Figure 1.8) (Bachvarov et Moss, 1991; Bazett-Jones et al., 1994; Stefanovsky et al., 2001-1). Les HMG-box sont connues pour lier l'ADN de façon non spécifique. Elles permettent d'ailleurs de moduler la structure de la chromatine en y effectuant une courbure. Pour cette raison, UBF est considéré comme étant un facteur architectural (Stefanovsky et Moss, 2008). UBF comprend un domaine de dimérisation en N-terminal et une queue acidique en C-terminal (Figure 1.8). Son domaine de dimérisation, tel que son nom l'indique, permet la dimérisation de la protéine (O'Mahony et al., 1992; Stefanovsky et al., 2001-1). Sa queue acidique, quant à elle, permet, lorsque phosphorylée, des interactions protéines-protéines avec certains complexes tels que le facteur SL1 (Tuan 1999).

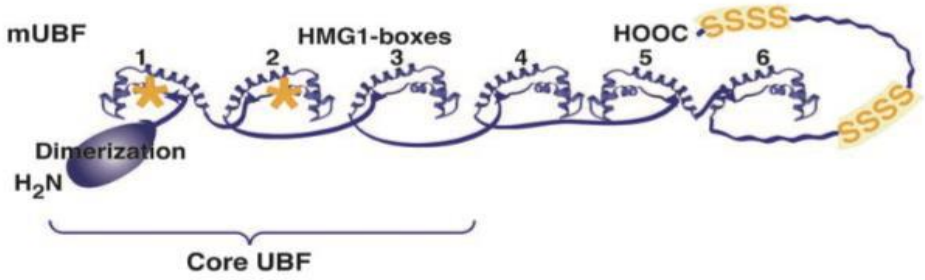

Figure 1.8 – Structure d'UBF. Il possède un domaine de dimérisation en N-terminal et une queue acidique en C-terminal ainsi que 6 HMG-box. Les sites de phosphorylation de ERK sont indiqués par des astérisques jaunes. Extrait de Moss et al. 2007 et reproduit avec permission.

UBF, aussi nommé UBF1, a un isoforme issu d'un épissage alternatif d'un gène, différant de 4.47 kDa, que l'on dénomme UBF2. Ces deux isoformes ont été caractérisés chez le rat, la souris, l'humain et le hamster (Figure 1.9) (O'Mahony et Rothblum, 1991). Malgré le fait que les deux isoformes sont présents en quantité équimolaire, seulement UBF1 est capable d'activer correctement la transcription ribosomique tant *in vitro* qu'*in vivo*; UBF2 est virtuellement inactif (Kuhn et al., 1994). En effet, UBF2 est tronqué de 37 acides aminés dans sa deuxième HMG-box et

celle-ci joue un rôle important dans la liaison spécifique d'UBF à l'ADNr. UBF2 a donc plus de difficulté à lier l'ADN, ce qui pourrait expliquer pourquoi il n'a pas d'effet activateur sur la transcription ribosomique.

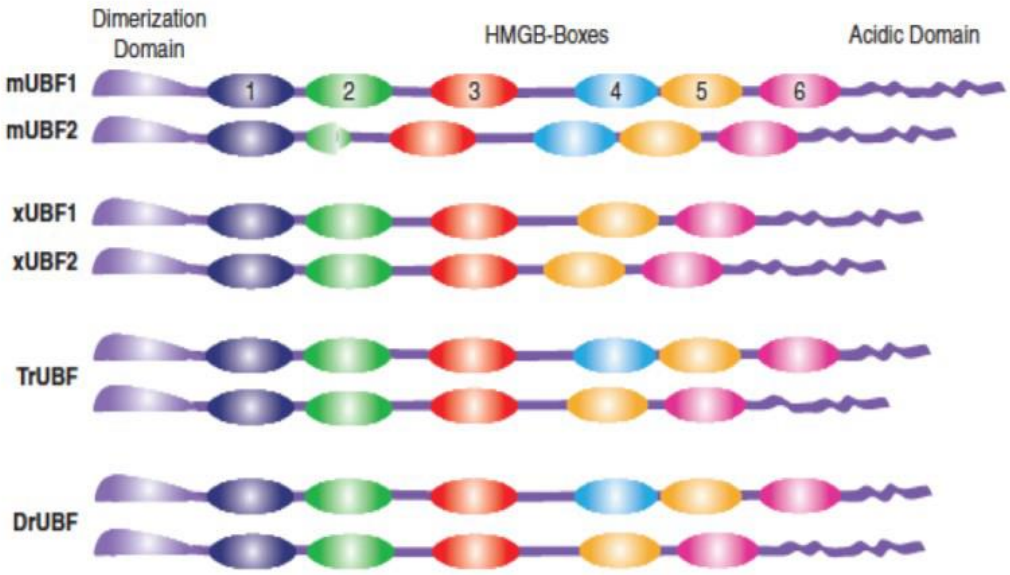

Figure 1.9 – UBF1 et UBF2 chez différentes espèces. Représentation des variants d'épissage chez différentes espèces. mUBF : souris, xUBF : Xenope, TrUBF : poisson-globe et DrUBF : poisson zèbre. Les couleurs similaires représentent une forte homologie. Extrait de Stefanovsky et Moss 2008 et reproduit avec permission.

UBF est connu pour former une structure que l'on nomme l'*enhancesome* (Figure 1.10) (Bazett-Jones, 1994; Stefanovsky et al., 2002). Cette dernière, similaire au nucléosome, est en fait une boucle de 360 degrés comprenant environ 140 pb que forme l'ADNr grâce à la liaison d'un dimère d'UBF. Cette structure fait le rapprochement de l'UCE et du CPE; cela permet la liaison du complexe SL1, responsable conjointement avec UBF de l'activation de la transcription, aux deux éléments du promoteur et ainsi favorise la formation et la stabilisation du complexe de préinitiation (*Preinitiation Complex*, PIC) (Bazett-Jones et al, 1994; Stefanovsky et al., 1996, Stefanovsky et al., 2001-1). Seulement les trois premières HMG-box sont nécessaires au bon repliement de l'*enhancesome*.

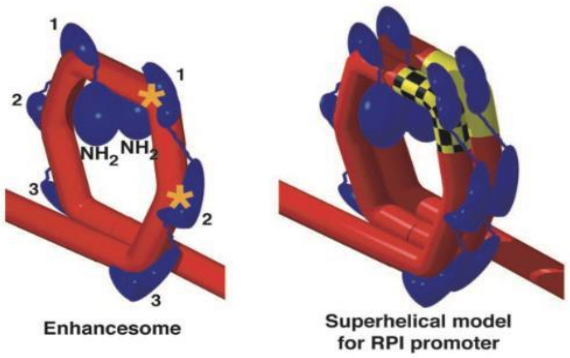

Figure 1.10 – *Enhancesome*. Sa structure permet de rapproche les éléments du promoteur. Les sites de phosphorylation par ERK sont indiqués par des astérisques jaunes. Extrait de Moss et al. 2007 et reproduit avec permission.

La première étape de l'initiation de la transcription ribosomique est considérée comme étant la liaison d'UBF à la fois à l'UCE et au CPE. Ensuite, SL1 interagit avec UBF et le complexe UBF-SL1 résultant est reconnu par RPI pour initier la transcription (Bell et al., 1988). UBF stimule donc la transcription ribosomique en formant un PIC stable.

Extracellular signal-regulated kinases (ERK) phosphoryle les HMG-box 1 et 2 d'UBF aux thréonines 117 et 201 (Figure 1.8, 1.10 et 1.11). Cela diminue l'affinité d'UBF pour l'ADN et empêche le repliement complet de l'*enhancesome* (Figure 1.11) (Stefanovsky et al., 2001-2). Ce faisant, le passage de la polymérase est facilité le long des gènes de l'ADNr. Il s'agit donc d'un mécanisme de régulation de la vitesse d'élongation de la transcription ribosomique (Moss et al., 2006). Les phosphorylations sur les thréonines ont un effet stimulateur de la transcription ribosomique. Cette dernière est donc sous le contrôle direct des signaux de croissance extracellulaires (Stefanovsky et al., 2001-2). En effet, ceux-ci ont pour effet de déclencher la cascade des *mitogen-activated protein kinases* (MAPK) qui aura ultimement pour effet la phosphorylation de ERK ce qui l'activera. Celui-ci pourra alors phosphoryler certains facteur tels que UBF dans le but d'accélérer la croissance cellulaire par l'entremise de la transcription ribosomique.

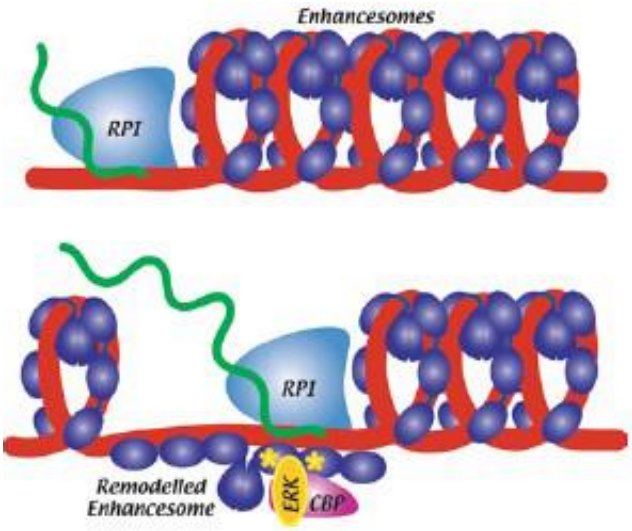

Figure 1.11 – Phosphorylation par ERK. La phosphorylation d'UBF par ERK permet le relâchement de l'ADN et facilite le passage de la polymérase. Les sites de phosphorylation par ERK sont indiqués par des astérisques jaunes. Extrait de Moss et al. 2006.

Des expériences d'immunoprécipitation de la chromatine suivie de séquençage haut débit (ChIP-seq) ont permis de découvrir qu'UBF n'est pas seulement présent aux promoteurs, mais aussi tout le long de la région transcrite de l'ADNr (Zentner et al., 2011; Herdman, Mars et al., 2017). Cela sous-entend qu'UBF occupe plus de fonctions que ce qui était auparavant suggéré. En effet, UBF est impliqué de plusieurs façons dans la transcription ribosomique. Il stimule l'initiation de la transcription, mais est aussi impliqué dans l'établissement d'une structure chromatinienne ouverte sur les gènes actifs. Sa liaison est également responsable de l'apparence des NORs actifs. Durant la mitose, UBF agit à titre de marque-page (*bookmark*) sur les NORs qui étaient actifs pendant l'anaphase. Cela permet d'assurer la réactivation rapide de la transcription (McStay, 2016).

UBF est présent chez tous les Métazoaires (à l'exception de *D. melanogaster* et de *C. elegans*) et on retrouve même son homologue, Hmo1, chez *S. cerevisiae* (McStay, 2016). La conservation de la séquence d'UBF entre les espèces suggère que sa fonction subsisterait dans l'évolution. L'isoforme UBF1 chez le rat est presque identique à son équivalent chez l'humain. En effet, seulement 13 résidus sur 764 ne sont pas identiques et ces différences sont des substitutions entre un

acide aspartique et un acide glutamique et vice-versa sur la queue C-terminale de la protéine. Étant donné la forte conservation de séquence entre UBF chez *H. sapiens* (hUBF) et chez *R. norvegicus* (rUBF), ces deux protéines sont interchangeables dans des essais de transcription *in vitro* (O'Mahony et Rothblum, 1991). La structure primaire prédite pour UBF chez *X. laevis* (xUBF) est identique à 73 % à celle prédite pour hUBF, ce qui correspond à une homologie significative et 50 % des changements d'acides aminés sont conservatifs ou semi-conservatifs. Les trois premières HMG-box d'UBF sont conservées chez hUBF et xUBF avec 94, 83 et 93 % d'identité de séquence respectivement. La conservation de la queue acidique d'UBF chez l'humain et la souris indique qu'ils ont des rôles fonctionnels conservés. xUBF et hUBF ne sont par contre pas interchangeables dans des essais de transcription *in vitro* (Bachvarov et Moss, 1991).

1.4.3 – SL1

Le complexe SL1 (ou TIF-IB chez la souris) a un poids moléculaire de 300 kDa et se compose de la *TATA-box binding protein* (TBP) et d'au minimum trois *TBP associated factors* (TAFs). D'après leur poids moléculaire, on les dénomme TAF48, TAF68 et TAF95 chez *M. musculus* et TAF48, TAF63 et TAF110 chez *H. sapiens* (aussi notés TAF1A, TAF1B et TAF1C) (Eberhard et al., 1993; Zomerdijk et al., 1994). D'autres TAFs semblent potentiellement faire partie du complexe. Par exemple, TAF41 (TAF1D) qui est toujours immunoprécipité avec SL1, est impliqué dans la transcription par RPI (Gorski et al., 2007). Également, TAF12, qui était considéré comme étant spécifique à RPII, s'est révélé impliqué dans la transcription par RPI (Denissov et al., 2007).

TBP a originellement été découvert au sein du complexe transcriptionnel de RPII, mais est également présent dans les complexes de RPI et RPIII. TBP permet de reconnaître les régions promotrices (Goodrich et Tijan, 1994).

Le complexe SL1 est essentiel à la reconnaissance par RPI du promoteur et est responsable de l'activation de la transcription conjointement avec la queue C-

terminale d'UBF (Tuan et al., 1999). SL1 lie le promoteur pour recruter RPI et son facteur associé Rrn3.

1.4.4 – RPI

L'ARN polymérase I (RPI, Pol I POLR1) est une polymérase qui est dédiée spécifiquement à la transcription de la majorité des ADNr, soit le 18S, le 5.8S et le 28S. Cela totalise au-delà de 50 % de l'ARN synthétisée dans une cellule (Russel et Zomerdijk, 2006). Le 5S, quant à lui, est transcrit par l'ARN polymérase III (RPIII) la deuxième des trois polymérases présentes chez les eucaryotes.

La cryo-microscopie électronique a permis de déterminer que RPI était une structure de 12 Å ayant un poids moléculaire de 589 kDa et composée de 14 sous-unités chez *S. cerevisiae* (Fernández-Tornero et al., 2013). Plusieurs sous-unités sont partagées entre les trois polymérases (Tableau 1). Par exemple, les sous-unités RPB5, 6, 8, 10 et 12 sont communes aux trois polymérases. Les deux grandes sous-unités A190 et A135 possèdent des régions homologues aux sous-unités Rpb1 et 2 de RPII respectivement. AC40 et 19 sont identiques dans RPIII et RPI; AC40 est homologue à la sous-unité Rbp3 de RPII et AC19 est homologue à la sous-unité Rbp11. L'hétérodimère A14/43 est lointainement relié à Rpb4/7 chez RPII et à C17-25 chez RPIII. A49/34.5 est homologue à TFIIF chez RPII et cette sous-unité est responsable de l'élongation de la transcription ribosomique. RPA12.2 est homologue à C11 chez RPIII et à Rbp9 dans RPII qui sont en charge de la relecture (*proofreading*) de l'ARN ribosomique et de son clivage en 3'. Le domaine C-terminal de RPA12.2, quant à lui, est homologue à TFIIS chez RPII et est responsable du clivage intrinsèque de l'ARNr (Figure 1.12) (Kuhn et al., 2007).

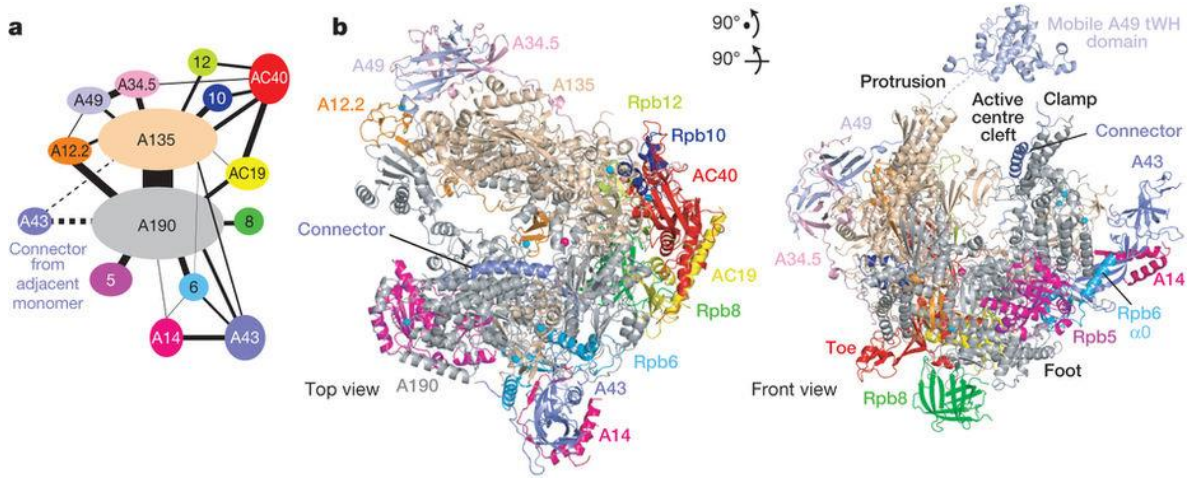

Figure 1.12 – Complexe de RPI et ses sous-unités. a) Représentation schématique du complexe des sous-unités de RPI. b) Représentation moléculaire tridimensionnelle de RPI à gauche la vue du dessus et à droite la vue de devant. Extrait de Engel et al., 2013 et reproduit avec permission.

RPA49 chez la levure est l'homologue de la *polymerase associated factor 53* (PAF53) chez la souris. PAF53 est un facteur associé à la polymérase tandis que RPA49 fait partie de la polymérase. Par contre, le fait que RPA49 est en mesure de se dissocier du complexe cœur de la polymérase pourrait indiquer que l'évolution a fait en sorte que la sous-unité se détache complètement du complexe cœur pour devenir un facteur associé. PAF53, lorsque acétylé ou déacétylé par la protéine SIRT7, module la liaison de la polymérase sur l'ADNr. Cela permet à la cellule de répondre à différentes conditions de stress en adaptant la transcription ribosomique à son état énergétique cellulaire (Chen et al., 2013). PAF53 est impliqué dans la formation du PIC au promoteur des gènes de l'ARNr en médiant les interactions entre RPI et UBF permettant la synthèse des ARNr (Hanada et al., 1996).

Tableau 1 – ARN polymérase I. Sous-unités de RPI chez *S. cerevisiae* et homologie chez *H. sapiens*. La plupart de ces sous-unités sont essentielle chez la levure, mais celles qui ne le sont pas sont marquées par : viables (^{v}) ou mutant conditionnel (^{c}). Extrait de Russell et Zomerdijk, 2006.

***S. cerevisiae* Pol I subunits**	**Unique or shared**	**Human Pol I subunit**	**Homologues**	**Interactions in mammalian PIC**
RPA190	I	hRPA190 (A194)	β'; RPB1 (B220)	–
RPA135	I	hRPA127	β; RPB2 (B140)	–
RPA49^{c}	I	PAF53	None	UBF
RPA43	I	hRPA43	RPB7	hRRN3
RPA40 (AC40)	I, III	hRPA40	α; RPB3 (B45)	–
RPA34.5^{v}	I	CAST (PAF49)	None	UBF and SL1
RPB5 (ABC27)	I, II, III	hRPB5	–	–
RPB6 (ABC23)	I, II, III	hRPB6	ω	–
RPA19 (AC19)	I, III	hRPA19	α; RPB11 (B12.5)	–
RPB8 (ABC14.5)	I, II, III	RPB8	–	–
RPA14^{v}	I	?	RPB4	–
RPA12^{c}	I	hRPA12.2	RPB9	–
RPB10(ABC10β)	I, II, III	RPB10	–	–
RPB12 (ABC10α)	I, II, III	RPB12	–	–

Il existe deux sous-populations de RPI : alpha (RPIα) et bêta (RPIβ). Seulement RPIβ est en mesure d'effectuer l'initiation de la transcription grâce à la liaison entre A43 de la sous-unité A14/A43 de RPI avec son facteur associé Rrn3. Cela représente moins de 10 % du RPI total présent dans la cellule (Miller et al., 2001). La liaison entre RPI et Rrn3 permet de former un complexe transcriptionnel fonctionnel au promoteur des gènes de l'ARNr. L'interaction entre A14/43 et Rrn3 est conservée de *S. pombe* à *H. sapiens* (Yuan et al., 2002; Imazawa et al., 2005). Le facteur Caséine Kinase II (CK2), qui module la transcription ribosomique, a été trouvé comme étant associé seulement à RPIβ. CK2 est recruté au promoteur des gènes de l'ARNr et effectue la phosphorylation d'UBF et de TAF110 du complexe SL1 régulant ainsi leur interaction (Lin et al., 2006).

Une étude a effectué l'inactivation (*knock-out*, KO), chez la souris, du gène codant pour l'ARN polymérase 1-2 (Rpo1-2), la deuxième plus grosse sous-unité de RPI (Chen et al., 2008). Cette étude a démontré que la perte de la transcription ribosomique induit la désorganisation de la structure nucléolaire et l'apoptose chez les embryons. En effet, la transcription de l'ADNr y est grandement diminuée. De plus, les embryons cessent leur développement au stade de morula (E3.5).

1.4.5 – RRN3

Tel que mentionné précédemment, *Transcription initiation factor IA* (TIF-IA, Rrn3) permettrait l'obtention d'un complexe de transcription compétent au promoteur des gènes de l'ARNr en s'associant avec une sous-population de l'ARN polymérase I, RPIβ (Miller et al., 2001; Cavanaugh et al., 2008). Rrn3 permet de faire le pont entre la sous-unité A43 de RPI et des sous-unités TAF68 et TAF110 chez l'humain (ou TAF95 chez la souris) du complexe SL1 (Miller et al., 2001). Cela permet le recrutement de RPI via SL1 au promoteur des gènes de l'ARNr. Une fois l'initiation de la transcription effectuée, Rrn3 se dissocie de la polymérase et celle-ci entre en élongation. La dissociation de Rrn3 est un prérequis pour l'élongation de la transcription (Bierhoff et al., 2008).

La kinase ERK phosphoryle Rrn3 aux sérines 633 et 649 (S633 et S649). Une étude a démontré que lorsque S649 est remplacé par une alanine, ERK ne peut plus effectuer sa phosphorylation et cela inactive Rrn3. Cela inhibe la synthèse de l'ARNr et retarde la croissance cellulaire (Zhao et al., 2003).

Malgré que Rrn3p soit essentiel chez la levure (Yamamoto et al., 1996), cela ne semble pas être le cas chez la souris. En effet, une étude a montré que les embryons où le KO de Rrn3 a été effectué peuvent se développer jusqu'à 9.5 jours (E9.5) (Yuan et al., 2005). Cela est étonnant puisque la contribution maternelle en ARN et en protéines provenant de l'ovocyte est généralement suffisante pour combler seulement le déficit de l'embryon jusqu'à ce qu'il atteigne le stade de deux cellules, là où il y a activation du génome fœtal (Pikó et Clegg, 1982). À l'opposé, le KO

d'autres facteurs impliqués dans la biogénèse des ribosomes ont révélé que les embryons étaient viables seulement jusqu'au stade morula (Chen et al., 2008; Hamdane et al., 2014). Dans des MEFs (*Mouse embryonic fibroblasts*) où le KO de Rrn3 a été effectué, on observe une perturbation du nucléole, un arrêt du cycle cellulaire, une activation de p53 et l'induction de l'apoptose (Yuan et al., 2005). Il a par contre été récemment montré que Rrn3 serait en fait essentiel, ce qui contredit les données précédemment présentées. En effet, en l'absence de Rrn3, les embryons arrêtent plutôt leur développement au stade morula (Herdman, Mars et al., 2017).

1.4.6 – TTF1

TTF1 est un facteur de terminaison de la transcription ribosomique; il permet d'arrêter la progression de l'ARN polymérase I. Il comprend 905 acides aminés chez l'humain et a une masse moléculaire de 103 kDa (selon la base de données *GeneCards*). Chez la levure, l'homologue de TTF1 se nomme Reb1p. Cette protéine permet d'arrêter la polymérase à la fin de la région codante, mais est incapable d'effectuer le relâchement du complexe ternaire (comprenant la polymérase, l'ARN naissant et la matrice d'ADN) au même titre que TTF1 (Mason et al., 1997).

TTF1 possède deux domaines de liaisons à l'ADN (*DNA binding domain*, DBD) faisant partie de la famille Myb ainsi qu'un domaine nommé *Negative regulatory domain* (NRD). Ce dernier, lorsque lié au DBD, inhibe la liaison à l'ADN offrant ainsi une forme d'autorégulation. Par contre, l'interaction avec *TTF-1 interactif protein 5* (TIP5), une sous-unité du complexe remodeleur de la chromatine nucléolaire (*nucleolar remodeling complex*, NoRC), permet de rétablir l'activité de liaison à l'ADN. Cela permet de recruter le complexe NoRC au promoteur des gènes de l'ARNr afin d'effectuer l'extinction (*silencing*) de la transcription de l'ADNr (Németh et al., 2004). TTF1 agit donc à titre d'activateur ainsi que de répresseur de la transcription ribosomique.

TTF1 lié au *proximal-promoter target site* (T_0) permet de recruter NoRC via TIP5 au promoteur. NoRC réorganise les nucléosomes et recrute des histones méthyltransférases, des ADN méthyltransférases et des histones déacétylases. Cela permet de modifier le statut des histones et de l'ADN afin d'éteindre les gènes de l'ARNr (Santoro et al., 2002).

TTF1 se lie au motif de 18 pb « AGGTCGACCAG(AT/TA)NTCCG » que l'on appelle Sal-box (Grummt et al., 1986). Ces Sal-box sont des sites terminateurs nommés T_{SP}, T_0 et T_1 à T_{10} (Németh et al., 2008) (Figure 1.13). Une mutation du terminateur T_0 ou son déplacement par aussi peu que 2 pb affecte sérieusement l'initiation de la transcription (Firek et al., 1989; McStay et Reeder, 1990). Des mutations dans les sites T_0 ou T_1 à T_{10} nuisent à la terminaison de la transcription (Grummt et al., 1986).

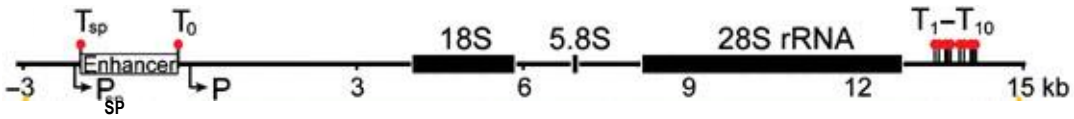

Figure 1.13 – Sites terminateurs. Représentation d'une répétition de l'ADNr. Les sites terminateurs sont en rouge. Extrait de Németh et al., 2008 et reproduit avec permission.

Une étude a montré que des niveaux non physiologiques de TTF1 inhibent la biogénèse des ribosomes en abolissant la synthèse des ARNr (Lessard et al., 2012). Cela suggère que des niveaux précis de TTF1 sont requis pour une biogénèse des ribosomes efficace.

La protéine ARF, un suppresseur de tumeurs, est responsable de la régulation négative de la biogénèse des ribosomes via TTF1. En effet, l'induction de ARF empêche la localisation de TTF1 au noyau en se liant à sa séquence de localisation nucléolaire (*Nucleolar Localisation Sequence*, NoLS). TTF1 voyage habituellement entre le nucléoplasme et le noyau; sa séquestration au nucléoplasme l'empêche donc de remplir son rôle dans la biogénèse des ribosomes. En temps normal, la protéine NMP (Nucléophosmine) permet la localisation nucléolaire de TTF1 en permettant potentiellement son déploiement à partir de son état auto-inhibé, ce qui exposerait son NoLS (Figure 1.14) (Lessard et al., 2010).

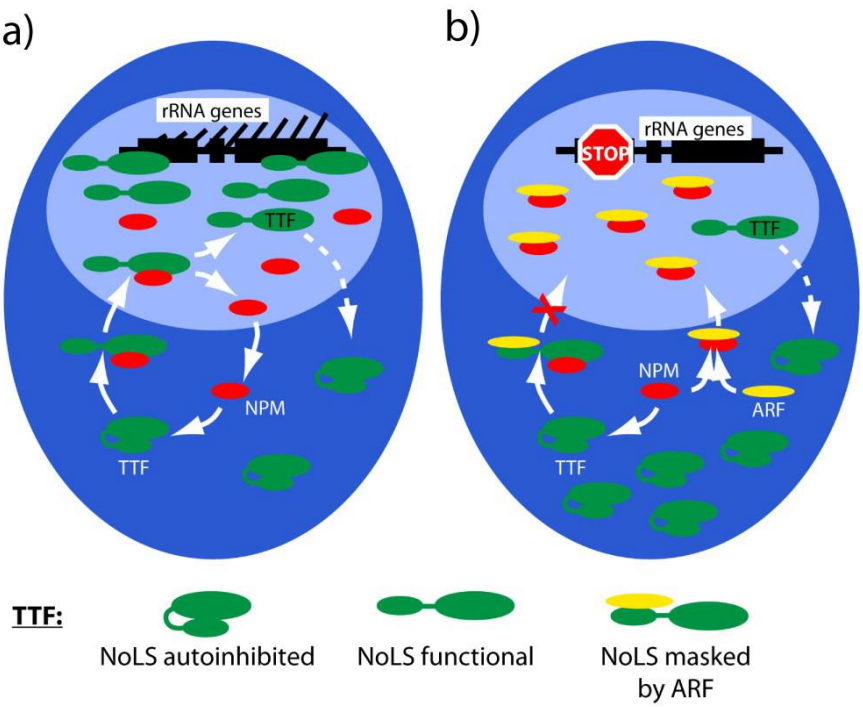

Figure 1.14 – Régulation de TTF1 par ARF. La zone en bleu foncé représente le noyau et la zone bleu clair représente le nucléole. a) NPM permet le déplacement de TTF1 au nucléole ce qui lui permet d'accomplir son rôle dans la régulation de la biogénèse des ribosomes. b) Après induction de ARF, la transcription ribosomique est inhibée par l'inactivation du NoLS de TTF1 ce qui empêche sa relocalisation au nucléole.

En plus de sa fonction dans la terminaison de la transcription, TTF1 agit aussi dans l'activité de la barrière des fourches de réplication (*Replication fork barrier*, RFB). Cette dernière permet d'éviter les collisions entre les machineries de réplication (réplisome) et de transcription en phase S, prévenant ainsi l'instabilité chromosomique (Gerber et al., 1997; Akamatsu et Kobayashi, 2015). Sur l'ADNr, les RFB se situent en aval du 47S où T_1 agit à titre de RFB polaire (unidirectionnel) tandis que T_4 et T_5 peuvent arrêter la fourche de réplication de façon bidirectionnelle. L'activité de la RFB à ces endroits dépend de TTF1, de RPI et d'un facteur du réplisome nommé *Timeless* (Akamatsu et Kobayashi, 2015). Des mutations au sein de l'élément terminateur ou une déplétion de TTF1 enlèvent l'activité de la RFB (Gerber et al., 1997). En l'absence de l'activité de RFB, la réplication est entravée par la transcription de l'ADNr (Akamatsu et Kobayashi, 2015).

Selon une étude de *chromosome conformation capture* (3C), TTF1 permet de rapprocher le promoteur et le site terminateur dans l'espace *in vivo*. Cela donne naissance au modèle « ribomoteur » dans lequel RPI pourrait être directement recyclé du site terminateur jusqu'au site d'initiation permettant ainsi une ré-initiation rapide. La majorité des gènes ayant ce type de boucle sont exempts de méthylation de l'ADN. Cela suggère que l'interaction médiée par TTF1 se retrouve uniquement sur la fraction des gènes actifs. La différence de topologie entre les gènes actifs et inactifs peut être causée par l'absence de TTF1 sur les gènes inactifs (Németh et al., 2008).

1.5 – La formation du complexe de préinitiation

1.5.1 – Mécanisme

Des expériences *in vitro* ont permis de mettre en avant un modèle séquentiel de formation du PIC. D'abord, UBF se lie aux séquences UCE et CPE du promoteur des gènes de l'ARNr en formant l'*enhancesome* (Bazett-Jones, 1994; Stefanovsky et al., 2002). Ensuite, SL1 se lie au promoteur par l'entremise d'UBF. Puis, Rrn3 interagirait avec à la fois avec SL1 et RPI afin d'amener RPI au site d'initiation de la transcription. TTF1, en se liant au T_0, remodèle la chromatine afin d'activer l'initiation de la transcription (Langst et al., 1997).

En contraste avec le modèle séquentiel classique, il existe une autre théorie, celle de l'holoenzyme. Dans des études faites chez la levure pour RPII, il a été montré que RPII était purifié avec plusieurs facteurs généraux de transcription (*general transcription factors*, GTF) en absence d'ADN (Koleske et Young. 1994). Cela suggère que la polymérase pourrait former un complexe avant de se lier à l'ADN pour initier la transcription. Une étude a montré que les composants requis pour la transcription ribosomique sont co-immunoprécipités avec un anticorps contre RPI en absence d'ADN. Cela indique que les facteurs tels qu'UBF, SL1 et Rrn3 sont capables d'interagir physiquement avec RPI sans avoir besoin de la présence de l'ADN (Seither et al., 1998). Dans une autre étude, ils ont trouvé que l'holoenzyme était plutôt formé de RPI, SL1, TTF1 et de facteurs de réplication et de réparation

tels que Ku70/80 et PCNA (Hannan et al., 1999). Cette étude semble plus convaincante puisque UBF ne fait pas partie de l'holoenzyme. En effet, UBF est requis pour le maintien de la structure ouverte du promoteur ainsi que pour former l'*enhancesome*. Il est alors contre intuitif de penser qu'UBF fasse partie de l'holoenzyme. Il est à noter, par contre, que ces études ont seulement été démontrées *in vitro*; il n'y a pas d'évidence d'un tel mécanisme *in vivo*.

1.5.2 – Initiation

La liaison d'UBF à l'UCE et au CPE est considérée comme la première étape de l'initiation de la transcription (Moss et Stefanovsky, 2002). La phosphorylation de la queue C-terminale d'UBF est reconnue par SL1, ce qui permet sa liaison au promoteur (Tuan et al., 1999). L'acétylation d'UBF permet son interaction avec l'hétérodimère PAF49/PAF53, une sous-unité de la sous-population du complexe RPI compétent pour l'initiation, ce qui permet de former le complexe d'initiation (Meraner et al., 2006). Le ciblage de PAF49/PAF53 par UBF pourrait induire un changement conformationnel de la polymérase permettant ainsi l'activation de la transcription (Panov et al., 2006). Par contre, une étude a montré que SL1 était en mesure d'interagir avec le promoteur et diriger la transcription de RPI en l'absence d'UBF ce qui remet la première théorie en question (Friedrich et al., 2005).

L'initiation de la transcription par RPI est dépendante de l'interaction directe entre Rrn3 et RPI (Schnapp et al., 1990). Par contre, pour que la transcription continue au-delà de l'initiation, RPI doit échapper au promoteur et se dissocier de Rrn3. Pour ce faire, il doit y avoir phosphorylation de Rrn3 par Caséine Kinase II (CK2) aux acides aminés Ser170 et Ser172 de Rrn3. Ces sérines sont conservées chez *H. sapiens*, *M. musculus*, *G. gallus* et chez *X. laevis*. Rrn3 serait ensuite déphosphorylé par FCP1, une sérine-thréonine phosphatase, ce qui permet la réassociation de Rrn3 à RPI et ainsi initier une autre ronde de transcription de l'ADNr (Bierhoff et al., 2008).

Il a été montré qu'une inhibition de la phosphorylation des sérines 170 et 172 ou encore l'attachement (*tethering*) de Rrn3 à la sous-unité RPA43 de la polymérase bloque la transcription ribosomique ce qui a pour conséquence une perturbation de la structure nucléolaire ainsi qu'un arrêt du cycle cellulaire (Bierhoff et al., 2008).

1.5.3 – Élongation

La biogénèse des ribosomes est un processus produisant une quantité importante d'ARNr. Pour y arriver, l'élongation de la transcription doit se faire de manière efficace. Une étude a montré que la phase d'élongation du gène de l'ARNr humain (13.3kb) prend, en tout, environ 140 secondes. Cela correspond à un taux de 95 nucléotides par seconde et d'environ 100 polymérases par gène (Dundr et al., 2002). Par contre chez la levure, un taux d'environ 60 nucléotides par seconde a été calculé, ce qui correspond à 50 ± 20 polymérases transcrivant simultanément une répétition du gène de l'ARNr (French et al., 2003).

Tel que mentionné précédemment, la vitesse d'élongation peut être modulée grâce à la phosphorylation d'UBF par ERK qui permet d'ouvrir l'*enhancesome* pour ainsi permettre le passage de la polymérase. L'étape d'élongation est importante pour le contrôle du taux de synthèse ainsi que du traitement de l'ARNr (Stefanovsky et al., 2006). Une étude suggère que l'élongation est l'étape limitante de la synthèse de l'ARNr (Hung et al., 2017).

1.5.4 – Terminaison

Le modèle proposé pour la terminaison de la transcription est le suivant. Premièrement, RPI est arrêté par TTF1 à la fin de la région transcrite. Ensuite, le facteur transactivateur *Pol I and transcription release factor* (PTRF) permet la dissociation du complexe ternaire de transcription menant ainsi au relâchement du transcrit et à la libération de la polymérase. Des séquences riches en thymidine en amont des sites terminateurs sont requis pour l'activité de PTRF (Jansa et al., 1998). Par contre, il a été montré que PTRF aurait un rôle important dans la formation et l'organisation des cavéoles, des invaginations de la membrane plasmique. En effet,

PTRF porte aussi le nom de cavin-1 pour *caveolae-associated protein 1* (Hayashi et al., 2009). De plus, des expériences de KO de cavin-1 chez la souris ont montré que cette délétion n'est pas létale dans ces organismes (Karbalaei et al., 2012). Il s'agit d'une constatation étonnante puisque le KO des divers facteurs impliqués dans la biogénèse des ribosomes a révélé que le développement embryonnaire est arrêté dès le stade de morula, tel que mentionné à la section 1.4.4. Cela, ainsi que le fait que les expériences sur PTRF ont été performées *in vitro* seulement, suggère que PTRF pourrait ne pas avoir de fonction dans le relâchement de la polymérase après tout.

1.5.5 – Réinitiation

Tel que mentionné précédemment, TTF1 permet de former une boucle rapprochant les éléments T_1 à T_{10} des éléments T_{SP} et T_0 (Németh et al., 2008). Cela permet un recyclage des facteurs de transcription ainsi que de la polymérase afin d'effectuer une réinitiation rapide.

1.6 – La bio-informatique

1.6.1 – Historique de la bio-informatique

La bio-informatique semble être une discipline relativement nouvelle aux yeux du grand public. Pourtant, l'histoire de la bio-informatique remonte aux années 1960 avec les travaux de la chimiste D^{re} Margaret Oakley Dayoff. Elle était la directrice associée du *National Biomedical Research Foundation*, une organisation encourageant le développement de logiciels informatiques et est considérée comme étant la mère et le père de la bio-informatique (Moody, 2004). Dayoff utilisait FORTRAN, le premier langage de programmation de haut niveau introduit par IBM en 1957, afin de résoudre des problèmes d'ordre biologique. Ses projets incluaient l'écriture d'une série de programmes FORTRAN afin de déterminer des séquences d'acides aminés des protéines. Elle a réussi à coder un programme ayant prédit la séquence d'une ribonucléase, une petite protéine, en l'espace de quelques minutes. Cette même tâche aura pris plusieurs mois à une équipe de recherche (Hagen, 2000). Dayoff fût responsable de la parution de l'*Atlas of Protein Sequence and*

Structure, une publication annuelle cataloguant toutes les séquences d'acides aminés connues des protéines (Hagen, 2000; Hogeweg, 2011). Il s'agissait de la première base de données en biologie moléculaire et devint rapidement une ressource indispensable pour la recherche computationnelle de l'époque.

L'ordinateur, point central de la bio-informatique, fût un outil important dans le domaine de la biologie moléculaire bien avant que le séquençage d'ADN soit rendu disponible, mais devint indispensable après cet avènement. Les ordinateurs furent rapidement ajoutés aux programmes de recherche dès les années 1960 (Hagen, 2000). L'évènement le plus important dans l'histoire de la bio-informatique fût sans doute le déploiement de l'internet. Cela permit d'améliorer l'accès aux données et aux publications ainsi que de transformer les bases de données (Kanehisa et Bork, 2003).

La définition de la bio-informatique a été introduite en 1970 dans un journal allemand par Ben Hesper. La traduction est la suivante : il s'agit de « l'étude des processus informatiques dans les systèmes biotiques » (Hogeweg, 2011). La bio-informatique est en fait un domaine multidisciplinaire regroupant les sciences informatiques, les statistiques, les mathématiques, l'ingénierie et la biologie. Son but est de développer des logiciels et des outils permettant l'analyse et la compréhension des données biologiques *in silico*, c'est-à-dire à l'aide d'un ordinateur. La bio-informatique regroupe entre autres les domaines suivants : la génomique, la protéomique, la phylogénétique, la modélisation moléculaire ainsi que la prédiction de structures secondaires et tertiaires des protéines.

Afin d'étudier le génome (gènes), le transcriptome (ARN messagers), le protéome (protéines), le métabolome (composants métaboliques) et l'interactome (interactions protéines-protéines), la bio-informatique a recours à plusieurs stratégies telles que la reconnaissance de motif, le *data mining*, l'intelligence artificielle (*machine learning*) et la visualisation. Les efforts de recherche sont concentrés sur l'alignement de séquences, la recherche de gènes, l'assemblage de génomes, le

design et la découverte des drogues, l'alignement des structures protéiques, la prédiction de structures protéiques, la prédiction de l'expression des gènes, les interactions protéines-protéines ainsi que les *genome-wide association study* (GWAS).

Durant les années 1980, les méthodes de séquençage d'ADN sont devenues largement disponibles (Kanehisa et Bork, 2003). C'est dans les années 1990 que la bio-informatique connu une croissance exponentielle avec l'avènement du séquençage de génomes entiers. Cela a commencé avec la bactérie *H. influenza* et la levure *S. cerevisiae*. Par la suite, des organismes plus complexes ont été séquencés tels que *C. elegans* et *D. melanogaster*. Ultimement, il eut le *Human Genome Project* visant à séquencer le génome complet de l'humain, qui fût achevé en 2003. La bio-informatique a donc été révolutionnée par les projets sur les génomes, mais aussi par les énormes bases de données, les super-ordinateurs et les ordinateurs personnels plus puissants.

Le développement rapide en biologie moléculaire et en informatique des dernières décennies va de pair avec l'accumulation d'une quantité énorme de données. Il est à noter, cependant, qu'une augmentation dans la quantité de données ne rime pas nécessairement avec un avancement des connaissances biologiques, à moins que cela soit accompagné de nouveaux outils bio-informatiques ou des améliorations d'outils existants. En effet, la bio-informatique permet l'extraction de résultats utiles à partir de grandes quantités de données.

Le GWAS permet de repérer les mutations responsables de maladies complexes telles que le cancer du sein (Véron et al., 2014), l'Alzheimer (Tosto et Reiz, 2013) et le diabète (Ionescu-Tîrgovişte et al., 2015) en cartographiant les mutations grâce aux données de séquençage de nouvelle génération.

Depuis les années 1980, plusieurs logiciels gratuits et *open source*, c'est-à-dire dont le code est accessible et peut être réutilisé et modifié par la communauté, ont fait

leur apparition et continuent de prospérer. Les différents logiciels utilisés en bio-informatique sont de types variés. Certains s'effectuent en ligne de commande tandis que d'autres possèdent une interface graphique plus conviviale (*user-friendly*). Il existe des plateformes permettant facilement l'analyse bio-informatique en y intégrant plusieurs logiciels de façon conviviale et rendant donc la bio-informatique accessible aux biologistes n'ayant pas nécessairement de compétences en programmation ou en informatique. La plateforme de ce genre la plus utilisée se nomme *Galaxy* (Afgan et al., 2016).

1.7 – Techniques de séquençage haut-débit

1.7.1 – Historique du séquençage

Au milieu des années 1970, la première méthode de séquençage a été développée par le Dr Frederick Sanger. Il s'agissait d'une méthode permettant la détermination de la séquence nucléotidique d'un fragment d'ADN simple brin par synthèse enzymatique (Sanger et Coulson, 1975). Cela leur a permis de séquencer le génome complet du bactériophage Φ X714, le premier génome à être reconstruit (Sanger et al., 1977). Pourtant, cette méthode est coûteuse, longue et demande beaucoup d'heures de travail en laboratoire (Metzker, 2005).

