Université de Montréal

Localisation de l'ARN polymérase II humaine à travers le génome en couplant double immunoprécipitation de la chromatine et clonage

par
Pierre Côte

Département de Biochimie
Faculté de Médecine

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade
de Maître ès sciences (M.Sc)
en biochimie

Janvier 2005

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :
Localisation de l'ARN Polymérase II humaine à travers le génome
en couplant double immunoprécipitation de la chromatine et clonage

Présenté par :

Pierre Côte

a été évalué par un jury composé des personnes suivantes :

Dr Jacques Archambault
président-rapporteur

Dr Benoit Coulombe
directeur de recherches

Dr Alain Nepveu
membre du jury

Résumé et mots-clés

La nanomachine moléculaire qui décode l'information contenue dans le génome humain, l'ARN polymérase II (ARN Pol II), transcrit les gènes en ARNm. Découverts puis caractérisés à l'échelle du gène, de nombreux facteurs impliqués dans la transcription (protéines seules ou complexes) nécessitent désormais une analyse à l'échelle du génome.

S'intégrant dans un vaste projet génomique/protéomique, ma contribution porte principalement sur la localisation de l'ARN Pol II à travers le génome humain *in vivo*. Dans cet objectif, nous avons développé une technique de localisation génomique combinant le système de purification par double affinité du TAP-tag à celui de l'immunoprécipitation de la chromatine. Utilisant cette méthode, nous avons pu localiser l'ARN Pol II sur des gènes d'intérêt (par PCR quantitative) ainsi qu'à travers le génome humain (par clonage). Le séquençage des fragments d'ADN immunoprécipité a révélé que 17.3% correspondaient à des promoteurs de gènes connus, 34.5% à des régions transcrites et 9.1% à des « fins de gène ». Comparées à leurs équivalents statistiques dans notre génome, respectivement 1.5, 24 et 1.5% (basé sur *NCBI human built 34*), ces valeurs mettent en évidence l'efficacité de la technique de localisation. Parallèlement à cette approche, une base de données référençant tous les sites de transcription humains a été créée (environ 17,000 sites). En croisant plusieurs sources de données, nous avons ramené cette base à un set de 2000 séquences consensus, dont 400 de très haute qualité. Parmi ces dernières, plusieurs ont été choisies afin de servir de modèles lors de la localisation de l'ARN Pol II. Très récemment, elles ont aussi été déposées sur une puce à ADN composée de promoteurs de gènes de classe II humains. Cette puce à ADN servira

entre autre à localiser, au niveau de promoteur, les différentes composantes de la machinerie transcriptionnelle de base.

Mots-clés : ARN; clonage; double affinité; génomique; immunoprécipitation de la chromatine; localisation; QPCR; ARN polymérase II; TAP-tag; transcription.

Abstract and key words

RNA Polymerase II (RNAPII) is the molecular nanomachine which deciphers the genome's protein encoding information by transcribing genes into mRNA. Previously discovered and then characterized at the gene level, numerous factors implicated in transcription (proteins or complexes) now need to be analyzed on a genome-wide scale.

As part of a larger genomic/proteomic project, my contribution to this work is to analyze the localization of RNAPII across the human genome *in vivo*. To achieve this goal, we have combined classical chromatin immunoprecipitation (ChIP) with the well-known TAP-tag purification approach. By using this method, RNAPII has been localized at a gene specific level (by quantitative PCR) and on a genome-wide level (by cloning). Sequencing reveals that 17.3% of immunoprecipitated DNA corresponds to promoters of known genes, 34.5% to transcribed regions and 9.1% to the 3'end of genes. Theoretically, these regions represent 1.5, 24 and 1.5%, respectively, of human genome (according to NCBI human built 34). Our location analysis technique, therefore, is successful in enriching for transcriptionally active regions of the human genome. By crossing different data sources, we have created an in-house transcription start site (TSS) database. Through regrouping all human sites (~ 17,000), we have reduced them to a set of 2000 consensus sites where ~400 possess a high quality consensus origin. A set of genes was chosen from this database and serves as a model for localizing the transcriptional machinery. Recently, the promoters of this set of genes have been spotted on a DNA microarray representing most of presently known class II promoters. This chip will be used to analyze, at a promoter specific level, the location of the components of the transcriptional machinery.

Key words: chromatin immunoprecipitation; cloning; double affinity; genomics; localization; PCR; promoter; RNA polymerase II; TAP-tag; transcription.