Ce n'est qu'en 1986 que le premier séquenceur automatique fait son entrée sur le marché; on parlera plus tard de séquenceur de première génération. Le AB370, développé par Applied Biosystem, permet d'identifier jusqu'à 500 kb par jour (Liu et al., 2012).

L'arrivée du séquenceur GS Titanium de la compagnie 454 Life Science marque l'apparition des séquenceurs de seconde génération (ou *Next-Generation Sequencing*, NGS). Cette machine génère des fragments de 400 nucléotides et produit 500 mégabases (Mb) par cycle de mesure (*run*) de 10h (Voelkerding et al., 2009).

Récemment, des appareils de séquençage de troisième génération ont vu le jour. Pacific Biosciences a commercialisé ce qu'on appelle le *single-molecule real time* (SMRT). Les appareils de troisième génération produisent les séquences les plus longues jusqu'à présent. Ils sont d'ailleurs plus rapides et demandent un moins gros effort dans la préparation des échantillons. Par contre, le taux d'erreurs de séquençage laisse place à l'amélioration (Bleidorn, 2015).

Les techniques de séquençage haut-débit ont influencé le domaine de la bio-informatique dans le sens qu'une nouvelle génération de logiciels a dû être développée afin d'analyser adéquatement les résultats de sortie (Al-Haggar et al., 2013). L'immense volume de résultats générés par le séquençage de seconde génération a d'ailleurs forcé les bio-informaticiens à créer des logiciels très performants. De plus, l'analyse de données a été révolutionnée, car le séquençage haut-débit a permis d'étudier le génome dans son ensemble.

1.7.2 – ILLUMINA HISEQ 2000

La machine HiSeq 2000 de la compagnie Illumina est la technologie de séquençage la plus utilisée à ce jour et a été employée pour produire les résultats de ce mémoire. Cette technologie, commercialisée en 2006, génère de courts fragments d'ADN, mais le fait à un très haut débit, ce qui en fait la machine de choix pour les scientifiques. C'est d'ailleurs la technologie la moins couteuse avec 0.02 USD par million de bases et est aussi celle qui génère le plus d'*output* (Liu et al., 2012).

Le fonctionnement de la machine HiSeq 2000 est le suivant. Les fragments d'ADN sont chargés sur une cellule à flux continue (*flow-cell*) contenant des adaptateurs spécifiques pour que les fragments d'ADN puissent y adhérer. Les fragments sont ensuite amplifiés en utilisant des ponts PCR. Les *flow-cell* contenant les fragments d'ADN amplifiés (*clusters*) sont par la suite séquencés par une technologie se nommant séquençage par synthèse. Les nucléotides marqués avec un fluorochrome sont incorporés au brin d'ADN par une polymérase. Le lien 3'OH est désactivé afin d'éviter l'incorporation de plus d'un nucléotide par cycle. La base est

identifiée par prise d'image. Le fluorochrome est ensuite enlevé et le 3'OH est réactivé afin de permettre un nouveau cycle d'incorporation (Liu et al., 2011).

1.7.3 – ChIP-seq

Le ChIP-seq est une technique permettant l'étude des interactions entre l'ADN et les protéines à la grandeur du génome. Elle utilise l'immunoprécipitation de la chromatine couplée au séquençage de nouvelle génération. Cette technique a été inventé en 2007 par le groupe de Keji Zhao (Barski et al., 2007).

Le protocole est le suivant. Les protéines liées sont fixées à l'ADN. Ensuite l'ADN est fragmenté à l'aide de la sonication. La protéine d'intérêt est alors immunoprécipitée, c'est-à-dire qu'un anticorps spécifique est utilisé afin de cibler la protéine permettant ainsi d'isoler l'ADN lié à celle-ci. Ensuite, l'ADN est purifié afin d'éliminer les protéines et l'échantillon est envoyé au séquençage haut-débit. Finalement, les fragments séquencés (*reads*) sont alignés sur le génome de référence (Figure 1.15).

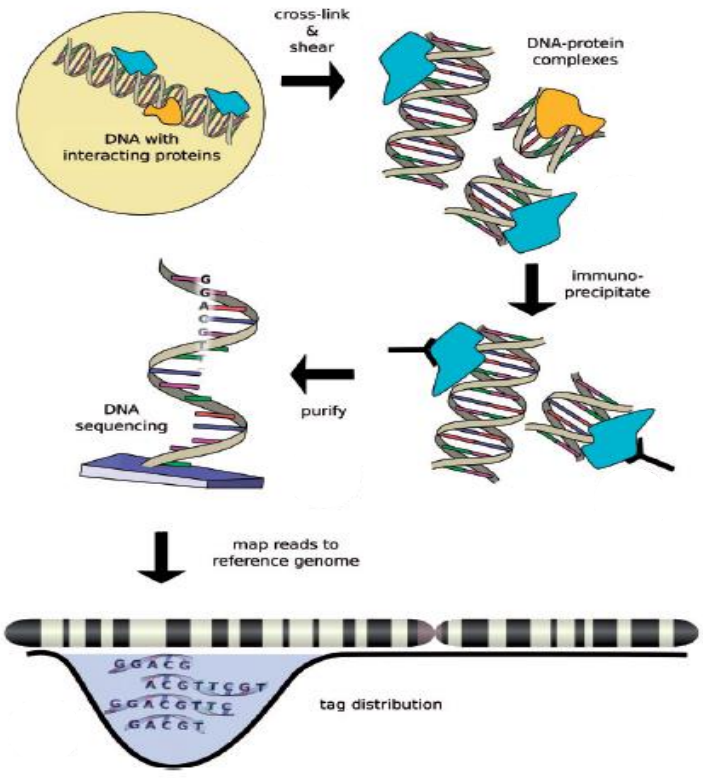

Figure 1.15 – Le ChIP-seq. Protocole expérimental de la technique de ChIP-seq. Extrait de Szalkowski et Schmid (2010) et reproduit avec permission.

Il est à noter que lors de l'alignement des *reads*, ceux-ci vont s'aligner de chaque côté du site de liaison. Afin de pallier à ce biais, il est important de faire une extension des *reads* (Figure 1.16). La longueur des fragments après l'extension est égale à la distance entre les brins Watson et Crick.

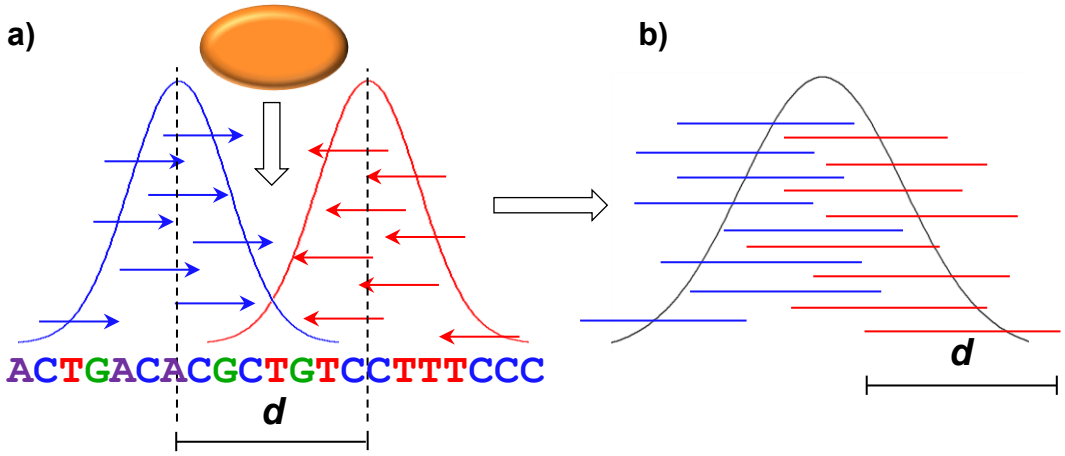

Figure 1.16 – Extension des *reads*. Il s'agit d'une étape importante dans le traitement des données ChIP-seq. a) La protéine en orange se lie au centre et les *reads forward* vont s'aligner à gauche tandis que les *reads reverse* vont s'aligner à droite. b) Après l'extension des *reads*, on obtient le vrai profil de liaison.

1.7.4 – DNase-seq

Le DNase-seq permet de visualiser les endroits accessibles du génome qui sont sensibles au clivage par la désoxyribonucléase I ou DNase I. La DNase I est une endonucléase qui clive les liens phosphodiesters de l'ADN (Madrigal et Krajewski, 2012). Premièrement, une extraction de la chromatine est performée. Deuxièmement, la DNase I effectue la digestion de l'ADN. Troisièmement, l'ADN est purifié afin de retirer les protéines ainsi que les nucléosomes. Quatrièmement, l'ADN est fractionné et la librairie est construite. Finalement, l'analyse bio-informatique, similaire à celle du ChIP-seq, est réalisée (Figure 1.17) (Zeng et Mortazavi, 2014).

Le DNase-seq complémente le ChIP-seq, car il permet d'identifier les endroits dépourvus de nucléosomes offrant donc la possibilité d'une liaison par un facteur de transcription. En effet, en combinant les deux techniques, il est possible de

distinguer les zones exemptes de facteurs de transcriptions de celles où ils sont présents.

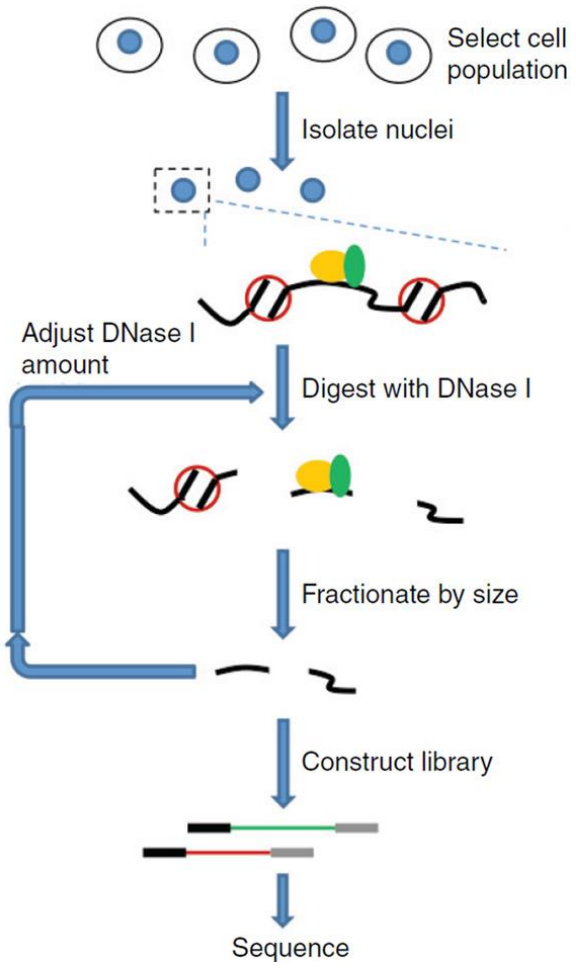

Figure 1.17 – Le DNase-seq. Protocole expérimental de la technique de DNase-seq. Extrait de Zeng et Mortazavi (2014) et reproduit avec permission.

1.8 – Hypothèses et objectifs

1.8.1 – Hypothèses du mémoire

Comme ce mémoire est composé de deux parties distinctes, il comprend deux hypothèses. En ce qui concerne la première, il a été remarqué que la couverture de séquence dans les échantillons de ChIP-seq est inégale sur l'ADNr, et ce, même pour les protéines ayant une distribution uniforme telle que RPI. Cela est dû majoritairement au fait que le séquençage ne permet pas un recouvrement égal des fragments séquencés. Il serait donc intéressant de développer un outil bio-

informatique permettant de palier à ce biais afin de révéler le vrai profil d'interaction protéique. Afin d'y arriver, les objectifs 1 à 3 ont été élaborés. En ce qui a trait à la deuxième hypothèse, des sites de liaison protéiques d'UBF ont été observés ailleurs que sur les gènes de l'ARNr. Puisque les rôles de cette protéine sont essentiellement associés aux gènes ribosomiques, il était étonnant de constater cette observation. L'identification des rôles potentiels d'UBF à la grandeur du génome suscitent un intérêt scientifique permettant d'approfondir nos connaissances sur cette protéine. La possibilité qu'UBF joue un rôle ailleurs dans le génome constitue donc la seconde hypothèse, développée dans l'objectif 4.

1.8.2 – Objectifs du mémoire

Les objectifs de ce mémoire sont les suivants :

1 – Développer un outil permettant la normalisation par déconvolution de données pouvant être utilisé pour traiter les résultats obtenu par expériences de ChIP-seq et de DNase-seq.

2 – Appliquer le procédé de déconvolution afin de normaliser tous les échantillons générés au laboratoire ainsi que certains présents dans les bases de données publiques.

3 – Utiliser la déconvolution afin de cartographier le *spacer promoter* chez l'humain et la souris.

4 – Déterminer les rôles potentiels d'UBF à l'échelle du génome.

CHAPITRE 2 – A DECONVOLUTION PROTOCOL FOR CHIP-SEQ REVEALS ANALOGOUS ENHANCER STRUCTURES ON THE MOUSE AND HUMAN RIBOSOMAL RNA GENES

Jean-Clément Mars[1,2]*, Marianne Sabourin-Félix[1,2]*, Michel G. Tremblay[1,2] and Tom Moss[1,2].

* Equal first authors. [1] Laboratory of Growth and Development, St-Patrick Research Group in Basic Oncology, Cancer Division of the Quebec University Hospital Research Centre. [2] Department of Molecular Biology, Medical Biochemistry and Pathology, Faculty of Medicine, Laval University.

Keywords

ChIP-Seq Deconvolution, RNA polymerase I (RPI, PolI, Polr1), Ribosomal RNA (rRNA) genes, Ribosomal DNA (rDNA), Upstream Binding Factor (UBF/UBTF), Selectivity Factor (SL1)

Cet article a été publié dans le journal G3: Genes | Genomes | Genetics. Jean-Clément Mars, Marianne Sabourin-Félix, Michel G. Tremblay and Tom Moss, *A deconvolution protocol for chip-seq reveals analogous enhancer structures on the mouse and human ribosomal RNA genes*, G3, January 2018.

Cet article est également disponible en ligne :

http://doi.org/10.1534/g3.117.300225

http://www.g3journal.org/content/8/1/303.long

PubMed: https://www.ncbi.nlm.nih.gov/pubmed/29158335

2.1 – AVANT-PROPOS

Le présent chapitre représente la version intégrale, sans modification, de l'article intitulé *A Deconvolution Protocol for ChIP-Seq Reveals Analogous Enhancer Structures on the Mouse and Human Ribosomal RNA Genes*. Celui-ci a été accepté pour publication le 15 novembre 2017 dans le journal Genes | Genomes | Genetics (G3) et a été publié dans l'édition de janvier 2018. Dans cet article, Jean-Clément Mars et moi-même sommes les deux premiers co-auteurs, le deuxième auteur est Michel G. Tremblay et le dernier auteur, soit le chercheur, est le Dr Tom Moss. Pour cet article, Michel G. Tremblay a produit les souris qui ont servies à générer les lignées cellulaires MEFs conditionnelles pour UBF. Jean-Clément Mars a effectué les expériences de ChIP-seq et de DNase-seq. Pour ma part, j'ai travaillé au développement d'une nouvelle méthode de normalisation des données de séquençage haut-débit appliquée au traitement des données ChIP-seq et DNase-seq. J'ai d'ailleurs effectué les analyses bio-informatique permettant l'obtention des données brutes prise comme première étape de la normalisation. Le Dr Tom Moss a généré les figures et rédigé le manuscrit. Tous les auteurs ont contribué à la révision et la correction du manuscrit. Cette publication est reproduite avec l'autorisation de tous les co-auteurs.

2.2 – RÉSUMÉ

Le ChIP-seq permet une meilleure compréhension de la structure des *enhancers* et de la chromatine. Toutefois, sa résolution est limitée par plusieurs facteurs. En appliquant le ChIP-seq pour l'étude des gènes de l'ARN ribosomique (ADNr), nous avons trouvé que la limitation majeure de la résolution réside dans la variabilité sous-jacente dans la couverture qui domine le profil d'interaction entre l'ADN et les protéines. Nous décrivons une approche de déconvolution qui corrige la variabilité et améliore la résolution des interactions déterminées par ChIP-seq. Cette approche a permis d'étudier l'organisation *in vivo* du complexe de préinitiation de RPI se formant aux promoteurs dans les gènes de l'ARN ribosomique chez l'humain et la souris, et a révélée une liaison en phase d'UBF le long de l'ADNr. De plus, les données ont permis d'identifier et de cartographier un *spacer promoter* et une polymérase arrêtée associée au sein de l'*Intergenic spacer* du gène ribosomique humain.

2.3 – ABSTRACT

The combination of Chromatin Immunoprecipitation and Massively Parallel Sequencing, or ChIP-Seq, has greatly advanced our genome-wide understanding of chromatin and enhancer structures. However, its resolution at any given genetic locus is limited by several factors. In applying ChIP-Seq to study the ribosomal RNA genes we found that a major limitation to resolution was imposed by the underlying variability in sequence coverage that very often dominates the protein-DNA interaction profiles. Here we describe a simple numerical deconvolution approach that in large part corrects for this variability and significantly improves both the resolution and quantitation of protein-DNA interaction maps deduced from ChIP-Seq data. This approach has allowed us to determine the *in vivo* organization of the RNA Polymerase I preinitiation complexes forming at the promoters and enhancers of the mouse and human ribosomal RNA genes, and reveal a phased binding of the key factor UBF across the rDNA. The data further identify and map a “Spacer Promoter” and associated stalled polymerase in the Intergenic Spacer of the human ribosomal RNA genes, and reveal a very similar Enhancer structure to that in rodents and lower vertebrates.

2.4 – INTRODUCTION

Data from Chromatin Immunoprecipitation (ChIP) combined with Massively Parallel DNA Sequencing (ChIP-Seq) can potentially provide high-resolution maps of transcription and chromatin factor interactions throughout the genome. The absolute resolution of these maps is determined by the size-range of chromatin fragments that are selected during the immunoprecipitation step. However, in practice several other factors limit the resolution achieved by the technique. These include the relative accessibility of the targeted protein-DNA complex (Teytelman et al. 2013), the efficiency of crosslinking, the combined effects of these limitations on complex recovery (Poorey et al. 2013) and the selectivity of the immunoprecipitation step. But a major limitation to mapping resolution is also imposed by the strong biases in DNA sequence coverage inherent in the Massively Parallel DNA Sequencing protocols. Sequence coverage biases have previously been noted for mitochondrial DNAs and shown to correlate with DNA composition and certain sequence motifs (Ekblom et al. 2014). Several data normalization approaches have been developed to correct for biases in sequence coverage maps (Park 2009; Kidder et al. 2011; Chen et al. 2012; Taslim et al. 2009), but are predominantly aimed at improving the reliability of the peak calling routines used to identify potential factor binding sites genome-wide and have had only limited success (Teytelman et al. 2013). However, when investigating details of factor binding at given sites within the genome, these approaches fail to correct for local biases in sequence coverage, and hence do little to improve mapping resolution of complexes at specific DNA sites.

Here we show that a simple numerical deconvolution approach successfully removes the sequencing biases introduced into ChIP-Seq data by massively parallel DNA Sequencing techniques and greatly improves the resolution of protein-DNA interaction maps. We have applied this approach to better understand the structure of the duplicated RNA Polymerase I (RPI/PolI) promoters, preinitiation complexes and Enhancers that form on the ribosomal RNA genes (rDNA) of mouse and human. Duplications of RPI promoters are found within the rDNA Intergenic Spacers (IGS) of insects, amphibia and rodents and are often referred to as “Spacer Promoters”.

They were first identified in the rDNA IGS of *Xenopus laevis* (Moss and Birnstiel 1979) and of *Drosophila melanogaster* (Coen and Dover 1983; Miller et al. 1983), but later were also found in other *Xenopus* and *Drosophila* species and in mouse, chinese hamster, rat and even plants (Bach et al. 1981; Murtif and Rae 1985; Kuhn and Grummt 1987; Tower et al. 1989; Cassidy et al. 1987; Doelling et al. 1993). These Spacer Promoters function as part of upstream transcriptional enhancer elements (Moss 1983; De Winter and Moss 1986, 1987; Paalman et al. 1995; Caudy and Pikaard 2002), and are often repeated several times within a given IGS, reviewed in (Moss et al. 1985; Moss and Stefanovsky 1995; Moss et al. 2007). More recently, the mouse Spacer Promoter has been suggested to be the source of a long non-coding RNA (lncRNA) that is responsible for in trans silencing and heterochromatinization of the rDNA at centric and pericentric chromosomal repeats (Guetg et al. 2010; Savic et al. 2014). But despite their demonstrated importance in transcription and silencing, the mouse and rat Spacer Promoters remain only partially mapped, while the existence of Spacer Promoters in other mammals and even in human is still largely a matter of speculation. Our deconvolution protocol revealed significant in vivo detail of the RNA Polymerase I (RPI or PolI) preinitiation complexes that form at the functional 47S ribosomal RNA (rRNA) Gene Promoters and the Spacer Promoters in mouse, and showed that they are indistinguishable despite the very poor homology between the underlying DNA sequences. The deconvolution protocol further identified and mapped a Spacer Promoter in the human rDNA IGS and showed that it exists in the context of an Enhancer complex closely resembling that occurring in mouse.

2.5 – Materiels and Methods

2.5.1 – Chromatin immunoprecipitation (ChIP)

Cells were fixed with 1% formaldehyde for 8 min at room temperature. Nuclei were isolated using Lysis Buffer (10 mM Tris pH 7.5, 10 mM NaCl, 3 mM MgCl2, 0.5% NP-40), transferred to Sonication Buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 2 mM EGTA, 4 mM EDTA, 0.1% SDS, 1% Triton X-100, 1% NP-40) and sonicated (Bioruptor, Diagenode) for 30 cycles of 30 sec on / 30 sec off at high intensity. Each

immunoprecipitation (IP) was carried out on the equivalent of 50 x 106 cells in IP Buffer (150 mM NaCl, 50 mM Tris-HCl pH 7.5, 5 mM EDTA, 0.5% NP-40, 1% Triton X-100) overnight at 4°C. The antibody slurry was prepared with 50 µl A-, 50 µl G-Dynabeads and 60 µg.ml-1 antibody per IP. Immunoprecipitated chromatin was treated with RNaseA and the DNA isolated using 2% Na SDS and 2mg.ml-1 Proteinase-K. Two or more biological replicates were analyzed for each antibody.

2.5.2 – Analysis of ChIP sample by massively parallel sequencing

ChIP DNA samples were quality controlled by qPCR as previously described (Herdman et al. 2017), before being sent for library preparation and 50 base single-end sequencing on an Illumina HiSeq 2000 by Genome Quebec (McGill University and Genome Quebec Innovation Centre).

2.5.3 – ChIP-Seq data alignment

The raw fastq.gz files from ChIP and input DNA were checked for quality using FastQC version 0.11.4 (Babraham Bioinformatics, S. Andrews). The data were then trimmed using Trimmomatic version 0.33 (Bolger et al. 2014) with the following parameters : LEADING:32, TRAILING:32, MINLEN:36, ILLUMINACLIP:TruSeq3-SE.fa:2:30:10. The resulting trimmed files were aligned to modified versions of the mouse and human genomes using Bowtie2 (Langmead and Salzberg 2012) with option -k 3. Alignment of the mouse data was to the mouse genome version GRCm38, to which a single copy of the rDNA repeat sequence (GenBank BK000964v3) was added as an extra chromosome. For convenience, the origin of the rDNA repeat was displaced to the EcoRI site at 30,493 such that the pre-rRNA initiation site now fell at nucleotide 14,815.

Alignment to the human rDNA proved a little more difficult using the same strategy due to the multiple rDNA sequences already present in version GRCh38. We therefore first searched the human *in silico* genome for regions most likely to interfere with alignment of rDNA sequences. The “canonical” rDNA repeat sequence (GenBank accession number U13369.1) was fragmented to generate 50 bp non-

overlapping pseudo-reads and these aligned on GRCh38 using Bowtie2 with the -k 10 option. This identified three major regions that would interfere with ChIP-Seq data alignment. The reference genome was, therefore, modified to remove these occurrences; the chromosomes chr22_KI270733v1_random and chrUn_GL000220v1 were removed and the rDNA sequence present on chromosome 21 was replaced with N (8,202,082-8,552,360). A single copy of the human rDNA repeat (GenBank accession number U13369.1) was then added as an extra chromosome. For convenience, the origin of the rDNA sequence was moved to the EcoRI site at 30,487 such that the pre-rRNA initiation site now fell at nucleotide 12,514.

2.5.4 – Deconvolution protocol

The rDNA chromosome was first extracted from the aligned file with the view command of SAMtools (Li et al. 2009). The rDNA data were then converted from BAM to BED6 format using the bamtobed command of the BEDTools suite version 2.25.0 (Quinlan and Hall 2010). Each read was extended 3' to the mean fragment length computed using the makeTagDirectory command of HOMER v4.3 (Heinz et al. 2010). Estimated fragment lengths fell between 75 and 125 dependent and so were standardized to the mean size of 100 bp. The coverage was then extracted with the genomecov command of BEDTools, smoothed using a 25 bp sliding window, and adjusted to reads per million (RPM). Data deconvolution was achieved by dividing the calculated sample DNA coverage by the appropriate input DNA coverage in order to remove the sequence coverage biases introduced by the sequencing protocol, as described in the main text. At positions where coverage in either data set was of low statistical significance, the deconvoluted data was set to 0 and ignored in subsequent interpretations. The resulting deconvoluted ChIP-Seq data was converted to BedGraph format and visualized using IGV (Integrative Genomics Viewer 2.3, Broad Institute). The manual for the deconvolution protocol and a corresponding Python script can be found at https://github.com/mariFelix/deconvoNorm. Gaussian curve fitting to rDNA promoter

sub-regions was perform using MagicPlot Pro (Magicplot Systems) on data extracted from the BedGraph files.

2.5.5 – Alignment of ChIP-nexus data

The 5' ends of reads from the ChIP-nexus datasets were mapped by first aligning sequences using Bowtie2 as above, but using the unique mapping -k 1 option. A Bedgraph of coverage for the 5’ position of each aligned read was then extracted using the genomecov command of BEDTools with the parameters -5, and -strand + (for forward reads) or -strand - (for reverse reads), and visualized using IGV.

2.5.6 – Data availability

Mouse strains are available from Jackson Laboratories (JAX Stock No. 029470, Ubtf<tm1.1Tmss>/J) and a very limited supply of derived cell lines may also be available upon request. Human cell lines are available from ATCC. The mouse mapping data can be found on ArrayExpress under the accession number E-MTAB-5839. The human data for UBF and RPI in K562 cells can be found on ArrayExpress under the accession number E-MTAB-6032. The HEK293T (UBF, RPI and input) and K562 (UBF and input) data from Zentner et al. (Zentner et al. 2011) can be found on the SRA database under the accession number SRP004897. The K562 data (UBF, TBP and input) from ENCODE can be found on the GEO DataSets database under the accession number GSE31477. The K562 data (CTCF and input) from ENCODE can be found on the GEO DataSets database under the accession numbers GSE29611 and GSE70764. The ChIP-exonuclease data for TBP can be found on the GEO DataSets database under the accession number GSE55306. A manual for the deconvolution protocol, a corresponding Python script and sample datasets can be found at https://github.com/mariFelix/deconvoNorm.

2.6 – Results

In order to better understand the *in vivo* functions of the RPI transcription factors, as part of an extensive study (Herdman et al. 2017), we performed ChIP analysis of wild type and conditional mouse embryonic fibroblasts using antibodies specific for

the various factors and subjected the resulting DNA fragments to massively parallel sequencing. The raw data was quality checked and trimmed and then aligned to the digital mouse genome that included a single rDNA repeat using Bowtie2, see Materials and Methods. Examples of the resulting factor binding profiles are shown in Figure 2.1A.

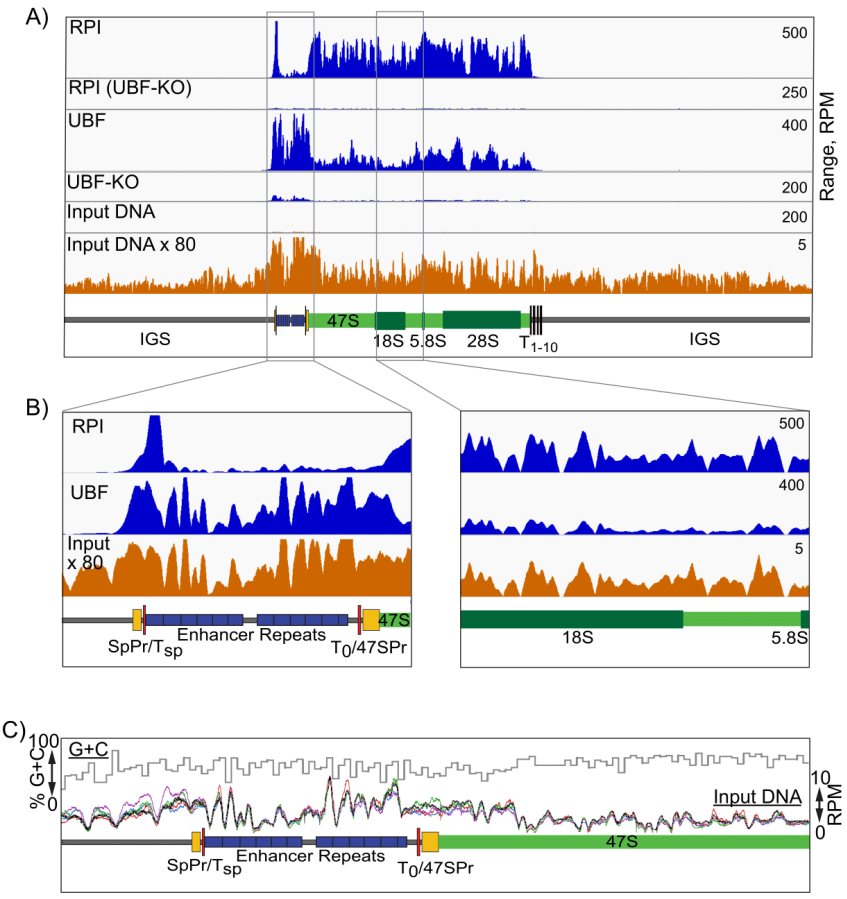

Figure 2.1 – Sequence coverage dominates the raw ChIP-Seq profiles for UBF and RPI. A) and B) Comparison of the ChIP-Seq profiles for RPI and UBF with the sequencing coverage for unselected "input" DNA. UBF and RPI ChIP-Seq data after UBF knock out "UBF-KO" is shown to demonstrate the specificity of the respective antibodies used. Diagrammatic maps of the rDNA are given below the mapping profiles, showing the 47S transcribed region and the 18S, 5.8S and 28S genes in green, the Enhancer repeats in blue, the extents of the 47S and known Spacer Promoters (47SPr, SpPr) in yellow and the TTF1 binding sites Tsp, T0 and T1-10¬ in red. Coverage across the complete rDNA repeat is shown in A and enlargements across the Enhancer and the central 47S transcribed regions in B. The vertical scales in A and B are given in Reads Per Million (RPM). C) A superimposition of sequence coverage in RPM for 5 biological replicas of unselected "input" DNA is shown below the percent G+C sequence composition across the upstream region of the rDNA.

When mapping RNA Polymerase I (RPI/PolI) engagement across the mouse rDNA gene body by ChIP-Seq, we expected to observe the dense, relatively even distribution of RPI seen in electron-microscope images of single mouse rRNA genes (Scheer and Benavente 1990). In contrast, the ChIP-Seq coverage maps suggested an extremely uneven distribution of RPI (Figure 2.1A), as had been previously noted in human (Zentner et al. 2011). This was even more surprising considering that the ChIP technique should reveal the summed RPI distribution across the several hundred active rRNA gene copies in each cell as averaged over a population of many millions of cells. Similarly, sequence coverage maps for the multi-HMGB-box factor UBF (UBTF) also suggested very variable occupancy across the gene (Figure 2.1A).

2.6.1 – ChIP-Seq profiles result from a convolution of the protein crosslinking and sequencing coverage profiles

ChIP of both UBF and RPI was extremely specific, since conditional inactivation of the floxed UBF gene (UBF-KO) in MEFs strongly suppressed sequence enrichment when using antibodies against either factor, RPI engagement being dependent on UBF (Hamdane et al. 2014; Herdman et al. 2017) (Figure 2.1A). Strikingly, both the RPI and UBF sequence coverage profiles displayed a strong similarity to the coverage distribution obtained for unselected (input) genomic DNA from the same chromatin preparations. This similarity was clearly apparent when sequence coverage was compared at higher resolution (Figure 2.1B). Both in the case of RPI and UBF, the ChIP-Seq profiles closely followed the input DNA sequence profiles over the same regions. Hence, the RPI and UBF interactions profiles were clearly superimposed on a pattern resulting from the unevenness of sequence coverage, and indeed this pattern dominated these interaction profiles. However, we noted that the pattern of input DNA sequence coverage was highly reproducible between biological preparations (Figure 2.1C). Thus, it was clearly a property intrinsic to the massively parallel sequencing protocol and did not result from variations in sample preparation. But, unlike the bias in sequence coverage observed for mitochondrial DNA (Ekblom et al. 2014), we saw little if any correlation with the local rDNA GC

content (Figure 2.1C). The coefficient of determination R2 between the mean input read profile of the 5 datasets shown and the GC content, both determined over 25 bp windows, was 0.07 for the full rDNA repeat and 0.002 for the 47S transcribed region.

2.6.2 – Deconvolution of ChIP-Seq data provides greatly improved resolution in protein-DNA interaction maps

The reproducibility of input sequence coverage profiles suggested that it should be possible to remove these sequencing biases by numerical deconvolution. However, despite average input DNA sequencing depths of well over 100, initial attempts at deconvolution by directly normalizing the raw sample to input (sample coverage / input coverage) for each base position gave an unacceptable level of noise in the mapping profile. To counter this without significantly affecting mapping resolution, we incorporated two steps prior to deconvolution (Figure 2.2A). Sequences were first extended to the predicted DNA fragment length, then sequence coverage was smoothed using a sliding window, see examples for RPI and UBF (tracks 1 to 3, Figure 2.2B and C). DNA fragment lengths were estimated using HOMER (Heinz et al. 2010) and found to consistently fall between 75 and 125bp. Thus, for convenience DNA fragment sizes of all sample and input data sets were standardized to the mean size of 100bp. We also investigated smoothing using three sizes of sliding window "w" (11, 25 or 51bp), such that;

$$\text{Smoothed base coverage } J = \frac{1}{w} \times \sum_{n-(w-1)/2}^{n-(w-1)/2} j_n$$

where, *j* = aligned raw coverage and *n* = base position.

We found a window of 25bp gave the best compromise between improved signal to noise and mapping resolution after deconvolution for our data sets. This said, we later found that for the datasets analyzed here smoothing did not give significant improvements in the final profile, but may still help in cases of low read density. See Materials and Methods for more detail.

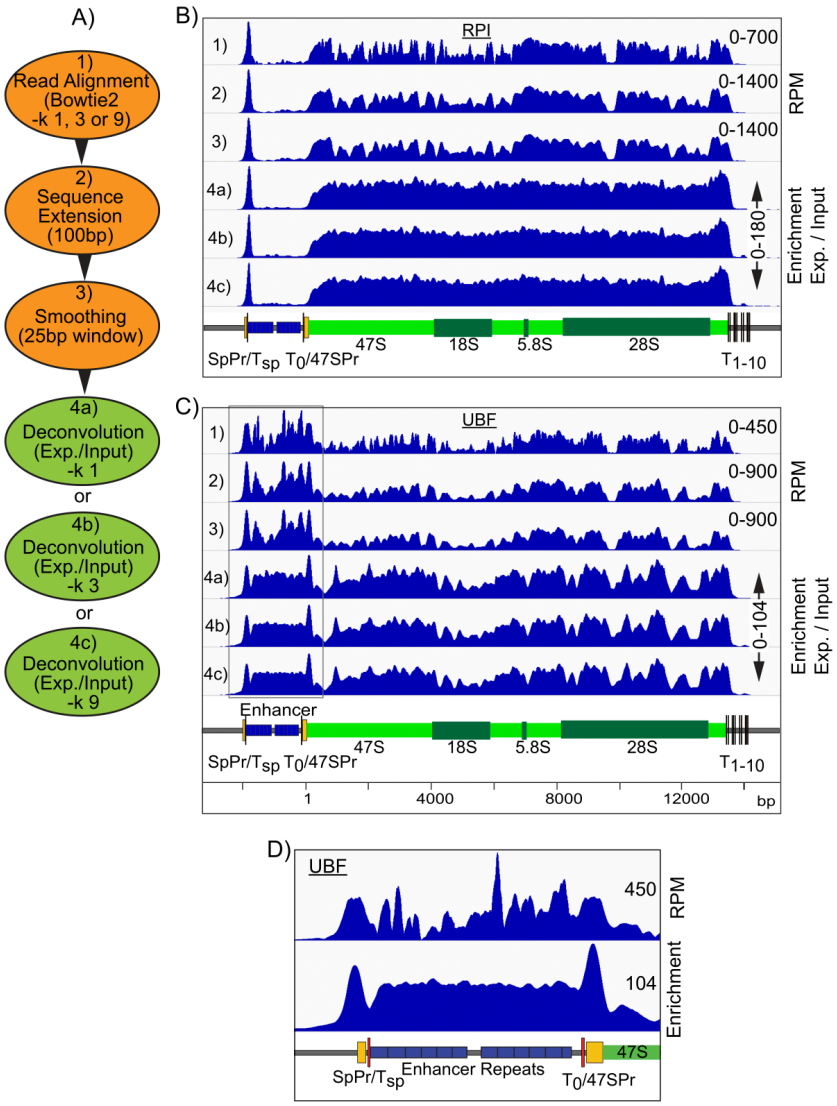

Figure 2.2 – Improved mapping with ChIP-Seq deconvolution. A) Summary of ChIP-Seq data handling, steps 1 to 3, and deconvolution, step 4. B) and C) Examples of the sequence coverage maps across the mouse rDNA at each step, 1 to 4, of data treatment respectively for RPI and UBF ChIP-Seq. D) Comparison of UBF mapping over the upstream gene region before (as in C lane 1) and after (as in C lane 4b) deconvolution. In B) through D) the vertical scales are given either in Reads Per Million (RPM) or as Enrichment relative to input DNA, and diagrammatic maps of the rDNA are given below the mapping profiles.

Given that the rDNA unit is present about 200 times in the biological mouse and human haploid genomes (Jackson et al. 2000; Henderson et al. 1974; Henderson et al. 1972), and several rDNA pseudogene fragments are present in the annotated mouse in silico genome, we investigated the effects of permitting Bowtie 2 to report multiple alignments for each sequence read. The -k Reporting Mode parameter in

Bowtie2 defines the number of genomic matches that are reported in the final alignment. We compared the alignments generated allowing only unique matches with those when up to 3 or 9 matches were allowed (- k 1 , 3, 9) (Figure 2.2A). Improvements in mapping between – k 1 and 3 were small (Figure 2.2B and C, tracks 4a, b and c), but in some regions of the rDNA, such as over the enhancer repeats, UBF mapping became more uniform, consistent with the expected binding of this factor (Putnam and Pikaard 1992; Hamdane et al. 2014). Increasing – k to 9, gave little further improvement. Since increasing the –k parameter in Bowtie2 also proportionately increased the computing time and the size of the resultant files, we set –k to 3 for all alignments.

The overall improvement in factor mapping using the deconvolution protocol can be qualitatively judged by comparing UBF binding across the Enhancer repeats as computed using Bowtie2 or the same alignment followed by the deconvolution protocol (Figure 2.2B-D). For example, a peak of UBF binding positioned over the Spacer and 47S promoters was only convincingly observed after deconvolution (Figure 2.2D).

2.6.3 – Reproducibility of deconvolution factor-binding profiles

To determine the degree of reproducibility of factor binding deduced using the deconvolution protocol, we compared the binding profiles obtained from different combinations of ChIP-Seq and input DNA biological replicates. Figure 2.3A shows each of two UBF ChIP-Seq replicates deconvoluted using sequence coverage obtained from three independent input DNA samples. Small variation in binding profile can be detected, but the overall distribution of UBF is essentially the same in all six calculations. This can be best judged when the Standard Deviation between these data sets is plotted against the mean binding profile from all six (Figure 2.3B). Here it can be seen that the variability between the profiles is no more than 10% and small enough that for most purposes it can be neglected.

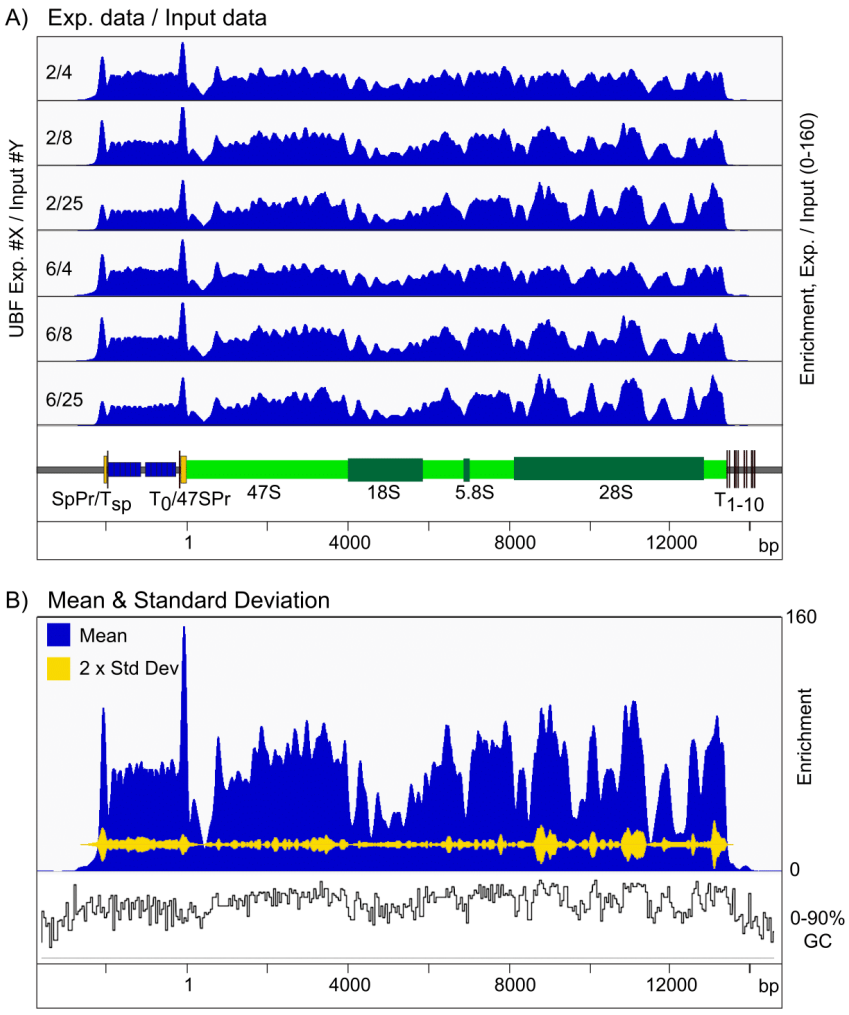

Figure 2.3 – ChIP-Seq deconvolution maps are highly reproducible. A) Comparison of two biological replicates of UBF ChIP-Seq data deconvoluted using data from three biological replicate "input" DNAs. B) Mean coverage from the 6 deconvolutions in A) is shown in blue and their standard deviation in yellow. The vertical scale in A) and B) gives the Enrichment relative to input DNA. A diagrammatic map of the rDNA is given below the mapping profiles in A). The percent G+C sequence composition across the rDNA is also shown in B).

2.6.4 – UBF POSITIONNING OVER THE 47S TRANSCRIBED REGION IS NOT RANDOM

UBF bound almost continuously throughout the 47S transcribed region, but even after deconvolution the interaction profile was much less uniform than that of RPI (compare Figure 2.2B and C), suggesting a non-random positioning of this factor. Over the 47S transcribed region the mean UBF profile followed the local GC content of the rDNA (Figure 2.3B), and the coefficient of determination R2 between these profiles of 0.47 indicated significant correlation (Figure 2.7A). This strongly

suggested that the peaks and troughs of the UBF interaction profile resulted at least in part from a preferential positioning of this factor. We counted around 74 peaks of UBF enrichment within the 47S transcribed region (Figure 2.7B), and these peaks displayed a mean spacing of 170 ± 58 bp. This was roughly consistent with the measured DNA contact length of a UBF dimer (Stefanovsky et al. 1996; Bazett-Jones et al. 1994), see Discussion.

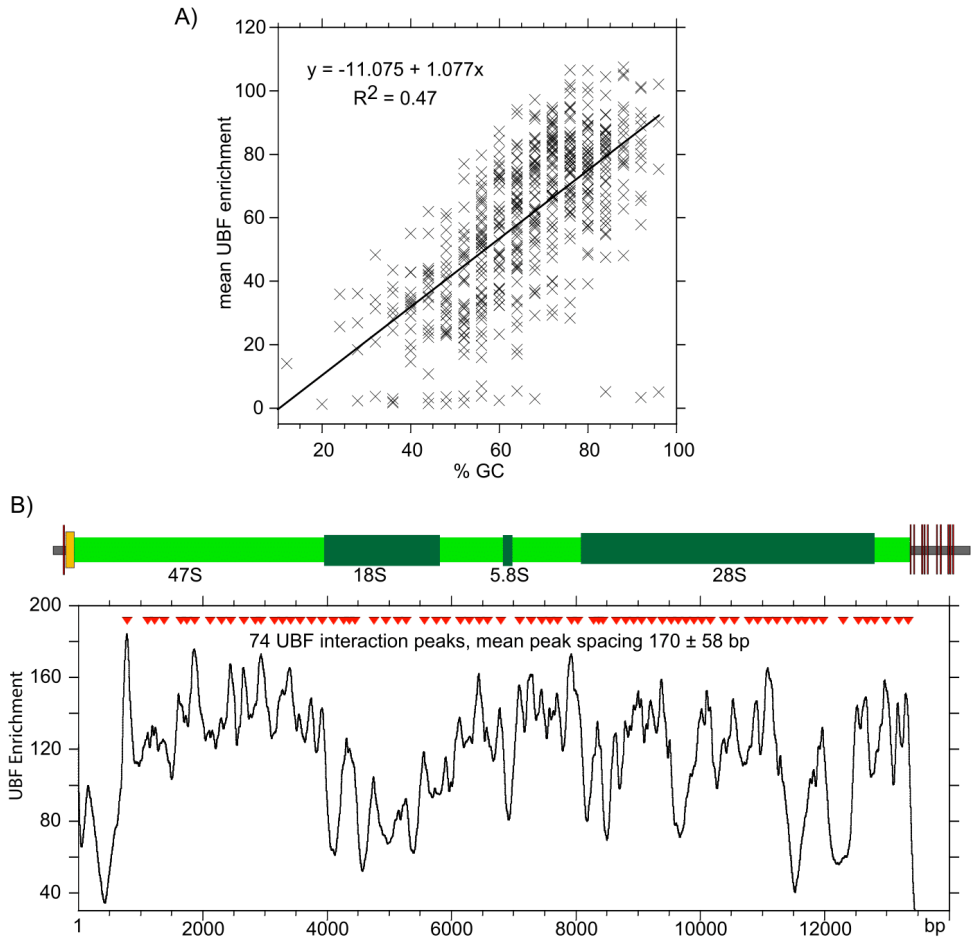

Figure 2.7 – Preferential positioning of UBF across the mouse rDNA. A) The mean UBF enrichment data from Figure 3B is shown plotted against the rDNA GC content using 25 bp data windows. The least squares linear best t and the corresponding coefficient of determination R2 are shown. B) Peak positions of UBF interaction across the 47S transcribed region were identified for a single deconvoluted UBF dataset (E-MTAB-5839, ChIP-seq_UBF_MEFs_UBFwt_4HT_Rep3.bedgraph). The mean spacing of peaks was calculated as 170 bp ± a standard deviation of 58 bp.

2.6.5 – Applying deconvolution ChIP-Seq to map the mouse rDNA Spacer Promoter

A functional Spacer Promoter was shown to lie within a 350bp region of the mouse IGS (-2279 to -1930bp relative to the 47S initiation site in GenBank BK000964v3) (Kuhn and Grummt 1987). In a cell-free assay the transcription initiation site was mapped to -1996bp adjacent to an imperfect 16bp homology with the 47S Promoter (Figure 2.4C). However, nothing further is known of the structure of this Spacer Promoter, nor is it known whether it has the bipartite structure common to all major RPI promoters. The improved resolution of deconvolution ChIP-Seq allowed us to ask if binding of the preinitiation complex factors at the 47S and Spacer Promoters were similar, and to use this information to better map the Spacer Promoter. We identified binding peaks for three components of the SL1 complex (TAF1B, -C and TBP) and for UBF at both promoters (Figure 2.4A and B). The SL1 components displayed highly reproducible and exactly overlapping peaks of binding, strongly suggesting that in vivo they indeed bound as a complex as was expected (Moss et al. 2007). Gaussian peak-fit analysis showed that SL1 binding at the 47S and Spacer Promoters was centred respectively at 60±1.2bp and 65±2.7bp upstream of the corresponding initiation sites (vertical dashed lines in Figure 2.4A and B). The position of the main peak of UBF interaction at each promoter was also highly reproducible and was centred respectively at 83±2.3 and 91±2.2 upstream of the 47S and Spacer initiation sites. Thus, the peak of UBF binding was shifted upstream of the peak of SL1 binding by close to 20bp at both promoters.

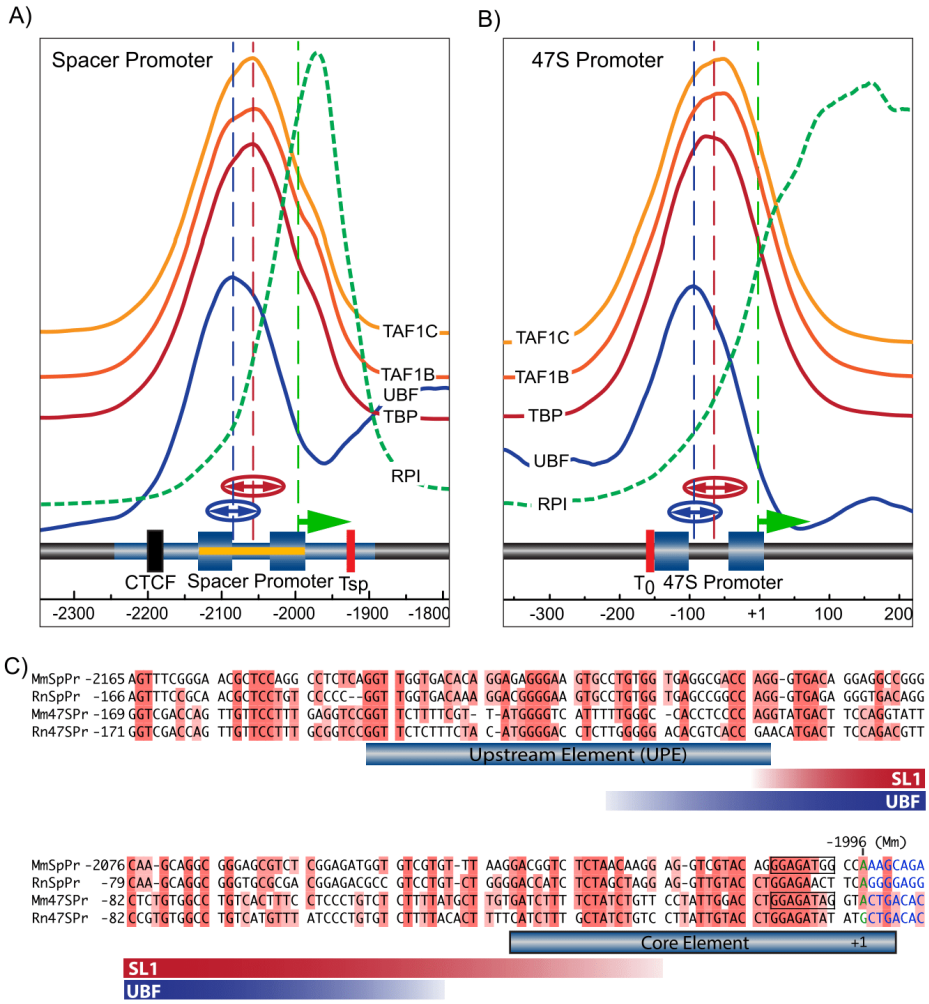

Figure 2.4 – Mapping of preinitiation complexes at the Spacer and 47S Promoters of MEFs. A) and B) Show the interaction profiles of the TAF1B, -C and TBP components of SL1, and of UBF and RPI across the Spacer and 47S Promoter regions in MEFs (MTAB-5893). The deconvoluted ChIP mapping profiles are shown stacked above a diagrammatic representation of the underlying rDNA sequence elements. The mapping profiles for each SL1 component (TAF1B, -C & TBP) are shown on the same vertical scale of enrichment in A) and B), indicating that they are recruited equally efficiently at both promoters. For convenience, the vertical scale of enrichment for RPI at the two promoter is, however, different, see Figure 2B for a quantitative comparison. The extent of the Spacer Promoter was predicted by analogy to the 47S Promoter indicated by the blue-shaded boxes corresponding to the mapped UPE and Core elements. The original identification of the mouse Spacer Promoter and Spacer initiation site at -1996bp (relative to the 47S initiation site, GenBank BK000964v3) (Kuhn and Grummt 1987), are indicated by blue shading band and an arrow (green). Functional mapping of the Spacer Promoter of rat (Smith et al. 1990), (-143 to +1bp relative to the initiation site and requiring sequences upstream of -90bp), is indicated in A) by a yellow band. The broken vertical blue and red lines in A) and B) indicate the mean centres and "⟷" the half-height half-widths of best-fit Gaussian distributions to the UBF and SL1-component mapping profiles obtained respectively from 5 and 8 independent biological replicas. C) Alignment of mouse (Mm) and rat (Rn) 47S and Spacer (SpPr) Promoters. The extent of SL1 components and UBF interactions are indicated by red and blue bands showing the mean half-height half-widths of the best-fit Gaussians to the mapping data, as in A) and B).

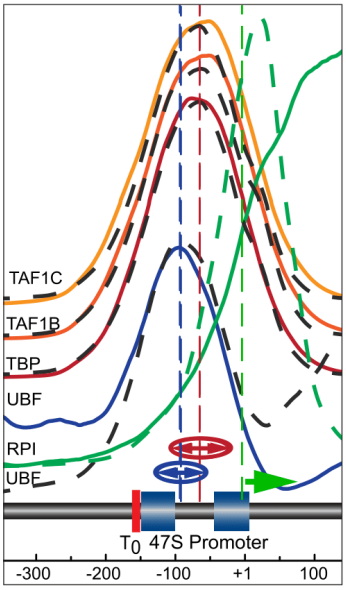

Figure 2.8 – Direct comparison of interaction proles of the TAF1B, -C and TBP components of SL1, and of UBF and RPI at the Spacer and 47S Promoter regions in MEFs (MTAB-5893). The data from Figure 4A and B are shown superimposed, and the 47S and Spacer Promoter initiation sites are aligned. The prole colouring is as in Figure 4, except that the Spacer Promoter proles corresponding to each SL1 component and to UBF are shown as dashed black lines. The RPI proles have been included in green, that at the Spacer Promoter as a dashed line. The broken vertical blue and red lines indicate the mean centres and "⟷" the half-height half-widths of best-t Gaussian distributions to the UBF and SL1-component mapping proles at both promoters and superimpose exactly.

The near identical positions of SL1 and UBF relative to the transcription initiation sites, see Figure 2.8 for an overlay, strongly argued that very similar if not identical preinitiation complexes formed at both 47S and Spacer Promoters. Further, the enrichment of each SL1 component and of UBF was found to be essentially identical at 47S and Spacer promoters, (note; the vertical enrichment scales are the same in Figure 2.4A and B). It was concluded that despite the extremely poor DNA base sequence homology between the two promoters (Figure 2.4C), UBF and SL1 must nonetheless recognize a common underlying promoter structure. Indeed, Marilley and Pasero (Marilley and Pasero 1996) predicted that rDNA promoters contain common features of curvature, twist and helix stability that could explain their specific recognition by the transcription machinery.

2.6.6 – Deconvolution ChIP-Seq also identifies a Spacer Promoter within the Human rDNA

Given its potential importance, it is surprising that a Spacer Promoter has not yet been identified in the human rDNA repeat, though references to its possible existence have been made in the literature, e.g. (Zentner et al. 2011; van de Nobelen et al. 2010). When we applied deconvolution ChIP-Seq to public datasets for UBF and RPI in human HEK cells a peak of UBF binding was resolved near the mapped 47S Promoter and at a site within the IGS ~800bp upstream of the 47S initiation site (Figure 2.5A and B). UBF binding at the human 47S promoter was centred ~90 bp upstream of the 47S initiation site and so mapped much as in mouse (Figure 2.4B). Assuming the human 47S and Spacer Promoters have similar organization, we were able to make an initial estimate of the position of the human Spacer Promoter as between -850 and -700bp relative to the 47S initiation site.

Deconvolution analysis of public and in-house ChIP-Seq data for RPI, TBP and UBF from human K562 cells further supported this Spacer Promoter mapping. Two peaks of TBP binding were observed on the rDNA, one at the 47S promoter and the other over the prospective Spacer Promoter site, and each TBP peak corresponded to a peak in the UBF binding profile (Figure 2.5C). At higher resolution, it was seen that each the TBP peak in fact mapped ~30bp downstream of the corresponding peak of UBF (e.g. Figure 2.5D), suggesting a very similar promoter organisation to that in mouse. Gaussian curve fitting to the binding profiles from both HEK and K562 cells placed the mean peak centres for TBP and UBF at the prospective Spacer Promoter at -758±12 and 789±8 respectively relative to the 47S initiation site, while at the 47S Promoter mean peak centres for TBP and UBF were -78±16 and -87±3. Assuming a similar positioning of TBP and UBF relative to the initiation sites at both promoters, this places the Spacer Promoter initiation site at -691±11. Alignment of the two promoter sequences shows a potential homology in this region, suggesting that the Spacer Promoter initiates transcription at or near -698bp (Figure 2.5E).

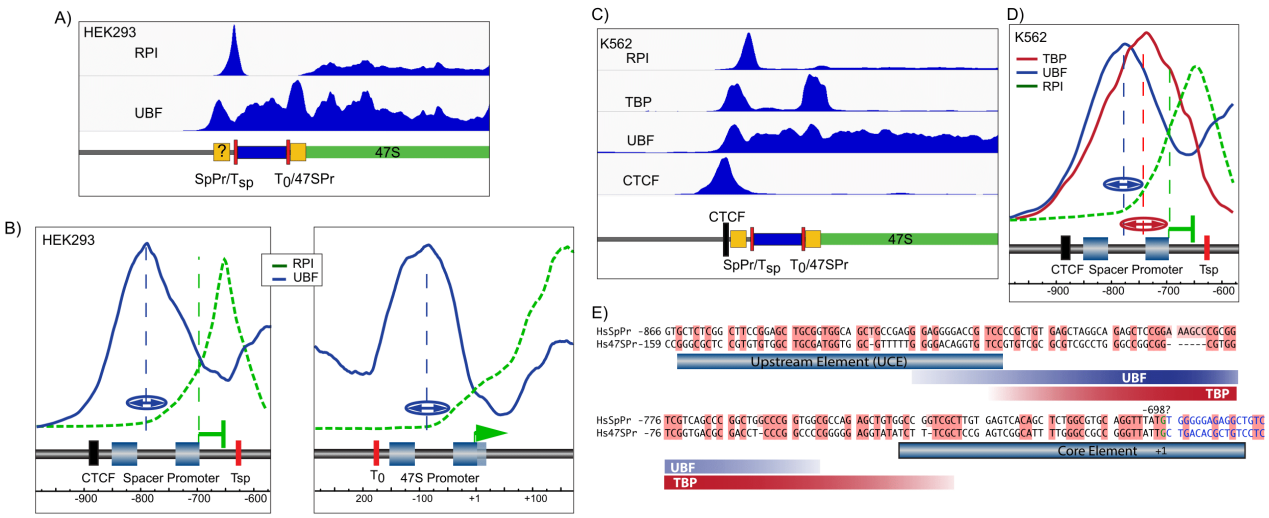

Figure 2.5 – Identification of a Spacer Promoter in the human rDNA. A) Deconvolution map of ChIP-Seq data for RPI and UBF (SRR087747, SRR087746, SRR087753) (Zentner et al. 2011) across the 47S rRNA start site of the HEK 293T cell line. B) High resolution plots of data in A) over the 47S and propesctive Spacer Promoter regions. A very similar arrangement to that in mouse is observed, with a peak of RPI lying ~40bp downstream of the predicted Spacer Promoter initiation site and ~20bp upstream of the adjacent TTF1 binding site motif "Tsp". The identified 47S Promoter sequence motifs (Haltiner et al. 1986), the probable extent of the Spacer Promoter, and positions of the CTCF and TTF1 binding sites are indicated diagrammatically. C) Realignment and deconvolution of ChIP-Seq data for RPI, TBP UBF and CTCF (data sets; SRR502378/9, SRR2096736/7, E-MTAB-6032) from the human K562 cell line. The mapped and predicted sequence motifs are shown diagrammatically below the sequence coverage maps. D) Detailed profiles of TBP, UBF and RPI mapping at the human Spacer Promoter in K562 cells, (data sets; SRR770743-5, E-MTAB-6032), show a very similar arrangement to those in mouse and in HEK293T cells. Here again a peak of RPI is detected downstream of the predicted initiation site and upstream of the adjacent TTF1 binding site motif "Tsp". The broken vertical blue and red lines in A) and C) indicate the mean centres and "⟷" the half-height half-widths of best-fit Gaussian distributions to the UBF and TBP mapping profiles obtained respectively from 3 and 2 independent biological replicas. E) Alignment of human (Hs) 47S and predicted Spacer (SpPr) Promoter sequences. The extent of TBP and UBF interactions are indicated by red and blue bands showing the mean half-height half-widths of the best-fit Gaussians to the mapping data, as in B) and D).

2.6.7 – The chromatin contexts of the human and mouse Spacer Promoters are closely similar

We previously found that in mouse, RPI transcription initiated at the Spacer Promoter is arrested about 40bp downstream adjacent to the binding site for the RPI Transcription Termination Factor TTF1 (Hamdane et al. 2014; Herdman et al. 2017) (Figure 2.4A). Strikingly, a peak of RPI was also observed just 50bp downstream of the probably human Spacer Promoter and immediately adjacent to a consensus

binding site (GGTCGACC) for TTF1 (Figure 2.5B). This striking similarity between the two systems strongly suggested that not only did the human rDNA possess an active Spacer Promoter, but that it was also regulated by TTF1 in a very similar manner. A further characteristic of the mouse Spacer Promoter was its position adjacent to a unique boundary complex consisting of CTCF and an upstream concentration of active chromatin marks (Herdman et al. 2017). Screening the sequenced human 43kbp rDNA repeat unit for likely CTCF binding sites using CTCFDSDBv2.0 (Ziebarth et al. 2013) revealed 4 potential sites with log-odd scores (Altschul et al. 2010) around 14, and one immediately upstream of the prospective Spacer Promoter (-896 to -876) with a log-odd score of over 19, (that is 80 x more likely than random). As previously shown (Zentner et al. 2011), alignment of public CTCF ChIP-Seq data from K562 cells revealed a single site of interaction corresponding to this best predicted CTCF site (Figure 2.5C). Thus, the chromatin and RPI factor contexts strongly suggest that we have not only accurately identified an active Spacer Promoter in the human rDNA, but also that it forms part of an entity analogous to the Enhancer Boundary Complex recently identified in mouse rDNA (Herdman et al. 2017).

2.6.8 – A COMMON MODE OF TBP-COMPLEX BINDING AT THE HUMAN SPACER AND 47S PROMOTERS

We took advantage of available ChIP-exonuclease mapping data for TBP in K562 (He et al. 2015) to better define SL1 complex interactions on the human rDNA. Realignment of the raw data revealed the potential 5’ and 3’ boundaries of the TBP-containing complexes (5’-top and 5’-bottom in Figure 2.6A). The data clearly identified complexes at both 47S and Spacer Promoters and suggested two DNA contact sites within each promoter. Strikingly, the sites corresponded closely to the mapped UPE (UCE) and Core promoter elements of the human 47S promoter (Haltiner et al. 1986), and suggested that the SL1 complex either contacts both promoter elements or that mammalian rDNA promoters, like the yeast rDNA promoter, recruit two distinct TBP associated complexes (Moss et al. 2007), see Discussion. The ChIP-exonuclease data further reinforce the notion that, despite the

poor primary sequence conservation, the 47S and Spacer Promoters have very similar binary structures.

2.6.9 – Identification of potential Enhancer Repeat in the human rDNA

The DNA lying immediately upstream of the major rRNA promoter in a wide range of eukaryotes has been found to include a variable number of short (~60 to 200bp) sequence repeats (Moss et al. 1985; Moss et al. 2007). In *Xenopus* and mouse these repeats possess enhancer or selector-like activities (Moss 1983; Labhart and Reeder 1984; De Winter and Moss 1986, 1987; Pape et al. 1989; Pikaard et al. 1990; Osheim et al. 1996; Moss et al. 2007). Our mapping of the human Spacer Promoter allowed us to investigate the organization of sequences within the region lying between it and the 47S Promoter. Though we found no clear evidence for near perfect "enhancer-like" repeats, a "DotPlot" search for homologies did reveal evidence for an underlying repetition of short highly GC-rich sequence homologies interspersed at roughly 100bp intervals by more complex sequence (Figure 2.6B). Alignment of these "repeat" units suggested that they possibly have a common evolutionary origin, and so may indeed be analogous to the enhancer repeats seen in other organisms (Figure 2.6C). Analysis of more recent rDNA sequences (GB Acc. AL3536449, AL592188, FP236383, KC876030) also suggest that unlike the rDNA of many other organisms, this region of the human rDNA is fully conserved, showing at most a 10 bp length difference with the most commonly referenced composite rDNA repeat sequence (GB Acc. U13369.1). This said, it should be noted that these newer sequences originate from Bacmids containing the rDNA Nucleolar Organiser Region (NOR) boundaries from specific chromosomes and so may not be representative of the bulk rDNA.

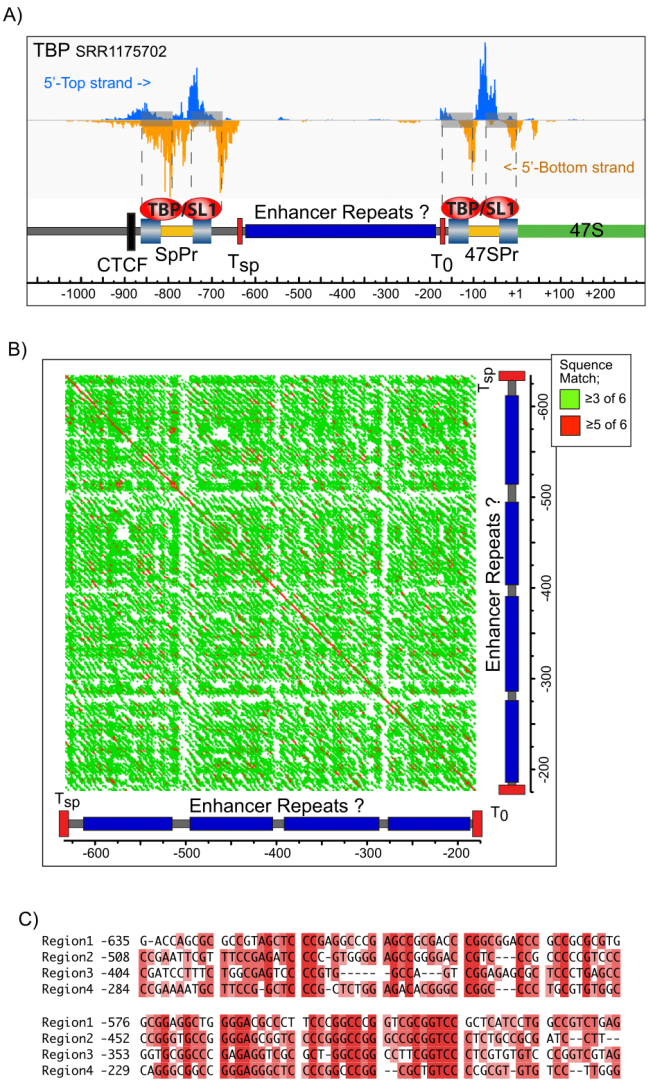

Figure 2.6 – Fine mapping of TBP complexes and potential enhancer repeat suggest functional parallels between the human and mouse rDNA. A) Realignment of TBP ChIP-exonuclease data from human K562 cells ((He et al. 2015), GEO Acc. GSE55306), onto the human rDNA reveals dual contact sites at both Spacer and 47S Promoters. Spacer and 47S Promoter Core and UPE are indicated by light blue shaded boxes and potential Tsp and T0 TTF1 sites by red boxes. B) "DotPlot" homology alignment of human rDNA sequences lying between the Spacer and 47S Promoters generated using the Gene Inspector software (Textco) and a sliding window of 6 bases. Red indicates ≥5 of 6 identically matching bases, and green ≥3 of 6 matches. C) Alignment of the four pseudo-repeats within the same region. In A) through C) sequence position is indicated relative to the 47S initiation site.

2.7 – Discussion

The potential for very significant improvements in ChIP-Seq mapping resolution afforded by our simple deconvolution protocol were recently demonstrated when the protocol was applied to map transcription factors and chromatin status across the

mouse rDNA (Herdman et al. 2017). Here, we provide a detailed deconvolution protocol, consider the effects of data smoothing and multiple site alignment, and demonstrate the reproducibility of the interaction maps generated. We show that given sufficient sequencing depth, variations in mapping profiles are small (±10%) and may in large part represent the variability introduced by the ChIP protocol and/or by biological variability between samples. In principle, our deconvolution protocol is applicable to any ChIP-Seq data for which sufficient sequencing depth is available. Based on our present studies, we estimate that the average number of reads across each base position of both input and ChIP datasets needs to be ≥100 in order for the deconvolved profiles to be statistically significant. Such a situation is easily attainable with present sequencing technologies.

When applied to ChIP-Seq data for the RPI polymerase, the deconvolution protocol revealed a near uniform recruitment across the 47S transcribed region of the mouse rDNA. In contrast, the recruitment of UBF across the same region displayed around 74 preferential positions spaced on average at 170 bp intervals. Closer inspection also revealed a correlation between UBF binding and the GC content of the underlying rDNA. Previous analyses have shown that UBF has a preference for GC-rich DNA (Copenhaver et al. 1994) and that a UBF dimer interacts with 110 to 160 bp of DNA, looping it into a single turn and leading to the suggestion that it may replace nucleosomal chromatin (Stefanovsky et al. 1996; Bazett-Jones et al. 1994; Herdman et al. 2017). Together, the data suggest that UBF dimers bind at preferential sites to form a semi-continuous pseudo-chromatin across the 47S transcribed region of the rDNA.

We have also applied the deconvolution protocol to fine map the 47S and Spacer Promoters of the mouse and human rDNA Intergenic Spacers (IGSs). Interestingly, the data suggest that, despite a complete lack of any significant homology at the level of the respective DNA sequences, the structure and the chromatin contexts of the human and mouse Spacer Promoters are very similar. We found that positioning of the preinitiation factors, UBF and the components of the RPI TBP complex SL1,

is nearly identical at the 47S and Spacer Promoters in both mouse and human. Further, we found that the ChIP enrichment of the known SL1 subunits at 47S and Spacer Promoters was within experimental error the same. Thus, all active rDNA units appear to recruit SL1 at both promoters with equal efficiency.

In contrast, the context of the Spacer Promoters, in being flanked immediately upstream by CTCF and Cohesin complexes and downstream by an arrested polymerase, is quite different from that of the 47S Promoter. As we recently demonstrated in mouse, the CTCF complex forms a boundary between the upstream chromatin and the transcriptionally active rDNA unit (Herdman et al. 2017). Loss of CTCF was also shown to eliminate UBF recruitment to the rDNA (van de Nobelen et al. 2010). Thus, the CTCF boundary most probably arrests the expansion of upstream repressive chromatin into the active rDNA unit. The recruitment of the Snf2h chromatin remodeller subunit at the CTCF site is probably important in this respect (Herdman et al. 2017). Recruitment of Cohesin to the CTCF boundary further suggests a role in chromatin looping and the spatial organization of the rDNA loci, see (Herdman et al. 2017) for further discussion.

The Spacer Promoter is also unique in being associated with a strong interaction peak of RPI. This peak is centred downstream of the initiation site and upstream of the adjacent TTF1 binding site, and suggests that transcription from this promoter is arrested after only 40 to 50 nucleotides in both mouse and human. Release of this arrested polymerase into active elongation would generate a long non-coding RNA (lncRNA) that has been suggested to control in trans rDNA silencing in mouse (Savic et al. 2014). It could potentially also regulate the activity of the mouse Enhancer Repeats lying downstream. Analysis of the sequences lying between the Spacer Promoter and 47S Promoter suggested that Enhancer Repeats may also exist in this region of the human rDNA and hence could quite possibly be analogous in function to the mouse and *Xenopus* Enhancers. But, a demonstration of this must await functional studies.

While the RPI promoters of different organisms from yeast to human show little or no DNA sequence conservation, they do conserve a common functional organisation of precisely spaced UPE and Core elements, suggesting a similar mode of recognition by the transcription machinery. In fact, we found that realignment of the ChIP-exonuclease (ChIP-nexus) data for TBP (He et al. 2015) revealed two distinct contact sites for SL1 that mapped closely to the UPE (UCE) and Core promoter elements of the human 47S promoter (Haltiner et al. 1986), see Figure 2.6A. This suggested either that a single SL1 complex contacts both promoter elements or that, as we have previously suggested, mammalian rDNA promoters, might recruit two SL1 complexes (Moss and Stefanovsky 2002). However, whether these contact sites would correspond to two identical SL1 complexes, or to two SL1 sub-complexes as seen in yeast where distinct TAF1 sub-complexes bind UPE and Core elements and are bridged by TBP (Moss et al. 2007), will require further study. It is relevant here to note that our present knowledge of the structure of mammalian SL1 is still incomplete (Gorski et al. 2007; Murano et al. 2014).

2.8 – Acknowledgments

We would like to thank Mark Robinson and Helen Lindsay (IMLS/SIB, University of Zürich) for their help and advice and for making their computing facilities available to us. This work was funded by operating grants from the Canadian Institutes of Health Research (CIHR, MOP12205/PJT153266) and the National Science and Engineering Council (NSERC) of Canada. The Research Centre of the Québec University Hospital Centre (CHU de Québec) is supported by the Fonds de Recherche du Québec - Santé (FRQS).

2.9 – Conflict of interest

On behalf of all authors, I declare that we have no financial or other conflicting or competing interests related to the work included in this manuscript.

2.10 – REFERENCES

Altschul, S.F., J.C. Wootton, E. Zaslavsky, and Y.K. Yu, 2010 The construction and use of log-odds substitution scores for multiple sequence alignment. PLoS Comput Biol 6 (7):e1000852.

Bach, R., B. Allet, and M. Crippa, 1981 Sequence organisation of the spacer in the ribosomal genes of Xenopus clivii and Xenopus borealis. Nucleic Acids Research 9:5311-5330.

Bazett-Jones, D.P., B. Leblanc, M. Herfort, and T. Moss, 1994 Short-range DNA looping by the Xenopus HMG-box transcription factor, xUBF. Science 264:1134-1137.

Bolger, A.M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30 (15):2114-2120.

Cassidy, B.G., H.F. Yang-Yen, and L.I. Rothblum, 1987 Additional RNA polymerase I initiation site within the nontranscribed spacer region of the rat rRNA gene. Molecular and Cellular Biology 7:2388-2396.

Caudy, A.A., and C.S. Pikaard, 2002 Xenopus ribosomal RNA gene intergenic spacer elements conferring transcriptional enhancement and nucleolar dominance-like competition in oocytes. J Biol Chem 277 (35):31577-31584.

Chen, Y., N. Negre, Q. Li, J.O. Mieczkowska, M. Slattery et al., 2012 Systematic evaluation of factors influencing ChIP-seq fidelity. Nat Methods 9 (6):609-614.

Coen, E.S., and G.A. Dover, 1983 Multiple polymerase I promoter sequences in rDNA of drosophila melanogaster. Nucleic Acids Research 10:7017-7026.

Copenhaver, G.P., C.D. Putnam, M.L. Denton, and C.S. Pikaard, 1994 The RNA polymerase I transcription factor UBF is a sequence-tolerant HMG-box protein that can recognize structured nucleic acids. Nucleic Acids Research 22:2651-2657.

De Winter, R.F., and T. Moss, 1986 Spacer promoters are essential for efficient enhancement of X. laevis ribosomal transcription. Cell 44 (2):313-318.

De Winter, R.F., and T. Moss, 1987 A complex array of sequences enhances ribosomal transcription in Xenopus laevis. J Mol Biol 196 (4):813-827.

Doelling, J.H., R.J. Gaudino, and C.S. Pikaard, 1993 Functional analysis of Arabidopsis thaliana rRNA gene and spacer promoters in vivo and by transient expression. Proceedings of the National Academy of Sciences of the United States of America 90:7528-7532.

Ekblom, R., L. Smeds, and H. Ellegren, 2014 Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. BMC Genomics 15:467.

Gorski, J.J., S. Pathak, K. Panov, T. Kasciukovic, T. Panova et al., 2007 A novel TBP-associated factor of SL1 functions in RNA polymerase I transcription. Embo J 26 (6):1560-1568.

Guetg, C., P. Lienemann, V. Sirri, I. Grummt, D. Hernandez-Verdun et al., 2010 The NoRC complex mediates the heterochromatin formation and stability of silent rRNA genes and centromeric repeats. Embo J 29 (13):2135-2146.

Haltiner, M.M., S.T. Smale, and R. Tjian, 1986 Two distinct promoter elements in the human rRNA gene identified by linker scanning mutagenesis. Molecular and Cellular Biology 6:227-235.

Hamdane, N., V.Y. Stefanovsky, M.G. Tremblay, A. Nemeth, E. Paquet et al., 2014 Conditional inactivation of Upstream Binding Factor reveals its epigenetic functions and the existence of a somatic nucleolar precursor body. PLoS Genetics 10 (8):e1004505.

He, Q., J. Johnston, and J. Zeitlinger, 2015 ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat Biotechnol 33 (4):395-401.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y.C. Lin et al., 2010 Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell 38 (4):576-589.

Henderson, A.S., E.M. Eicher, M.T. Yu, and K.C. Atwood, 1974 The chromosomal location of ribosomal DNA in the mouse. Chromosoma 49 (2):155-160.

Henderson, A.S., D. Warburton, and K.C. Atwood, 1972 Location of ribosomal DNA in the human chromosome complement. Proc Natl Acad Sci U S A 69 (11):3394-3398.

Herdman, C., J.C. Mars, V.Y. Stefanovsky, M.G. Tremblay, M. Sabourin-Felix et al., 2017 A Unique Enhancer Boundary Complex on the Mouse Ribosomal RNA Genes persists after loss of Rrn3 or UBF and the Inactivation of RNA Polymerase I Transcription. PLoS Genetics 13 (7):e1006899.

Jackson, D.A., A. Pombo, and F. Iborra, 2000 The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells. Faseb J 14 (2):242-254.
Kidder, B.L., G. Hu, and K. Zhao, 2011 ChIP-Seq: technical considerations for obtaining high-quality data. Nat Immunol 12 (10):918-922.

Kuhn, A., and I. Grummt, 1987 A novel promoter in the mouse rDNA spacer is active in vivo and in vitro. EMBO Journal 6:3487-3492.

Labhart, P., and R.H. Reeder, 1984 Enhancer-like properties of the 60/81 bp elements in the ribosomal gene spacer of Xenopus laevis. Cell 37:285-289.

Langmead, B., and S.L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat Methods 9 (4):357-359.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan et al., 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25 (16):2078-2079.

Marilley, M., and P. Pasero, 1996 Common DNA structural features exhibited by eukaryotic ribosomal gene promoters. Nucleic Acids Research 24:2204-2211.

Miller, J.R., D.C. Hayward, and D.M. Glover, 1983 Transcription of the non-transcribed spacer of Drosophila melanogaster rDNA. Nucleic Acids Research 11:11-19.

Moss, T., 1983 A transcriptional function for the repetitive ribosomal spacer in Xenopus laevis. Nature 302:223-228.

Moss, T., and M.L. Birnstiel, 1979 The putative promoter of a Xenopus laevis ribosomal gene is reduplicated. Nucleic Acids Research 6:3733-3743.

Moss, T., F. Langlois, T. Gagnon-Kugler, and V. Stefanovsky, 2007 A housekeeper with power of attorney: the rRNA genes in ribosome biogenesis. Cell Mol Life Sci 64 (1):29-49.

Moss, T., K. Mitchelson, and R. de Winter, 1985 The promotion of ribosomal transcription in eukaryotes. Oxf Surv Eukaryot Genes 2:207-250.

Moss, T., and V.Y. Stefanovsky, 1995 Promotion and Regulation of Ribosomal Transcription in Eukaryotes by RNA Polymerase I., pp. 25-66 in Progress in Nucleic Acids and Molecular Biology, edited by W.E. Cohn and K. Moldave. Academic Press, Inc., San Diego.

Moss, T., and V.Y. Stefanovsky, 2002 At the center of eukaryotic life. Cell 109 (5):545-548.

Murano, K., M. Okuwaki, F. Momose, M. Kumakura, S. Ueshima et al., 2014 Reconstitution of human rRNA gene transcription in mouse cells by a complete SL1 complex. J Cell Sci 127 (Pt 15):3309-3319.

Murtif, V.L., and P.M.M. Rae, 1985 In vivo transcription of rDNA spacers in Drosophila. Nucleic Acids Research 13:3221-3240.

Osheim, Y.N., E.B. Mougey, J. Windle, M. Anderson, M. O'Reilly et al., 1996 Metazoan rDNA enhancer acts by making more genes transcriptionally active. Journal of Cell Biology 133:943-954.

Paalman, M.H., S.L. Henderson, and B. Sollner-Webb, 1995 Stimulation of the mouse rRNA gene promoter by a distal spacer promoter. Mol.Cell Biol 15:4648-4656.

Pape, L.K., J.J. Windle, E.B. Mougey, and B. Sollner-Webb, 1989 The Xenopus ribosomal DNA 60- and 81-base-pair repeats are position-dependent enhancers that function at the establishment of the preinitiation complex: Analysis in vivo and in an enhancer-responsive in vitro system. Molecular and Cellular Biology 9:5093-5104.

Park, P.J., 2009 ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10 (10):669-680.

Pikaard, C.S., L.K. Pape, S.L. Henderson, K. Ryan, M.H. Paalman et al., 1990 Enhancers for RNA polymerase I in mouse ribosomal DNA. Molecular and Cellular Biology 10:4816-4825.

Poorey, K., R. Viswanathan, M.N. Carver, T.S. Karpova, S.M. Cirimotich et al., 2013 Measuring chromatin interaction dynamics on the second time scale at single-copy genes. Science 342 (6156):369-372.

Putnam, C.D., and C.S. Pikaard, 1992 Cooperative binding of the Xenopus RNA polymerase I transcription factor xUBF to repetitive ribosomal gene enhancers. Molecular and Cellular Biology 12:4970-4980.

Quinlan, A.R., and I.M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26 (6):841-842.

Savic, N., D. Bar, S. Leone, S.C. Frommel, F.A. Weber et al., 2014 lncRNA Maturation to Initiate Heterochromatin Formation in the Nucleolus Is Required for Exit from Pluripotency in ESCs. Cell Stem Cell 15 (6):720-734.

Scheer, U., and R. Benavente, 1990 Functional and dynamic aspects of the mammalian nucleolus. Bioessays 12 (1):14-21.

Smith, S.D., E. Oriahi, H.-F. Yang-Yen, W. Xie, C. Chen et al., 1990 Interaction of RNA polymerase I transcription factors with a promoter in the nontranscribed spacer of rat ribosomal DNA. Nucleic Acids Research 18:1677-1685.

Stefanovsky, V.Y., D.P. Bazett-Jones, G. Pelletier, and T. Moss, 1996 The DNA supercoiling architecture induced by the transcription factor xUBF requires three of its five HMG-boxes. Nucleic Acids Research 24:3208-3215.

Taslim, C., J. Wu, P. Yan, G. Singer, J. Parvin et al., 2009 Comparative study on ChIP-seq data: normalization and binding pattern characterization. Bioinformatics 25 (18):2334-2340.

Teytelman, L., D.M. Thurtle, J. Rine, and A. van Oudenaarden, 2013 Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc Natl Acad Sci U S A 110 (46):18602-18607.

Tower, J., S.L. Henderson, K.M. Dougherty, P.J. Wejksnora, and B. Sollner-Webb, 1989 An RNA polymerase I promoter located in the CHO and mouse ribosomal DNA spacers: functional analysis and factor and sequence requirements. Molecular and Cellular Biology 9:1513-1525.

van de Nobelen, S., M. Rosa-Garrido, J. Leers, H. Heath, W. Soochit et al., 2010 CTCF regulates the local epigenetic state of ribosomal DNA repeats. Epigenetics Chromatin 3 (1):19.

Zentner, G.E., A. Saiakhova, P. Manaenkov, M.D. Adams, and P.C. Scacheri, 2011 Integrative genomic analysis of human ribosomal DNA. Nucleic Acids Research 39 (12):4949-4960.

Ziebarth, J.D., A. Bhattacharya, and Y. Cui, 2013 CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. Nucleic Acids Research 41 (Database issue):D188-194.

3.1 – AVANT-PROPOS

Ce chapitre présente une partie des travaux effectués au cours de ma maîtrise. Il porte sur la localisation ainsi que les rôles potentiels d'UBF à l'échelle du génome murin. Encore une fois pour ces jeux de données, Michel G. Tremblay a contribué à la production des souris qui ont permis de générer les lignées cellulaires MEFs conditionnelles pour UBF et Jean-Clément Mars a procédé aux expériences de ChIP-seq, de DNase-seq et de *microarray*. En ce qui me concerne, j'ai traité bio-informatiquement les différents jeux de donnés. J'ai aussi effectué toutes les analyses présentées dans ce chapitre.

3.2 – RÉSUMÉ

UBF est un facteur de transcription associé à l'ADNr; il permet la transcription des ARNr 18S, 5.8S et 28S. Pourtant, des sites d'interactions protéiques d'UBF ont été découverts à la grandeur du génome. À l'aide d'expériences de ChIP-seq, DNase-seq et de *microarray* effectuées dans des cellules MEFs $UBF^{fl/fl}$ $ERCre^{+/+}$, des rôles potentiels d'UBF ont pu être mis en lumière. L'analyse de ces données a révélé plusieurs particularités d'UBF. Premièrement, UBF se retrouve près des sites d'initiation de la transcription, ce qui pourrait indiquer qu'il agit à titre de facteur de transcription pour ces gènes tel qu'il le fait sur ADNr. Deuxièmement, UBF colocalise avec les marques d'histones actives. Cela indique qu'UBF est associé aux gènes transcrits et potentiellement transcrits. Finalement, UBF se retrouve majoritairement à des sites sensibles au clivage à la DNase I. Il pourrait donc servir de protecteur de l'ADN dépourvu de nucléosomes dans ces régions.

3.3 – ABSTRACT

UBF is a transcription factor associated to the rDNA; it allows the transcription of the 18S, 5.8S and 28S rRNAs. However, UBF interaction sites with de DNA throughout the genome has been discovered. Using ChIP-seq, DNase-seq and microarray experiments in MEFs $UBF^{fl/fl}$ $ERCre^{+/+}$ cells, potential roles of UBF have been highlighted. This data analysis revealed several features of UBF. First, UBF is found near transcription start sites, which could indicate that it acts as a transcription factor for these genes as it does on rDNA. Secondarily, UBF colocalize with actives histones marks. This indicates that UBF is associated with transcribed or potentially transcribed genes. Finally, UBF is mainly found at DNase I sensitivity sites. It could thus serve as a protector for DNA lacking nucleosomes in these regions.