Table des matières

Liste des tableaux

Liste des figures

Abréviation des sigles utilisés

3'end	fin du gène
A	Adénine (base nucléique)
ADN	Acide désoxyribonucléique
ARN	Acide ribonucléique
ARNm	ARN messager
ARN Pol	ARN polymérase
ATG	Codon d'initiation de la traduction, codant pour une méthionine
ATP	Adénosine triphosphate
av	Aval
BLAT	*BLAST like alignment tools*
BRE	*TFIIB recognition element*
C	Cytosine (base nucléique)
CBP	Calmodulin binding peptide
Cdk7	*Cyclin dependant kinase 7*
CDS	Coding DNA sequence
cDNA	ADN complémentaire
CG	Paire dinucléique CG
C+G	Proportion de C et de G dans une séquence d'ADN
CGI	*CpG island*
ch	Chromosome
ChIP	Immunoprécipitation de la chromatine
chip	*Microarrays*
CTD	*Carboxy Terminus repeat Domain*
C-term	Carboxy-terminal
DBTSS	*Database transcription start site*
dNTP	Désoxyrinucléotide triphosphate
DPE	*Downstream promoter element*
EKC	*Embryonic kidney cells*
FA	Formaldehyde

fl-RNA	*Full-length RNA*
G	Guanine
Gb	Gigabase
GTF	Facteur généraux de transcription.
HAT	*Histone acetyl transferase*
HDAC	*Histones deactetylase*
Inr	Initiateur
IP	Immunoprécipitation
kb	Kilobase
kDa	Kilodalton
LCR	*Locus Control Region*
MAR	*Matrix Attachment Region*
Mb	Mégabase
MGC	*Mammalian gene collection*
MTE	*Motif ten element*
nb	Nombre
N-term	Amino-terminal
nt	Nucléotides
NTP	Ribonucléotide triphosphate
ORF	*Open reading frame*
pb	Paire de base
PCR	*Polymerase chain reaction*
QPCR	*Quantitative PCR*
RT-PCR	*Reverse transcriptase PCR*
PIC	Complexe de préinitiation
Pol II	ARN Polymérase II
Poly(A)	Polyadénilation
pr	Promoteur
Pu	Purine
Py	Pyrimidine
Rpb	RNA Polymérase B

SIT	Site d'Initiation de la Transcription
SAGE	*Serial Analisis of Gene Expression*
SP1	Transcription factor Sp1
SWI/SNF	*Switching mating-type / Sucrose non fermenting*
T	Thymine
TAF	*TBP Associated Factor*
TATA	Boîte TATA
TAP	*Tandem affinity peptide*
TBP	*Tata Binding Protein*
TDS	*Transcribed DNA Sequence*
TF	Transcription Factor
TFII	*Transcription factor associated with RNA polymerase II* (voir GTF)
TLF	*TBP like factor*
TRF	*TBP related factor*
UAS	*Upstream activating factor*
UTR	*Untranslated Region*
VP16	*Herpes simplex virion protein 16*
WCE	*Whole Cell Extract*
x	Fois (nombre de)
X	Chromosome X
XP	*Xerodema pigmentosum*

Dédicace

"You step into the Road, and if you don't keep your feet,
there is no knowing where you might be swept off to."

- Bilbo Baggins

Remerciements

C'est mon directeur de recherche, le Dr Benoit Coulombe, que je tiens à remercier en premier, pour m'avoir permis de débuter dans le monde de la recherche ainsi que de m'avoir supporté financièrement durant ces deux années.

Ensuite, je remercie tout naturellement les membres du laboratoire pour leur aide et leur soutien.

J'aimerais remercier particulièrement : Célia Jéronimo, Marie-France Langelier et Vincent Trinh pour les échanges que nous avons eu, tant professionnels que culturels; Merilena Cojocaru pour sa collaboration tout au long du projet; Annie Bouchard pour les cellules qu'elle m'a si souvent préparée ; le Dr Dominique Bergeron pour son aide, ses suggestions et ses très bons sites d'analyse bioinformatique ; Diane Bourque pour son aide en infographie ; Maria-Teresa Dicenza, pour m'avoir si souvent aidé à passer de la langue de Molière à celle de Shakespeare ;

J'aimerais aussi remercier Dorothée Begin et Sylvie Beauchemin pour m'avoir facilité la vie, du point de vue administratif, durant toute ma maîtrise.

Enfin, et non les moindre, je tiens à remercier Adeline Martin, pour m'avoir supporté, m'avoir écouté et m'avoir aidé à corriger les fautes d'orthographe (à des heures impossibles).

À tous, merci d'avoir eu de la patience avec moi lorsque mes réserves étaient depuis longtemps épuisées.

I) Introduction

L'information servant à la synthèse des protéines est encodée dans le génome, dont une succession de quatre bases d'acide désoxyribonucléique forme le code. Le code génétique est, dans un premier temps, transcrit par l'ARN Pol II en ARN puis celui-ci est traduit en polypeptides par les ribosomes. Ce dogme, ADN→ARN→Protéine, est la base de la biologie moléculaire.

I.1. L'ADN, support de l'information génétique

Que ce soit dans une cellule eucaryote ou procaryote, l'information génétique nécessaire à la croissance, au développement et à la division des cellules est contenue dans la cellule elle-même : dans son génome.