CHAPITRE 3 – ÉTUDE DU RÔLE DU FACTEUR ARCHITECTURAL UBF : UNE ÉTUDE À LA GRANDEUR DU GÉNOME

3.4 – INTRODUCTION

Tel que mentionné au chapitre 1, UBF a tendance à se lier à l'ADN de façon non spécifique (Bazett-Jones et al., 1994; Stefanovsky et Moss, 2008). D'après des études de ChIP-seq, UBF se retrouve sur les régions promotrices, mais aussi le long de la région transcrite des gènes de l'ARN ribosomique (Figure 3.1) (Zentner et al., 2011; Herdman, Mars et al., 2017).

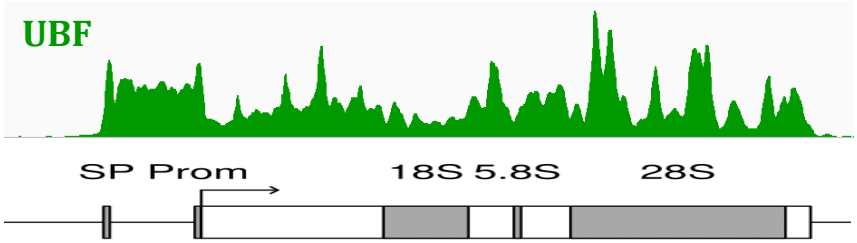

Figure 3.1 – UBF sur le gène de l'ARN ribosomique. L'échelle horizontale représente la position le long du gène de l'ARNr tandis que l'échelle verticale représente l'enrichissement. UBF est présent sur la région promotrice mais aussi le long de la région transcrite.

Par contre, UBF se retrouve aussi ailleurs dans le génome de la souris (Figure 3.2). Comme il s'agit d'une protéine pour laquelle on ne connait d'autres fonctions qu'aux gènes de l'ARN ribosomique, il est curieux de voir un tel profil de liaison. Nous en sommes venus à nous demander quel rôle UBF pouvait avoir à ces endroits.

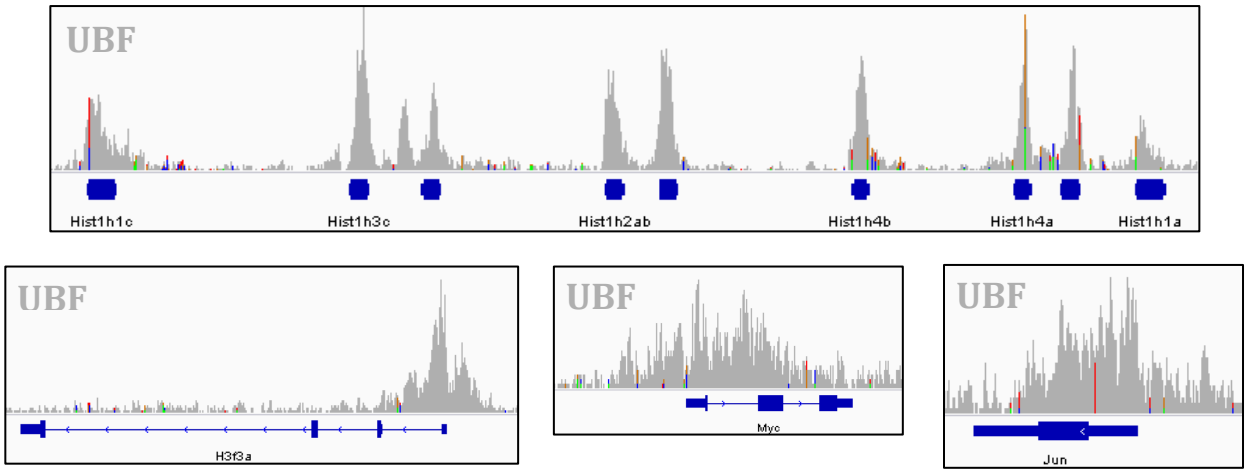

Figure 3.2 – UBF à l'échelle du génome. UBF se retrouve à plusieurs endroits dans le génome murin.

Afin de déterminer si un site d'interaction protéique (pic) est réel ou non, le KO d'UBF a été utilisé. Si le même pic était présent avant et après l'induction du KO, celui-ci ne sera pas pris en considération dans les expériences subséquentes.

3.5 – MATÉRIELS ET MÉTHODES

Afin de répondre à cette question, plusieurs logiciels bio-informatiques ont été utilisés. Ces derniers sont décrits dans les sections suivantes.

3.5.1 – DESCRIPTION ET UTILISATION DE BOWTIE2

Bowtie2 est un logiciel permettant l'alignement des fragments séquencés sur un génome de référence de façon *single-end* ou *paired-end*. Ce logiciel fonctionne en quatre étapes. Premièrement, Bowtie2 extrait ce qu'on appelle un *seed* soit une partie du *read* et sa séquence complémentaire inversée (*reverse complement*). Deuxièmement, les *seeds* sont alignés sur le génome de référence sans y insérer d'indel (insertion/délétion). Troisièmement, la position des *seeds* est calculée à partir d'un index. Quatrièmement, les *seeds* sont étendus jusqu'à l'alignement complet en performant la programmation dynamique SIMD-accélérée (Single-Instruction Multiple-Data) (Langmead et Salzberg, 2012).

Afin d'aligner les données du laboratoire ainsi que les données publiques, j'ai utilisé la commande présentée à la Figure 3.3. Où $refgenome représente le génome de référence soit GRCh38 (hg20/hg38) pour l'humain ou GRCm38 (mm10) pour la souris, $trimFile représente le fichier d'entrée et où $alnFile représente le fichier de sortie aligné. La variable p représente le nombre de fils d'exécution (*threads*) utilisés tandis que la variable k indique le nombre d'alignements autorisés par *reads*. Il est à noter que -k 3 a aussi été utilisé dans le chapitre 2. Il faut « piper » (|) l'*output* de Bowtie2 dans Samtools, car la sortie de Bowtie2 est un fichier SAM (*sequence alignment map*) imprimé directement dans le terminal. Il faut donc le convertir en fichier compressé de type *BAM* (*binary sequence alignment map)*.

```
bowtie2 -p 16 -k 1 -x $refGenome -U $trimFile | samtools view
-F 4 -bS - > $alnFile
```

Figure 3.3 – Utilisation de Bowtie2. Commande utilisée pour aligner les données de ChIP-seq.

3.5.2 – DESCRIPTION ET UTILISATION DE MACS2

Le logiciel *Model-based Analysis of ChIP-Seq data* (MACS2) permet de détecter les sites de liaison protéiques (pics) de facteurs de transcription qui ont été traités avec le ChIP-seq.

MACS2 a été créé pour améliorer la résolution des sites de liaison en combinant l'information du séquençage et de la position et de l'orientation des *reads*. MACS2 peut être utilisé avec ou sans input, mais l'utilisation d'un input améliore la véracité des pics obtenus. MACS2 peut être utilisé tant pour les facteurs de transcription (*narrowPeak*) qu'avec les marques d'histones (*broadPeak*).

MACS2 utilise un algorithme qui permet le déplacement des *reads* d'une valeur de ± d/2, *d* étant la distance entre les pics Watson et Crick ou encore la longueur des fragments pris pour le séquençage. MACS2 fournit plusieurs information pour chaque pics tels que les coordonnées génomiques, la valeur p, le *False Discovery Rate* ( ), le *fold enrichment* et le centre du pic (Zhang et al., 2008).

La commande présentée à la Figure 3.4 a été utilisée afin d'effectuer le *peakcalling*, c'est-à-dire la détection des pics à la grandeur du génome. Où $IPfile représente le fichier de l'immunoprécipitation, $inputFile représente le fichier de l'input et $outputDir représente le dossier de sortie. Dans cette expérience, le fichier servant d'input est le fichier $UBF^{fl/fl}$ KO, car cela permet d'éliminer les faux positifs, c'est-à-dire les pics présents tant avant qu'après l'induction du KO. L'option -f désigne le type de fichier d'entrée, ici, il s'agit de *BAM*. L'option -g indique la taille du génome de référence alignable (*mappable*), ici celui de la souris (mm10) correspond à 1.87e9 pb. L'option -q correspond au seuil minimal de FDR. Dans cette expérience, un seuil de 0.05 a été utilisé afin d'être moins strict puisque le KO d'UBF a été utilisé comme *input*. L'option --nomodel commande de ne pas générer de modèle du pic aux TSS. L'option --extsize indique que les *reads* seront étendus jusqu'à atteindre 100 pb. L'option -B permet de générer des fichiers *bedGraph*. L'option --SPMR ajuste la

couverture en *Reads Per Million* (RPM). Finalement, l'option --keep-dup all permet de conserver tous les *reads* y compris les réplicats d'amplification PCR.

```
macs2 callpeak -t $IPfile -c $inputFile -n $name --outdir
$outputDir -f BAM -g mm -q 0.05 --nomodel --extsize 100 -B
--SPMR --keep-dup all
```

Figure 3.4 – Utilisation de MACS2. Commande utilisée pour effectuer le *peakcalling*.

3.5.3 – Description et utilisation de BEDOPS

BEDOPS est un logiciel bio-informatique ayant un large éventail d'opérations. Par contre, la seule commande utilisée dans ce mémoire se nomme « *bedops* ». Cette dernière permet d'effectuer des chevauchements (*overlaps*) entre des jeux de données de pics détectés par MACS2. Cela permet de connaître la colocalisation deux facteurs donnés (Neph et al., 2012).

À la Figure 3.5 se retrouve la commande utilisée afin d'effectuer le chevauchement entre deux facteurs où les variables $first et $second sont des fichiers BED dont on veut faire le chevauchement. L'option -e signifie « *element of* » et garde les pics originaux de $first lors d'un chevauchement avec $second (Figure 3.6).

```
bedops -e $first $second > $result
```

Figure 3.5 – Utilisation de BEDOPS. Commande utilisée pour effectuer le chevauchement de deux facteurs.

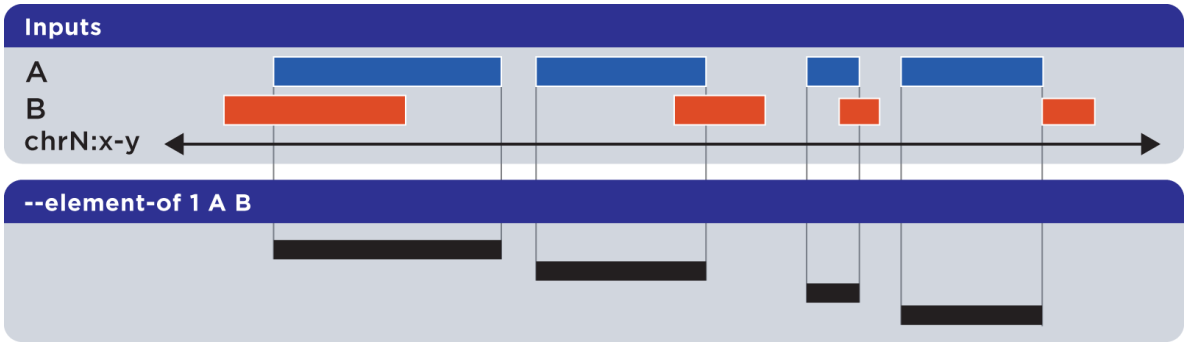

Figure 3.6 – Chevauchement avec BEDOPS. Ce logiciel garde seulement les pics originaux du premier fichier entré. Extrait du site web de BEDOPS.

3.5.4 – DESCRIPTION ET UTILISATION DE GREAT

Le *Genomic Regions Enrichment of Annotations Tool* (GREAT) est une application en ligne permettant l'annotation fonctionnelle des régions génomiques identifiées par MACS2.

Le fonctionnement de GREAT est simple, il suffit de télécharger le fichier de positions génomiques des pics sur le site web. L'*output* comprend plusieurs résultats; entre autres la position des pics par rapport aux TSS, les fonctions moléculaires, les processus biologiques et les composants cellulaires (McLean et al., 2010).

3.5.5 – DESCRIPTION ET UTILISATION DE R

R est un environnement logiciel pour l'informatique statistique et les graphiques. Il permet la création de fonctions maison ou l'utilisation de paquets (*packages*) comprenant des fonctions déjà établies. R permet aussi la création d'une vaste gamme de graphiques (R Development Core Team, 2008).

La Figure 3.6 montre le script utilisé pour créer l'histogramme du chevauchement des marques d'histone avec UBF présenté dans la section Résultats. La commande *barplot* permet de créer l'histogramme, le titre et l'étiquette (*label*) verticale. La commande *box* dessine une boîte autour des données. La commande *axis* permet de mettre l'axe vertical et finalement la commande *text* permet de mettre les noms des modifications d'histones.

```
barplot(c(88.45,83.99,76.27,88.57,82.21,4.31,11.93,3.59),
        ylim = c(0,100), yaxt="n", border=NA, col =
        c("darkgreen", "darkgreen", "darkgreen", "darkgreen",
        "darkgreen", "red","red","red","red"), main =
        "Overlap of histones modification",
        ylab = "% ChIP-seq peaks", xaxt= "n")
box()

axis(2, at = seq(0, 100, by = 20))

text(seq(1,10, by = 1.2), labels = c("H2AZ", "H2AZac",
     "H3K4me2", "H3K4me3", "H4ac", "H3K79me2", "H3K27me3",
     "H3K9me3"),par("usr")[3] -5, srt=45, pos = 2, xpd = TRUE)
```

Figure 3.7– Utilisation de R. Commande utilisée pour générer l'histogramme du chevauchement des marques d'histones avec UBF.

3.5.6 – DESCRIPTION ET UTILISATION DE HOMER

Hypergeometric Optimization of Motif EnRichment (HOMER) est un logiciel écrit en Perl et C++ permettant d'analyser des données de ChIP-seq, GRO-seq, RNA-seq, DNase-seq et Hi-C (Heinz et al., 2010). La fonction utilisée dans le cas de ce mémoire se nomme *getDifferentialPeaks*. Cette fonction permet de déterminer quels pics sont différentiellement exprimés dans certaines conditions telles qu'avant et après un KO. La commande *makeTagDirectory* doit être effectuée en premier lieu.

Afin d'effectuer cette procédure, les commandes présentées à la Figure 3.8 ont été utilisées. Dans la commande *makeTagDirectory*, $name représente le nom du dossier de sortie et l'option *-keep all* garde tous les alignements du fichier *BAM* (même les duplicats). $UBF_wtKO et $UBF_flKO représente les fichiers d'alignements au format *BAM*. Pour la commande *getDifferentialPeaks*, $UBF_wtKO est le fichier *narrowPeak* des pics retrouvés à la grandeur du génome pour les cellules $UBF^{wt/wt}$ KO et même chose pour $UBF_flKO, mais cette fois pour les cellules $UBF^{fl/fl}$ KO. Les variables *_tag représente les dossiers créés par *makeTagDirectory*. L'option -F 2 détermine un seuil de 2 *fold enrichment* au-dessus du *background tag* (troisième paramètre dans la commande).

```
makeTagDirectory $name -format bowtie -keepAll $UBF_floxKO
makeTagDirectory $name -format bowtie -keepAll $UBF_wtKO

getDifferentialPeaks $UBF_wtKO $UBF_wtKO_tag $UBF_flKO_tag -F 2
> $output

getDifferentialPeaks $UBF_flKO $UBF_flKO_tag $UBF_wtKO_tag -F 2
> $output
```

Figure 3.8 – Utilisation de HOMER. Commande utilisée pour l'expression différentielle des pics d'UBF avant et après le KO.

HOMER fonctionne de manière à identifier les pics plus exprimés dans une condition que dans l'autre (Figure 3.9). Par exemple, dans un cas où l'on a deux conditions, soit avant ou après un KO, il détecte les pics plus enrichis dans la condition placée en premier. Par exemple, la commande illustrée à la quatrième ligne de la Figure 3.8 détecte les pics plus enrichis dans la condition $UBF^{fl/lf}$ KO tel que le montre le deuxième cas de la Figure 3.9.

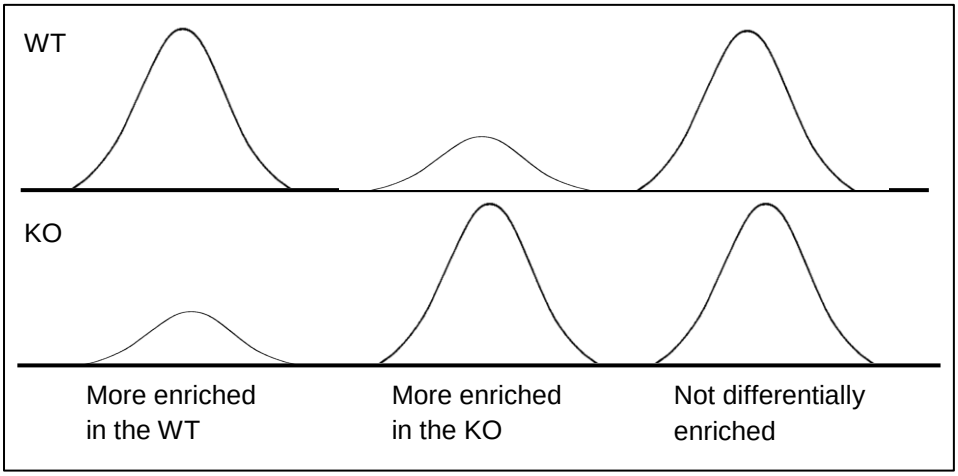

Figure 3.9 – Enrichissement différentiel des pics. Les pics sont soit plus enrichis dans le WT (avant induction du KO) soit plus enrichis dans le KO (après l'induction du KO). Dans les autres cas, il n'y a pas d'enrichissement différentiel.

3.6 – RÉSULTATS

L'utilisation des logiciels présentés dans la partie Matériels et méthodes a permis de générer les résultats de cette section.

Les pics présentés dans cette section sont issus de cellules MEFs $UBF^{fl/fl}$ $p53^{-/-}$ $ERCre^{+/+}$, c'est-à-dire qu'elles sont conditionnelles pour le gène UBF et qu'elles expriment la ERCre (Cre recombinase fusionnée à un récepteur à l'estrogène). La ERCre est activée par le 4-hydroxytamoxifène (4HT), la forme active du tamoxifène, ce qui induit la relocalisation de la Cre au noyau. Grâce à son activité recombinase, la Cre va enlever les exons 3, 4 et 5 d'UBF qui ont été préalablement floxés, c'est-à-dire que des sites loxP ont été ajoutés de part et d'autre de cette région du gène. Trois jours après l'induction du KO, les expériences de ChIP-seq ont été effectuées (pour les détails de l'expérience, voir la section Matériels et Méthodes de l'Annexe II).

Les fichiers *fastq.gz* obtenus du séquenceur Illumina HiSeq 2000 ont été alignés sur le génome de référence en utilisant Bowtie2. La recherche de pics à la grandeur du génome (*peakcalling*) a ensuite été effectuée à l'aide de MACS2 en utilisant le KO d'UBF à titre de contrôle afin d'éliminer les faux positifs, c'est-à-dire les pics qui se retrouvent tant avant qu'après l'induction du KO. Un exemple de faux positif se retrouve à la figure 3.10 (Figure 3.10a). En tout, 4041 pics d'UBF ont été retenus pour les expériences subséquentes.

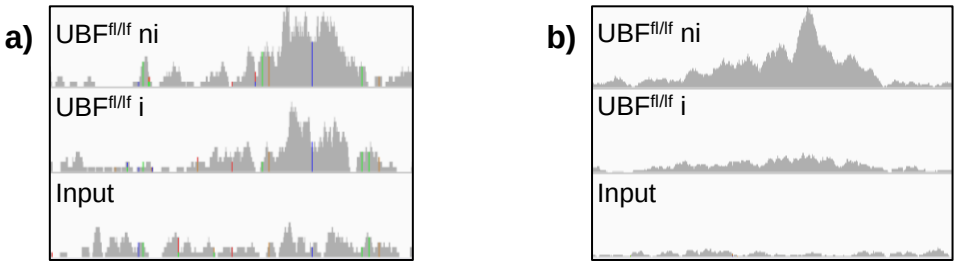

Figure 3.10– Faux positifs des pics d'UBF. a) Exemple d'un faux positif c'est-à-dire qu'il se retrouve tant dans le fichier $UBF^{fl/fl}$ dont le KO n'est pas induit (ni) que dans le fichier où le KO est induit (i), mais pas dans l'input. Sans avoir utilisé le fichier $UBF^{fl/fl}$ induit, ce pic aurait été sélectionné. b) Exemple d'un vrai positif. Le pic se retrouve seulement dans le fichier $UBF^{fl/fl}$ non induit.

3.6.1 – Distance des pics d'UBF par rapport aux sites d'initiation de la transcription

Après avoir identifié les sites de liaison d'UBF à travers le génome, GREAT a été utilisé afin de déterminer la distance de ces pics par rapport au site d'initiation de la

transcription. Il se trouve que la majorité des pics d'UBF se situent à une distance inférieure ou égale à 5 kb en aval des TSS en général (Figure 3.11).

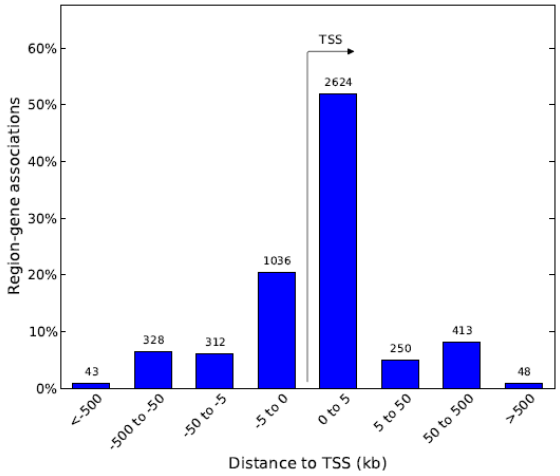

Figure 3.11– Distance des pics d'UBF par rapport aux TSS. La majorité des pics d'UBF se retrouvent à moins de 5 kb du site d'initiation de la transcription.

3.6.2 – Chevauchement avec les marques d'histones actives et répressives

En utilisant BEDOPS pour effectuer les chevauchements entre les pics d'UBF et des marques d'histones tant actives que répressives, force est de constater que les marques d'histones actives (en vert sur la Figure 3.12) sont davantage présentes aux pics d'UBF en comparaison aux marques répressives (en rouge). En effet, en moyenne 83.90 % des pics d'UBF colocalisent avec les marques d'histones actives tandis que seulement 6.61 % en moyenne des pics de marques d'histones répressives en font de même (Figure 3.12).

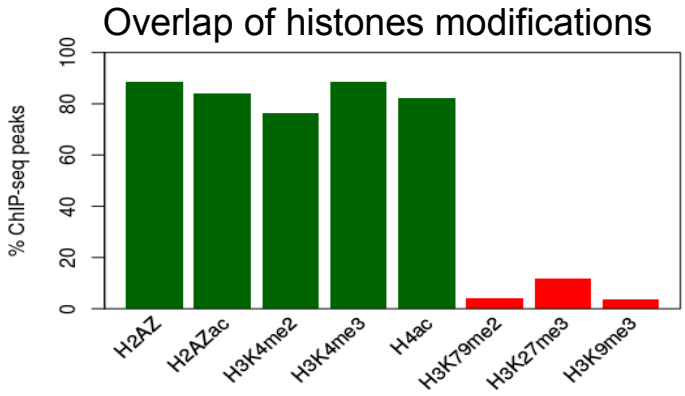

Figure 3.12 – Chevauchement des pics d'UBF et de marques d'histones. Il y a un plus grand nombre de chevauchements avec les marques actives (en vert) qu'avec les marques répressives (en rouge).

3.6.3 – Chevauchement avec les régions sensibles à la DNase I

L'utilisation de BEDOPS, afin d'effectuer le chevauchement des pics d'UBF et des pics de sensibilité à la DNase I, a permis de mettre en évidence que la majorité (85.5 %) des pics d'UBF chevauchent des endroits accessibles au clivage à la DNase I. Par contre, il s'agit seulement d'une minorité (6.3 %) des pics de DNase-seq (Figure 3.13).

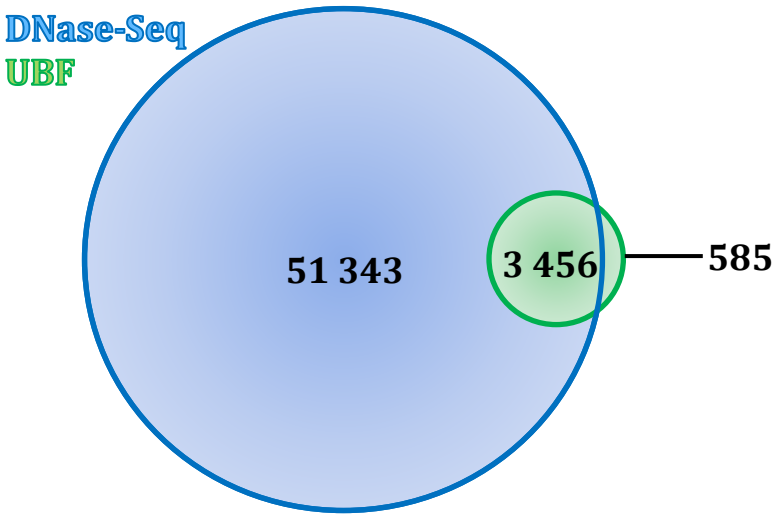

Figure 3.13 – Chevauchement des pics d'UBF et de DNase-seq. La majorité des pics d'UBF se retrouvent dans des endroits du génome sensibles au clivage à la DNase I.

3.6.4 – Perte d'accessibilité à la DNase I lors du KO d'UBF

L'utilisation du logiciel HOMER afin de déterminer si les pics de DNase-seq étaient différentiellement enrichis a montré que 69.58 % des pics ne l'étaient pas entre les deux conditions, soit $UBF^{WT/WT}$ KO (mimant un $UBF^{fl/fl}$ avant le KO) et $UBF^{fl/fl}$ KO. Cela a d'ailleurs démontré que 25.44 % des pics étaient plus enrichis lorsque le KO d'UBF était induit sur des cellules de type sauvage (phénotype similaire aux cellules $UBF^{fl/fl}$ avant l'induction du KO) tandis que 4.98 % des pics étaient plus enrichis lors de l'induction du KO d'UBF sur des cellules floxées (Figure 3.14).

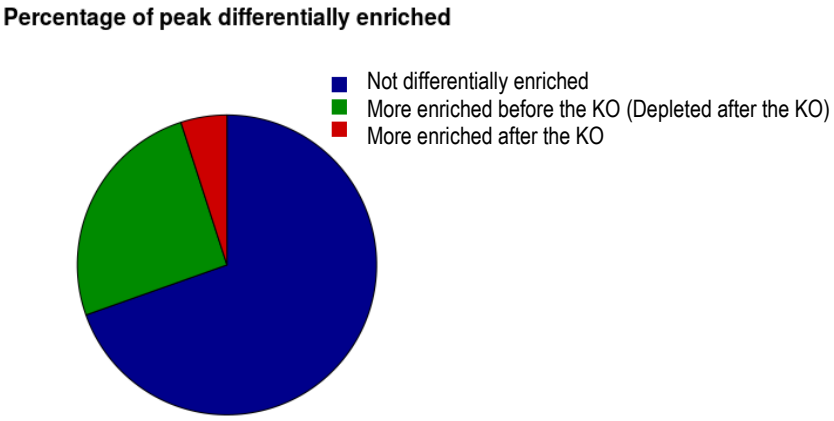

Figure 3.14 – Pourcentage de pics différentiellement enrichis. Une grande proportion des pics ne sont pas différentiellement enrichis.

BEDOPS a été utilisé afin de faire le chevauchement des pics d'UBF et des pics de DNase-seq différentiellement enrichis. La majorité soit 2651 pics d'UBF se retrouvent aux endroits où les pics de DNase-seq ne sont pas différentiellement enrichis entre $UBF^{WT/WT}$ KO et $UBF^{fl/fl}$ KO. De plus, 887 pics d'UBF colocalisent avec des régions où les pics de DNase-seq sont plus enrichis dans $UBF^{WT/WT}$ KO tandis que 257 pics colocalisent avec des régions où les pics de DNase-seq sont plus enrichis dans $UBF^{fl/fl}$ KO. Les pics d'UBF restants se retrouvent à des endroits où la chromatine n'est pas accessible à la DNase I (Figure 3.15).

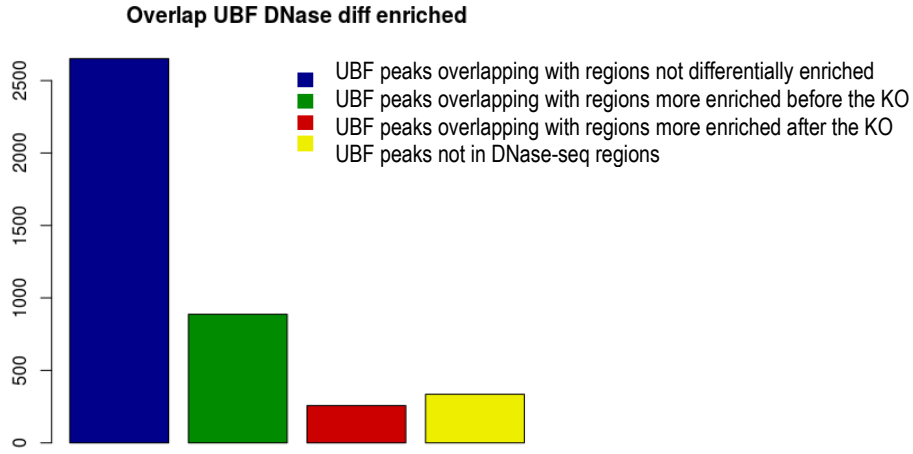

Figure 3.15 – Chevauchement des pics de DNase-seq différentiellement enrichis avec les pics d'UBF. La majorité des pics d'UBF sont dans des régions où il n'y a pas d'enrichissement différentiel des pics de DNase-seq.

3.6.5 – Colocalisation des gènes dérégulés dans les expériences de microarray avec les pics d'UBF plus enrichis avant le KO d'UBF

Les 887 pics d'UBF qui colocalisent avec les régions de DNase-seq plus enrichis dans la condition $UBF^{WT/WT}$ KO ont été corrélés avec les TSS et de ceux-ci, 675 pics se retrouvent à moins de 2 kb des sites d'initiation de la transcription. Parmi ces 675 pics, 26 corrèlent avec des gènes dérégulés à plus de 100 % dans les expériences de *microarray*, c'est-à-dire un changement de plus de 2 dans les valeurs de log (Figure 3.16).

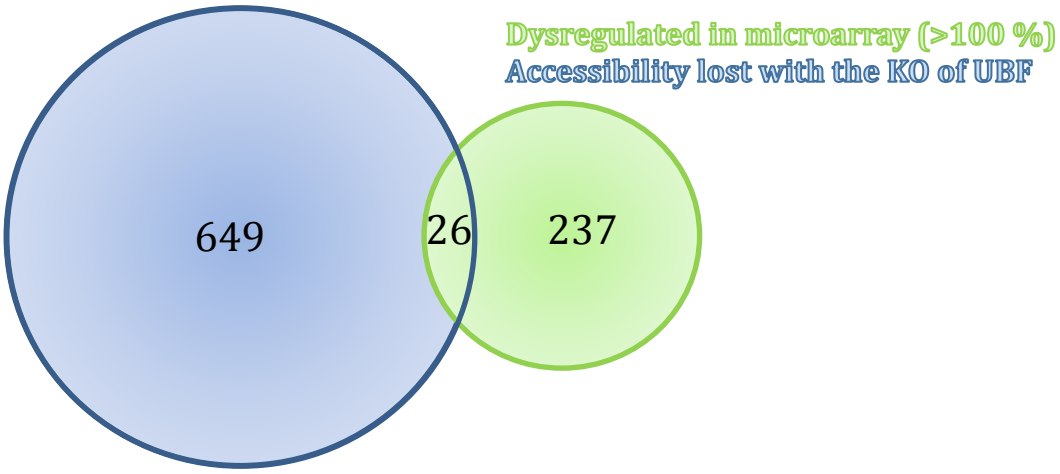

Figure 3.16 – Chevauchement des pics d'UBF corrélant avec des régions où les pics de DNase-seq sont plus enrichis dans les cellules $UBF^{WT/WT}$ KO et les gènes dérégulés dans les expériences de *microarray*. Une faible proportion des pics se chevauchent.

3.7 – Discussion

3.7.1 – UBF se retrouve près des sites d'initiation de la transcription

L'histogramme présenté à la Figure 3.11 montre que les pics d'UBF se retrouvent principalement près des TSS. Cela suggère qu'il agit à titre de facteur d'initiation de la transcription pour ces gènes ou encore qu'il agit à titre de protecteur des zones dépourvus de nucléosomes. En effet, les promoteurs se retrouvent près des TSS et sont connus comme étant de l'ADN sans nucléosome. Cela est dû au fait que les nucléosomes et les facteurs de transcription compétitionnent pour la liaison à l'ADN; les promoteurs étant un site de recrutement des facteurs de transcription (Ozonov et al., 2013).

3.7.2 – UBF COLOCALISE AVEC LES MARQUES D'HISTONES ACTIVES

UBF colocalise davantage avec les marques d'histones actives que les marques d'histones répressives. Cela indique qu'UBF est en lien avec les gènes transcrits ou potentiellement transcrits. Il pourrait agir à titre de facteur d'initiation de la transcription comme il le fait sur les gènes de l'ARNr.

3.7.3 – UBF SE RETROUVE AUX ENDROITS ACCESSIBLES À LA DNASE I

La Figure 3.13 montre que la majorité des pics d'UBF (85.5 %) colocalisent avec des endroits accessibles à la DNase I cela pourrait indiquer qu'UBF permet de protéger l'ADN dépourvus de nucléosomes. Pourtant, ces pics chevauchants représentent une minorité de l'ensemble des pics de DNase-seq (6.3 %). Cela pourrait donc dire qu'UBF agit de cette façon seulement sur un sous-ensemble de gènes particuliers, par exemple des gènes hautement transcrits.

3.7.4 – IL Y A PEU DE PERTE D'ACCESSIBILITÉ À LA DNASE I APRÈS LE KO D'UBF

Parmi les pics d'UBF, 887 sont plus enrichis dans les expériences de DNase-seq pour les cellules $UBF^{WT/WT}$ KO que pour les cellules $UBF^{fl/fl}$ KO. C'est donc dire que pour seulement 21.48 % des pics, l'accessibilité à la DNase I est perdue après le KO d'UBF aux emplacements où se trouve UBF. À ces endroits, UBF permettrait donc de protéger l'ADN des dommages en remplaçant probablement les nucléosomes. En effet, en son absence, l'ADN est plus compacte et alors moins accessible au clivage par la DNase I. Cet effet a aussi été observé sur l'ADNr où l'accessibilité est perdue avec le KO d'UBF (Herdman, Mars et al., 2017). Cela pourrait indiquer que les gènes où ce phénomène est observé sont hautement transcrits comme c'est le cas avec les gènes de l'ARNr.

3.7.5 – LA PERTE D'ACCESSIBILITÉ À LA DNASE I APRÈS LE KO D'UBF N'EST PAS EN LIEN AVEC LA DÉRÉGULATION DES GÈNES

Seulement 9.89 % des 263 gènes dérégulés à plus de 100 % lors du KO d'UBF dans les expériences de *microarray* voient leur accessibilité à la DNase I perdue suite au KO d'UBF. Cela représente seulement 3.85 % de tous les endroits où l'accessibilité

à la DNase I est perdue après le KO d'UBF. Cela indique qu'une faible minorité des gènes semble être régulée par UBF.

CHAPITRE 4 – DISCUSSION ET CONCLUSION

4.1 – LA PROCÉDURE DE DÉCONVOLUTION

Des ChIP-seq ont été effectués sur des MEFs conditionnelles pour UBF en utilisant des anticorps spécifiques pour plusieurs facteurs. Les immunoprécipitations d'UBF et de RPI étaient très spécifiques étant donné que l'inactivation d'UBF supprime l'enrichissement des deux facteurs. En utilisant le ChIP-seq pour RPI, la distribution de celui-ci n'était pas homogène contrairement à ce qui était attendu. Les profils obtenus tant d'UBF que de RPI montraient une ressemblance frappante avec le profil de l'*input*. L'inégalité de la couverture semble donc masquer le vrai profil d'interactions. Étant donné que le profil de l'*input* est reproductible, on peut considérer que l'inégalité de la couverture n'est pas dépendante de la préparation des échantillons, mais bien du protocole de séquençage en lui-même. Une déconvolution numérique semblait donc tout indiquée afin de retirer les biais de séquençage. La procédure de déconvolution comprend trois étapes : premièrement l'extension des *reads*, deuxièmement le lissage et troisièmement la division du signal de l'immunoprécipitation par celui de l'*input*. Le code de cette déconvolution est implémenté en Python (van Rossum, 1995) et génère des fichiers intermédiaires pour chaque étape.

Un exemple notable de l'efficacité de la déconvolution se retrouve au niveau des promoteurs. En effet, des interactions spécifiques avec UBF à ces endroits sont visibles seulement après déconvolution.

L'écart-type de la moyenne de six calculs (deux jeux de données d'UBF déconvolués avec trois *input* différents) montre que les variations entre les profils semblent négligeables biologiquement parlant.

4.2 – UTILISATION DE LA DÉCONVOLUTION DANS LE BUT DE CARTOGRAPHIER LE SPACER PROMOTER

Grâce à la déconvolution, la position exacte du *spacer promoter* murin a pu être mise à jour. En effet, SL1 se lie directement sur les promoteurs tandis qu'UBF se lie à 20 pb en amont. Le fait qu'UBF et SL1 se lient de manière similaire tant au *spacer promoter* qu'au 47S promoteur suggère que le même type de complexe de préinitiation s'y forme et qu'ils reconnaissent une certaine structure malgré la très faible homologie de séquence.

Dans l'humain, un *spacer promoter* n'a pas encore été cartographié. À l'aide du ChIP-seq des facteurs tels que TBP et UBF, cela a été rendu possible. Le pic d'UBF est placé à environ 30 pb en amont du pic de TBP. De plus, RPI se retrouve à 50 pb en aval du *spacer promoter* et directement adjacent à un site consensus de TTF1. Enfin, le site probable du *spacer promoter* est flanqué en amont par l'*Enhancer Boudary Complex* récemment identifié (Herdman, Mars et al., 2017). Tout cela suggère que le *spacer promoter* dans l'humain est régulé de la même façon que chez la souris.

4.3 – UTILISATION DE LA DÉCONVOLUTION DANS LE BUT DE CARTOGRAPHIER LES DIFFÉRENTS FACTEURS DE TRANSCRIPTION

La procédure de déconvolution a été appliquée à plusieurs jeux de données générés au laboratoire ainsi qu'aux jeux de données retrouvés dans les banques de données publiques telles que *GEO* (*Gene Expression Omnibus*) *DataSets database* de NCBI (*National Center for Biotechnology Information*). Cela a permis d'élucider le profil d'interaction protéique de toutes ces protéines. La Figure 4.1 montre le profil de RPI, Rrn3, UBF, TAF1B (TAF68) et TTF1 (Figure 4.1). Le profil de RPI est distribué de façon homogène le long de la région transcrite ce qui est en accord avec le modèle classique de *Miller Spread* présenté à la section 1.3.2. Le profil de Rrn3 est de forme exponentielle sur la région en aval du 47S promoteur. Cela indique que cette protéine est relâchée de façon stochastique de RPI, c'est-à-dire qu'elle ne dépend

d'un mécanisme en particulier, elle est relâchée dès que la polymérase entre en élongation. On remarque qu'il y a plus d'enrichissement d'UBF là où le pourcentage en GC est plus élevé. Donc, le profil d'UBF n'est pas homogène contrairement à RPI. Quant au profil de TAF68 et de TTF1, le double pic aux promoteurs, tel qu'illustré à la figure 1.15, est éliminé par la procédure de l'extension des *reads*.

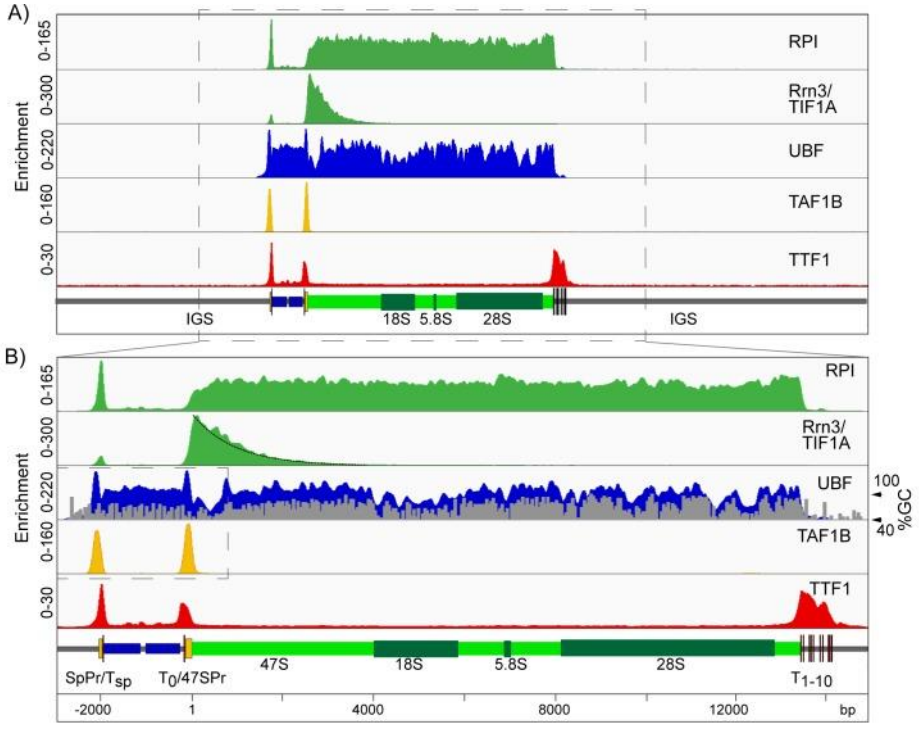

Figure 4.1 – Cartographie des différents facteurs de transcription. A) De haut en bas on retrouve RPI, Rrn3, UBF, TAF68 et TTF1. B) Agrandissement de A) avec le pourcentage en GC en gris superposé au profil d'UBF. La procédure de déconvolution a permis d'élucider la forme des profils de liaison.

4.4 – Rôle d'UBF à l'échelle du génome

Les analyses effectuées dans le chapitre 3 ont montré qu'UBF se retrouve principalement aux TSS, qu'il colocalise avec les marques d'histones actives et avec les endroits accessibles à la DNase I. De plus, l'accessibilité à la DNase I est perdue après le KO d'UBF sur une minorité d'emplacements où se trouve UBF. Tout cela semble indiquer qu'UBF agirait soit à titre de facteur de transcription soit à titre de protecteur de l'ADN en remplacement aux nucléosomes sur un petit sous-ensemble

de gènes par exemple certains gènes activement transcrits comme c'est le cas pour l'ADNr.

Cependant, la perte d'UBF entraîne l'arrêt de la transcription, de la réplication et de la division cellulaire. Ces vastes changements dans le métabolisme cellulaire pourraient potentiellement masquer les fonctions d'UBF sur les sites qui ne sont pas ribosomique. UBF est une protéine relativement abondante dans la cellule et a une faible affinité caractérisée par un haut taux de renouvellement (*turnover*) dû à un taux élevé de relâchement (*off-rate*) sur ses sites cibles de l'ADNr. Ainsi, pour maintenir les niveaux de liaison à l'ADNr, un grand réservoir (*pool*) d'UBF serait nécessaire pour mener à sa liaison sur l'ADNr. Il est donc possible d'imaginer que la liaison d'UBF à travers le génome est un moyen d'assurer sa disponibilité. Donc, la liaison d'UBF à l'ADNr est en échange avec un plus grand réservoir d'UBF lié dans tout le génome. Dans ce cas, nous pourrions nous attendre à ce qu'UBF se lie à tous les sites d'ADN disponibles, avec une très faible affinité, de sorte que les pics identifiés par ChIP-seq peuvent représenter seulement une petite fraction de la distribution d'UBF à la grandeur du génome.