I.1.1. Organisation de l'ADN génomique dans les cellules eucaryotes

Partant de ce principe, la première difficulté que rencontra la nature fut que le volume brut du contenu (l'ADN) était bien plus grand que le volume du contenant (la cellule). Mais le problème s'avéra encore plus complexe chez les eucaryotes où le génome est confiné dans un espace encore plus petit : le noyau. La nature a donc du trouver des stratégies pour pallier ce manque d'espace. Pour cela, le génome des eucaryotes, en association avec certaines protéines (les histones), forme des complexes nucléoprotéiques : les nucléosomes. Ces unités de bases servent à leur tour à former des structures de plus en plus complexes aboutissant à un volume inférieur à celui du noyau permettant ainsi son emprisonnement.

I.1.1.1. Structure chromatinienne

Il y a de cela maintenant 30 ans Oudet *et al.* mirent en évidence l'organisation particulière des génomes eucaryotes (Oudet et al., 1975). Initialement représentée comme une structure uniforme ponctuée de particules en « collier de perles », l'image que nous en avons actuellement est plus complexe, possédant différents niveaux d'organisation : d'ADN en chromatine, puis en fibre et enfin en chromosome mitotique (figure 1, tirée de Felsenfeld and Groudine, 2003 ; voir aussi Olins and Olins, 2003).

I.1.1.1.1. Structure de base : le nucléosome.

Unité fondamentale répétée sur tout le génome eucaryote, le nucléosome est composé par une partie en acide nucléique (ADN double brin) et une partie protéique (histones). Du point de vue structural, les histones forment un octamère cylindrique de deux fois deux paires d'hétérodimères, H2A avec H2B et H3 avec H4 et l'ADN, 146 paires de bases (pb), s'enroulent autour du cylindre d'histones (Richmond and Davey, 2003). Une séquence d'environ cinquante pb sépare deux nucléosomes le long de la chromatine (figure 1). La cristallographie (2.8Å) du nucléosome publiée en 1997 (Luger et al., 1997) constitue une avancée marquante dans l'étude des interactions protéine/ADN au niveau du nucléosome. Cette cristallographie a aussi mis en évidence l'organisation particulière des queues des histones qui, pointant hors du cœur du nucléosome, permettraient des interactions directement entre nucléosomes, contribuant ainsi à l'organisation en superstructure (Carruthers and Hansen, 2000). Du point de vue protéique, il est intéressant de noter que les quatre histones composant l'octamère sont les protéines les mieux conservées du monde eucaryote.

Toutefois, les histones principales peuvent être substituées au profit de variants impliquant des modifications fonctionnelles pour la chromatine. L'histone H2AZ, en se substituant à H2A sur certains sites précis dans le génome, réduit la stabilité du nucléosome (Redon et al., 2002). H2AX, distribué au hasard dans le génome en remplacement de H2A, est ciblé par la phosphorylation accompagnant la réparation de l'ADN (Redon et al., 2002). H3.3, une variante de l'histone H3, peut-être incorporée dans la chromatine des cellules quiescentes, et préférentiellement au niveau des gènes transcrits (Ahmad and Henikoff, 2002). Le remplacement des histones par leurs variants constitue l'un des trois types de modification des histones jouant un rôle majeur dans l'organisation du génome.

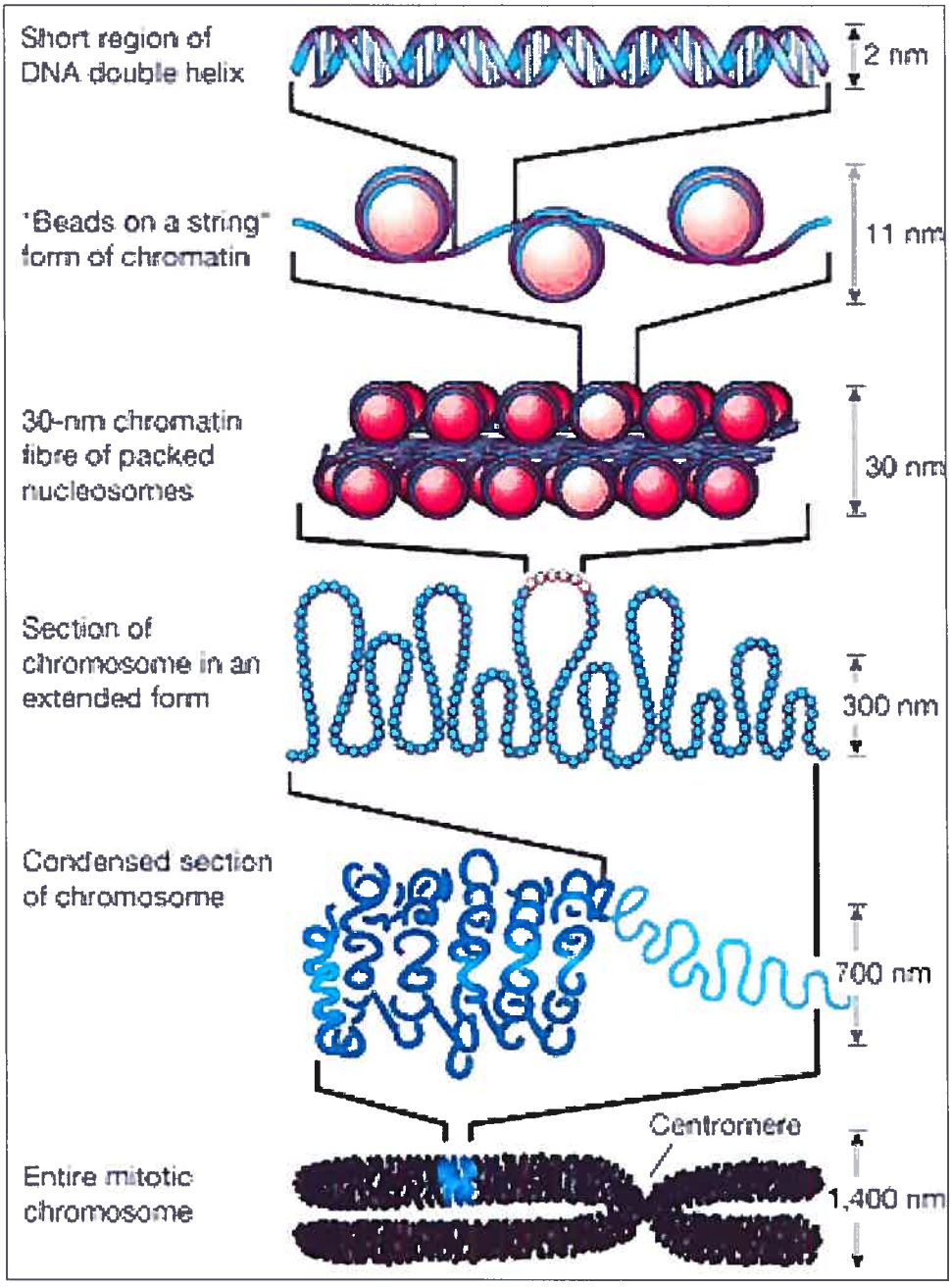


Figure 1. Organisation de la chromatine au sein du génome humain. (tiré de Felsenfeld and Groudine, 2003)

I.1.1.1.2. Superstructure : fibres et chromosomes

Les nucléosomes en « collier de perles » ne constituent que la base de l'organisation de la chromatine au sein du noyau. Le filament de nucléosome s'enroule sur lui-même afin de créer une fibre de plus grande densité encore : la fibre « de 30nm » (figure 1). Cette organisation est possible grâce aux histones H1 (histones *linker*) et aux queues N-terminales des histones interagissant entre elles (Carruthers and Hansen, 2000 ; revu par Hayes and Hansen, 2001). En terme de chiffre, si la formation d'un nucléosome a permis une condensation de l'ADN de l'ordre de 5 à 10 fois, le super-enroulement en fibres « de 30nm » permet d'obtenir une condensation supplémentaire d'environ 5 fois. Lors de la mitose, cette fibre doit être encore plus compactée (environ 200 fois de plus) pour aboutir aux chromosomes mitotiques, alors visibles au microscope photonique (revu par Felsenfeld and Groudine, 2003). En résumé, un chromosome eucaryote n'est fait que d'une seule molécule d'ADN qui, lors de la mitose, peut être extrêmement condensée (de 10,000 à 50,000 fois). Mais il est possible d'aller encore plus loin. Lors de la spermatogénèse, les histones sont remplacées par des protamines qui permettent une condensation maximale du génome (Govin et al., 2004). Une étude récente propose que les protamines dériveraient des histones H1 (Lewis et al., 2004), ce qui pourrait expliquer à la fois leur affinité pour l'ADN et leur fort potentiel de compaction.

I.1.1.1.3. Modifications covalentes des histones

Les histones ne sont pas des acteurs passifs dans l'architecture de la chromatine. Elles peuvent subir de nombreuses modifications chimiques qui servent de signaux de reconnaissance pour la fixation de facteurs spécifiques. Ces nombreuses modifications, post-traductionnelles apparaissent principalement sur la queue N-terminale des histones

et constituent un deuxième type de modification de la structure chromatinienne du génome (Richmond and Davey, 2003). Parmi les modifications connues à ce jour, il est possible de citer l'acétylation des lysines, la méthylation des lysines et arginines, la phosphorylation des sérines et thréonines, l'ubiquitination des lysines, la sumoilation des lysines et l'ADP-ribosilation des acides glutamiques. Ces modifications et les enzymes les effectuant ont été revues récemment (Khorasanizadeh, 2004). Elles reflètent localement un état particulier du génome. Par exemple, la chromatine au voisinage des gènes transcrits est enrichie en histones H3 acétylées sur les lysines 4 (revu par Kurdistani and Grunstein, 2003) alors que la méthylation des lysines 9 des mêmes histones est associée à de la chromatine silencieuse (dite condensée) (revu par Lachner et al., 2003). Toutes ces modifications peuvent s'affecter mutuellement et beaucoup corrèlent, positivement ou négativement, entre elles. Les complexes *Histones Acetyl Transferase* (HAT, qui rime avec déstabilisation des nucléosomes donc accessibilité de la chromatine) et *Histones Deacetylase* (HDAC, qui rime avec l'inverse) illustrent le dynamisme des modifications covalentes sur les histones. Cette réversibilité semble valable pour toutes les modifications (Khorasanizadeh, 2004), incluant la méthylation (Shi et al., 2004). Collectivement, ces modifications dynamiques ont donné naissance à l'hypothèse d'un code relatif aux événements survenus localement sur la chromatine : le code des histones Strahl and Allis, 2000. Comprendre complètement ce code reste un défi à relever.