4.5 – CONCLUSION

En conclusion, ce mémoire a décrit une procédure de déconvolution permettant la normalisation des données de séquençage de nouvelle génération telles que le ChIP-seq et le DNase-seq. Cette méthode de normalisation pourrait éventuellement être appliquée à la grandeur du génome pour des gènes à copie unique si la profondeur de séquençage est égale ou supérieure à 100 *reads* à chaque paire de base du génome, c'est-à-dire en condition de séquençage ultra profond (*ultra-deep sequencing*). Cela pourrait donc permettre une meilleure estimation du profil de liaison d'une multitude de facteurs de transcription et même des marques d'histones. Par contre, pour ce faire, il faudrait créer un logiciel de détection des sites d'interactions protéiques utilisant les fichiers au format *BED* comme fichier d'entrée (*input*).

Ce mémoire a d'ailleurs permis de mettre en lumière les potentiels rôles d'UBF à l'échelle du génome à l'aide d'expériences de ChIP-seq, de DNase-seq et de *microarray*.

BIBLIOGRAPHIE

Afgan E., Baker D., van den Beek M., Blankenberg D., Bouvier D., Čech M., Chilton J., Clements D., Coraor N., Eberhard C., Grüning B., Guerler A., Hillman-Jackson J., Von Kuster G., Rasche E., Soranzo N., Turaga N., Taylor J., Nekrutenko A., et Goecks J. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research 44: W3-W10.

Ahmad Y., Boisvert F.M., Gregor P., Cobley A. et Lamond AI. (2009). NOPdb: Nucleolar Proteome Database--2008 update. Nucleic Acids Research 37 (Database issue): D181-D184.

Akamatsu Y. et Kobayashi T. (2015). The human RNA polymerase I transcription terminator complex acts as a replication fork barrier that coordinates the progress of replication with rRNA transcription activity. Molecular and Cellular Biology 35: 1871-1881.

Al-Haggar M.M.S., Khair-Allaha B.A., Islam, M.M., Mohamed A.S.A. (2013). Bioinformatics in High Throughput Sequencing: Application in Evolving Genetic Diseases. Journal of data mining in genomics & proteomics 4: 131.

Bachellerie J.P., Cavaillé J. et Hüttenhofer A. (2002). The expanding snoRNA world. Biochimie 8: 775-790.

Bachvarov D., et Moss T. (1991). The RNA polymerase I transcription factor xUBF contains 5 tandemly repeated HMG homology boxes. Nucleic acids research 19: 2331-2335.

Barski A., Cuddapah S., Cui K., Roh T.Y., Schones D.E., Wang Z., Wei G., Chepelev I., Zhao K. (2007). High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837.

Bazett-Jones D.P., Leblanc B., Herfort M., et Moss T. (1994). Short-range DNA looping by the Xenopus HMG-box transcription factor, xUBF. Science 264: 1134-1137.

Bell S.P., Learned R.M., Jantzen H.M., Tjian R. (1988). Functional cooperativity between transcription factors UBF1 and SL1 mediates human ribosomal RNA synthesis. Science 41: 1192-1197.

Bierhoff H., Dundr M., Michels A.A., et Grummt I. (2008). Phosphorylation by casein kinase 2 facilitates rRNA gene transcription by promoting dissociation of TIF-IA from elongating RNA polymerase I. Molecular and cellular biology 28: 4988-4998.

Bird A. (2002) DNA methylation patterns and epigenetic memory. Genes & Development 16: 6-21.

Bird A.P., Taggart M.H., Nicholls R.D. et Higgs D.R. (1987). Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. EMBO Journal 6: 999-1004.

Boisvert F., van Koningsbruggen S., Navascués J. et Lamond A.I. (2007). The multifunctional nucleolus. Nature Reviews Molecular Cell Biology 8: 574-585.

Boulon S., Westman B.J., Hutten S., Boisvert F.M., et Lamond A.I. (2010). The nucleolus under stress. Molecular cell 40: 216-227.

Caudy A.A. et Pikaard C.S. (2002). Xenopus ribosomal RNA gene intergenic spacer elements conferring transcriptional enhancement and nucleolar dominance-like competition in oocytes. Journal of Biological Chemistry 35: 31577-31584.

Cavanaugh A.H., Evans A., et Rothblum L.I. (2008). Mammalian Rrn3 is required for the formation of a transcription competent preinitiation complex containing RNA polymerase I. Gene expression 14: 131-147.

Chédin S., Laferté A., Hoang T., Lafontaine D.L., Riva M. et Carles, C. (2007). Is ribosome synthesis controlled by polI transcription? Cell cycle 1: 11-15.

Chen H., Li Z., Haruna K., Li Z., Li Z., Semba K., Araki M., Yamamura K., et Araki K. (2008). Early pre-implantation lethality in mice carrying truncated mutation in the RNA polymerase 1-2 gene. Biochemical and biophysical research communications 365: 636-642.

Chen S., Seiler, J., Santiago-Reichelt M., Felbel K., Grummt I., et Voit R. (2013). Repression of RNA polymerase I upon stress is caused by inhibition of RNA-dependent deacetylation of PAF53 by SIRT7. Molecular cell 52: 303-313.

Denissov S., van Driel M., Voit R., Hekkelman M., Hulsen T., Hernandez N., Grummt I., Wehrens R. et Stunnenberg H. (2007). Identification of novel functional TBP-binding sites and general factor repertoires. EMBO Journal 26: 944-954.

Derenzini M., Trerè D., Pession A., Govoni M., Sirri V. et Chieco P. (2000). Nucleolar size indicates the rapidity of cell proliferation in cancer tissues. Journal of pathology 191: 181-186.

De Winter R.F.J. et Moss T. (1986). Spacer promoters are essential for efficient enhancement of X. laevis ribosomal transcription. Cell 44: 313-318.

De Winter R.F.J. et Moss T. (1987). A complex array of sequences enhances ribosomal transcription in Xenopus laevis. Journal of Molecular Biology 196: 813-827.

Dundr M., Hoffmann-Rohrer U., Hu Q., Grummt I., Rothblum L.I., Phair R.D. et Misteli T. (2002). A kinetic framework for a mammalian RNA polymerase in vivo. Science 298: 1623-1626.

Eberhard D., Tora L., Egly J.M. et Grummt I. (1993). A TBP-containing multiprotein complex (TIF-IB) mediates transcription specificity of murine RNA polymerase I. Nucleic Acids Research 21: 4180-4186.

Engel C., Sainsbury S., Cheung A.C., Kostrewa D. et Cramer P. (2013). RNA polymerase I structure and transcription regulation. Nature 502: 650-655.

Fernández-Tornero C., Moreno-Morcillo M., Rashid U.J., Taylor N.M., Ruiz F.M., Gruene T., Legrand P., Steuerwald U., Müller C.W. (2013). Crystal structure of the 14-subunit RNA polymerase I. Nature 502: 644-649.

Firek S., Read C., Smith D.R. et Moss T. (1989). The Xenopus laevis ribosomal gene terminator contains sequences that both enhance and repress ribosomal transcription. Molecular and Cellular Biology 9: 3777-3784.

French S.L., Osheim Y.N., Cioci F., Nomura M. et Beyer A.L. (2003). In exponentially growing Saccharomyces cerevisiae cells, rRNA synthesis is determined by the summed RNA polymerase I loading rate rather than by the number of active genes. Molecular and Cellular Biology 23: 1558-1568.

Friedrich J. K., Panov K.I, Cabart P., Russell J. et Zomerdijk, J.C. (2005). TBP-TAF complex SLI directs RNA polymerase 1 pre-initiation complex formation and stabilizes upstream binding factor at the rDNA promoter. Journal of biology and chemistry 280: 29551-29558.

Galamb O., Kalmár A., Barták B.K., Patai Á.V., Leiszter K., Péterfia B., Wichmann B., Valcz G., Veres G., Tulassay Z. et Molnár B. (2016). Aging related methylation influences the gene expression of key control genes in colorectal cancer and adenoma. World Journal of gastroenterology 22: 10325-10340.

Gerber J.K., Gögel E., Berger C., Wallisch M., Müller F., Grummt I. et Grummt F. (1997). Termination of mammalian rDNA replication: polar arrest of replication fork movement by transcription termination factor TTF-I. Cell 90: 559-567.

Gonzalez I.L. et Sylvester J.E. (1995). Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. Genomics 2: 320-328

Goodrich J.A. et Tjian R. (1994). TBP-TAF complexes: selectivity factors for eukaryotic transcription. Current Opinion in Cell Biology 6: 403-409.

Gorski J.J., Pathak S., Panov K., Kasciukovic T., Panova T., Russell J., Zomerdijk J.C.B.M. (2007). A novel TBP-associated factor of SL1 functions in RNA polymerase I transcription. EMBO Journal 26: 1560-1568.

Granneman S., et Baserga S.J. (2005). Crosstalk in gene expression: coupling and co-regulation of rDNA transcription, pre-ribosome assembly and pre-rRNA processing. Current opinion in cell biology 17: 281-286.

Grozdanov P., Georgiev O. et Karagyozov L. (2003). Complete sequence of the 45-kb mouse ribosomal DNA repeat: analysis of the intergenic spacer. Genomics 6: 637-643.

Grummt I., Rosenbauer H., Niedermeyer I., Maier U., et Ohrlein A. (1986). A repeated 18 bp sequence motif in the mouse rDNA spacer mediates binding of a nuclear factor and transcription termination. Cell 45: 837-846.

Hagen J.B. (2000). The origins of bioinformatics. Nature Review Genetics 1: 231-236.

Hamdane N., Stefanovsky V.Y., Tremblay M.G., Németh A., Paquet E., Lessard F., Sanij E., Hannan R., Moss T. (2014). Conditional inactivation of Upstream Binding Factor reveals its epigenetic functions and the existence of a somatic nucleolar precursor body. PLoS Genetics 10:e1004505.

Hanada K., Song C.Z., Yamamoto K., Yano K., Maeda Y., Yamaguch, K. et Muramatsu M. (1996). RNA polymerase I associated factor 53 binds to the nucleolar transcription factor UBF and functions in specific rDNA transcription. The EMBO journal 15: 2217-2226.

Hannan K.M., Sanij E., Rothblum L.I., Hannan R.D., et Pearson R.B. (2013). Dysregulation of RNA polymerase I transcription during disease. Biochimica et biophysica acta 1829: 342-360.

Hannan R.D., Cavanaugh A., Hempel W.M., Moss T., and Rothblum L. (1999). Identification of a mammalian RNA polymerase I holoenzyme containing components of the DNA repair/replication system. Nucleic acids research 27: 3720-3727.

Hayashi Y.K., Matsuda C., Ogawa M., Goto K., Tominaga K., Mitsuhashi S., Park Y.E., Nonaka I., Hino-Fukuyo N., Haginoya K., Sugano H. et Nishino I. (2009). Human PTRF mutations cause secondary deficiency of caveolins resulting in muscular dystrophy with generalized lipodystrophy. Journal of clinical investigation 119: 2623-2633.

Heinz S., Benner C., Spann N., Bertolino E., Lin Y.C., Laslo P., Cheng J.X., Murre C., Singh H., Glass C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell 38: 576-589.

Henderson S. et Sollner-Webb B. (1986). A transcriptional terminator is a novel element of the promoter of the mouse ribosomal RNA gene. Cell 47: 891-900.

Herdman C., Mars J.C., Stefanovsky V.Y., Tremblay M.G., Sabourin-Felix M., Lindsay H., Robinson M.D., Moss T. (2017). A unique enhancer boundary complex on the mouse ribosomal RNA genes persists after loss of Rrn3 or UBF and the inactivation of RNA polymerase I transcription. PLoS Genetics 13: e1006899.

Hernandez-Verdun D. et Louvert E. (2004). Le nucléole: structure, fonctions et maladies associées. Médecine/sciences 20: 37-44.

Hogeweg P. (2011). The roots of bioinformatics in theoretical biology. PLoS Computational Biology 7: e1002021.

Hung S.S., Lesmana A., Peck A., Lee R., Tchoubrieva E., Hannan K.M., Lin J., Sheppard K.E., Jastrzebski K., Quinn L.M., Rothblum L.I., Pearson R.B., Hannan R.D., Sanij E. (2017). Cell cycle and growth stimuli regulate different steps of RNA polymerase I transcription. Gene 15: 36-48.

Imazawa Y., Hisatake K., Mitsuzawa H., Matsumoto M., Tsukui T., Nakagawa K., Nakadai T., Shimada M., Ishihama A., et Nogi Y. (2005). The fission yeast protein Ker1p is an ortholog of RNA polymerase I subunit A14 in Saccharomyces cerevisiae and is required for stable association of Rrn3p and RPA21 in RNA polymerase I. The Journal of biological chemistry 280: 11467-11474.

Ionescu-Tîrgovişte C., Gagniuc P.A., Guja C. (2015). Structural properties of gene promoters highlight more than two phenotypes of diabetes. PLoS One 10: e0137950.

Jansa P., Mason S.W., Hoffmann-Rohrer U., et Grummt I. (1998). Cloning and functional characterization of PTRF, a novel protein which induces dissociation of paused ternary transcription complexes. The EMBO journal 17: 2855-2864.

Jantzen H. M., Admon A., Bell S.P. et Tjian R. (1990). Nucleolar transcription factor hUBF contains a DNA-binding motif with homology to HMG proteins. Nature 344: 830-836.

Kanehisa M. et Bork P. (2003). Bioinformatics in the post-sequence era. Nature Genetics 33: Suppl. 305-310.

Karbalaei M.S., Rippe C., Albinsson S, Ekman M., Mansten A., Uvelius B. et Swärd K. (2012). Impaired contractility and detrusor hypertrophy in cavin-1-deficient mice. European Journal of Pharmacology 15: 179-185.

Klose R.J. et Bird A.P. (2006). Genomic DNA methylation: the mark and its mediators. Trends in biochemical sciences 31: 89-97.

Koleske A.J., and Young R.A. (1994). An RNA polymerase II holoenzyme responsive to activators. Nature 368: 466-469.

Kuhn A. et Grummt I. (1987). A novel promoter in the mouse rDNA spacer is active in vivo and in vitro. The EMBO Journal 11: 3487-3492.

Kuhn A., Voit R., Stefanovsky V., Evers R., Bianchi M. et Grummt I. (1994). Functional differences between the two splice variants of the nucleolar transcription factor UBF: the second HMG box determines specificity of DNA binding and transcriptional activity. EMBO Journal 13: 416-424.

Kuhn C.D., Geiger S.R., Baumli S., Gartmann M., Gerber J., Jennebach S., Mielke T., Tschochner H., Beckmann R. et Cramer P. (2007). Functional architecture of RNA polymerase I. Cell 131: 1260-1272.

Langmead B. et Salzberg S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357-359.

Langst G., Blank T.A., Becker P.B., et Grummt I. (1997). RNA polymerase I transcription on nucleosomal templates: the transcription termination factor TTF-I induces chromatin remodeling and relieves transcriptional repression. The EMBO journal 16: 760-768.

Law J.A. et Jacobsen S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. 11: 204-220.

Learned R.M., Learned T.K., Haltiner M.M. et Tjian R.T. (1986). Human rRNA transcription is modulated by the coordinate binding of two factors to an upstream control element. Cell. 45: 847-857.

Lessard F., Morin F., Ivanchuk S., Langlois F., Stefanovsky V., Rutka J. et Moss T. (2010). The ARF tumor suppressor controls ribosome biogenesis by regulating the RNA polymerase I transcription factor TTF-I. Molecular Cell 38: 539-550.

Lessard F., Stefanovsky V., Tremblay M.G., Moss T. (2012). The cellular abundance of the essential transcription termination factor TTF-I regulates ribosome biogenesis and is determined by MDM2 ubiquitinylation. Nucleir Acids Research 40: 5357-5367.

Lin C.Y., Navarro S., Reddy S. et Comai L. (2006). CK2-mediated stimulation of Pol I transcription by stabilization of UBF–SL1 interaction. Nucleic Acids Research 34: 4752-4766.

Liu L., Hu N., Wang B., Chen M., Wang J., Tian Z., He Y. et Lin D. (2011). A brief utilization report on the Illumina HiSeq 2000 sequencer. Mycology 2: 169-191.

Liu L., Li Y., Li S., Hu N., He Y., Pong R., Lin D., Lu L. et Law M. (2012). Comparison of next-generation sequencing systems. Journal of Biomedicine & biotechnology 251364.

Madrigal P. et Krajewski P. (2012). Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. Frontiers in genetics 31: 230.

Martindill D. et Riley P.R. (2008). Cell cycle switch to endocycle: the nucleolus lends a hand. Cell cycle 7: 17-23.

Mason S.W., Wallisch M. et Grummt I. (1997). RNA polymerase I transcription termination: similar mechanisms are employed by yeast and mammals. Journal of Molecular Biology 268: 229-234.

Mayer C., Schmitz K.M., Li J., Grummt I., et Santoro, R. (2006). Intergenic transcripts regulate the epigenetic state of rRNA genes. Molecular cell 22: 351-361.

McLean C.Y., Bristor D., Hiller M., Clarke S.L., Schaar B.T., Lowe C.B., Wenger A.M., Bejerano G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnologies 28: 495-501.

McStay B. (2016) Nucleolar organizer regions: genomic 'dark matter' requiring illumination. Genes & Development 30: 1598-1610.

McStay B., et Grummt I. (2008). The epigenetics of rRNA genes: from molecular to chromosome biology. Annual review of cell and developmental biology 24: 131-157.

McStay B. et Reeder R.H. (1990). An RNA polymerase I termination site can stimulate the adjacent ribosomal gene promoter by two distinct mechanisms in Xenopus laevis. Genes & Development 4: 1240-1251.

Meehan R.R. et Stancheva I. (2001). DNA methylation and control of gene expression in vertebrate development. Essays in Biochemistry 37: 59-70.

Meraner J., Lechner, M, Loidl A, Goralik-Schramel M, Voit R. Grummt I. et Loidl P. (2006). Acetylation ofUBF changes during the cell cycle and regulates the interaction of UBF with RNA polymerase 1. Nucleic acids research 34: 1798-1806.

Metzker M.L. (2005). Emerging technologies in DNA sequencing. Genome research 15: 1767-1776.

Miller G., Panov K.I., Friedrich J.K., Trinkle-Mulcahy L., Lamond A.I., Zomerdijk J.C.B.M. (2001). hRRN3 is essential in the SL1-mediated recruitment of RNA Polymerase I to rRNA gene promoters. EMBO Journal 20: 1373-1382.

Moody G. (2004). Digital Code of Life : How bioinformatics is revolutionizing science, medicine and business.

Moss T. (1983). A transcriptional function for the repetitive ribosomal spacer in Xenopus laevis. Nature 302: 223-228.

Moss T. (2011). DNA methyltransferase inhibition may limit cancer cell growth by disrupting ribosome biogenesis. Epigenetics 6: 128-133.

Moss T., Langlois F., Gagnon-Kugler T. et Stefanovsky V. (2007). A housekeeper with power of attorney: the rRNA genes in ribosomal biogenesis. Cellular and molecular life sciences 64: 29-49.

Moss T., Stefanovsky V.Y. (2002). At the center of Eucaryotic life. Cell 109: 545-548.

Moss T., Stefanovsky V., Langlois F., et Gagnon-Kugler T. (2006). A new paradigm for the regulation of the mammalian ribosomal RNA genes. Biochemical Society transactions 34: 1079-1081.

Mullineux S.T., et Lafontaine D.L. (2012). Mapping the cleavage sites on mammalian pre-rRNAs: where do we stand? Biochimie 94: 1521-1532.

Nazar, R.N. (2004). Ribosomal RNA processing and ribosome biogenesis in eukaryotes. IUBMB life 56: 457-465.

Németh A., Guibert S., Tiwari V.K., Ohlsson R., et Langst G. (2008). Epigenetic regulation of TTF-I-mediated promoter-terminator interactions of rRNA genes. The EMBO journal 27: 1255-1265.

Németh A., Strohner R., Grummt I., et Langst G. (2004). The chromatin remodeling complex NoRC and TTF-I cooperate in the regulation of the mammalian rRNA genes in vivo. Nucleic acids research 32: 4091-4099.

Neph S., Kuehn M.S., Reynolds A.P., Haugen E., Thurman R.E., Johnson A.K.,1 Rynes E., Maurano M.T., Vierstra J., Thomas S., Sandstrom R., Humbert R., and Stamatoyannopoulos J.A. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics 28: 1919-1920.

O'Brown Z.K., Greer E.L. (2016). N6-Methyladenine: A Conserved and Dynamic DNA Mark. Advances in experimental medicine and biology 945: 213-246.

Olson M., Hingorani K. et Szebeni A. (2002). Conventional and nonconventional roles of the nucleolus. International review of cytology 219: 199-266.

O'Mahony D.J. et Rothblum L.I. (1991). Identification of two forms of the RNA polymerase I transcription factor UBF. Proceedings of the National Academy of Sciences of the United States of America 88: 3180-3184.

O'Mahony D.J., Smith S.D., Xie W., et Rothblum L.I. (1992). Analysis of the phosphorylation, DNA-binding and dimerization properties of the RNA polymerase I transcription factors UBF1 and UBF2. Nucleic acids research 20: 1301-1308.

Ozonov E.A., van Nimwegen E. (2013). Nucleosome free regions in yeast promoters result from competitive binding of transcription factors that interact with chromatin modifiers. PLoS Computational Biology 9: e1003181.

Paalman M.H., Henderson S.L. et Sollner-Webb B. (1995). Stimulation of the mouse rRNA gene promoter by a distal spacer promoter. Molecular and Cellular Biology 8: 4648-4656.

Panov K.I., Panova T.B., Gadal O., Nishiyama K., Saito T., Russell J. et Zomerdijk J.C. (2006). RNA polymerase I-specific subunit CAST/hPAF49 has a role in the activation of transcription by upstream binding factor. Molecular and Cellular Biology 26: 5436-5448.

Pietrzak M., Rempala G.A., Nelson P.T. et Hetman M. (2016). Non-random distribution of methyl-CpG sites and non-CpG methylation in the human rDNA promoter identified by next generation bisulfite sequencing. 858: 35-43.

Pikó L. et Clegg K.B. (1982). Quantitative changes in total RNA, total poly(A), and ribosomes in early mouse embryos. Developmental Biology 89: 362-378.

R Development Core Team. (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Raska, I. (2003). Oldies but goldies: searching for Christmas trees within the nucleolar architecture. Trends in Cell Biology 13: 517-525.

Rubbi C. et Milner J. (2003). Disruption of the nucleolus mediates stabilization of p53 in response to DNA damage and other stesses. The european molecular biology organization journal 22: 6068-6077.

Russell J. et Zomerdijk J.C.B.M. (2005). RNA-polymerase-I-directed rDNA transcription, life and works. TRENDS in biochemichal sciences 30: 87-96.

Russel J. et Zomerdijk J.C.B.M. (2006) The RNA polymerase I transcription machinery. Biochemistry Society Symposium 73: 203-2016.

Sanger F., Air G.M., Barrell B.G., Brown N.L., Coulson A.R., Fiddes C.A., Hutchison C.A., Slocombe P.M. et Smith M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265: 687-695.

Sanger F. et Coulson A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. 94: 441-448.

Santoro R., Li J et Grummt 1. (2002). The nucleolar remodeling complex NoRC mediates heterochromatin formation and silencing of ribosomal gene transcription. Nature genetics 32: 393-396.

Schnapp A., Pfleiderer C., Rosenbauer H. et Grummt I. (1990). A growth-dependent transcription initiation factor (TIF-IA) interacting with RNA polymerase I regulates mouse ribosomal RNA synthesis. EMBO Journal 9: 2857-2863.

Sirri V., Urcuqui-Inchima S, Roussel P et Hernandez-Verdun D. (2008). Nucleolus: the fascinating nuclear body. Histochem Cell Biol 129: 13-31.

Stefanovsky V.Y. et Moss T. (2008). The splice variants of UBF differentially regulate RNA polymerase I transcription elongation in response to ERK phosphorylation. Nucleic Acids Research 36: 5093-5101.

Stefanovsky V.Y., Langlois F., Gagnon-Kugler T., Rothblum L.I. et Moss T. (2006). Growth factor signaling regulates elongation of RNA polymerase I transcription in mammals via UBF phosphorylation and r-chromatin remodeling. Molecular Cell 21: 629-639.

Stefanovsky V.Y., Pelletier G., Bazett-Jones D.P., Crane-Robinson C., et Moss T. (2001-1). DNA looping in the RNA polymerase I enhancesome is the result of non-cooperative in-phase bending by two UBF molecules. Nucleic acids research 29: 3241-3247.

Stefanovsky V.Y., Pelletier G., Hannan R., Gagnon-Kugler T., Rothblum L.I. et Moss T. (2001-2). An immediate response of ribosomal transcription to growth factor stimulation in mammals is mediated by ERK phosphorylation of UBF. Molecular Cell 8: 1063-1073.

Stults D.M., Killen M.W., Pierce H.H. et Pierce A.J. (2008). Genomic architecture and inheritance of human ribosomal RNA gene clusters. Genome Research 18: 13-18.

Stults D.M., Killen M.W., Williamson E.P., Hourigan J.S., Vargas H.D., Arnold S.M., Moscow J.A. et Pierce A.J. (2009). Human rRNA gene clusters are recombinational hotspots in cancer. Cancer Research 23: 9096-9104.

Szalkowski A.M. et Schmid C.D. (2010). Rapid innovation in ChIP-seq peak-calling algorithm is outdistancing benchmarking efforts. Briefings in Bioinformatics 12: 626-633.

Tosto G. et Reitz C. (2013). Genome-wide association studies in Alzheimer's disease: a review. Current Neurology and Neuroscience Reports 13: 381.

Tsai R. Y. et McKay R.D. (2002). A nucleolar mechanism controlling cell proliferation in stem cells and cancer cells. Genes & Development 16: 2991-3003.

Tuan J., Zhai W.G., Comai L. (1999). Recruitment of TATA-binding protein-TAF complex SL1 to the human ribosomal DNA promoter is mediated by the carboxyterminal activation domain of upstream binding factor (UBF) and is regulated by UBF phosphorylation. Molecular cell biology 19: 2872-2879.

van Riggelen J., Yetil A. et Felsher D.W. (2010). MYC as a regulator of ribosome biogenesis and protein synthesis. Nature Reviews Cancer 10: 301-309.

van Rossum G. (1995) Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI).

Véron A., Blein S. et Cox D.G. (2014). Genome-wide association studies and the clinic: a focus on breast cancer. Biomarkers in medicine 8: 287-296.

Visintin R. et Amon A. (2000). The nucleolus: the magician's hat for cell cycle tricks. Current opinion in cell biology 12: 372-377.

Voelkerding K.V., Dames S.A. et Durtschi J.D. (2009) Next-generation sequencing: from basic research to diagnostics. Clinical chemistry 55: 641-658.

Yamamoto R.T., Nogi Y., Dodd J.A. et Nomura M. (1996). RRN3 gene of Saccharomyces cerevisiae encodes an essential RNA polymerase I transcription factor which interacts with the polymerase independently of DNA template. EMBO Journal 15: 3964-3973.

Yuan X., Zhao J., Zentgraf H., Hoffmann-Rohrer U., et Grummt I. (2002). Multiple interactions between RNA polymerase I, TIF-IA and TAF(I) subunits regulate preinitiation complex assembly at the ribosomal gene promoter. EMBO reports 3: 1082-1087.

Yuan X., Zhou Y., Casanova E., Chai M., Kiss E., Gröne H.J., Schütz G. et Grummt I. (2005). Genetic inactivation of the transcription factor TIF-IA leads to nucleolar disruption, cell cycle arrest, and p53-mediated apoptosis. Molecular Cell 19: 77-87.

Zeng W., Mortazavi A. (2012). Technical considerations for functional sequencing assays. Nature immunology 13: 802-807.

Zentner G.E., Saiakhova A., Manaenkov P., Adams M.D. et Scacheri P.C. (2011) Integrative genomic analysis of human ribosomal DNA. Nucleic Acids Research 39: 4949-4960.

Zhang X., Hu M., Lyu X., Li C., Thannickal V.J., Sanders Y.Y. (2017). DNA methylation regulated gene expression in organ fibrosis. Biochimica et Biophysica acta 4439: 30147

Zhang Y, Liu T., Meyer C.A., Eeckhoute J, Johnson D.S., Bernstein B.E., Nusbaum C., Myers R.M., Brown M., Li W. et Shirley Liu X.S. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome biology 9: R137.

Zhao J., Yuan X., Frodin M., et Grummt I. (2003). ERK-dependent phosphorylation of the transcription initiation factor TIF-IA is required for RNA polymerase I transcription and cell growth. Molecular cell 11: 405-413.

Zomerdijk J.C.B.M., Beckmann H., Comai L., Tjian R. (1994). Assembly of transcriptionally active RNA polymerase I initiation factor SL1 from recombinant subunits. Science 266: 2015-2018.

ANNEXE I – SCRIPT DECONVONORM

Ce script, écrit en Python, permet d'effectuer la déconvolution des données ChIP-seq présentées au Chapitre 2.

```
#!/usr/bin/python
# encoding: utf-8

# deconvoNorm
# author: Marianne S. Felix
# marianne.sabourin-felix.1@ulaval.ca
# Version : 1.0
# 2017-02-08
#
# Tested on : Ubuntu 14.04 LTS, 16.04 LTS
#      with : Python 2.7.10, 2.7.11


"""
deconvoNorm module
"""

import os
import sys
import argparse
from subprocess import call
import shutil
from math import ceil
from itertools import izip


def transition():
    """
    Transition between steps
    """
    print "-" * 60


###################################
#                                 #
#####  Variable validation   #####
#                                 #
###################################

def _validFile(path):
    """
    Intern function that raises an error if filename
    doesn't exists or is not in BAM format

    Return path
    """

    if not os.path.isfile(path):
        err = "The input file doesn't exist."
        raise argparse.ArgumentTypeError(err)
    else:
```

```
        try:
            open(path, 'r')
        except:
            err = "The input file can't be opened in reading mode."
            raise argparse.ArgumentTypeError(err)

        if not path.endswith(".bam"):
            err = "The input file must be in BAM format."
            raise argparse.ArgumentTypeError(err)
    return path


def _validDirectory(path):
    """
    Intern function that raises an error if filename
    doesn't exists or is not in narrowPeak format

    Return path
    """

    if not os.path.isdir(path):
        err = "The IP folder is not a directory."
        raise argparse.ArgumentTypeError(err)
    else:
        if not any(fname.endswith('.bam') for fname in
os.listdir(path)):
            err = "The IP folder must contain bam files."
            raise argparse.ArgumentTypeError(err)
        else:
            return path


def _validFragmentLength(value):
    """
    Intern function that raises an error if
    the fragment length is below 1

    Return fragment length
    """

    try:
        value = int(value)
    except:
        raise argparse.ArgumentTypeError("The fragment length must be a
int.")

    if value <= 0:
        raise argparse.ArgumentTypeError("The fragment length must be
over 0.")

    return value
```

2

```
def _validWindowSize(value):
    """
    Intern function that raises an error if the
    window size is below 1 and an even number

    Return window size
    """

    try:
        value = int(value)
    except:
        raise argparse.ArgumentTypeError("The window size must be a
int.")

    if value <= 0:
        raise argparse.ArgumentTypeError("The window size must be over
0.")

    if value % 2 == 0:
        raise argparse.ArgumentTypeError(
            "The window size must be an odd number.")

    return value


def _validThreshold(value):
    """
    Intern function that raises an error if the
    threshold value is below 0

    Return threshold
    """

    try:
        value = float(value)
    except:
        raise argparse.ArgumentTypeError("The threshold must be a
float.")

    if value < 0:
        raise argparse.ArgumentTypeError("The threshold must be at least
0.")

    return value


def parseArgv(argv=None):
    """
    This function allows to parse the command line input options
```

```
    Return parsed input parameters
    """

    ######################
    ### Parse arguments ###
    ######################
    parser = argparse.ArgumentParser(
        formatter_class=argparse.RawDescriptionHelpFormatter,
        description="  This script allows to use deconvolution"
                  + " to normalize sequencing data.",
        epilog="""
example 1 (one file):
  python deconvoNorm.py -i input.bam -f ip.bam -c MmrDNA -l 100 -w 25 -t
10 -o output

example 2 (many files):
  python deconvoNorm.py -i input.bam -d ipFolder -c MmrDNA -l 100 -w 25
-t 10 -o output
  """)

    ### OPTIONAL ARGUMENTS ###
    #parser.add_argument("--listchr", type=_validFile, metavar=('FILE'),
                        #help="List chromosomes from BAM file")

    ### FILES ###
    group1 = parser.add_argument_group("required files")

    # Input DNA file
    group1.add_argument("-i", "--input", type=_validFile, required=True,
                        help="Input DNA file (BAM format)")

    # IP file
    options = group1.add_mutually_exclusive_group(required=True)
    options.add_argument("-f", "--ipfile", type=_validFile,
                        help="Immunoprecipitation file (BAM format)")
    options.add_argument("-d", "--ipdirectory", type=_validDirectory,
                        help="Directory with many immunoprecipitation
files")

    ### OPTIONS ###
    group2 = parser.add_argument_group("options")

    # Chromosome name
    group2.add_argument("-c", "--chrname", type=str, required=True,
                        help="Chromosome of interest (to normalize on)")

    # Fragment length
    group2.add_argument("-l", "--fragmentlength", default=100,
                        type=_validFragmentLength,
                        help="Sequenced fragment length (Default =
```

```
100)")

    # Window size
    group2.add_argument("-w", "--windowsize", default=25,
                        type=_validWindowSize,
                        help="Smoothing window size (Default = 25)")

    # Threshold
    group2.add_argument("-t", "--threshold", default=10,
                        type=_validThreshold,
                        help="Coverage threshold (Default = 10)")

    # Output name
    group2.add_argument("-o", "--outputname", type=str, required=True,
                        metavar=('OUTPUT'), help="Output folder name " +
                        "(if folder already exists, it will be
overwritten)")

    # Keep intermediate files
    group2.add_argument("-k", "--keepfiles", action="store_true",
                        help="Keep intermediates files (Default =
False)")

    # No RPM ratio
    group2.add_argument("-r", "--norpm", action="store_true",
                    help="Don't adjust in Read Per Million (Default =
False)")

    return parser.parse_args(argv)


##################################
#                                #
##########  Functions   ##########
#                                #
##################################

def listChr(filename):
    """
    This function allows to list the chromosomes
    present in a BAM file

    Return the list of chromosomes
    """

    ### File validation ###
    if not os.path.isfile(filename):
        print("error : The input file doesn't exist.")
        sys.exit()
    try:
        open(filename, 'r')
```

```
    except IOError:
        print("error : The input file can't be opened in reading mode.")
        sys.exit()

    if not filename.endswith(".bam"):
        print("error : The input file must be in BAM format.")
        sys.exit()

    ### Create index if doesn't exist ###
    createIndex(filename)

    ### List chromosomes ###
    print "List of chromosomes for file {} :".format(filename)
    #out = check_output([])

    # Store idxstats in a temporary file
    with open("idxstats_deconvoNorm.txt", 'w') as f:
        call("samtools idxstats {}".format(filename), stdout=f,
shell=True)

    # Read first column of idxstats
    with open("idxstats_deconvoNorm.txt", 'r') as r:
        for line in r:
            print(line.split()[0])

    os.remove("idxstats_deconvoNorm.txt")


def removeFiles(boolean, listOfFiles):
    """
    This function removes temporary
    files if -k option is OFF
    """

    ### If k is not present, removes the temporary files ###
    if not boolean:
        for oldFile in listOfFiles:
            os.remove(oldFile)


def createIndex(filename):
    """
    This function creates an index file
    if it doesn't already exists
    """

    basename = os.path.basename(filename)
    name = os.path.splitext(filename)[0]
    index = name + ".bai"

    # Check if file.bai or file.bam.bai exist
```

```
    if not os.path.isfile(index) and not os.path.isfile(filename +
".bai"):
        print "Creating index for file {}...".format(basename)
        call("samtools index {}".format(filename), shell=True)


def extractChr(filename, chrom, outFolder):
    """
    This function creates a bam file
    with only the chromosome of interest

    Return chromosome file names
    """

    ### Validation of chromosome name ###
    with open("idxstats_deconvoNorm.txt", 'w') as f:
        call("samtools idxstats {}".format(filename), stdout=f,
shell=True)

    formattedChrom = "{}\t".format(chrom)
    if formattedChrom not in open("idxstats_deconvoNorm.txt",
'r').read():
        print("error : Chromosome {} not found in file {}"
              .format(chrom, filename))
        #print("Please, choose a chromosome in the list.")
        #print("To display the list of chromosomes names do :" +
        #      "python deconvoNorm.py --listChr filename.bam")
        sys.exit()

    os.remove("idxstats_deconvoNorm.txt")

    ### Creation of output name ###
    basename = os.path.basename(filename)
    newname = os.path.splitext(basename)[0] + "-{}.bam".format(chrom)
    outname = "{}/{}".format(outFolder, newname)

    ### Extraction of the chromosome of interest ###
    print("Extracting chromosome {} from file {}...".format(chrom,
basename))
    call("samtools view -b {} {} > {}".format(filename, chrom, outname),
         shell=True)
    print("Temporary chromosome file created !")

    return outname


def bamtobed(filename):
    """
    This function creates a bed file
    with only the chromosome of interest
```

```
    Return bed file names
    """

    ### Creation of output name ###
    basename = os.path.basename(filename)
    outname = os.path.splitext(filename)[0] + ".bed"

    ### Bam to bed conversion ###
    print("Converting {} to bed format...".format(basename))
    call("bedtools bamtobed -i {} > {}".format(filename, outname),
shell=True)
    print("Temporary bed file created !")

    return outname


def extReadLength(filename, fragmSize, chrLength):
    """
    This function extend read
    length from bed files

    Return extended bed file names
    """

    ### Creation of output name ###
    basename = os.path.basename(filename)
    outname = os.path.splitext(filename)[0] + "-
l{}.bed".format(fragmSize)

    ### Extending read length ###
    print("Extending read length for file {}...".format(basename))

    # Read the bed file line by line
    with open(filename, 'r') as f, open(outname, 'w') as o:
        for line in f:
            # Assignation parameters
            chrom, start, end, name, score, strand = line.split()

            # Don't exceed the chromosome end
            if strand == "+":
                #newEnd = (chrLength) if (start + fragmSize > chrLength)
                #else (start + fragmSize)

                if (int(start) + fragmSize > int(chrLength)):
                    newEnd = chrLength
                else:
                    newEnd = int(start) + fragmSize

                o.write("{}\t{}\t{}\t{}\t{}\t{}\n"
                        .format(chrom, start, newEnd, name, score,
strand))
```

8

```
            # Don't exceed the chromosome start
            else:
                #newStart = (0) if (end - fragmSize < 0) else (end -
fragmSize)

                if (int(end) - fragmSize < 0):
                    newStart = 0
                else:
                    newStart = int(end) - fragmSize

                o.write("{}\t{}\t{}\t{}\t{}\t{}\n"
                        .format(chrom, newStart, end, name, score,
strand))

    print("Temporary extended bed file created !")

    return outname


def genomecoverage(filename, chromInfo):
    """
    This function extract the
    coverage from bed files

    Return coverage bed file names
    """

    ### Creation of output name ###
    basename = os.path.basename(filename)
    outname = os.path.splitext(filename)[0] + "-cov.bed"

    ### Extending read length ###
    print("Extracting coverage for file {}...".format(basename))

    call("bedtools genomecov -i {} -g {} -d > {}"
         .format(filename, chromInfo, outname), shell=True)

    print("Temporary coverage bed file created !")

    return outname


def smooting(filename, winSize, threshold, chrLength):
    """
    This function smooths the coverage
    profile with a sliding window
    and replace by zero when the coverage
    falls below the threshold

    Return smoothed coverage bed file names
```

```
    """
    ### Creation of output name ###
    basename = os.path.basename(filename)
    outname = os.path.splitext(filename)[0] + "-w{}-
t{}.bed".format(winSize, threshold)

    ### Smoothing coverage profile ###
    print("Smoothing coverage for file {}...".format(basename))

    slidingWindow = ceil(float(winSize) / 2)

    # Read the bed file line by line
    with open(filename, 'r') as f, open(outname, 'w') as o:
        lineNo = 0
        average = []
        for line in f:
            lineNo += 1
            # Assignation parameters
            chrom, pos, count = line.split()

            # Add the count to the end of the array
            average.append(float(count))

            # The first floor(winSize/2) lines will have a mean of zero
            if lineNo < slidingWindow:
                o.write("{}\t{}\t{}\n".format(chrom, pos, 0))

            # Once the average array contain winSize element, print the
mean
            if len(average) == int(winSize):
                newPos = int(pos) - int(slidingWindow) + 1

                #newCount = sum(average) / len(average)

                #if newCount < threshold:
                #    newCount = 0

                mean = sum(average) / len(average)

                if float(mean) < float(threshold):
                    newCount = 0
                else:
                    newCount = mean

                o.write("{}\t{}\t{}\n".format(chrom, newPos, newCount))

                # Remove the first entry of the array
                del average[0]

            # The last floor(winSize/2) lines will have a mean of zero
            if lineNo == int(chrLength):
```

10

```
                newPos = int(pos) - int(slidingWindow) + 2
                for i in range(newPos, int(chrLength) + 1):
                    o.write("{}\t{}\t{}\n".format(chrom, i, 0))

    print("Temporary smoothed bed file created !")

    return outname


def scaleToRPM(filename, rpm):
    """
    This function scales the coverage
    in per million reads

    Return scaled coverage bed file names
    """

    ### Creation of output name ###
    basename = os.path.basename(filename)
    outname = os.path.splitext(filename)[0] + "-rpm.bed"

    print("Scaling to RPM for file {}...".format(basename))

    with open(filename, 'r') as f, open(outname, 'w') as o:
        for line in f:
            chrom, pos, count = line.split()

            newCount = float(count) * float(rpm)

            o.write("{}\t{}\t{}\n".format(chrom, pos, newCount))

    print("Temporary rpm bed file created !")

    return outname


def divisionOfInput(ipfile, inputfile):
    """
    This function divide the ip
    signal by the input signal

    Return division bed file names
    """

    basename = os.path.basename(ipfile)
    outname = os.path.splitext(ipfile)[0] + "_norm.bed"

    print("Division by input for file {}...".format(basename))

    # Scan the ip file and the input file simultaneously
    with open(ipfile, 'r') as ip, open(inputfile, 'r') as inp,
```

11

```
open(outname, 'w') as o:
        for lineIp, lineInput in izip(ip, inp):
            # Get parameters
            chromIP, posIP, countIP = lineIp.split()
            chromInput, posInput, countInput = lineInput.split()

            # Avoid division by zero error
            if float(countInput) == 0.0:
                division = 0
            else:
                division = float(countIP) / float(countInput)

            o.write("{}\t{}\t{}\n".format(chromIP, posIP, division))

    print("Temporary divided bed file created !")

    return outname


def bedtobedgraph(filename):
    """
    This function convert bed file
    into bedgraph format
    """

    basename = os.path.basename(filename)
    outname = os.path.splitext(filename)[0] + ".bedgraph"

    print("Final conversion to bedgraph format for file
{}...".format(basename))

    with open(filename, 'r') as f, open(outname, 'w') as o:

        # Header of bedgraph file
        o.write("track type=bedgraph visibility=full color=0,0,204
altColor=204,0,0 autoScale=on maxHeightPixels=128:128:11
viewLimits=0.0:25.0 yLineMark=0.0 windowingFunction=maximum\n")

        for line in f:
            chrom, end, count = line.split()

            start = int(end) - 1

            o.write("{}\t{}\t{}\t{}\n".format(chrom, start, end, count))

    print("Final bedgraph file created !")


###################################
#                                 #
##########  M A I N  #############
```

```
#                                 #
###################################

def main():
    """
    Main function

    """

    ## List chromosomes ###
    #if sys.argv[1] == "--listChr":
        #listChr(sys.argv[2])
        #sys.exit
    #else:
        ## Parse argv
        #parseArgv()

    ### Parse argv ###
    argv = parseArgv()

    ### List immunoprecipitation files in an array ###
    listIPfiles = []
    if argv.ipdirectory:
        # Put the bam files in a list (except from input if present)
        for bamFile in os.listdir(argv.ipdirectory):
            path = "{}/{}".format(argv.ipdirectory, bamFile)
            if bamFile.endswith(".bam") and path != argv.input:
                listIPfiles.append(path)
        if len(listIPfiles) == 0:
            print("error : Only input present in
{}".format(argv.ipdirectory))
            sys.exit()
    elif argv.ipfile:
        listIPfiles.append(argv.ipfile)

    ### New list with all the files ###
    listFiles = list(listIPfiles)
    listFiles.append(argv.input)

    ### Create index if necessary ###
    for bamFile in listFiles:
        createIndex(bamFile)

    ### Create output folder (overwrite if exists) ###
    out = argv.outputname
    if os.path.exists(out):
        shutil.rmtree(out)
    os.makedirs(out)

    ### Extract chromosome of interest ###
    transition()
```

```
    chrFiles = []
    for bamFile in listFiles:
        chrFiles.append(extractChr(bamFile, argv.chrname, out))

    ### Bam to bed conversion ###
    transition()
    bedFiles = []
    for chrFile in chrFiles:
        bedFiles.append(bamtobed(chrFile))

    # Remove temporary chromosomes files
    removeFiles(argv.keepfiles, chrFiles)

    ### Extension of read length ###
    # Extract chromosome length
    # Store idxstats in a temporary file
    with open("idxstats_deconvoNorm.txt", 'w') as f:
        call("samtools idxstats {}".format(argv.input), stdout=f,
shell=True)

    # Find length of the chromosome of interest
    with open("idxstats_deconvoNorm.txt", 'r') as r:
        for line in r:
            chrom, length, aligned, notaligned = line.split()
            if chrom == argv.chrname:
                chrLength = length
                break

    # Extend read length
    transition()
    extFiles = []
    for bedFile in bedFiles:
        extFiles.append(extReadLength(bedFile, argv.fragmentlength,
chrLength))

    # Remove temporary bed files
    removeFiles(argv.keepfiles, bedFiles)

    ### Extraction of coverage ###
    # Create chromInfo file
    with open("idxstats_deconvoNorm.txt", 'r') as r,
open("chromInfo_deconvoNorm.txt", 'w') as o:
        for line in r:
            chrom, length, aligned, notaligned = line.split()
            if chrom == argv.chrname:
                o.write("{}\t{}\t{}\t{}"
                        .format(chrom, length, aligned, notaligned))
                break

    os.remove("idxstats_deconvoNorm.txt")
```

14

```
    # Extract coverage
    transition()
    covFiles = []
    for extFile in extFiles:
        covFiles.append(genomecoverage(extFile,
"chromInfo_deconvoNorm.txt"))

    os.remove("chromInfo_deconvoNorm.txt")

    # Remove temporary extended bed files
    removeFiles(argv.keepfiles, extFiles)

    ### Smoothing coverage files ###
    transition()
    smoothFiles = []
    for covFile in covFiles:
        smoothFiles.append(smooting(covFile, argv.windowsize,
argv.threshold, chrLength))

    # Remove temporary coverage bed files
    removeFiles(argv.keepfiles, covFiles)

    ### Scaling to RPM ###
    # Get RPM ratio associated with each file

    if not argv.norpm:
        dictRpm = {}
        for originalFile in listFiles:
            # Associate originalFile with extFile
            basename = os.path.basename(originalFile)
            new = os.path.splitext(basename)[0] + "-{}-l{}-cov-w{}-
t{}.bed".format(argv.chrname, argv.fragmentlength, argv.windowsize,
argv.threshold)
            outname = "{}/{}".format(argv.outputname, new)

            # Store idxstats in a temporary file
            with open("idxstats_{}.txt".format(basename), 'w') as f:
                call("samtools idxstats {}".format(originalFile),
                     stdout=f, shell=True)

            # Compute RPM ratio
            mappedReads = 0
            with open("idxstats_{}.txt".format(basename), 'r') as r:
                for line in r:
                    chrom, length, aligned, notaligned = line.split()
                    mappedReads += int(aligned)

                rpm = float(1000000) / float(mappedReads)

            # Add file and rpm to dictionnary
            dictRpm[outname] = rpm
```

15

```
            os.remove("idxstats_{}.txt".format(basename))

        # Scale to RPM
        transition()
        rpmFiles = []
        for smoothFile in smoothFiles:
            rpmFiles.append(scaleToRPM(smoothFile, dictRpm[smoothFile]))

        removeFiles(argv.keepfiles, smoothFiles)

    ### Division of IP files by input DNA file ###
    basename = os.path.basename(argv.input)

    if argv.norpm:
        new = os.path.splitext(basename)[0] + "-{}-l{}-cov-w{}-
t{}.bed".format(argv.chrname, argv.fragmentlength, argv.windowsize,
argv.threshold)
    else:
        new = os.path.splitext(basename)[0] + "-{}-l{}-cov-w{}-t{}-
rpm.bed".format(argv.chrname, argv.fragmentlength, argv.windowsize,
argv.threshold)

    newInput = "{}/{}".format(argv.outputname, new)

    transition()
    divFiles = []
    for ipfile in listIPfiles:

        basename = os.path.basename(ipfile)

        if argv.norpm:
            new = os.path.splitext(basename)[0] + "-{}-l{}-cov-w{}-
t{}.bed".format(argv.chrname, argv.fragmentlength, argv.windowsize,
argv.threshold)
        else:
            new = os.path.splitext(basename)[0] + "-{}-l{}-cov-w{}-t{}-
rpm.bed".format(argv.chrname, argv.fragmentlength, argv.windowsize,
argv.threshold)

        newIp = "{}/{}".format(argv.outputname, new)

        divFiles.append(divisionOfInput(newIp, newInput))

    # Remove temporary rpm or smooth bed files
    if argv.norpm:
        removeFiles(argv.keepfiles, smoothFiles)
    else:
        removeFiles(argv.keepfiles, rpmFiles)

    ### Final bedgraph file ###
```

```
    transition()
    for divFile in divFiles:
        bedtobedgraph(divFile)

    # Remove temporary coverage bed files
    removeFiles(argv.keepfiles, divFiles)

#################################

if __name__ == '__main__':

    main()

#################################
```

17

ANNEXE II – ARTICLE PUBLIÉ 4E AUTEUR

Cette annexe contient la version intégrale de l'article intitulé *A unique enhancer boundary complex on the mouse ribosomal RNA genes persists after loss of Rrn3 or UBF and the inactivation of RNA polymerase I transcription* qui a été publié le 17 juillet 2017 dans PLoS genetics. Il est reproduit dans ce mémoire avec la permission des coauteurs. Ma contribution dans cet article peut se voir aux figures 1, 2, 3, 6, 7 et 8 dans lesquelles j'ai analysé les données de ChIP-seq et effectué la procédure de déconvolution décrite au chapitre 2 et dans la section *Analysis of massively parallel sequence data* du Matériels et méthodes de ce présent manuscrit.