I.1.1.1.4. Modification de l'ADN

En plus des modifications des histones, l'ADN peut lui aussi être sujet à des modifications covalentes. Chez les plantes et les mammifères, la méthylation des

cytosines est un mécanisme de régulation important et conduit à la répression de la transcription (revu par Bird, 2002). Dans le cas des mammifères, la méthylation semble être limitée aux cytosines impliquées dans la séquence dinucléotidique CG. Des études chez l'humain ont montré une corrélation entre la méthylation de l'ADN et l'inactivation du chromosome X. De plus, certaines maladies humaines (ex : syndrome de Rett ou syndrome du X Fragile) résultent de mutations dans les facteurs régulant la méthylation de l'ADN. Une fois méthylé, l'ADN peut être reconnu et fixé de manière spécifique. MeCP2 est une protéine possédant un domaine de fixation méthyle-CpG qui, une fois fixé sur l'ADN, peut réprimer la transcription en recrutant Sin3 du complexe HDAC. Il semble aussi que MeCP2 puisse réprimer la transcription en favorisant la méthylation de la lysine 9 des histones H3, établissant ainsi un lien entre méthylation de l'ADN et méthylation des histones (Fuks et al., 2003). Toutefois, toutes les paires CG ne sont pas méthylées dans le génome humain et des régions particulières sont, à l'inverse, sous-méthylées et constituent ce qui est appelé des îlots CpG (GCI) (Fazzari and Greally, 2004). L'ADN méthylé est un élément épigénitique de première importance, conservé (hérité) lors de la réplication de l'ADN (revu par Bird, 2002).

I.1.1.2. Le génome humain

Si l'organisation structurale des génomes eucaryotes représente leurs points communs, sa composition représente leurs différences. Chacun a ses propriétés physiques, ses caractéristiques cytogénétiques, sa composition génétique, son nombre de chromosomes propre (tableau I).

Tableau I. Comparaison de certaines propriétés physiques des génomes de certains organismes

Espèce	règne	C value	Nb de chr.	Nb gènes classe II	Publication de la séquence
ϕ-X 174	virus	5.386	0	10	1977, 125, 687 -Nature
Escherichia coli	procaryote	$4.6\ 10^6$	1	4,377	1997, 277,1453 -Science
Schizosaccharomyces pombe	Eukaryote (levure)	$12.4\ 10^6$	3	4,929	2002, 415, 871 -Nature
Saccharomyces cerevisae	Eukaryote (levure)	$12.4\ 10^6$	16	5,770	1996, 274, 546 –Science
Caenorhabditis elegans	Eukaryote (annélide)	$100\ 10^6$	6	19,000	1998, 282, 2012 –Science
Arabidopsis thaliana	Eucaryote (plante)	$115\ 10^6$	5	25,498	2000, 408, 796 –Nature
Drosophila melanogaster	Eucaryote (invertébré)	$122\ 10^6$	5	13,472	2000, 287, 2185 –Science
Tetraodon nigroviridis	Eucaryote (poisson)	$340\ 10^6$	21	27,918	2004, 431, 946 -Nature
Fugu rubripes	Eucaryote (poisson)	$365\ 10^6$	21	De 20 à 33,000[a]	2002, 297, 1301 –Science
Oryza sativa (riz)	Eucaryote (plante)	$466\ 10^6$	12	de 45 à 55,000	2002, 296, 79 –Science
Mus musculus	Eucaryote (mammifère)	$2.9\ 10^9$	21	~28,000	2002, 420, 520 –Nature
Rattus norvegicus	Eucaryote (mammifère)	$2.5\ 10^9$	22	~24,000	2004, 428, 493 –Nature
Homo sapiens	Eucaryote (mammifère)	$3.3\ 10^9$	24	De 20 à 25,000[b]	2001, Nature & Science
Psilotum nudum	Eucaryote (plante)	$250\ 10^9$	?	?	aucune

Légende :

a : Suite à la publication du génome du Tétraodon, le chiffre de 38,000 a été revu à la baisse

b : Suite à la publication de (International Human Genome Sequencing Consortium, 2004), le chiffre de ~30,000 a été revu à la baisse

I.1.1.2.1. Propriétés physiques

Le séquençage du génome humain représentait, de par sa taille et sa complexité, un véritable défi tant sur le plan technique que logistique. Il aura fallu 10 ans de travail pour arriver à la version « brouillon » de notre génome (Venter et al., 2001; Lander et al., 2001). La publication de ces résultats constitue une étape majeure de la science moderne. Le « boom » du séquençage n'a pas laissé en reste les autres espèces, particulièrement celles dites « modèles », qui elles aussi ont désormais leur génome séquencé (tableau I, pour plus de détails consulter NCBI, section Genome*). Plusieurs surprises sortirent de l'analyse de notre génome. L'exemple le plus frappant fût l'estimation revue à la baisse du nombre possible de gènes encodés dans notre génome, passant de 100,000 à 30,000, puis récemment entre 20 et 25,000 (International Human Genome Sequencing Consortium, 2004). Il est aussi intéressant de voir que notre génome ne semble ni contenir le plus de gènes (comparé au riz) ni avoir la plus grande taille (comparé à *P. nudum* et à de nombreux amphibiens). Il est aussi intéressant de noter que l'augmentation de la taille d'un génome (*C-value*) ne corrèle ni avec la quantité de gènes dans ce génome ni avec l'évolution de l'organisme (*C-value paradox*) (revu par Hartl, 2000). Ainsi par rapport à notre génome, celui du fugu est très riche et très compact, alors que celui de *P. nudum* est bien plus pauvre et plus vaste. Il est surprenant de voir une telle chose pour cette plante évolutivement moins avancée (ni feuille, ni fleur) que *A. thaliana* et qui pourtant, possède un génome 2000 fois plus grand. Avant d'étudier en profondeur le génome humain, il est important de le concevoir tel qu'il est dans sa conformation naturelle, à l'intérieur du noyau sous la forme de chromosomes indépendants

---

* : http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome

I.1.1.2.2. Les chromosomes humains

Comme précédemment indiqué (tableau I), le génome des eucaryotes est compartimenté en chromosomes, un ou plusieurs, qui possèdent eux aussi leurs caractéristiques propre (taille, position du centromère, densité de gène, caractéristiques cytogénétiques). Afin d'illustrer plus précisément ces différences chez l'humain, les principales données chromosomiques connues à ce jour ont été rassemblées dans un tableau (tableau II) (basées sur la version *build.34d* du génome humain daté de mai 2004[¶]). Outre leur taille, une des différences les plus marquantes, sinon des plus visibles, à propos des chromosomes, apparaît lors de leur coloration au Giemsa (Bickmore and Sumner, 1989; Miklos and John, 1979). Les chromosomes apparaissent alors zébrés et les bandes refléteraient l'état de condensation de la chromatine, du moins condensé (euchromatine) au plus condensé (hétérochromatine) (tableau II, du clair au foncé). Ces bandes ont été nommées en fonction de leur composition nucléique et génique présumée. Ainsi, la bande C serait constituée d'hétérochromatine constitutive (représentant 20% du génome) et les bandes G-, R- et T- auraient un pourcentage de G+C et une densité de gènes croissants (Bickmore and Sumner, 1989; Miklos and John, 1979). L'analyse nucléique du génome, après son séquençage (Lander et al., 2001; Venter et al., 2001), ne révèle pas de frontière aussi nette dans la répartition des gènes que ne le fait la coloration au Giemsa. Une autre hypothèse est l'organisation du génome en « isochores » (Bernardi, 2000).

[¶] : basé sur http://www.ensembl.org/Homo_sapiens/34dbuild.html

Tableau II. Les propriétés cytogénétiques et génomiques des chromosomes humains.

Ch.	Taille Mb	Gènes trouvés	Densité par Mb	Représentation schématique des chromosomes humains
1	246048	2311	9.39	
2	243416	1573	6.46	
3	199289	1242	6.23	
4	191722	944	4.92	
5*	181015	1110	6.13	
6*	170912	1239	7.25	
7*	158546	1165	7.35	
X	153692	957	6.23	
8	146309	842	5.75	
9*	136372	924	6.78	
10*	135037	898	6.65	
11	134483	1508	11.21	
12	132018	1169	8.85	
13*	113028	425	3.76	
14*	105261	780	7.41	
15	100257	746	7.44	
16	90037	997	11.07	
17	81740	1287	15.74	
18	76115	338	4.44	
19*	63807	1426	22.35	
20*	63692	676	10.61	
Y*	50287	117	2.23	
22*	49374	563	11.40	
21*	46976	292	6.22	
Random	130000			
Total	3199435	23529	7.35	

Légende :
Mb : Mégabase
* : Chromosome dont le séquençage complet a été réalisé depuis 2001. Voir Human Genome Gateway http://www.nature.com/nature/focus/humangenome/
: chromosome riche en gènes
: chromosome pauvre en gènes
: densité croissante de gènes par régions chromosomiques (noir pour faible et blanc pour forte)

Les isochores se définissent par rapport à leur composition nucléique (G+C) sur des fragments de chromatine supérieure à 300kb. Mais là encore, le modèle est valide pour certains chromosomes (ceux riches en gènes : 17,19,22 et ceux pauvres en gènes : X,4,13,18,Y) mais pas tous (ch8 et 15). Récemment, une étude (Gilbert et al., 2004b) est venue relancer la question de l'activité transcriptionnelle dans la chromatine (eu- ou hétéro-), en démontrant que l'hétérochromatine (bande C) n'est pas forcement un lieu où la transcription est absente et inversement, tous les gènes ne sont pas activement transcrits dans l'euchromatine.