Cet article est également disponible en ligne :
http://doi.org/10.1371/journal.pgen.1006899
http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006899
PubMed: https://www.ncbi.nlm.nih.gov/pubmed/28715449

Annexe II – RÉSUMÉ

La transcription de plusieurs centaines de gènes de l'ARN ribosomique chez l'humain et la souris représente la majorité de la synthèse des ARN dans le noyau de la cellule et est le déterminant de l'abondance des ribosomes cytoplasmiques, un facteur clé dans la régulation de l'expression des gènes. Les gènes de l'ARNr, désignés globalement comme l'ADNr, sont regroupés en répétitions directes dans les *Nucleolar Organizer Regions*, NORs, de plusieurs chromosomes et dans de nombreuses cellules. Les répétitions actives sont transcrites à des niveaux près de la saturation. L'ADNr est également un *hotspot* de recombinaison et de point de rupture des chromosomes, et donc comprendre son contrôle a une vaste importance. Malgré la nécessité d'un niveau élevé de transcription d'ADNr, typiquement, seule une fraction de l'ADNr est transcriptionnellement active et certains NORs sont éteints (*silenced*) en permanence par la méthylation des CpG. Divers complexes de remodelage de la chromatine permettent de contrecarrer le *silencing* afin de maintenir l'activité de l'ADNr. Cependant, la structure de la chromatine de la portion active des ADNr est encore vague. Ici, nous avons combiné un protocole ChIP-Seq haute résolution ainsi qu'une inactivation conditionnelle des facteurs clefs fondamentaux pour mieux comprendre ce qui détermine la chromatine active de l'ADNr. Les données élucident les questions concernant l'interdépendance des facteurs de transcription fondamentaux, montrent que la formation du complexe de préinitiation est dirigée par le facteur d'architecture UBF (UBTF) indépendamment de la transcription et que la terminaison et la libération de RPI correspondent au site de liaison de TTF1. Ils révèlent d'autre part l'existence d'un *Enhancer Boudary Complex* asymétrique formé par CTCF et Cohésine et flanqué en amont par des nucléosomes en phase et en aval par un complexe de l'ARN polymérase I arrêté. Nous avons trouvé que l'*Enhancer Boudary Complex* est le seul site de modification des histones actives sur la répétition de 46 kb de l'ADNr. De manière frappante, elle délimite non seulement chaque gène fonctionnel de l'ARNr, mais aussi est maintenue de manière stable après l'inactivation du gène et le rétablissement de la chromatine répressive environnante. Nos données définissent un état en équilibre (*poised*) de la chromatine de l'ADNr et placent l'*Enhancer*

Boudary Complex comme point d'entrée probable pour les complexes de remodelage de la chromatine.

PLOS | GENETICS

RESEARCH ARTICLE

A unique enhancer boundary complex on the mouse ribosomal RNA genes persists after loss of Rrn3 or UBF and the inactivation of RNA polymerase I transcription

Chelsea Herdman[1,2‡], Jean-Clement Mars[1,2‡], Victor Y. Stefanovsky[1], Michel G. Tremblay[1], Marianne Sabourin-Felix[1,2], Helen Lindsay[3,4], Mark D. Robinson[3,4], Tom Moss[1,2*]

1 Laboratory of Growth and Development, St-Patrick Research Group in Basic Oncology, Cancer Division of the Quebec University Hospital Research Centre, Québec, Canada, **2** Department of Molecular Biology, Medical Biochemistry and Pathology, Faculty of Medicine, Laval University, Québec, Canada, **3** Institute of Molecular Life Sciences, University of Zürich, Zürich, Switzerland, **4** SIB Swiss Institute of Bioinformatics, University of Zürich, Zürich, Switzerland

‡ These authors share first authorship on this work.
* Tom.Moss@crhdq.ulaval.ca

Citation: Herdman C, Mars J-C, Stefanovsky VY, Tremblay MG, Sabourin-Felix M, Lindsay H, et al. (2017) A unique enhancer boundary complex on the mouse ribosomal RNA genes persists after loss of Rrn3 or UBF and the inactivation of RNA polymerase I transcription. PLoS Genet 13(7): e1006899. https://doi.org/10.1371/journal.pgen.1006899

Editor: Brian McStay, National University of Ireland Galway, IRELAND

Received: March 20, 2017

Accepted: June 27, 2017

Published: July 17, 2017

Data Availability Statement: All ChIP-Seq and DNase-Seq data, including biological replicas, have been deposited in the ArrayExpress database at EMBL-EBI (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-5839.

Funding: The study was funded by an operating grant from the Canadian Institutes of Health Research (CIHR, MOP12205) and a CIHR Frederick Banting and Charles Best Canada Graduate

Abstract

Transcription of the several hundred of mouse and human Ribosomal RNA (rRNA) genes accounts for the majority of RNA synthesis in the cell nucleus and is the determinant of cytoplasmic ribosome abundance, a key factor in regulating gene expression. The rRNA genes, referred to globally as the rDNA, are clustered as direct repeats at the Nucleolar Organiser Regions, NORs, of several chromosomes, and in many cells the active repeats are transcribed at near saturation levels. The rDNA is also a hotspot of recombination and chromosome breakage, and hence understanding its control has broad importance. Despite the need for a high level of rDNA transcription, typically only a fraction of the rDNA is transcriptionally active, and some NORs are permanently silenced by CpG methylation. Various chromatin-remodelling complexes have been implicated in counteracting silencing to maintain rDNA activity. However, the chromatin structure of the active rDNA fraction is still far from clear. Here we have combined a high-resolution ChIP-Seq protocol with conditional inactivation of key basal factors to better understand what determines active rDNA chromatin. The data resolve questions concerning the interdependence of the basal transcription factors, show that preinitiation complex formation is driven by the architectural factor UBF (UBTF) independently of transcription, and that RPI termination and release corresponds with the site of TTF1 binding. They further reveal the existence of an asymmetric Enhancer Boundary Complex formed by CTCF and Cohesin and flanked upstream by phased nucleosomes and downstream by an arrested RNA Polymerase I complex. We find that the Enhancer Boundary Complex is the only site of active histone modification in the 45kbp rDNA repeat. Strikingly, it not only delimits each functional rRNA gene, but also is stably maintained after gene inactivation and the re-establishment of surrounding repressive chromatin. Our data define a poised state of rDNA chromatin and place the Enhancer Boundary Complex as the likely entry point for chromatin remodelling complexes.

Scholarship Doctoral Award to CH (CIHR CGS-D). The Research Centre of the Québec University Hospital Centre (CHU de Québec) is supported by the Fonds de recherche du Québec - Santé (FRQS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

The ability to translate genetic messages into functional proteins is an absolute necessity for life forms on our planet, and is achieved by the largest known enzyme, the ribosome. The mammalian ribosome is a 4 MDa complex of 4 catalytic RNA molecules and more than 80 proteins that uses amino-acid charged transfer RNAs (tRNAs) to decode the genome via the intermediary of the messenger RNAs (mRNAs) [1]. The efficiency of mRNA translation is a key determinant of gene expression. Expression is of course contingent on gene transcription, on RNA processing, transport and degradation, but changes in translational capacity not only control the rate of protein synthesis, but also the spectrum of mRNAs that are translated. In this way translation plays an important role in the regulation of gene expression independently of gene transcription [2–4]. This necessarily also feeds back onto genome programming by determining the available spectrum of transcription and epigenetic factors, as well as on the full spectrum of nuclear and cytoplasmic organelle functions. The importance of differential mRNA translation is evident in the control of cell growth, tumour suppressors, oncogenes and the differentiation of stem cells [4–6]. In short, the ability to translate mRNAs is a key determinant of cellular phenotype, and the central factor in this process is the ribosome. It is therefore essential that we understand the factors that control ribosome synthesis and assembly.

Since a mammalian cell contains up to 10 million ribosomes, e.g. [7, 8], their synthesis in proliferating cells is a major limitation for growth and occupies up to 50% of all gene transcription and a significant proportion of translation [1, 9, 10]. Indeed, ribosome synthesis directly determines cell proliferation, and a doubling of proliferation rate requires a four-fold increase in the ribosome synthesis rate [11]. The major components of the mammalian ribosome, the 18S and 28S ribosomal RNAs (rRNAs) are synthesized along with the 5.8S rRNA as a single 47S rRNA precursor. This precursor rRNA is assembled into pre-ribosomal particles in the nucleolus with the aid of several hundred accessory proteins and hundreds of small guide RNAs, before being processed into the mature rRNAs and transported to the cytoplasm [1, 12, 13].

In human and mouse, the 47S rRNA is encoded on some 200 haploid gene copies that are organized in direct repeats on the short arms of five chromosomes [9, 14, 15]. These large ribosomal DNA (rDNA) loci constitute the Nucleolar Organiser Regions (NORs), and their active transcription is responsible for the assembly of the nucleoli, the largest subnuclear organelle. The 47S rRNA is transcribed by RNA polymerase I (RPI/Polr1/PolI) and a set of basal transcription factors that are dedicated to this task. Three basal RPI factors have been identified, the "Selectivity" complex SL1, containing TBP and TAF1A through D, and the HMG1-box architectural Upstream Binding Factor (UBF) are responsible for forming the pre-initiation complex. The third factor, Rrn3-TIF1A, associates with RPI and is required for recognition of the UBF/SL1 complex at the promoter. Transcription Termination Factor 1 (TTF1), a Reb-homology DNA binding factor related to yeast Reb1 and Nsi1, is required for termination of the 47S rRNA transcript, but also plays a role in determining silencing of the rDNA [16, 17].

Transcription of the rDNA is stimulated in response to nutrients, growth factors and a wide range of cellular stresses, and both Rrn3 and UBF are direct targets of growth regulatory pathways [1, 18–20]. However, a fraction of the rDNA is transcriptionally silent in most somatic cells and cell lines, and some NORs are heterochromatic, probably corresponding to a heavily methylated rDNA subfraction. Various chromatin-remodelling complexes have been implicated in rDNA activity and have been proposed to act by displacing nucleosomes, in particular at the 47S rRNA promoter site. However, the chromatin structure of the active rDNA fraction is still far from clear. Here we have combined a high-resolution ChIP-Seq protocol

with conditional inactivation of two key regulated basal factors, Rrn3/TIF1A and UBF to better understand what determines active rDNA chromatin. The data resolve many questions concerning the interdependence and functions of the basal transcription factors and show that UBF defines the chromatin structure of the actively transcribed rDNA. They also answer doubts as to the essential nature of Rrn3/TIF1A *in vivo*. Perhaps most importantly, the data reveal the existence of an Enhancer Boundary Complex formed by CTCF and Cohesin and three or four phased nucleosomes lying immediately adjacent to the Spacer Promoter-Enhancer repeat unit and an arrested RPI elongation complex. We find that this boundary complex is the only significant site of active histone modifications in the whole 45kbp rDNA repeat. Strikingly, the Enhancer Boundary Complex not only clearly delimits the functional rRNA gene unit, but is stably maintained after inactivation of RPI transcription and the replacement of UBF by nucleosomal chromatin. Our data potentially places the Enhancer Boundary Complex as the key entry point for the remodelling complexes that activate rDNA transcription and defines the poised state of rDNA chromatin.

Results

In order to better understand the factors that determine rRNA gene activity, we first needed to establish high resolution, low noise interaction maps for the components of the RPI transcription machinery. We achieved this using a combination of Chromatin Immuno-Precipitation, massively parallel DNA sequencing (ChIP-Seq) and normalization to sequence coverage obtained on control (input) DNA, see Materials and Methods for details.

A high-resolution map of basal factors across the mouse rDNA

The normalized ChIP-Seq data from Mouse Embryonic Fibroblasts (MEFs) precisely delineated UBF binding, the SL1 pre-initiation complexes (PICs), engagement of initiation competent and elongating forms of RPI, and the sites of termination factor TTF1 binding (Fig 1). Engagement of RPI was found to be near uniform throughout the 47S transcribed region, strongly suggesting that in MEFs there are no major sites of pausing at which polymerase accumulates. RPI engagement was also found to end abruptly at the downstream TTF1 sites (Fig 1A and 1B), see Discussion. In contrast, the RPI associated initiation factor Rrn3-TIF1A showed an interaction only over the first 2 to 3 kb of the transcribed region, consistent with *in vitro* data showing that Rrn3 is released sometime after initiation [21, 22]. The Rrn3-RPI interaction decayed exponentially with increasing distance from the initiation site, suggesting that its release was stochastic (Fig 1B). Assuming this, and a mean elongation rate of 60 $nt.sec^{-1}$ [23], we could estimate that the half-life of the elongating RPI-Rrn3 complex was around 15 s in MEFs. Thus, Rrn3 acts mechanistically much like the Sigma factors of eubacteria that target the polymerase to promoters but are released during elongation [24–26]. Indeed, like bacterial Sigma, Rrn3 was recently shown to contain a DNA interaction domain that was required for RPI initiation [27]. This said, the recent structural data for RPI-Rrn3 and the RPI-Rrn3-Core Factor complex suggest Rrn3 functions by modulating the RPI structure and dimerization, and that contact between Rrn3 and the RPI promoter DNA is unlikely to occur within the initiation complex [28, 29].

UBF binding precisely delimits to the functional rDNA unit

Various *in vitro* studies have implicated UBF in the formation of the RPI preinitiation complex [1, 10, 30, 31]. But it has also been implicated in regulating RPI elongation [23] and was shown to bind over a wide region of the rDNA repeat, suggesting a role more in line with the formation of active chromatin [32–35]. Our data now show that in fact UBF is in a position to

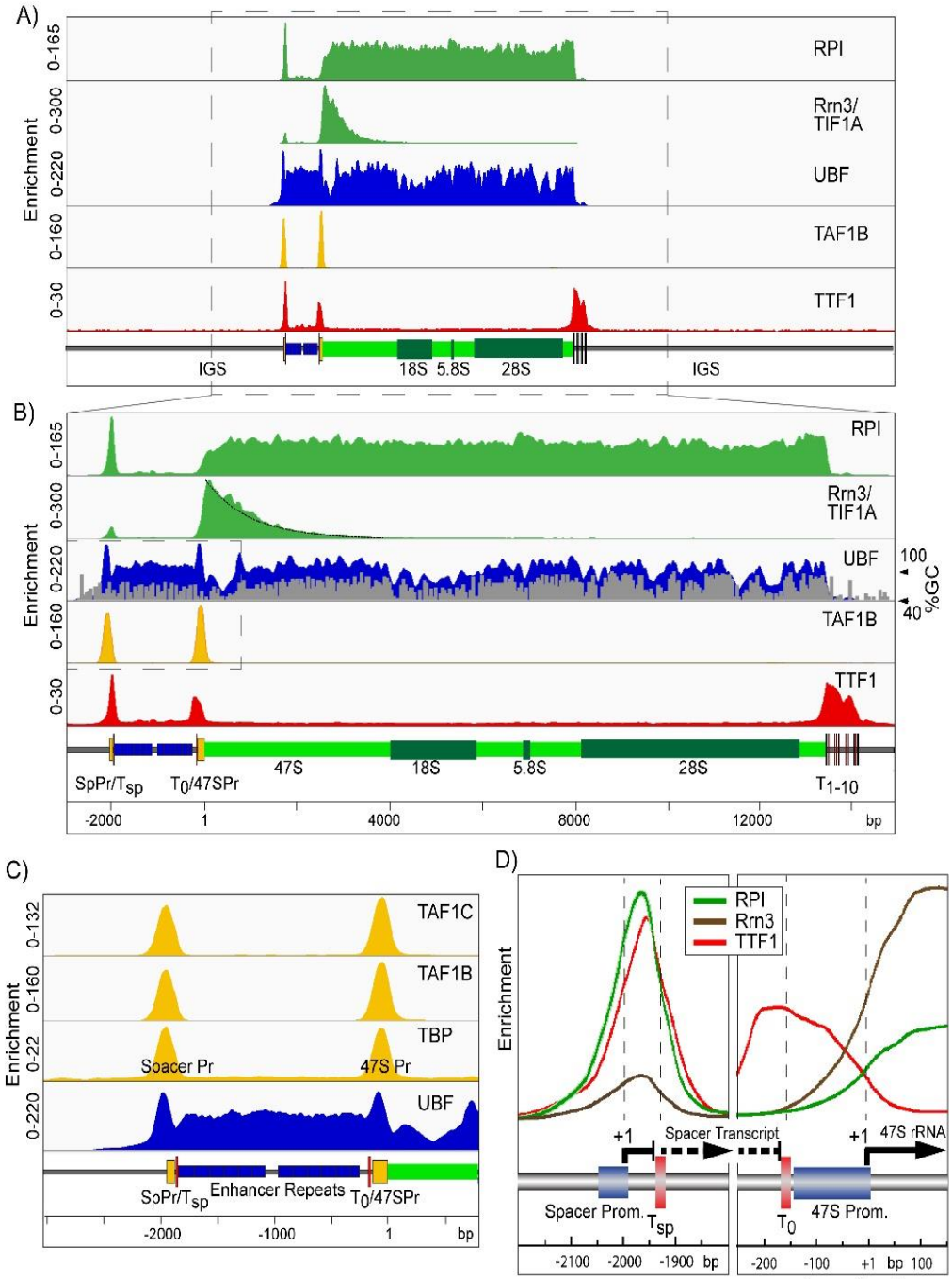

Fig 1. High resolution ChIP-Seq maps of RPI, the RPI basal factors and TTF1 across the mouse rDNA repeat unit of *Ubf*$^{+/+}$/*Rrn3*$^{+/+}$*p53*$^{-/-}$ control MEFs. A) and B) show the ChIP enrichment profile maps for RPI, Rrn3, UBF, the SL1 component TAF1B and TTF1, A) over the full mouse rDNA repeat and, B) at higher resolution for just the functional rRNA gene unit (boxed region in A). The UBF enrichment profile in B) is overlayed with

the GC-content profile of the rDNA sequence, see also S1 Fig, and an exponential curve fit to the Rrn3 enrichment profile (black) is shown downstream of the 47S initiation site (+1). C) An enlargement of the enrichment profiles (boxed region in B) for the SL1 components TAF1C, TAF1B and TBP in comparison with that for UBF across the Spacer Promoter, Enhancer Repeat and 47S Promoter. D) Shows the superimposed enrichment profiles of RPI, Rrn3 and TTF1 at the Spacer and 47S Promoters. The enrichment scale for each factor is the same in left and right panels. In A) to D) a scale map of the rDNA sequence elements is given below each panel. ChIP enrichment for each factor is given as; ChIP-Seq reads per million (RPM)/Input DNA RPM.

https://doi.org/10.1371/journal.pgen.1006899.g001

perform all these roles. UBF was mapped throughout the functional rRNA gene unit and was bounded by two flanking DNA elements, upstream by the Spacer Promoter (SpPr) and downstream by the transcription termination sites (Sal-boxes) bound by Transcription Termination Factor 1 (TTF1). Consistent with its very low sequence selectivity [36–38], UBF bound almost continuously throughout the 47S transcribed region, but displayed a modulation that closely followed the G+C (above 50%) profile of the DNA, (Fig 1B and S1 Fig). This modulation of binding probably represented the true UBF occupancy and was not due to cross-linking or sequencing biases, since it was not observed for RPI. Further, the correlation with high G+C did not hold for the 47S and Spacer Promoter sequences. Both promoters displayed UBF binding that peaked in low G+C sequences and overlapped the SL1/TIF1B TBP-complex binding (TAF1B, -1C and TBP, Fig 1C), defining the promoter PIC. Further, no UBF at all (<1% of 47S region) was detected in the relatively G+C neutral Intergenic Spacer (IGS) (Fig 1A and S1 Fig). Thus, it was very unlikely that recruitment of UBF to the rRNA genes was determined by DNA sequence selectivity. Though once recruited, exact UBF positioning may be affected by underlying DNA sequences and/or the recruitment of other factors such as SL1 and TTF1.

A stalled RPI transcription complex lies near the upstream UBF boundary

Our previous ChIP data showed a very significant peak of RPI mapping to the Spacer Promoter [35]. The much higher resolution of our ChIP-Seq data revealed that this peak was in fact centred not over the promoter, but 24 bp downstream of the initiation site and just 13 bp upstream of the peak of TTF1 binding associated with the Spacer Termination site T_{sp} (Fig 1D). RPI at this site was also associated with 10 times less Rrn3 than RPI immediately adjacent to the 47S Promoter, and hence was in large part in an initiation incompetent state. Both these observations strongly suggested that transcription complexes initiated at the Spacer Promoter were arrested by TTF1 very early in elongation. The Spacer Promoter was previously shown to be important for the function of the Enhancer repeats [39, 40], which are also thought to be the main entry points for UBF binding [41]. Thus, it was possible that polymerase stalling at the Spacer Promoter played some role in UBF recruitment. Alternatively, it could play a regulatory part in TTF1 induced DNA looping [42] or, by analogy with RNA polymerase II genes, as part of an insulator complex [43].

Rrn3-TIF1A is an essential factor in mouse

Given the essential nature of UBF in rDNA activity [35], we sought to better understand what defined the extent of its binding and first asked whether rDNA transcription itself was necessary. Rrn3-TIF1A associates with mouse RPI to form the initiation competent form of the polymerase [1, 21, 22] (Fig 1), hence its elimination should specifically inactivate rDNA transcription. However, while the yeast Rrn3 ortholog is essential for rDNA transcription [44], there was some doubt whether this was also the case in mouse [45]. Mouse embryos lacking Rrn3-TIF1A were reported not to arrest development until E9.5, by which time zygotic transcription normally increases rRNA levels by over 1000 fold [46, 47]. Thus, it was possible that

mouse Rrn3-TIF1A may not be essential for rDNA transcription. When we analyzed the Rrn3-null mice we found that in fact they arrested development during the early cleavage stages (S2A, S2B and S3A Figs), as previously observed for UBF and RPI deletions. We therefore concluded that mouse Rrn3 was indeed an essential and non-redundant part of the RPI transcription machinery, and hence was an ideal target to induce specific inhibition of rDNA transcription, see legend to S2 Fig for further details.

RPI transcription is not required to maintain actively poised rDNA

MEFs conditional for Rrn3 (*Rrn3*$^{fl/fl}$/*ER-Cre*$^{+/+}$/*p53*$^{-/-}$) and isogenic control MEFs (*Rrn3*$^{+/+}$/*ER-Cre*$^{+/+}$/*p53*$^{-/-}$) were isolated from embryos carrying the *Rrn3-TIF1A*^{flox} allele [45] (S3A Fig, Materials and Methods), and used to determine the rDNA status before and after Rrn3 loss. Treatment of these MEFs with multiple low doses of 4-hydroxytamoxifen (4-HT) inactivated the Rrn3 gene and reduced Rrn3 protein levels by >90%, while the same treatment had no effect on Rrn3 levels in the *Rrn3*$^{+/+}$ isogenic control cells (S3B Fig and Materials and Methods). Concomitantly with Rrn3 depletion, *de novo* synthesis of the 47S precursor RNA, as determined by metabolic labelling [48], was strongly suppressed (S3F Fig).

High resolution mapping of factor binding across the rDNA repeat in *Rrn3*$^{fl/fl}$/*ER-Cre*$^{+/+}$/*p53*$^{-/-}$ MEFs was indistinguishable from that seen in the wild type MEFs, (compare Figs 1B and 2A). Near complete loss of Rrn3 protein after 4-HT treatment essentially eliminated its engagement with the 5' of the 47S transcribed region and strongly suppressed recruitment of RPI throughout the 47S region, as well as at the Spacer Promoter (SpPr) (Fig 2A and 2B). In contrast, Rrn3 loss had no effect on the maintenance of UBF binding either within the 47S region, the enhancer repeats, or at the 47S and Spacer Promoters. Binding of the SL1 TBP complex (TAF1B and TBP) at both Promoters was also unaffected by Rrn3 loss, in agreement with data showing Rrn3 is not required to maintain UAF and Core Factor binding at the yeast RPI promoter [49]. Further, TTF1 remained bound at the adjacent T_{sp} and T_0 sites, as well as at the T_1-T_{10} 47S termination sites after Rrn3 loss. Parallel 4-HT treatment of isogenic *Rrn3*$^{+/+}$/*ER-Cre*$^{+/+}$/*p53*$^{-/-}$ MEFs had no effect on either Rrn3 or RPI recruitment, or indeed on any of the chromatin factors in this study, see below. Thus, Rrn3 and RPI engagement on the rDNA appeared not to be required for the establishment of the preinitiation complexes at the Spacer and 47S Promoters, or for the normal pattern of UBF binding. In support of this, colony forming assays (S3G Fig) strongly suggested that the remnant engagement of RPI and Rrn3 detected after 4-HT treatment in Fig 2 resulted from a small percentage of cells that had retained a functional Rrn3 gene, and did not represent a low level of transcription in all cells.

These data demonstrated that the potentially active state of the rRNA genes, as defined by SL1-UBF preinitiation complexes at both Spacer and 47S promoters and UBF binding throughout the Enhancer Repeats and 47S gene body, was stably maintained through 48h of very low total Rrn3 levels and 24h of significant transcriptional repression (S3 Fig, panels B and F). This long-term stability of UBF binding was surprising given its low DNA binding constant (Kd ~ 10nM) [50] and high *in vivo* off-rate ($t_{1/2}$ 9 to 25 s) and its inability to compete with nucleosome formation [51]. Thus, the data suggested that a transcription independent mechanism may exist to maintain UBF binding and the potentially active state of the rRNA genes.

UBF is essential for the recruitment of the RPI transcription machinery

Loss of UBF in SV40 transformed conditional MEFs was previously shown to strongly repress RPI transcription, [35], and this we found also to be the case in the p53$^{-/-}$ immortalized (*Ubf*$^{fl/fl}$/*ER-Cre*$^{+/+}$/*p53*$^{-/-}$) MEFs used in the present study (S3D–S3F Fig). Using these MEFs,

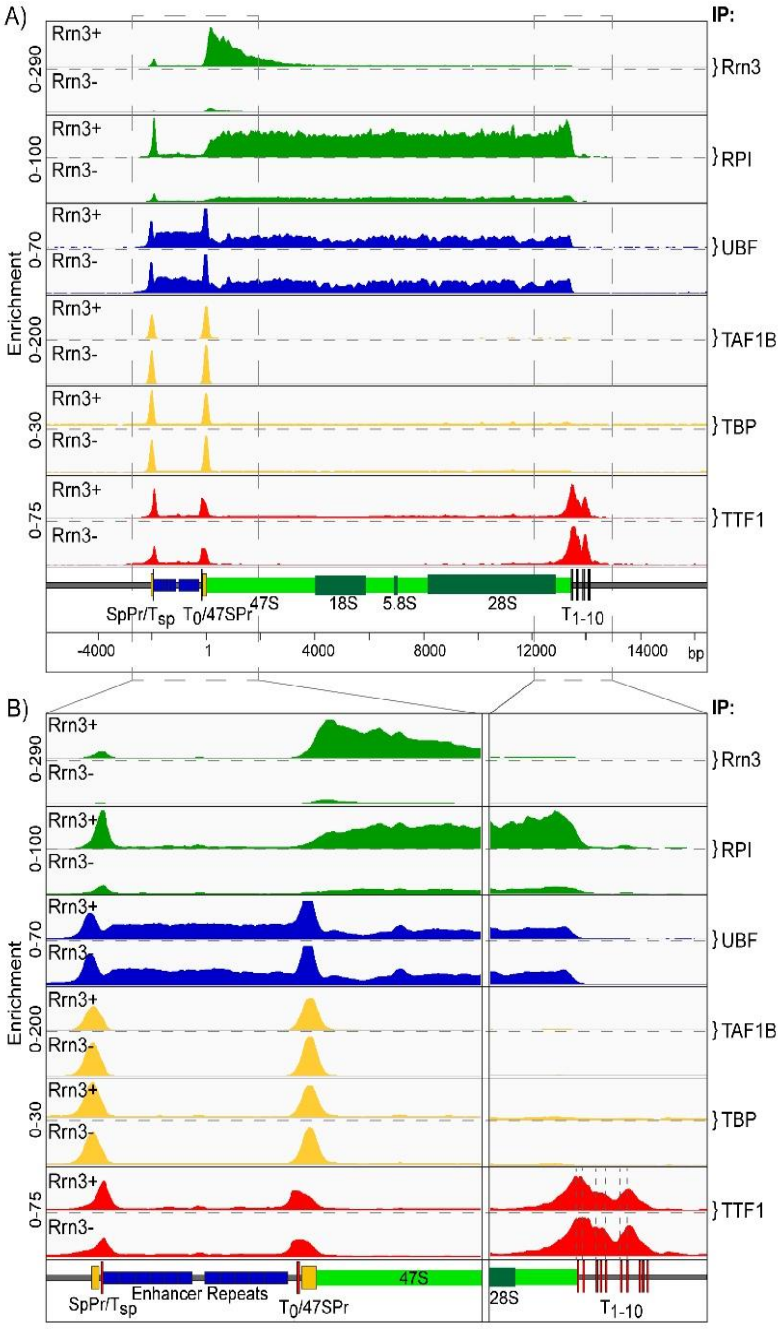

Fig 2. Conditional deletion of the Rrn3 gene inhibits RPI transcription but does not affect pre-initiation complex formation. As in Fig 1., A) and B) show the ChIP enrichment profile maps before (Rrn3+) and after (Rrn3-)

Rrn3 gene inactivation (72h 4-HT time point, S3B and S3F Fig). The binding profiles (IP) for Rrn3, RPI, UBF, the SL1 components TAF1B and TBP, and TTF1 are shown in A) over the functional mouse rRNA gene unit and, B) at higher resolution for the upstream Enhancer and Promoter elements and the 47S termination site, (boxed regions in A). As in Fig 1, a scale map of the rDNA sequence elements is given below each panel, and ChIP enrichments for each factor are given as; ChIP-Seq RPM/Input DNA RPM.

https://doi.org/10.1371/journal.pgen.1006899.g002

and the ChIP-Seq approach, we found that not only was RPI and Rrn3 recruitment eliminated by UBF loss, but also SL1 (e.g. TAF1C) binding at both Spacer and 47S promoters (Fig 3A and 3B). Further, TTF1 binding at the upstream T_{sp} and T_0 sites was suppressed and there was a small but clear shift in its binding preferences at the 47S termination region, the T_1 and T_2 sites being reduced in favour of binding at the downstream sites. This suggested that the domain of UBF binding modulated accessibility of TTF1 to the T_1 and T_2 sites. We also consistently observed a significant but low level of TTF recruitment throughout the 47S transcribed region that was significantly reduced on inactivation of RPI transcription whether by loss of Rrn3 or UBF (S4 Fig). This suggested a cycling of TTF1 between the upstream and downstream sites that would be consistent with its rapid dynamics noted previously by Fluorescence Recovery After Photo-bleaching (FRAP) [52]

UBF determines psoralen accessibility and nucleosome exclusion

Chromatin accessibility to psoralen cross-linking has long been used to differentiate active and inactive rRNA genes, a technique believed to be based on the absence or presence of nucleosomes within the 47S transcribed region [53]. The psoralen technique revealed that in the $p53^{-/-}$ MEFs 64 ± 4% of rDNA migrated in the low mobility (psoralen hyper-accessible) active (a) fraction (Fig 4A and 4B). Suppression of RPI transcription by 4-HT induced inactivation of the UBF gene in the $Ubf^{fl/fl}/ER\text{-}Cre^{+/+}/p53^{-/-}$ MEFs eliminated this low mobility band and enhanced the higher mobility band of the inactive genes (Fig 4A), consistent with our previous data [35, 54]. In contrast, suppression of transcription by inactivation of the Rrn3 gene in the $Rrn3^{fl/fl}/ER\text{-}Cre^{+/+}/p53^{-/-}$ MEFs did not affect the psoralen crosslinking pattern (Fig 4B). Thus, the enhanced psoralen accessibility of the active rRNA genes was independent of RPI recruitment or active transcription, and corresponded with the binding of UBF. This is in agreement with data from yeast where the probable UBF ortholog Hmo1 is also sufficient for enhanced psoralen accessibility [55]. It is also an important point when interpreting previous data, since psoralen accessibility can no longer be used as an indicator of active RPI transcription. As will be argued below it is more likely an indicator of the absence of core histones, as was originally suggested [53], and their replacement by UBF.

We further asked if loss of UBF also led to the reformation of nucleosomal chromatin on the previously active rRNA genes. As expected, the rDNA IGS in both conditional and wild type MEFs displayed a "ladder" of Micrococcal Nuclease (MNase) inter-nucleosomal cleavage characteristic of nucleosomal chromatin (S5B and S5C Fig). In contrast, the 47S region displayed a near continuum of MNase cleavage typical of nucleosome-free and actively transcribed DNA. Loss of UBF in $Ubf^{fl/fl}/ER\text{-}Cre^{+/+}/p53^{-/-}$ MEFs (72h post 4-HT) led to the establishment of a nucleosomal cleavage ladder in the 47S region, which now resembled the IGS (Figs 4D and S5B). However, loss of Rrn3 ($Rrn3^{fl/fl}/ER\text{-}Cre^{+/+}/p53^{-/-}$ MEFs, 72h post 4-HT) did not have this effect, the 47S region remaining non-nucleosomal (Figs 4E and S5C). Taken together with the psoralen accessibility analysis (and DNase-Seq, see below), these data showed that UBF was sufficient to exclude nucleosomes from the 47S region of the rDNA. We previously argued that the "Enhancesome" UBF-DNA nucleoprotein structure is incompatible with nucleosomes [56]. However, as discussed above, the low binding constant and the high *in vivo*

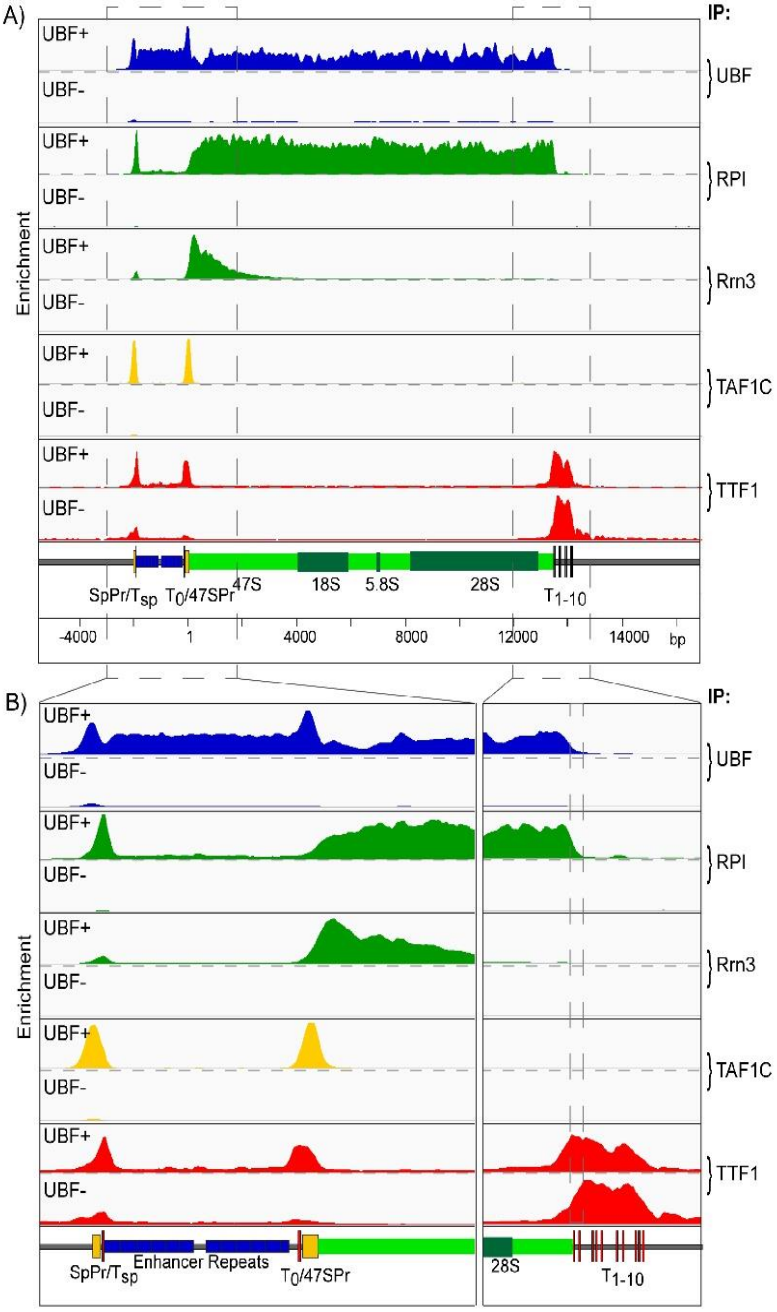

Fig 3. Conditional deletion of the UBF gene not only ablates RPI transcription but also prevents preinitiation complex formation. A) and B) show the ChIP enrichment profile maps before (UBF+) and after (UBF-) UBF gene inactivation (72h 4-HT time point, S3E and S3F Fig). The binding profiles (IP) for UBF RPI, Rrn3, the SL1 component

TAF1C, and TTF1 are shown; A) over the functional mouse rRNA gene unit and, B) at higher resolution for the upstream Enhancer and Promoter elements and the 47S termination site. As in Fig 1, a scale map of the rDNA sequence elements is given below each panel, and ChIP enrichments for each factor are given as; ChIP-Seq RPM/Input DNA RPM.

https://doi.org/10.1371/journal.pgen.1006899.g003

off-rate both argue that UBF alone would not be able to prevent the incursion of nucleosomes. Rather, the data, suggest the existence of a specific, transcription-independent mechanism for UBF deposition, one perhaps involving chromatin modifying and remodelling complexes [57–59].