I.2. Le gène type de classe II et son organisation

Au cours des dernières décennies, le modèle théorique d'un gène s'est étoffé, passant d'une simple séquence transcrite (ou pas) à une séquence génomique possédant de multiples régions spécifiques (régions régulatrices, promoteur, site d'initiation de la transcription, exons, introns, 3' UTR).

I.2.1. Régions régulatrices

« Comment les activateurs régulent-ils la transcription génique ?», voici une question qui fut vigoureusement débattue pendant plus d'une décennie (Ptashne, 1988). Classées dans la catégorie des éléments fixes de régulation, de courtes séquences d'ADN spécifiques disséminées dans le génome modulent la transcription en servant de point d'encrage aux facteurs transcriptionnels (TF), éléments mobiles de régulation. Elles permettent la régulation de la transcription à un lieu plus ou moins précis du génome.

I.2.1.1. Le promoteur basal

Le promoteur basal est défini comme étant la région minimale, partant de -35 à +35 pb autour du Site d'Initiation de la Transcription (SIT), permettant l'initiation de la transcription par la machinerie transcriptionnelle de base (figure 2).

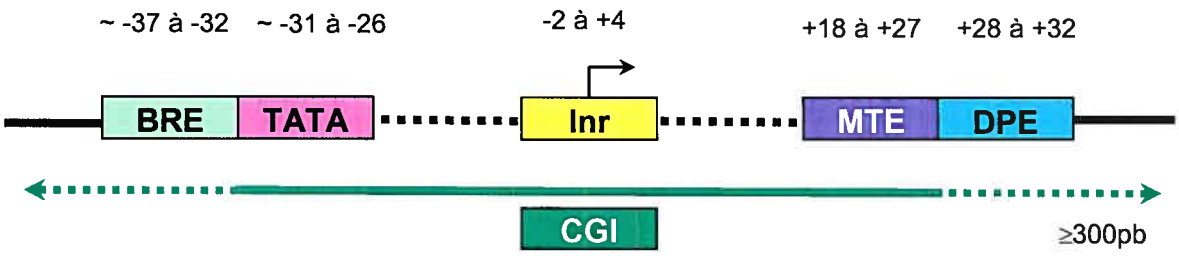


Figure 2. Organisation du promoteur basal eucaryote pour un gène de classe II.

Le promoteur basal est composé d'une combinaison de ces éléments. Les distances présentées sont relatives au +1 de transcription, symbolisé par la flèche. L'ADN double brin est symbolisé par la bande noire alors que la verte représente un îlot CpG potentiel. BRE, *TFIIB recognation element* ; TATA, boite TATA ; Inr, initiateur ; MTE, *Motif ten elements* ; DPE, *downstream promoter element* ; CGI, îlots CpG.

I.2.1.1.1. La boite TATA

Ce fut le premier élément du promoteur basal découvert (Breathnach and Chambon, 1981). Elle peut être comparée à la boite « Pribnow » des procaryotes tout en n'étant pas son homologue. Localisée chez l'homme à environ 30 nucléotides en amont du SIT, sa place peut être plus variable chez d'autres organismes (ex : chez *S. cerevisae* sa position varie de 40 à 100 pb en amont du SIT). Sa séquence, consensus pour TATAAA, est reconnue par TBP (*Tata-box Binding Protein*) (Bushnell et al., 2004). D'abord imaginé comme étant un élément indispensable du promoteur, elle est actuellement trouvée dans une fraction seulement des gènes transcrits par l'ARN Pol II. Chez l'humain sa présence, parmi les gènes de classe II, est estimée à environ 32% (Suzuki et al., 2001), alors que chez la drosophile, elle serait de l'ordre des 43% (Kutach and Kadonaga, 2000). Faisant partie du mégacomplexe TFIID (13 à 15 sous-unités, plus de 1,3 MDa), TBP est le principal peptide à fixer la boite TATA, mais il n'est pas le seul (Muller and Tora, 2004). Récemment mis en évidence, les *TBP Related Factor* (TRF,Hansen et al., 1997), *TBP Like Factor* (TLF, Dantonel et al., 2000), et TFIID-TBPless interviennent aussi dans la transcription au niveau des promoteurs avec ou sans boite TATA (respectivement $TATA^+$ ou $TATA^-$). Des études à grande échelle devraient permettre de caractériser, au niveau génomique, l'implication de ces nouveaux complexes (Muller and Tora, 2004).

I.2.1.1.2. Le BRE

Nommée BRE pour *TFIIB Recognition Element*, cette courte séquence découverte en premier chez l'humain est reconnue et fixée de manière spécifique par TFIIB (Lagrange et al., 1998). Possédant une séquence consensus du type G/C-G/C-G/A-C-G-C-C, elle se situe en amont immédiat de la boîte TATA. Toutefois, elle n'est retrouvée que dans 12% des promoteurs avec boite TATA. Sa (ses) fonction(s) précise(s) reste (nt) à déterminer. Des études tendent à démontrer qu'elle favoriserait l'initiation de la transcription en facilitant le recrutement de TFIIB (Lagrange et al., 1998) alors que d'autres observent un effet négatif sur la transcription, en supprimant son niveau d'expression basal (Evans et al., 2001).