Rrn3-loss causes partial rDNA chromatin collapse but does not disrupt the nucleolus

We previously demonstrated that deletion of the UBF gene led to the complete disassembly of the nucleolus [35, 60]. In the absence of UBF, the RPI transcription machinery and at least one early rRNA processing factor were shown to form a compact somatic nucleolar body (e.g. see Fig 5A), and the rDNA loci shown to scattered throughout the nucleus, suggesting that they collapsed back onto their chromosomal loci, see [35]. When we performed a similar analysis after Rrn3 deletion, we found quite a different nucleolar behavior (Fig 5B). Before deletion, UBF and fibrillarin displayed the expected speckled nucleolar pattern that coincided with the sites of rRNA synthesis (EU incorporation). After Rrn3 deletion, rRNA synthesis was ablated (loss of EU) and UBF and RPI both collapsed into denser, more discrete foci. But, these foci remained immediately proximal to fibrillarin foci and their spatial distribution and number were consistent with the prior sites of active nucleoli. Since under these conditions the rRNA genes remained bound by UBF and SL1 (TAF1s/TBP) (Fig 2), nucleosome-free (Fig 4) and therefore potentially active, the data suggested that the rDNA remained associated with remnant nucleolar structures, and had simply contracted to the heterochromatic edges of the nucleoli, much as seen for transcription inhibition by Actinomycin D [61].

UBF may replace core histones on the active rDNA

While our data had suggested that the rDNA IGS was nucleosomal, the 47S region of active genes lacked a nucleosomal cleavage pattern, suggesting a highly disorganized chromatin structure (Figs 4D & 4E and S5B & S5C). However, it was estimated from psoralen crosslinking that 64 ± 4% of the rDNA in MEFs was active and 35 ± 4% inactive (Fig 4A & 4B). The inactive repeats might be expected to be nucleosomal throughout, much as we observed after UBF loss (Fig 4D). Consistent with this, histone H3 and the heterochromatic marker H3K9me3 were detected throughout the rDNA, but both were reduced to roughly 40% of the IGS value over the Enhancer and 47S gene regions (Fig 6A). The data therefore suggested that H3 and H3K9me3 were predominantly absent from the Enhancer and 47S gene regions of active genes, but occupied the full rDNA repeat of inactive genes as well as the IGS of the active genes. This suggested that the gene bodies of the active rDNA may be histone-free and instead occupied by UBF. This situation would be reminiscent of the chromatin status of the yeast rDNA, where Hmo1 replaces the histones on active genes [62].

The domain of UBF recruitment is delineated by a unique upstream Enhancer Boundary Complex

Despite around 60% of the rRNA genes being active in the MEFs, our ChIP-Seq normalization procedure revealed a complete lack of active or potentially active chromatin marks across the

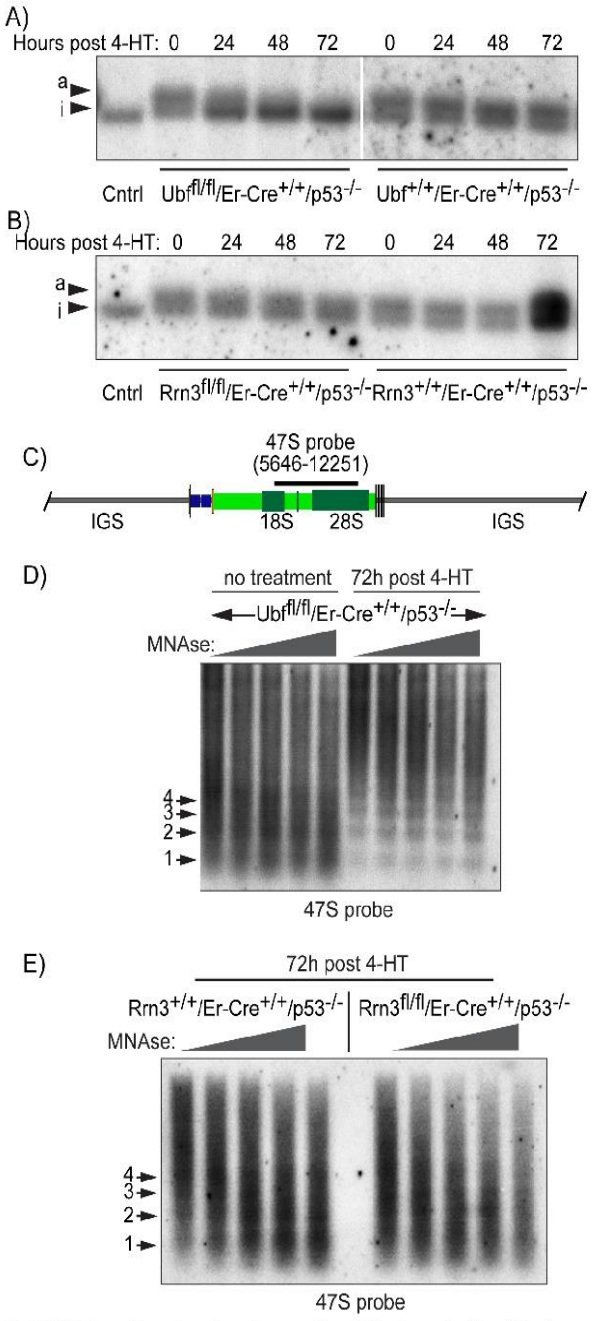

Fig 4. UBF determines the activated state of the rDNA chromatin. A) and B) show psoralen accessibility analyses during the time course of UBF and Rrn3 gene deletion (see 4HT time course S3 Fig). The high accessibility so-called active "a" gene fraction, and the low accessibility inactive. "i" gene fraction were

detected using the 47S probe. C) The mapped position of the hybridization probe relative to the start of the 47S rRNA is indicated above a diagram of the rDNA repeat. D) The profiles of increasing MNase cleavage of chromatin from UBF conditional MEFs either before or after (72h post 4-HT) inactivation of the UBF gene. E) The profiles of increasing MNase cleavage of chromatin from Rrn3 wild type and conditional (floxed) MEFs after 72h of treatment with 4-HT. The data in D) and E) was obtained using the 47S hybridization probe shown in C), and the positions of mono- (1), di- (2), etc nucleosomes are indicated. More complete datasets of the MNase analyses are given in S5 Fig.

https://doi.org/10.1371/journal.pgen.1006899.g004

Spacer and 47S promoters, the Enhancer repeats, the 47S gene body and all but a very small region of the IGS (Fig 6A). This was very surprising considering the extensive literature suggesting active histone modifications and nucleosome sliding at the 47S promoter regulate its

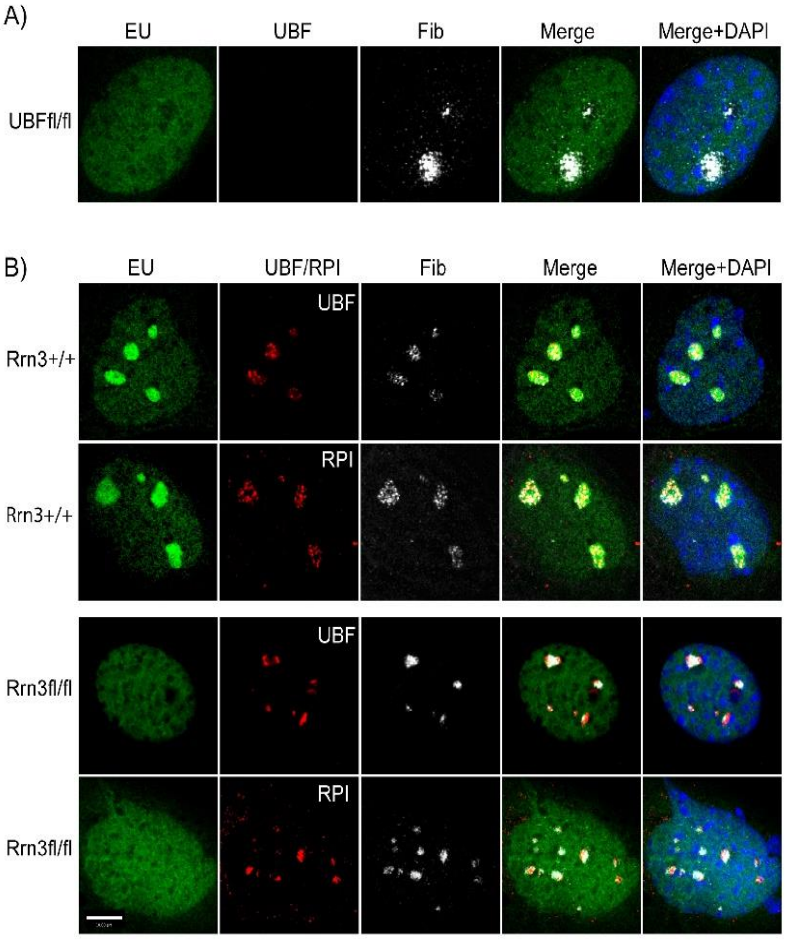

Fig 5. Comparative nucleolar structures before and after UBF or Rrn3 gene inactivation. A) In situ RNA labeling (EU) and UBF and fibrillarin (Fib) staining of $Ubf^{fl/fl}/ER\text{-}Cre^{+/+}/p53^{-/-}$ MEFs 72h post 4-HT treatment. B) In situ RNA labeling (EU), and fibrillarin (Fib) and UBF or RPI staining of $Rrn3^{wt/wt}/ER\text{-}Cre^{+/+}/p53^{-/-}$ and $Rrn3^{fl/fl}/ER\text{-}Cre^{+/+}/p53^{-/-}$ MEFs 72h post 4-HT treatment. Cultures were also counterstained with DAPI.

https://doi.org/10.1371/journal.pgen.1006899.g005

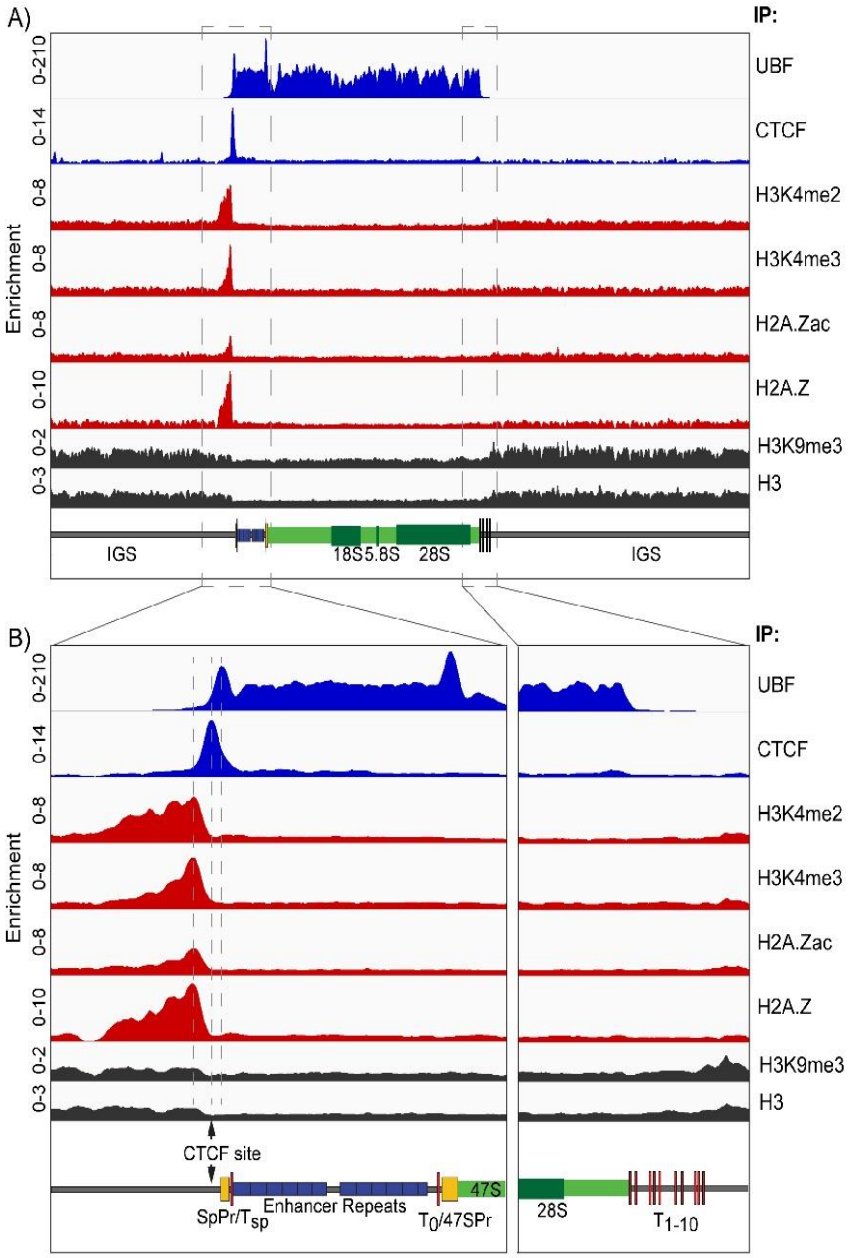

Fig 6. The UBF binding domain is delineated by an Enhancer Boundary Complex. A) The ChIP enrichment profile maps for UBF, CTCF, H3K4me2, H3K4me3, H2A.Zac, H2A.Z, H3K9me3 and H3 across the full rDNA repeat unit. B) Higher resolution maps for the upstream Enhancer and Promoter elements and the 47S termination site (boxed in A). The position of the predicted CTCF binding site [67] is indicated. As in Fig 1, a scale map of the rDNA sequence elements is given below each panel, and ChIP enrichments for each factor are given as; ChIP-Seq RPM/Input DNA RPM.

https://doi.org/10.1371/journal.pgen.1006899.g006

activity, reviewed in [63–65]. The only exception was a site immediately 5' of the Spacer Promoter at which we observed strong and overlapping occupancy of H3K4me2, H3K4me3, H2A.Z and H2A.Zac (K4ac/K7ac/K11ac) [66]. Our realignment of publicly available ChIP-Seq data also revealed unique and overlapping peaks of H3K4me1/me2/me3, H3K36me3, H3K9ac and H3K27ac (ENCODE GSE32218) at the same site in mouse embryonic stem cells (mESC), (S6 Fig), and a similar observation for H3K4me2/me3 was made by Zentner et al. [34] by realignment of public data sets (GSE11172, GSE8024) also from mESCs. This concentration of activating marks was flanked immediately 3' by a unique site occupied by the genome architecture and boundary CCCTC-binding factor CTCF [67, 68], which in turn immediately flanked the upstream boundary of UBF binding (Fig 6A & 6B). Thus, the active Enhancer and 47S transcribed regions of the rDNA repeats were flanked on their 5' side by a unique chromatin boundary complex that contained multiple marks of active and poised chromatin. The lack of active chromatin marks within the functional rRNA gene is consistent with a near total lack of core histone in this UBF-bound region. As the unique exception to this, the Enhancer Boundary Complex appeared likely to be a key structure in the control of gene activity.

The Enhancer Boundary Complex is maintained in the absence of UBF

We found that the Enhancer and 47S regions of the rDNA were highly sensitive to DNase I, again suggesting a very open structure consistent with UBF binding. This hypersensitivity was eliminated when UBF was deleted (Fig 7A), and corresponded with the loss of psoralen accessibility and the formation of a nucleosome ladder (Fig 4). In contrast, the site occupied by the Enhancer Boundary Complex remained DNase hypersensitive even after UBF deletion. We therefore asked if the components of this Complex were also maintained.

CTCF binding to the unique Spacer Promoter adjacent site was predominantly maintained or somewhat enhanced after Rrn3 or UBF loss and H2A.Z binding immediately upstream of this site was unaffected by the loss of UBF (S7 Fig and Fig 7B). Surprisingly, the H3K4me3 active chromatin mark was not only maintained, but was enhanced by the loss of either Rrn3 or UBF (Fig 7B and S7 Fig), and H2A.Zac was also significantly enhanced after UBF loss (Fig 7). These active chromatin marks followed a 200bp repeat pattern that was reiterated in the DNase-Seq pattern of fragment release, suggesting that they lay within three or four phased nucleosomes immediately adjacent to the CTCF site (indicated by asterisks in Fig 7B, see also Fig 8B). These phased nucleosomes clearly displayed reducing degrees of H3K4me3 and H2A.Zac modification with increasing distance upstream from CTCF and from the 5' UBF boundary.

Cohesin is also recruited to the Enhancer Boundary Complex

CTCF is known to recruit the Cohesin complex via it C-terminal domain, and indeed in many cases Cohesin is required for the maintenance of CTCF binding [68]. Our ChIP-Seq analyses revealed the presence of the Cohesin subunits Smc1 and Rad21 exactly overlapping the peak of CTCF binding (Fig 8A). As for CTCF and the active histone marks, Cohesin recruitment was in predominantly independent of UBF and hence also of gene activity. Given the stability of not only CTCF and Cohesin at the Enhancer boundary through many hours of complete gene inactivation, but also of the maintenance of the active chromatin marks despite the reestablishment of surrounding repressive nucleosomal chromatin (see again Figs 4 and 7), we suggest that the formation of the Enhancer Boundary Complex must be an early event in rDNA activation. In support of this, realignment of public ChIP-Seq data further revealed that the Enhancer Boundary is the major site of binding for the central SWI/SNF chromatin remodeller subunit Snf2h/SMARC5 and the histone acetyltransferase (HAT) CPB/p300, known to

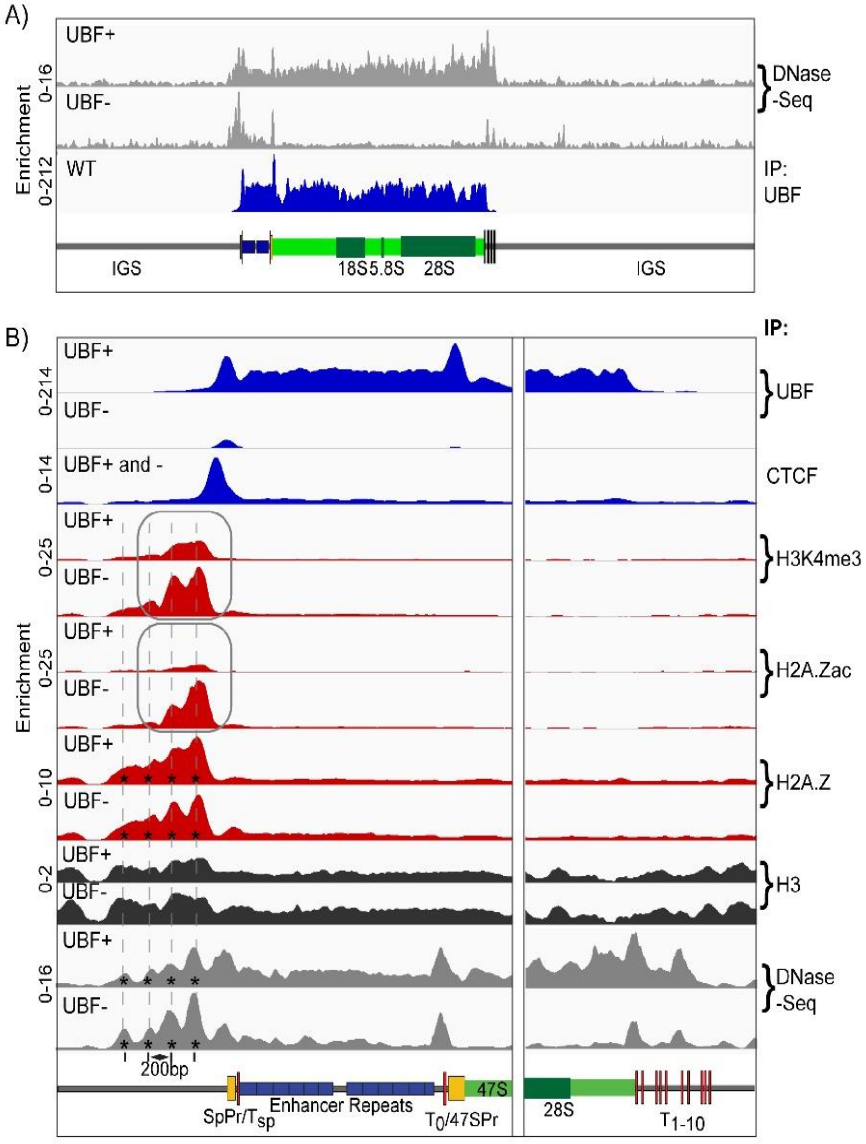

Fig 7. The Enhancer Boundary Complex exists independently of UBF and rDNA activity. A) DNase-Seq analysis of *Ubf*$^{wt/wt}$/*ER-Cre*$^{+/+}$/*p53*$^{-/-}$ (UBF+) and *Ubf*$^{fl/fl}$/*ER-Cre*$^{+/+}$/*p53*$^{-/-}$ (UBF-) MEFs 72h post 4-HT treatment across the full rDNA repeat unit as compared to the UBF ChIP-Seq profile of *Ubf*$^{wt/wt}$/*ER-Cre*$^{+/+}$/*p53*$^{-/-}$ (WT) MEFs also 72h post 4-HT treatment. B) Higher resolution ChIP enrichment profile maps for UBF, CTCF, H3K4me3, H2A.Zac, H2A.Z and H3 across the upstream Enhancer and Promoter elements and the downstream 47S termination site of *Ubf*$^{fl/fl}$/*ER-Cre*$^{+/+}$/*p53*$^{-/-}$ MEFs either untreated (UBF+) or 72h post 4-HT (UBF-). Binding of CTCF (UBF+ and -) was unaffected by UBF loss. Higher resolution DNase-Seq profiles are also given in panel B) as are the positions of phased nucleosomes (*). Scale maps of the rDNA sequence elements are provided below each panel. Enrichments for each track are given as; Sample DNA RPM/Input DNA RPM.

https://doi.org/10.1371/journal.pgen.1006899.g007

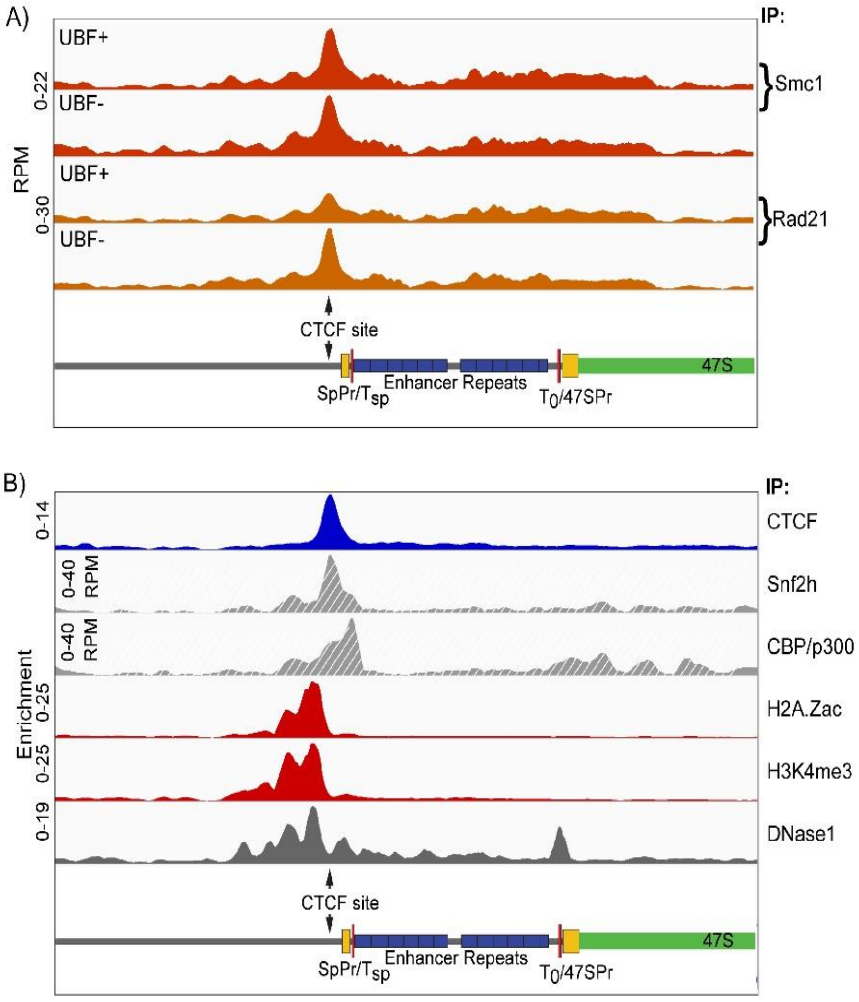

Fig 8. Cohesin subunits and chromatin remodelers also map to the Enhancer Boundary. A) Distribution of Cohesin subunits Smc1 and Rad21 across the rDNA repeat unit of *Ubf*$^{fl/fl}$*/ER-Cre*$^{+/+}$*/p53*$^{-/-}$ MEFs either untreated (UBF+) or 72h post 4-HT (UBF-). B) Comparative distribution of CTCF, H2A.Zac, H3K4me3 and DNase1 cleavage across the mouse rDNA from the present study and Snf2h (GSE53583) [91] and CPB/p300 (GSE54453) [92]. Scale maps of the rDNA sequence elements are provided below each panel. Enrichments for each track are given either as; Sample DNA RPM/Input DNA RPM, or directly in RPM as indicated.

https://doi.org/10.1371/journal.pgen.1006899.g008

directly bind and acetylate UBF [57] (Fig 8B). Thus, the assembly of the Enhancer Boundary Complex is likely to be the dominant determinant of rDNA activity.

Discussion

Using a novel ChIP-Seq normalization protocol we have been able to generate *in vivo* high-resolution maps of RPI transcription factor and chromatin interactions across the full mouse

rDNA repeat in both the presence and absence of the two key regulatory factors Rrn3 and UBF. The level of mapping detail provides an in-depth understanding of active rDNA chromatin, Spacer Promoter function, RPI transcription and termination, and the interdependence of basal transcription factors. The data further reveal the existence of a unique Enhancer Boundary Complex that marks the rDNA repeats even in the complete absence of active transcription and the basal factors, and suggest that this complex could be a key entry site for factors that drive and/or maintain rDNA activation.

We found that UBF is present throughout the 47S gene body and across the Spacer Promoter and Enhancer repeats of active genes. The UBF binding domain is immediately flanked on the upstream side by a unique Enhancer Boundary Complex of CTCF and Cohesin and then by nucleosomal chromatin. Three or four phased nucleosomes lie adjacent to CTCF/Cohesin and form the sole site in the rDNA marked by active chromatin modifications (H3K4me2/3, H2A.Z, H2A.Zac and probably H3K9ac, H3K27ac and H3K36me3), the rest of the rDNA repeat including both RPI promoters and the UBF-bound domain completely lacking active histone marks. On the downstream side, the UBF domain stops abruptly at the first two TTF1-bound termination sites, and these are followed further downstream by nucleosomal chromatin. These data are summarized in Fig 9, and we will return to consider their

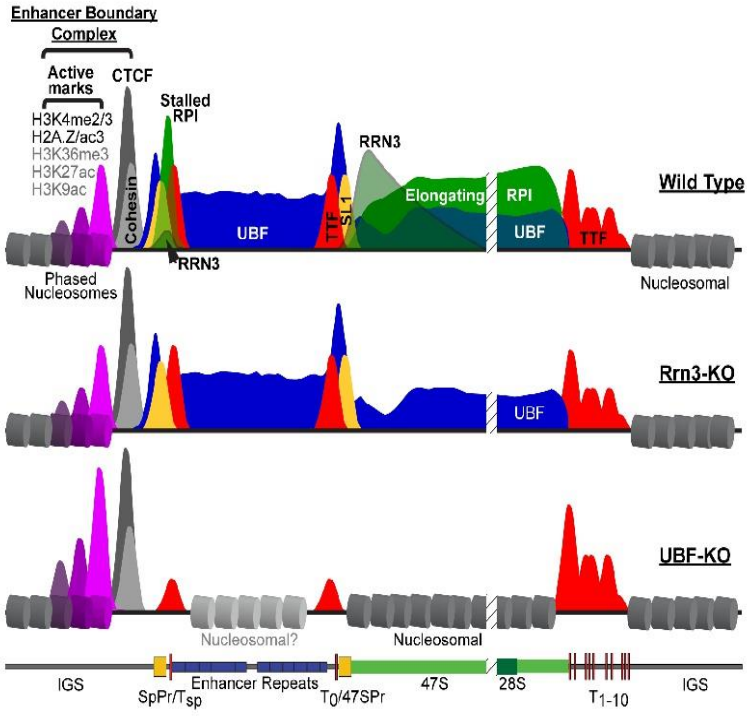

Fig 9. Diagrammatic summary of factor binding and chromatin structure across the active mouse rDNA repeat (Wild Type) and after deletion of Rrn3 (Rrn3-KO) or UBF (UBF-KO). UBF is indicated in blue, RPI in green, Rrn3 in grey/green, TTF1 in red, SL1 in yellow, CTCF in dark grey and Cohesin in light grey. The CTCF adjacent phased nucleosomes are indicated in magenta and their degree of modification indicated by the height of corresponding peaks.

https://doi.org/10.1371/journal.pgen.1006899.g009

importance after first discussing the mapping of the RPI transcription machinery and the effects of Rrn3 and UBF deletion on rDNA activity and chromatin.

Clear peaks of UBF binding are found overlapping the mapped 47S and Spacer Promoter sequences, and correspond to sites bound by the SL1 preinitiation complex. RPI maps throughout the 47S gene body, but association of Rrn3 with RPI reduces exponentially starting from the site of initiation, and is consistent with stochastic release of Rrn3 from the elongation complex. The RPI termination factor TTF1 maps to the $T_{1\text{-}10}$ termination sites with a preference for the T_1 and T_2 sites, but also maps to the predicted 47S Promoter and Spacer Promoter adjacent T_0 and T_{sp} sites. Interaction of RPI with the 47S gene body ends abruptly at the $T_{1\text{-}10}$ cluster of sites, and no RPI ($\leq$ 2%) is detected further downstream. This suggests that TTF1 bound at $T_{1\text{-}10}$ is sufficient to arrest the RPI elongation complex and promote its disruption, in agreement with recent findings in yeast [69]. RPI elongation complexes display a very uniform profile across the 47S gene body (Fig 1A & 1B, and "Elongating RPI" in Fig 9), excluding significant sequence-specific pausing. In contrast, RPI elongation complexes initiating at the Spacer Promoter are arrested at the adjacent Tsp site, most probably by TTF1 (Figs 1D and 9). It has been proposed that long non-coding RNAs (lncRNAs) initiated from the Spacer Promoter are processed to generate the pRNA required for rDNA silencing [16, 17, 70, 71]. Thus, TTF1 bound at the Tsp site controls synthesis of the pRNA precursor and hence the potential for rDNA silencing. Consistent with this, TTF1 abundance and subnuclear localization were previously shown to regulate rDNA activity [52, 72].

Normalized ChIP-Seq in conditional MEFs showed loss of Rrn3 blocked RPI initiation at both 47S and Spacer Promoters and prevented synthesis of the 47S rRNA. It did not, however, have any effect on SL1 preinitiation complex formation at either promoter. Further, Rrn3-loss did not affect the profile of UBF binding, nor psoralen accessibility of the rDNA. Thus, in agreement with data from yeast [55], RPI transcription appears not to be required to maintain UBF binding and the psoralen accessible status of the rDNA in MEFs. The data also infer that elongating RPI must be able to traverse the UBF-DNA complex without displacing it, and is consistent with the role of ERK and CBP modification of UBF in regulating RPI passage [18, 23, 57, 73]. As spin-off from our Rrn3-inactivation study, we were able to allay doubts as to the essential nature of the mouse *Rrn3(Tif1a)* gene (S2 Fig).

In contrast to Rrn3-loss, deletion of UBF eliminated not only all interaction of RPI and Rrn3 with the rDNA, but also SL1 preinitiation complex formation at both 47S and Spacer promoters. Thus, as predicted from the earliest in vitro studies [74], SL1 recruitment depends on the presence of UBF. However, our data do not speak to any cooperativity between UBF and SL1 recruitment as has also been suggested [75], and this must await the results of ongoing studies of SL1 inactivation. UBF, but not Rrn3 deletion, reduced binding of TTF1 to its promoter associated sites and increased binding at the termination sites. Further, both Rrn3 and UBF loss affected the low level TTF1 interaction throughout the 47S gene body, suggesting a dynamic RPI-dependent cycling between these sites.

Psoralen accessibility is used to distinguish active and inactive rDNA repeats [53], and has long been assumed to detect active RPI transcription. However, we found that deletion of Rrn3 had no effect on psoralen accessibility, while UBF deletion eliminated it. Thus, psoralen accessibility is clearly a property of UBF bound rDNA and not an indicator of active RPI transcription. Consistent with this, the UBF domain of the rDNA was DNase hypersensitive and lacked nucleosomal chromatin, while deletion of UBF led to the reformation of nucleosomes and the loss of this DNase hypersensitivity. Our data also suggest that not only nucleosomes, but also histones may be absent from the regions bound by UBF, potentially explaining the complete lack of activating histone modifications over the rRNA gene unit.

Strikingly, the Enhancer Boundary Complex, consisting of CTCF, Cohesin and the adjacent phased nucleosomes, was not perturbed by UBF deletion and ChIP-QPCR analyses showed this was also the case for Rrn3 deletion. (summarized in Fig 9). This boundary complex was all the more remarkable for being the sole site of activating histone modifications (H3K4me2 and H3K4me3, H2A.Z and H2A.Zac, and probably H3K36me3, H3K9ac and H3K27ac) in the whole rDNA repeat, and showed itself to be highly stable, remaining many hours after UBF deletion and the release of all RPI transcription factors. Indeed, the associated H3K4me3 and H2A.Zac modifications were even enhanced by UBF deletion. The Enhancer Boundary Complex, therefore, represents a stable marker of the potentially active or poised rDNA fraction.

The Enhancer Boundary Complex also displays a striking asymmetry, on its upstream side CTCF being flanked by phased nucleosomes containing the near full range of activating histone modifications and the H2A.Z histone variant, and on its downstream side by an a SL1/UBF PIC and a transcriptionally engaged RPI complex. CTCF binding sites are in themselves asymmetric, and this is believed to enable 180deg. DNA looping-out via CTCF interactions between distant palindromic sites [68] or the formation of 360deg. circular loops by interactions between tandemly oriented sites, e.g. see [76]. It was previously noted that the unique CTCF site in the mouse rDNA is oriented away from the gene body [67]. Thus, CTCF interactions between adjacent gene repeats would be unlikely to induce 180deg. looping-out unless they involved heterologous interactions, e.g. with TTF1 [42]. They would, however, permit 360deg. circular looping analogous to that suggested for the SL1 complex [77], though at the scale of the whole rDNA repeat. Alternatively, it was noted that in human the tandem repeated rDNA loci are interspersed with inverted, often incomplete, rDNA units [78]. If such aberrant repeats are present in mouse, by presenting inverted CTCF sites, they might induce the formation of 180deg. loops containing multiple rDNA units. In any of these scenarios, the role of Cohesin at the site of the Enhancer Boundary Complex would be expected to stabilize looping via this site [68, 76], and could play an important role in defining active rDNA arrays, and/or controlling inter-repeat recombination. Indeed, data from yeast and human has already suggested Cohesin plays an important role in the activity and stability of the rDNA arrays [79].

The observation that no active histone marks occur within the mouse rRNA gene body, the RPI promoters or the Enhancers conflicts with several studies suggesting an importance of histone modification, particularly at the 47S Promoter, in activating gene transcription, see [63] for review. It is, however, difficult to imagine how, as has been suggested, nucleosomes could co-exist with the preinitiation complex and UBF at the 47S promoter of actively transcribed genes. Indeed, the data from yeast has shown that little or no histone can be detected at the 35S rRNA promoter [62] and so is in full agreement with our findings. Further, public ChIP-Seq data also failed to detect significant active histone modifications over the functional rRNA gene. Together the data suggest that active histone modifications could only play a transient role, perhaps during the displacement nucleosomes and their replacement by UBF. As we have shown, once it is recruited UBF remains stably bound across the functional rRNA gene even in the absence of active transcription, leaving little chance for histone modifications to play a regulatory role.

We previously showed that the rDNA exists in not just two, but three distinct states, heterochromatin and CpG methylated, a poised state defined at the time by low psoralen accessibility but completely lacking DNA methylation and an actively transcribed state defined by high psoralen accessibility [80, 81]. Since we now show that the psoralen assay classifies active and inactive genes based solely on UBF-binding, we can suggest a refinement to our definition of the poised rDNA state as one that is nucleosomal throughout, but is marked by the Enhancer Boundary Complex. Preliminary data suggest that this boundary complex is also present on the human rDNA. We suggest that the Enhancer Boundary Complex is the initial site from

which rDNA activity is induced. This is supported by the presence of the ATPase chromatin remodeling factor Snf2h (SMARCA5) and probably the acetyltransferase CBP/p300 at this boundary. It is also in agreement with data showing that loss of CTCF reduces the recruitment of UBF throughout the gene body as well as the recruitment of RPI to the Spacer Promoter [67]. These authors also showed that binding of CTCF to the Spacer Promoter adjacent site occurs predominantly on the unmethylated and so potentially active rDNA fraction. Taken together, the data strongly suggest a model for the active rDNA repeats in which UBF replaces histone chromatin from the Spacer Promoter through to the TTF1 bound $T_{1\text{-}10}$ termination sites (Fig 9), and that this situation is established or maintained by chromatin modifying activities emanating from the upstream Enhancer Boundary Complex.

Materials and methods

Isolation and culturing of MEFs

The generation of conditional *Ubf*$^{fl/fl}$*ER-Cre*$^{+/+}$ and control mouse lines was previously described (Hamdane et al. 2014). *Ubf*$^{fl/fl}$*/ER-Cre*$^{+/+}$*/p53*$^{-/-}$ and *Ubf*$^{+/+}$*/ER-Cre*$^{+/+}$*/p53*$^{-/-}$ isogenic mice were generated by introducing the p53-null allele from strain *129-Trp53*^{tm1Tyj}*/J* (Jackson Laboratory #002080) [82]. To generate *Rrn3*$^{+/-}$, *Rrn3*$^{fl/fl}$*/ER-Cre*$^{+/+}$*/p53*$^{-/-}$ and isogenic *Rrn3*$^{+/+}$ *ER-Cre*$^{+/+}$*/p53*$^{-/-}$ mice, the following mouse strains from Jackson Laboratory were used; B6.Cg-Tg(Sox2-cre)1Amc/J (#008454), B6;129-*Gt(ROSA)26Sor*$^{tm1(cre/ERT)Nat}$/J (#004847) and B6.Cg-Tg(UBC-cre/ERT2)1Ejb/2J (#008085). C57BL/6NCrl wild-type mice for backcrossing were from Charles River. Primary mouse embryonic fibroblasts (MEFs) were generated from E14.5 *Ubf*$^{fl/fl}$*/ER-Cre*$^{+/+}$*/p53*$^{-/-}$, *Rrn3*$^{fl/fl}$*/ER-Cre*$^{+/+}$*/p53*$^{-/-}$, and equivalent *Ubf*$^{+/+}$ and *Rrn3*$^{+/+}$ embryos as previously described [35, 83]. MEFs were cultured in Dulbecco's modified Eagle medium (DMEM)-high glucose (Life Technologies), supplemented with 10% fetal bovine serum (Wisent), L-glutamine (Life Technologies) and Antibiotic/Antimycotic (Wisent).

Embryo collection and genotyping

Heterozygous *Rrn3*$^{+/-}$ mice were inter-crossed and embryos isolated from pregnant females at 3.5, 6.5, 7.5, 8.5, 9.5 and 10.5dpc. DNA from 3.5 dpc embryos was amplified using the REPLI-g Mini kit (QIAGEN). Genotyping on DNA from all embryo stages was performed by PCR using the same primers as for cell lines, see below.

Inactivation of *Ubf* or of *Rrn3* in cell culture, and analysis of genotype, rRNA synthesis and proteins

As previously described (Hamdane et al. 2014), cells were initially plated in 6 cm petri dishes ($0.8x10^6$ cells each) and cultured for 18 hours in DMEM, high glucose, 10% fetal bovine serum. To initiate *Ubf* inactivation, 4-hydroxytamoxifen (4-HT) was added to both *Ubf*$^{fl/fl}$ and *Ubf*$^{+/+}$ cell cultures to a final concentration of 50nM (the 0h time point for analyses), and after 4 hr incubation the medium replaced with fresh medium without 4-HT [35]. Cultures were then subjected to analysis at various subsequent time points. In the case of *Rrn3*, cells were treated with 50nM 4-HT for 4h, passaged and diluted 1:2 and retreated 5h later with 50nM 4-HT for 15h, and this treatment protocol immediately repeated. Finally, the cells were replated at 80% confluency (the 0h time point for analyses), treated with 50nM 4-HT for 15h, and the medium replaced with fresh medium without 4-HT, before analysis at various subsequent time points. Comparative colony forming assays showed that generally less than 0.01% of *Ubf*$^{fl/fl}$ cells formed colonies after 4-HT treatment, while a few percent of *Rrn3*$^{fl/fl}$ cells were still able to form colonies and hence still carried a functional *Rrn3* allele. Analyses

systematically included genotyping and UBF and Rrn3 protein level determination. Cells were genotyped by PCR using the following primers, for *Ubf*: A; 5'TGATCCCTCCCTTTCTGATG, B; 5'TGGGGATAGGCCTTAGAGAGA, C; 5'CACGGGAAAACAAGGTCACT and for *Rrn3*: A; 5'-GATCTTAATGGAGGGCAGCA, B; 5'-TGGATCCTGCAACTTTTTCC, C; 5' TCCCAACCCTGACCTATCAC. To determine UBF or Rrn3 protein levels, cells were washed with cold phosphate buffered saline (PBS), scraped into PBS, centrifuged 2 min at 2000 r.p.m., then resuspended in SDS–polyacrylamide gel electrophoresis (SDS-PAGE) [84] loading buffer. After fractionation on 8% SDS-PAGE, cell extracts were analysed by standard Western blotting procedures. Metabolic labelling of RNA was carried out immediately before cell harvesting by the addition of 10 μCi [^{3}H]-uridine (PerkinElmer) to the culture medium and incubation for a further 3h. RNA was extracted using Trizol (Life Technologies) according to the manufacturer's protocol, analyzed by gel electrophoresis and fluoroimaging (Enhance, PerkinElmer) and RNA species quantitated by scintillation counting as previously described [18, 48, 73].

Colony forming assays

Rrn3$^{fl/fl}$-, and *Ubf*$^{fl/fl}$/*ER-Cre*$^{+/+}$/*p53*$^{-/-}$ and matched *Rrn3*$^{+/+}$-, and *Ubf*$^{+/+}$/*ER-Cre*$^{+/+}$/*p53*$^{-/-}$ MEFs were treated with 4-HT as described above, and 48h later each culture was replated in duplicate at dilutions of 10 000, 50 000, 100 000 cells per 60 mm petri. 144h post 4-HT treatment, petri dishes were fixed for 5 mins with 4% paraformaldehyde/PBS and stained with 0.05% crystal violet in distilled water (filtered) for 30 mins. Dishes were then rinsed 3 times with water and left inverted to dry. Crystal violet staining was quantified by adding 1ml of methanol to each dish to solubilize the dye, the methanol recovered and its optical density at 540nm determined after appropriate dilution.

Antibodies for western blot, immunofluorescence and ChIP

Rabbit polyclonal antibodies against mouse UBF, RPI large subunit (RPA194/Polr1A), TTF1, TAF1B (TAF68) and RRN3 were generated in the laboratory. The UBF, RPI and TTF1 antibodies have been previously described [52]. Rabbit antisera were raised against TAF68 aa 54–175 and RRN3 aa 464–656, expressed in E. coli and peptides were purified using the guanidinium chloride-urea denaturation method. Rabbit antibody against TAF1C (TAF95) was a gift from Ingrid Grummt. Anti-H2A.Z, anti-H2A.Zac and ant-H3K4me2 were gifts from Colyn Crane-Robinson. All other antibodies were obtained commercially; TBP (#ab818 Abcam), anti-Tubulin (#T5168 Sigma), anti-Fibrillarin (#905001 BioLegend), anti-CTCF (#07–729 Millipore), anti-SMC1 (#A300-055A Bethyl), anti-Rad21 (#ab992 Abcam) anti-H3K4me3 (#ab1012 Abcam), anti-H2A.Z (#ab4174 Abcam), anti-H3K9me3 (#17–625 Millipore), anti-H3 (#ab1791 Abcam).

Chromatin immunoprecipitation (ChIP)

Cells were fixed with 1% formaldehyde for 8 min at room temperature. Nuclei were isolated using Lysis Buffer (10 mM Tris pH 7.5, 10 mM NaCl, 3 mM MgCl2, 0.5% NP-40), transferred to Sonication Buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 2 mM EGTA, 4 mM EDTA, 0.1% SDS, 1% Triton X-100, 1% NP-40) and sonicated (Bioruptor, Diagenode) for 30 cycles of 30 sec on / 30 sec off at high intensity. Each immunoprecipitation (IP) was carried out on the equivalent of 50 x 10^6 cells in IP Buffer (150 mM NaCl, 50 mM Tris-HCl pH 7.5, 5 mM EDTA, 0.5% NP-40, 1% Triton X-100) overnight at 4˚C. The antibody slurry was prepared with 50 μl A-, 50 μl G-Dynabeads and 60 μg.ml^{-1} antibody per IP. Immunoprecipitated chromatin was treated with RNaseA and the DNA isolated using 2% Na SDS and 2mg.ml^{-1} Proteinase-K. Two or more biological replicates were analyzed for each antibody.

Analysis of ChIP samples by massively parallel sequencing

ChIP DNA samples were quality controlled by qPCR before being sent for library preparation and 50 base single-end sequencing on an Illumina HiSeq 2000 (McGill University and Génome Québec Innovation Centre). For qPCR analysis, reactions (20 μl) were performed in triplicate using 2.5 μl of sample DNA, 20 pmol of each primer, and 10 μl of Quantitect SYBR Green PCR Master Mix (QIAGEN). Forty reaction cycles of 10 s at 95˚C and 30 s at 58˚C were carried out on a Multiplex 3005 Plus (Stratagene/Agilent). The amplicon coordinates relative to the 47S rRNA initiation site (BK000964v3) were as follows: IGS3, 42646–42903; SpPr, 43089–43253; Tsp, 43267–43421; 47SPr, 45133–40; 47S, 159–320; ETS, 3078–3221; 28S, 10215–10411; T1–3, 13412–13607. Data was analyzed using the MxPro software (Agilent). The relative occupancy of each factor was determined by comparison with a standard curve of amplification efficiency for each amplicon using a range of input DNA amounts and generated in parallel with each qPCR run.

DNase-Seq

DNase-Seq analysis of MEFs was carried out following the published protocol [85], the only modifications being adjustment of digestion level to obtain the required fragmentation. DNA samples were quality controlled by qPCR before size selection following the published protocol. The resultant DNA was sent for library preparation and single-end sequencing on an Illumina HiSeq 2000 (McGill University and Génome Québec Innovation Centre).

Analysis of massively parallel sequence data

The data analysis pipeline was developed to take best advantage of the sequencing depth achievable across the rDNA. We noted that sequence coverage was strongly, but reproducibly biased by the underlying DNA sequences regardless of sample type and devised a simple approach to correct for this bias. Briefly, the raw sequence data from experimental and input DNA samples was checked for quality using FastQC version 0.11.4 (Babraham Bioinformatics, S. Andrews). The data was then trimmed using Trimmomatic version 0.33 [86] and the resulting quality filtered files were aligned to the mouse genome version MmGRCm38 to which a single copy of the mouse rDNA repeat sequence (GenBank BK000964v3) was added as an extra chromosome using Bowtie2 [87]. For convenience, the origin of the rDNA repeat was displaced to the EcoRI site at 30,493 such that the pre-rRNA initiation site now fell at nucleotide 14,815. Aligned reads were extended to 100bp (fragment sizes estimated using HOMER fell between 75 and 125bp [88]), the coverage calculated using BEDtools (Quinlan Lab, University of Utah), and smoothed using a window of 25 bp as:

$$J = \frac{1}{25} \bullet \sum_{n+12}^{n-12} j_n \quad (1)$$

where J = smoothed base coverage, j = aligned coverage and n = base position).

The data was then converted to reads per million (RPM) and the sample DNA coverage normalized to the input DNA coverage. The normalized sample data J_{norm} was calculated for each base position using AWK in Ubuntu as:

$$J_{norm} = {}^{J_{chip}}/_{J_{input}} \quad (2)$$

where J_{ChIP} = smoothed sample DNA coverage, and J_{input} = smoothed input DNA coverage.

The resulting normalized BED file was converted to BEDgraph format and visualized using IGV (Integrative Genomics Viewer 2.3, Broad Institute).

The ChIP-Seq and DNase-Seq data have been deposited in the ArrayExpress database at EMBL-EBI (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-5839.

Psoralen crosslinking accessibility

The psoralen crosslinking accessibility assay was performed on cells grown in 60 mm petri dishes as previously described [53, 89], using the 6.7kb 47SrRNA gene EcoRI fragment (pMr100) [53] and the 6.2 kb EcoRI IGS fragment [90]. The ratio of "active" to "inactive" genes was estimated by analyzing the intensity profile of low and high mobility bands revealed by phospho-imaging on a Fuji FLA-5100 (FUJIFILM Life Science) using a Gaussian peak fit generated with MagicPlotPro (MagicPlot Systems).

MNase digestion of chromatin

After trypsinizing and washing with PBS the cells were resuspended in buffer A (60 mM KCl, 15 mM NaCl, 15 mM TrisHCl pH 7.6) plus 0.25 M sucrose, 1 mM EDTA, 0.1 mM EGTA, 0.5 mM spermine, 0.15 mM spermidine, 0.1 mM PMSF and protease inhibitors. Triton X100 was added to final concentration 0.25% and the cells homogenized in a Dounce homogenizer using 10 piston strokes and centrifuged 5 min at 2000 rpm. The pellet was washed once by resuspending in buffer A plus 0.34M glucose and then centrifuged for 5 min at 2000 rpm. It was then resuspended in 1 ml buffer A without sucrose, recentrifuged and resuspended in buffer A containing 60 units MNase per 10^7 cells. After incubating for 2 min on ice, 1 mM $CaCl_2$ was added and incubation continued on ice. 200 μl aliquots were taken after 5, 10, 15, 20, and 30 min and the reaction was stopped by adding EDTA to final concentration of 10 mM, SDS to 0.5% and Proteinase K to 1 $mg.ml^{-1}$. After overnight incubation at 55˚C, 0.1 $mg.ml^{-1}$ RNase A and 3 units RNase T1 were added and incubated for 1 h at 37˚C. This was followed by the addition of 1 $mg.ml^{-1}$ Proteinase K for 2 h at 55˚C. The DNA was then extracted with phenol-chloroform, precipitated and dissolved in 200–400 μl TE pH8.3. 10 μl DNA was mixed with 2 x gel loading buffer containing 0.1% SDS and 0.1 $mg.ml^{-1}$ RNase A without dye. After 30 min incubation at 37˚C the samples were resolved on 1.5% agarose gel, the gel transferred onto a Biodyne B (Pall) membrane and nucleosome ladders detected by DNA hybridization using the same probes as for the psoralen analysis.

In situ immunofluorescence, EU labelling and microscopy

$Rrn3^{wt/wt}$-, $Rrn3^{fl/fl}$-, and $UBF^{fl/fl}/ER\text{-}Cre^{+/+}/p53^{-/-}$ MEFs were grown on poly-lysine coated coverslips and treated with 4-HT to induce full inactivation of floxed genes (72h time point), and gene inactivation monitored in parallel by genotyping and measurement of protein levels. 1mM 5-ethynyl uridine (EU, Molecular Probes / Thermo Fisher) was added to the culture medium and cultures incubated for 1h at 37˚C. Cells were then washed with PBS, fixed in 4% paraformaldehyde (PFA) in PBS for 15 minutes and permeabilized with 0.5% Triton in PBS for 5 minutes. EU incorporation was revealed following the manufacturers protocol (Click-iT RNA HCS, Alexa488). Cells were then incubated with the indicated primary antibodies for 1h in PBS plus 5% BSA or Maxblock (Active Motif) and subsequently with AlexaFluor 568/647 conjugated anti-rabbit or -mouse secondary antibodies (Molecular Probes / Thermo Fisher) and counterstained with DAPI. After mounting in 50% glycerol/50% 0.2 M Na-glycine, 0.3 M NaCl, image stacks were acquired using a Leica SP5 II scanning confocal microscope and formatted with Volocity software (Perkin Elmer Improvision).

Ethics statement

All animal care and animal experiments were conducted in accordance with the guidelines provided by the Canadian Council for Animal Protection, under the surveillance and authority of the institutional animal protection committees of Laval University and the CHU de Québec. The specific studies described were performed under protocol #2011–054, 2014–100, and 2014–101 examined and accepted by the "Comité de protection des animaux du CHU de Québec". This ensured that all aspects of the work were carried out following strict guidelines to ensure careful, consistent and ethical handling of mice.

Supporting information

S1 Fig. GC content of the rDNA and UBF binding. The normalized map profile of UBF binding across the full rDNA repeat unit and the percent G+C content of the rDNA calculated over 50bp non-overlapping sliding windows [93]. A scale map of the rDNA sequence elements is given below the panel.
(PDF)

S2 Fig. Deletion of the *Rrn3-Tif1a* gene arrests mouse development during early cleavage divisions. A) Genotyping of embryos and live pups from matings of *Rrn3-Tif1a*$^{+/-}$ mice either before or after (BL6) backcrossing to the C57BL/6 background. B) Example images of E3.5 embryos obtained from the matings. C) Double heterozygous *Ubf*$^{+/-}$/*Rrn3-Tif1a*$^{+/-}$ mice are both viable and generated with the expected Mendelian frequency in matings between *Ubf*$^{+/-}$ *and Rrn3-Tif1a*$^{+/-}$ mice. It was previously found that homozygous deletion of the mouse Fibrillarin (*Fbl*), RPI second largest subunit (*Rpa135/Rpo1-2/Polr1b*) or Upstream Binding Factor (*Ubf/Ubtf*) genes all cause developmental arrest during the cleavage divisions and well before the blastula stage [35, 94, 95]. This is consistent with the activation of the rRNA genes at or soon after the 2-cell stage [60, 96]. In contrast, *Rrn3-Tif1a*$^{-/-}$ mouse embryos were reported not to arrest development until E9.5 at which point they clearly displayed axis formation, tissue differentiation and the beginnings of organogenesis [45]. By E9.5 zygotic transcription would normally be expected to have increased rRNA levels over 1000 fold [46, 47]. Thus, these data suggested that Rrn3-TIF1A might either not strictly be essential or be partly redundant with some other factor. While establishing *Rrn3-Tif1a* conditional cell lines carrying the *Tif1a*^{flox} allele created by Yuan et al., we also generated mice carrying the same *Rrn3-Tif1a-null* allele studied by these authors. When progeny from the *Rrn3-TIF1A*$^{+/-}$ mice were analyzed we found that null embryos in fact arrested during the cleavage divisions as un-compacted morulae. The same result was obtained after extensive backcrossing to C57BL/6, the mouse strain used in the original publication. Since our *Rrn3-Tif1a*$^{-/+}$ mouse lines were extensively backcrossed to remove any transgenes used in recombining the *Rrn3-Tif1a*^{flox} allele, we presently have no explanation for the discrepancy with the previous study. We concluded that, despite previous data to the contrary, Rrn3-TIF1a, like UBF and RPI, is essential in mouse soon after the normal onset of rRNA gene activity. This strongly argues that the mouse Rrn3 is indeed an essential and non-redundant part of the RPI transcription machinery. It further suggests that the fertilised oocyte does not contain a significant amount of maternal Rrn3 message or protein or, as has been suggested for UBF [97], that these are subject to rapid degradation during the first few cleavage divisions.
(PDF)

S3 Fig. Functional analysis of UBF and Rrn3 inactivation in conditional MEFs. A) and D) show maps of the wild type (wt), conditional (flox) and deleted (Δ) Rrn3 and UBF gene alleles. B) and E) show typical time courses (Hours post 4-HT) of Rrn3 and UBF gene deletion

determined by PCR genotyping (Genotype), in parallel with corresponding protein levels for each factor. C) The anti-Rrn3 antibody generated in our laboratory and used for ChIP analyses revealed a single endogenous Rrn3 polypeptide that corresponded in mobility with the known Rrn3 species (Genbank XP_156394 and NP_001034610) expressed by transient transfection (Exogenous). F) Time courses of relative 47S rRNA synthesis rates in conditional and wild type Rrn3 and UBF MEFs post 4-HT treatment as determined by metabolic labeling (Materials and Methods). G) Colony forming assay for $Rrn3^{fl/fl}$-, and $UBF^{fl/fl}/ER\text{-}Cre^{+/+}/p53^{-/-}$ (*flox/flox*) and matched $Rrn3^{+/+}$-, and $UBF^{+/+}/ER\text{-}Cre^{+/+}/p53^{-/-}$ (*wt/wt*) MEFs. Cultures were standardly treated with 4-HT, replated 48h later and crystal violet staining of resulting cell colonies determined at 144h post 4-HT treatment. The assay showed that around 4% of $Rrn3^{fl/fl}$ cells, but less than 1% of $Ubf^{fl/fl}$ cells, were able to form colonies and hence retained a functional *Rrn3* or *Ubf* after the 4-HT treatment.
(PDF)

S4 Fig. TTF1 may bind at low level throughout the 47S gene body during active transcription. The normalized ChIP-Seq profiles for TTF1 before and after (72h post 4-HT) Rrn3 or UBF gene deletion and compared to the same data for RPI. The data are similar to those in Figs 2 and 3, but the vertical scale has been magnified to reveal the low level enrichments.
(PDF)

S5 Fig. UBF determines the activated state of the rDNA chromatin. A) The mapped position of the hybridization probes relative to the start of the 47S rRNA are indicated above a diagram of the rDNA repeat. B) The profiles of increasing MNase cleavage of chromatin from UBF conditional MEFs before or after (72h post 4-HT) inactivation of the gene, and C) from Rrn3 wild type and conditional (floxed) MEFs after 72h of treatment with 4-HT. The total DNA cleavage ladders were revealed by ethidium bromide (EtBr) staining, and the cleavage ladders within the IGS and the 47S gene body were revealed by hybridization with the corresponding probes shown in A). The positions of mono- (1), di- (2), etc nucleosomes are indicated.
(PDF)

S6 Fig. H3K9ac, H3K36me3 and H3K27ac histone modifications also map to the Enhancer Boundary Complex. The RPM profiles for H3K9ac, H3K36me3 and H3K27ac realigned from public data (ENCODE GSE32218) over the Promoter and Enhancer region of the mouse rDNA are shown in comparison with the CTCF and H3K4me3 enrichment profiles established in the present study.
(PDF)

S7 Fig. Comparative ChIP-QPCR analyses show Enhancer Boundary Complex components are retained after Rrn3 and UBF deletion. A) Positions of QPCR amplicons relative to rDNA sequence motifs. B) and C) ChIP-QPCR analyses at 3 days post 4HT treatment respectively for $Rrn3^{fl/fl}$- and $Rrn3^{+/+}/ER\text{-}Cre^{+/+}/p53^{-/-}$ and $Ubf^{fl/fl}$- and $Ubf^{+/+}/ER\text{-}Cre^{+/+}/p53^{-/-}$ MEFs. The data show the results for Histone H3, H3-K4me3, CTCF and either Rrn3 or UBF, the factor targeted by deletion.
(PDF)

Acknowledgments

We wish to thank Dr Ross Hannan for providing the $Rrn3\text{-}Tif1a^{fl/fl}$ mice and Drs Ingrid and Gunter Schutz for making these mice available. We also thank Drs I. Grummt, and C. Read and C. Crane-Robinson for providing antibodies respectively to TAF1C and to H2A.Z and

H2A.Zac, Dr Lucie Jeannotte for making the $p53^{+/-}$ mice available, and Drs Ross Hannan, Lucie Jeannotte and Jacques Coté for advice and helpful discussions.

Author Contributions

Conceptualization: Chelsea Herdman, Jean-Clement Mars, Victor Y. Stefanovsky, Tom Moss.

Data curation: Victor Y. Stefanovsky, Marianne Sabourin-Felix, Tom Moss.

Formal analysis: Marianne Sabourin-Felix, Helen Lindsay, Mark D. Robinson, Tom Moss.

Funding acquisition: Tom Moss.

Investigation: Chelsea Herdman, Jean-Clement Mars, Victor Y. Stefanovsky, Michel G. Tremblay.

Methodology: Chelsea Herdman, Jean-Clement Mars, Victor Y. Stefanovsky, Michel G. Tremblay, Marianne Sabourin-Felix, Tom Moss.

Project administration: Michel G. Tremblay, Tom Moss.

Resources: Chelsea Herdman, Jean-Clement Mars, Victor Y. Stefanovsky, Michel G. Tremblay, Marianne Sabourin-Felix, Helen Lindsay, Mark D. Robinson, Tom Moss.

Software: Marianne Sabourin-Felix, Helen Lindsay, Mark D. Robinson, Tom Moss.

Supervision: Tom Moss.

Validation: Chelsea Herdman, Jean-Clement Mars, Victor Y. Stefanovsky, Marianne Sabourin-Felix, Tom Moss.

Visualization: Chelsea Herdman, Jean-Clement Mars, Marianne Sabourin-Felix, Tom Moss.

Writing – original draft: Chelsea Herdman, Jean-Clement Mars, Tom Moss.

Writing – review & editing: Chelsea Herdman, Jean-Clement Mars, Victor Y. Stefanovsky, Marianne Sabourin-Felix, Helen Lindsay, Mark D. Robinson, Tom Moss.

References

1. Moss T, Langlois F, Gagnon-Kugler T, Stefanovsky V. A housekeeper with power of attorney: the rRNA genes in ribosome biogenesis. Cell Mol Life Sci. 2007; 64(1):29–49. https://doi.org/10.1007/s00018-006-6278-1 PMID: 17171232.
2. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. Science. 2016; 352(6292):1413–6. Epub 2016/06/18. https://doi.org/10.1126/science.aad9868 PMID: 27313038.
3. Loreni F, Mancino M, Biffo S. Translation factors and ribosomal proteins control tumor onset and progression: how? Oncogene. 2014; 33(17):2145–56. Epub 2013/05/07. https://doi.org/10.1038/onc.2013.153 PMID: 23644661.
4. Brar GA, Weissman JS. Ribosome profiling reveals the what, when, where and how of protein synthesis. Nat Rev Mol Cell Biol. 2015; 16(11):651–64. Epub 2015/10/16. https://doi.org/10.1038/nrm4069 PMID: 26465719.
5. Ingolia NT, Lareau LF, Weissman JS. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. Cell. 2011; 147(4):789–802. https://doi.org/10.1016/j.cell.2011.10.002 PMID: 22056041
6. Werner A, Iwasaki S, McGourty CA, Medina-Ruiz S, Teerikorpi N, Fedrigo I, et al. Cell-fate determination by ubiquitin-dependent regulation of translation. Nature. 2015; 525(7570):523–7. Epub 2015/09/25. https://doi.org/10.1038/nature14978 PMID: 26399832; PubMed Central PMCID: PMCPMC4602398.
7. Duncan R, Hershey JW. Identification and quantitation of levels of protein synthesis initiation factors in crude HeLa cell lysates by two-dimensional polyacrylamide gel electrophoresis. J Biol Chem. 1983; 258(11):7228–35. Epub 1983/06/10. PMID: 6853516.

8. Wolf SF, Schlessinger D. Nuclear metabolism of ribosomal RNA in growing, methionine-limited, and ethionine-treated HeLa cells. Biochemistry. 1977; 16:2783–91. PMID: 889788
9. Jackson DA, Pombo A, Iborra F. The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells. Faseb j. 2000; 14(2):242–54. Epub 2000/02/05. PMID: 10657981.
10. Moss T, Stefanovsky VY. At the center of eukaryotic life. Cell. 2002; 109(5):545–8. PMID: 12062097
11. Condon C, Squires C, Squires CL. Control of rRNA transcription in Escherichia coli. Microbiol Rev. 1995; 59(4):623–45. PMID: 8531889.
12. Tschochner H, Hurt E. Pre-ribosomes on the road from the nucleolus to the cytoplasm. Trends Cell Biol. 2003; 13(5):255–63. PMID: 12742169.
13. Fatica A, Tollervey D. Making ribosomes. Curr Opin Cell Biol. 2002; 14(3):313–8. PMID: 12067653
14. Henderson AS, Eicher EM, Yu MT, Atwood KC. The chromosomal location of ribosomal DNA in the mouse. Chromosoma. 1974; 49(2):155–60. https://doi.org/10.1007/BF00348887 PMID: 4448113
15. Henderson AS, Warburton D, Atwood KC. Location of ribosomal DNA in the human chromosome complement. Proc Natl Acad Sci U S A. 1972; 69(11):3394–8. PMID: 4508329.
16. Savic N, Bar D, Leone S, Frommel SC, Weber FA, Vollenweider E, et al. lncRNA Maturation to Initiate Heterochromatin Formation in the Nucleolus Is Required for Exit from Pluripotency in ESCs. Cell Stem Cell. 2014; 15(6):720–34. Epub 2014/12/07. https://doi.org/10.1016/j.stem.2014.10.005 PMID: 25479748.
17. Santoro R, Grummt I. Epigenetic mechanism of rRNA gene silencing: temporal order of NoRC-mediated histone modification, chromatin remodeling, and DNA methylation. Mol Cell Biol. 2005; 25 (7):2539–46. https://doi.org/10.1128/MCB.25.7.2539-2546.2005 PMID: 15767661.
18. Stefanovsky VY, Pelletier G, Hannan R, Gagnon-Kugler T, Rothblum LI, Moss T. An immediate response of ribosomal transcription to growth factor stimulation in mammals is mediated by ERK phosphorylation of UBF. Mol Cell. 2001; 8(5):1063–73. PMID: 11741541
19. Zhao J, Yuan X, Frodin M, Grummt I. ERK-Dependent Phosphorylation of the Transcription Initiation Factor TIF-IA Is Required for RNA Polymerase I Transcription and Cell Growth. Mol Cell. 2003; 11 (2):405–13. PMID: 12620228.
20. Grummt I. The nucleolus-guardian of cellular homeostasis and genome integrity. Chromosoma. 2013. Epub 2013/09/12. https://doi.org/10.1007/s00412-013-0430-0 PMID: 24022641.
21. Schnapp A, Schnapp G, Erny B, Grummt I. Function of the growth-regulated transcription initiation factor TIF-IA in initiation complex formation at the murine ribosomal gene promoter. Molecular and Cellular Biology. 1993; 13:6723–32. PMID: 8413268
22. Milkereit P, Tschochner H. A specialized form of RNA polymerase I, essential for initiation and growth-dependent regulation of rRNA synthesis, is disrupted during transcription. Embo J. 1998; 17(13):3692–703. https://doi.org/10.1093/emboj/17.13.3692 PMID: 9649439.
23. Stefanovsky V, Langlois F, Gagnon-Kugler T, Rothblum LI, Moss T. Growth factor signaling regulates elongation of RNA polymerase I transcription in mammals via UBF phosphorylation and r-chromatin remodeling. Mol Cell. 2006; 21(5):629–39. Epub 2006/03/02. https://doi.org/10.1016/j.molcel.2006.01.023 PMID: 16507361.
24. Bar-Nahum G, Nudler E. Isolation and characterization of sigma(70)-retaining transcription elongation complexes from Escherichia coli. Cell. 2001; 106(4):443–51. PMID: 11525730.
25. Mukhopadhyay J, Kapanidis AN, Mekler V, Kortkhonjia E, Ebright YW, Ebright RH. Translocation of sigma70 with RNA polymerase during transcription: Fluorescence resonance energy transfer assay for movement relative to DNA. Cell. 2001; 106(4):453–63. PMID: 11525731
26. Paget MS, Helmann JD. The sigma70 family of sigma factors. Genome Biol. 2003; 4(1):203. PMID: 12540296.
27. Stepanchick A, Zhi H, Cavanaugh AH, Rothblum K, Schneider DA, Rothblum LI. DNA binding by the ribosomal DNA transcription factor rrn3 is essential for ribosomal DNA transcription. J Biol Chem. 2013; 288(13):9135–44. https://doi.org/10.1074/jbc.M112.444265 PMID: 23393135; PubMed Central PMCID: PMC3610986.
28. Engel C, Gubbey T, Neyer S, Sainsbury S, Oberthuer C, Baejen C, et al. Structural Basis of RNA Polymerase I Transcription Initiation. Cell. 2017; 169(1):120–31.e22. Epub 2017/03/25. https://doi.org/10.1016/j.cell.2017.03.003 PMID: 28340337.
29. Pilsl M, Crucifix C, Papai G, Krupp F, Steinbauer R, Griesenbeck J, et al. Structure of the initiation-competent RNA polymerase I and its implication for transcription. Nat Commun. 2016; 7:12126. Epub 2016/07/16. https://doi.org/10.1038/ncomms12126 PMID: 27418187; PubMed Central PMCID: PMCPMC4947174.

30. Grummt I. Life on a planet of its own: regulation of RNA polymerase I transcription in the nucleolus. Genes Dev. 2003; 17(14):1691–702. https://doi.org/10.1101/gad.1098503R PMID: 12865296.

31. Goodfellow SJ, Zomerdijk JC. Basic mechanisms in RNA polymerase I transcription of the ribosomal RNA genes. Subcell Biochem. 2012; 61:211–36. Epub 2012/11/15. https://doi.org/10.1007/978-94-007-4525-4_10 PMID: 23150253.

32. O'Sullivan AC, Sullivan GJ, McStay B. UBF binding in vivo is not restricted to regulatory sequences within the vertebrate ribosomal DNA repeat. Mol Cell Biol. 2002; 22(2):657–68. https://doi.org/10.1128/MCB.22.2.657-668.2002 PMID: 11756560

33. Sanij E, Diesch J, Lesmana A, Poortinga G, Hein N, Lidgerwood G, et al. A novel role for the Pol I transcription factor UBTF in maintaining genome stability through the regulation of highly transcribed Pol II genes. Genome Res. 2015; 25(2):201–12. https://doi.org/10.1101/gr.176115.114 PMID: 25452314.

34. Zentner GE, Balow SA, Scacheri PC. Genomic characterization of the mouse ribosomal DNA locus. G3 (Bethesda). 2014; 4(2):243–54. Epub 2013/12/19. https://doi.org/10.1534/g3.113.009290 PMID: 24347625; PubMed Central PMCID: PMCPMC3931559.

35. Hamdane N, Stefanovsky VY, Tremblay MG, Nemeth A, Paquet E, Lessard F, et al. Conditional inactivation of Upstream Binding Factor reveals its epigenetic functions and the existence of a somatic nucleolar precursor body. PLoS Genetics. 2014; 10(8):e1004505. https://doi.org/10.1371/journal.pgen.1004505 PMID: 25121932; PubMed Central PMCID: PMC4133168.

36. Bazett-Jones DP, Leblanc B, Herfort M, Moss T. Short-range DNA looping by the Xenopus HMG-box transcription factor, xUBF. Science. 1994; 264:1134–7. PMID: 8178172

37. Copenhaver GP, Putnam CD, Denton ML, Pikaard CS. The RNA polymerase I transcription factor UBF is a sequence-tolerant HMG-box protein that can recognize structured nucleic acids. Nucleic Acids Res. 1994; 22:2651–7. PMID: 8041627

38. Stefanovsky VY, Bazett-Jones DP, Pelletier G, Moss T. The DNA supercoiling architecture induced by the transcription factor xUBF requires three of its five HMG-boxes. Nucleic Acids Res. 1996; 24:3208–15. PMID: 8774902

39. Paalman MH, Henderson SL, Sollner-Webb B. Stimulation of the mouse rRNA gene promoter by a distal spacer promoter. MolCell Biol. 1995; 15:4648–56.

40. De Winter RFJ, Moss T. Spacer promoters are essential for efficient enhancement of X laevis ribosomal transcription. Cell. 1986; 44:313–8. PMID: 3943126

41. Mais C, Wright JE, Prieto JL, Raggett SL, McStay B. UBF-binding site arrays form pseudo-NORs and sequester the RNA polymerase I transcription machinery. Genes Dev. 2005; 19(1):50–64. https://doi.org/10.1101/gad.310705 PMID: 15598984.

42. Nemeth A, Guibert S, Tiwari VK, Ohlsson R, Langst G. Epigenetic regulation of TTF-I-mediated promoter-terminator interactions of rRNA genes. Embo J. 2008; 27(8):1255–65. Epub 2008/03/21. https://doi.org/10.1038/emboj.2008.57 PMID: 18354495.

43. Chopra VS, Cande J, Hong JW, Levine M. Stalled Hox promoters as chromosomal boundaries. Genes Dev. 2009; 23(13):1505–9. https://doi.org/10.1101/gad.1807309 PMID: 19515973; PubMed Central PMCID: PMCPMC2704471.

44. Yamamoto RT, Nogi Y, Dodd JA, Nomura M. RRN3 gene of Saccharomyces cerevisiae encodes an essential RNA polymerase I transcription factor which interacts with the polymerase independently of DNA template. EMBO J. 1996; 15:3964–73. PMID: 8670901

45. Yuan X, Zhou Y, Casanova E, Chai M, Kiss E, Grone HJ, et al. Genetic Inactivation of the Transcription Factor TIF-IA Leads to Nucleolar Disruption, Cell Cycle Arrest, and p53-Mediated Apoptosis. Mol Cell. 2005; 19(1):77–87. https://doi.org/10.1016/j.molcel.2005.05.023 PMID: 15989966.

46. Nagy A, Gertsenstein M, Vintersten K, Behringer R. Isolating Total RNA from Mouse Embryos or Fetal Tissues. Cold Spring Harbor Protocols. 2007; 2007(9):pdb.prot4773. https://doi.org/10.1101/pdb.prot4773 PMID: 21357161

47. Lee J. Expected Recovery of Total RNA from Mouse Embryos, Extraembryonic Tissues, and Fetal and Adult Tissues. Cold Spring Harbor Protocols. 2007; 2007(9):pdb.ip38. https://doi.org/10.1101/pdb.ip38 PMID: 21357156

48. Stefanovsky VY, Moss T. Metabolic Labeling in the Study of Mammalian Ribosomal RNA Synthesis. Methods Mol Biol. 2016; 1455:133–45. Epub 2016/09/01. https://doi.org/10.1007/978-1-4939-3792-9_11 PMID: 27576716.

49. Goetze H, Wittner M, Hamperl S, Hondele M, Merz K, Stoeckl U, et al. Alternative chromatin structures of the 35S rRNA genes in Saccharomyces cerevisiae provide a molecular basis for the selective recruitment of RNA polymerases I and II. Mol Cell Biol. 2010; 30(8):2028–45. Epub 2010/02/16. https://doi.org/10.1128/MCB.01512-09 PMID: 20154141; PubMed Central PMCID: PMC2849473.

50. Leblanc B, Read C, Moss T. Recognition of the Xenopus ribosomal core promoter by the transcription factor xUBF involves multiple HMG box domains and leads to an xUBF interdomain interaction. EMBO J. 1993; 12:513–25. PMID: 8440241

51. Chen D, Huang S. Nucleolar components involved in ribosome biogenesis cycle between the nucleolus and nucleoplasm in interphase cells. J Cell Biol. 2001; 153(1):169–76. PMID: 11285283.

52. Lessard F, Morin F, Ivanchuk S, Langlois F, Stefanovsky V, Rutka J, et al. The ARF tumor suppressor controls ribosome biogenesis by regulating the RNA polymerase I transcription factor TTF-I. Mol Cell. 2010; 38(4):539–50. Epub 2010/06/02. https://doi.org/10.1016/j.molcel.2010.03.015 PMID: 20513429.

53. Conconi A, Widmer RM, Koller T, Sogo JM. Two different chromatin structures coexist in ribosomal RNA genes throughout the cell cycle. Cell. 1989; 57:753–61. PMID: 2720786

54. Sanij E, Poortinga G, Sharkey K, Hung S, Holloway TP, Quin J, et al. UBF levels determine the number of active ribosomal RNA genes in mammals. J Cell Biol. 2008; 183(7):1259–74. Epub 2008/12/24. https://doi.org/10.1083/jcb.200805146 PMID: 19103806.

55. Wittner M, Hamperl S, Stockl U, Seufert W, Tschochner H, Milkereit P, et al. Establishment and maintenance of alternative chromatin states at a multicopy gene locus. Cell. 2011; 145(4):543–54. Epub 2011/05/14. https://doi.org/10.1016/j.cell.2011.03.051 PMID: 21565613.

56. Stefanovsky VY, Pelletier G, Bazett-Jones DP, Crane-Robinson C, Moss T. DNA looping in the RNA polymerase I enhancesome is the result of non-cooperative in-phase bending by two UBF molecules. Nucleic Acids Res. 2001; 29(15):3241–7. PMID: 11470882

57. Pelletier G, Stefanovsky VY, Faubladier M, Hirschler-Laszkiewicz I, Savard J, Rothblum LI, et al. Competitive recruitment of CBP and Rb-HDAC regulates UBF acetylation and ribosomal transcription. Mol Cell. 2000; 6(5):1059–66. Epub 2000/12/07. PMID: 11106745.

58. Strohner R, Nemeth A, Nightingale KP, Grummt I, Becker PB, Langst G. Recruitment of the Nucleolar Remodeling Complex NoRC Establishes Ribosomal DNA Silencing in Chromatin. Mol Cell Biol. 2004; 24(4):1791–8. https://doi.org/10.1128/MCB.24.4.1791-1798.2004 PMID: 14749393.

59. Vintermist A, Bohm S, Sadeghifar F, Louvet E, Mansen A, Percipalle P, et al. The chromatin remodelling complex B-WICH changes the chromatin structure and recruits histone acetyl-transferases to active rRNA genes. PLoS ONE. 2011; 6(4):e19184. Epub 2011/05/12. https://doi.org/10.1371/journal.pone.0019184 PMID: 21559432; PubMed Central PMCID: PMC3084792.

60. Hamdane N, Tremblay MG, Dillinger S, Stefanovsky VY, Nemeth A, Moss T. Disruption of the UBF gene induces aberrant somatic nucleolar bodies and disrupts embryo nucleolar precursor bodies. Gene. 2016. Epub 2016/09/11. https://doi.org/10.1016/j.gene.2016.09.013 PMID: 27614293.

61. Floutsakou I, Agrawal S, Nguyen TT, Seoighe C, Ganley AR, McStay B. The shared genomic architecture of human nucleolar organizer regions. Genome Res. 2013; 23(12):2003–12. https://doi.org/10.1101/gr.157941.113 PMID: 23990606; PubMed Central PMCID: PMC3847771.

62. Merz K, Hondele M, Goetze H, Gmelch K, Stoeckl U, Griesenbeck J. Actively transcribed rRNA genes in S. cerevisiae are organized in a specialized chromatin associated with the high-mobility group protein Hmo1 and are largely devoid of histone molecules. Genes Dev. 2008; 22(9):1190–204. Epub 2008/05/03. https://doi.org/10.1101/gad.466908 PMID: 18451108.

63. Birch JL, Zomerdijk JC. Structure and function of ribosomal RNA gene chromatin. Biochem Soc Trans. 2008; 36(Pt 4):619–24. Epub 2008/07/18. https://doi.org/10.1042/BST0360619 PMID: 18631128.

64. Grummt I, Langst G. Epigenetic control of RNA polymerase I transcription in mammalian cells. Biochim Biophys Acta. 2013; 1829(3–4):393–404. Epub 2012/10/16. https://doi.org/10.1016/j.bbagrm.2012.10.004 PMID: 23063748.

65. McStay B, Grummt I. The epigenetics of rRNA genes: from molecular to chromosome biology. Annu Rev Cell Dev Biol. 2008; 24:131–57. Epub 2008/07/12. https://doi.org/10.1146/annurev.cellbio.24.110707.175259 PMID: 18616426.

66. Bruce K, Myers FA, Mantouvalou E, Lefevre P, Greaves I, Bonifer C, et al. The replacement histone H2A.Z in a hyperacetylated form is a feature of active genes in the chicken. Nucleic Acids Res. 2005; 33(17):5633–9. https://doi.org/10.1093/nar/gki874 PMID: 16204459; PubMed Central PMCID: PMCPMC1243646.

67. van de Nobelen S, Rosa-Garrido M, Leers J, Heath H, Soochit W, Joosen L, et al. CTCF regulates the local epigenetic state of ribosomal DNA repeats. Epigenetics Chromatin. 2010; 3(1):19. Epub 2010/11/10. https://doi.org/10.1186/1756-8935-3-19 PMID: 21059229; PubMed Central PMCID: PMC2993708.

68. Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet. 2014; 15(4):234–46. https://doi.org/10.1038/nrg3663 PMID: 24614316; PubMed Central PMCID: PMCPMC4610363.

69. Merkl P, Perez-Fernandez J, Pilsl M, Reiter A, Williams L, Gerber J, et al. Binding of the termination factor Nsi1 to its cognate DNA site is sufficient to terminate RNA polymerase I transcription in vitro and to

induce termination in vivo. Mol Cell Biol. 2014; 34(20):3817–27. Epub 2014/08/06. https://doi.org/10.1128/MCB.00395-14 PMID: 25092870; PubMed Central PMCID: PMCPMC4187712.

70. Mayer C, Schmitz KM, Li J, Grummt I, Santoro R. Intergenic Transcripts Regulate the Epigenetic State of rRNA Genes. Mol Cell. 2006; 22(3):351–61. https://doi.org/10.1016/j.molcel.2006.03.028 PMID: 16678107.

71. Mayer C, Neubert M, Grummt I. The structure of NoRC-associated RNA is crucial for targeting the chromatin remodelling complex NoRC to the nucleolus. EMBO Rep. 2008. Epub 2008/07/05. https://doi.org/10.1038/embor.2008.109 PMID: 18600236.

72. Lessard F, Stefanovsky V, Tremblay MG, Moss T. The cellular abundance of the essential transcription termination factor TTF-I regulates ribosome biogenesis and is determined by MDM2 ubiquitinylation. Nucleic Acids Res. 2012; 40(12):5357–67. Epub 2012/03/03. https://doi.org/10.1093/nar/gks198 PMID: 22383580; PubMed Central PMCID: PMC3384320.

73. Stefanovsky VY, Langlois F, Pelletier G, Bazett-Jones DP, Moss T. ERK modulates DNA bending and Enhancesome Structure by phosphorylating HMG1-boxes 1 and 2 of the RNA polymerase I transcription factor UBF. Biochemistry. 2006; 45(11):3626–34. https://doi.org/10.1021/bi051782h PMID: 16533045.

74. Bell SP, Learned RM, Jantzen HM, Tjian R. Functional cooperativity between transcription factors UBF1 and SL1 mediates human ribosomal RNA synthesis. Science. 1988; 241:1192–7. PMID: 3413483

75. Friedrich JK, Panov KI, Cabart P, Russell J, Zomerdijk JC. TBP-TAF complex SL1 directs RNA polymerase I pre-initiation complex formation and stabilizes upstream binding factor at the rDNA promoter. J Biol Chem. 2005; 280(33):29551–8. https://doi.org/10.1074/jbc.M501595200 PMID: 15970593.

76. Skibbens Robert V. Cell Biology: Cohesin Rings Leave Loose Ends. Curr Biol. 2015; 25(3):R108–R10. https://doi.org/10.1016/j.cub.2014.12.015 PMID: 25649818

77. Denissov S, Lessard F, Mayer C, Stefanovsky V, van Driel M, Grummt I, et al. A model for the topology of active ribosomal RNA genes. EMBO Rep. 2011; 12(3):231–7. Epub 2011/02/19. https://doi.org/10.1038/embor.2011.8 PMID: 21331097.

78. Caburet S, Conti C, Schurra C, Lebofsky R, Edelstein SJ, Bensimon A. Human ribosomal RNA gene arrays display a broad range of palindromic structures. Genome Res. 2005; 15(8):1079–85. Epub 2005/07/19. https://doi.org/10.1101/gr.3970105 PMID: 16024823; PubMed Central PMCID: PMC1182220.

79. Bose T, Lee KK, Lu S, Xu B, Harris B, Slaughter B, et al. Cohesin proteins promote ribosomal RNA production and protein translation in yeast and human cells. PLoS Genet. 2012; 8(6):e1002749. https://doi.org/10.1371/journal.pgen.1002749 PMID: 22719263; PubMed Central PMCID: PMCPMC3375231.

80. Moss T. DNA methyltransferase inhibition may limit cancer cell growth by disrupting ribosome biogenesis. Epigenetics. 2011; 6(2):128–33. Epub 2010/10/12. https://doi.org/10.4161/epi.6.2.13625 PMID: 20935488.

81. Gagnon-Kugler T, Langlois F, Stefanovsky V, Lessard F, Moss T. Loss of human ribosomal gene CpG methylation enhances cryptic RNA polymerase II transcription and disrupts ribosomal RNA processing. Mol Cell. 2009; 35(4):414–25. Epub 2009/09/01. https://doi.org/10.1016/j.molcel.2009.07.008 PMID: 19716787.

82. Hamdane N, Herdman C, Mars JC, Stefanovsky V, Tremblay MG, Moss T. Depletion of the cisplatin targeted HMGB-box factor UBF selectively induces p53-independent apoptotic death in transformed cells. Oncotarget. 2015; 6(29):27519–36. Epub 2015/09/01. https://doi.org/10.18632/oncotarget.4823 PMID: 26317157; PubMed Central PMCID: PMCPMC4695006.

83. Giroux S, Tremblay M, Bernard D, Cadrin-Girard JF, Aubry S, Larouche L, et al. Embryonic death of Mek1-deficient mice reveals a role for this kinase in angiogenesis in the labyrinthine region of the placenta. Curr Biol. 1999; 9:369–72. PMID: 10209122

84. Laemmli UK. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature. 1970; 227:680–5. PMID: 5432063

85. He HH, Meyer CA, Hu SS, Chen MW, Zang C, Liu Y, et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nat Methods. 2014; 11(1):73–8. https://doi.org/10.1038/nmeth.2762 PMID: 24317252; PubMed Central PMCID: PMCPMC4018771.

86. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404; PubMed Central PMCID: PMCPMC4103590.

87. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9(4):357–9. https://doi.org/10.1038/nmeth.1923 PMID: 22388286; PubMed Central PMCID: PMCPMC3322381.

88. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38(4):576–89. Epub 2010/06/02. https://doi.org/10.1016/j.molcel.2010.05.004 PMID: 20513432; PubMed Central PMCID: PMCPMC2898526.

89. Stefanovsky VY, Moss T. Regulation of rRNA synthesis in human and mouse cells is not determined by changes in active gene count. Cell Cycle. 2006; 5(7):735–9. https://doi.org/10.4161/cc.5.7.2633 PMID: 16582637.

90. Grozdanov P, Georgiev O, Karagyozov L. Complete sequence of the 45-kb mouse ribosomal DNA repeat: analysis of the intergenic spacer. Genomics. 2003; 82(6):637–43. PMID: 14611805.

91. Morris SA, Baek S, Sung MH, John S, Wiench M, Johnson TA, et al. Overlapping chromatin-remodeling systems collaborate genome wide at dynamic chromatin transitions. Nat Struct Mol Biol. 2014; 21 (1):73–81. https://doi.org/10.1038/nsmb.2718 PMID: 24317492; PubMed Central PMCID: PMCPMC3947387.

92. Kasper LH, Qu C, Obenauer JC, McGoldrick DJ, Brindle PK. Genome-wide and single-cell analyses reveal a context dependent relationship between CBP recruitment and gene expression. Nucleic Acids Res. 2014; 42(18):11363–82. https://doi.org/10.1093/nar/gku827 PMID: 25249627; PubMed Central PMCID: PMCPMC4191404.

93. Ekblom R, Smeds L, Ellegren H. Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. BMC Genomics. 2014; 15:467. https://doi.org/10.1186/1471-2164-15-467 PMID: 24923674; PubMed Central PMCID: PMCPMC4070552.

94. Newton K, Petfalski E, Tollervey D, Caceres JF. Fibrillarin is essential for early development and required for accumulation of an intron-encoded small nucleolar RNA in the mouse. Molecular and cellular biology. 2003; 23(23):8519–27. Epub 2003/11/13. PubMed Central PMCID: PMC262675. https://doi.org/10.1128/MCB.23.23.8519-8527.2003 PMID: 14612397

95. Chen H, Li Z, Haruna K, Semba K, Araki M, Yamamura K, et al. Early pre-implantation lethality in mice carrying truncated mutation in the RNA polymerase 1–2 gene. Biochem Biophys Res Commun. 2008; 365(4):636–42. Epub 2007/11/21. https://doi.org/10.1016/j.bbrc.2007.11.019 PMID: 18023416.

96. Engel W, Zenzes MT, Schmid M. Activation of mouse ribosomal RNA genes at the 2-cell stage. Hum Genet. 1977; 38(1):57–63. PMID: 903155.

97. Fulka H, Langerova A. The maternal nucleolus plays a key role in centromere satellite maintenance during the oocyte to embryo transition. Development. 2014; 141(8):1694–704. https://doi.org/10.1242/dev.105940 PMID: 24715459.