I.2.1.1.3. L'initiateur

Découvert dans les années 80 dans de nombreux organismes eucaryotes, l'Initiateur (Inr) est défini comme étant un élément discret du promoteur basal (Smale and Baltimore, 1989). Il se trouve aussi bien dans les promoteurs avec ou sans boite TATA. Sa séquence consensus, variable dépendamment des espèces (ex : Py-Py-A_{+1}-N-T/A-Py-Py pour les mammifères), englobe le SIT, qui est communément désigné par A_{+1}. L'Inr serait reconnu par de nombreux facteurs. Des études *in vitro* indiquent que TFIID (via TAF1 et TAF2, Chalkley and Verrijzer, 1999) et l'ARN Pol II (Weis and Reinberg, 1997) seraient capable de le reconnaître. Des facteurs spécifiques (TFII-I, YY1 respectivement Grueneberg et al., 1997 et Weis and Reinberg, 1997) semblent capables de s'y fixer et de moduler l'expression des gènes associés. Une étude récente menée chez les eucaryotes primitifs a montré qu'une protéine, IBP39, serait capable de reconnaître et de fixer l'Inr (seul élément présent du promoteur basal) et de recruter l'ARN Pol II au

SIT (Schumacher et al., 2003).

I.2.1.1.4. Le DPE

Initialement découvert chez la drosophile lors de la purification de TFIID, le *Downstream Promoter Element* (DPE) a été depuis découvert chez d'autres espèces. Généralement présent dans le promoteur TATA$^-$, sa présence est aussi répandue que celle de la boite TATA : sur environ 30% des promoteurs de drosophiles testés (Kutach and Kadonaga, 2000). Sa séquence consensus est du type A/G_{+28}-G-A/T-C/T-G/A/C mais elle peut aussi être relativement dégénérée. L'analyse, chez la drosophile, des promoteurs DPE-dépendants a mis en évidence l'importance de l'Inr (toujours présent) et la distance très précise séparant le DPE du SIT (variant de 28 à 32 paires de bases, Kutach and Kadonaga, 2000).

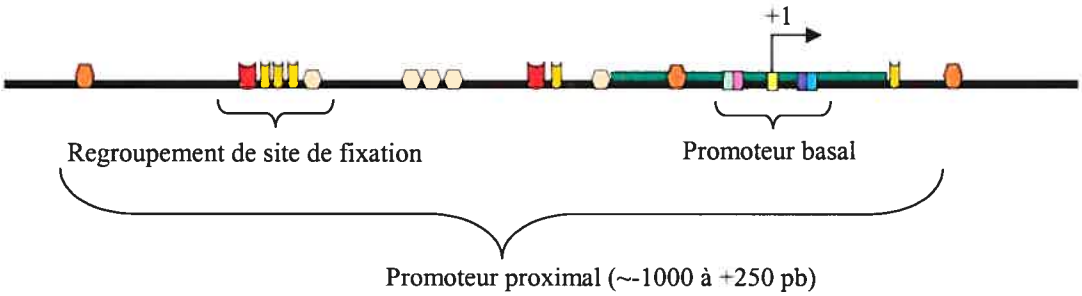
I.2.1.1.5. Le MTE

Découvert cette année (Lim et al., 2004), le *Motif Ten Element* (MTE) est un élément de régulation du promoteur basal de l'ARN Pol II qui semble conservé de la drosophile à l'humain. Positionné en aval du A_{+1} de transcription et couvrant les bases de +18 à +27, il est capable, en présence d'un Inr, d'initier la transcription et ce même sur un promoteur TATA$^-$/DPE$^-$. Toutefois, Lin *et al.* ne présentaient leurs résultats que sur des promoteurs TATA$^-$, il serait donc intéressant d'analyser la fréquence et le rôle de MTE dans le promoteur basal d'une vaste population de gènes.

I.2.1.1.6. Les îlots CpG

Il existe dans le génome des vertébrés, des séquences d'ADN anormalement riches en CG qui ont la particularité d'être sous méthylées par rapport aux CG situés ailleurs dans le génome : ce sont les îlots CpG (Bird, 1986). Ces îlots sont définis comme étant des régions d'ADN supérieures à 300 pb possédant une quantité de G+C supérieure à 50% et un ratio de CG « observé contre attendu » supérieur ou égal à 0.6 (Weinmann et al., 2002). Localisés préférentiellement au site d'initiation des gènes, il a été observé que la plupart des gènes de maintenance (*house keeping gene*) en possède. Compte tenu de leur taille, les CGI englobent généralement un (ou plusieurs) promoteur(s) (Adachi and Lieber, 2002), qui semble d'ailleurs être dépourvu de boite TATA et de DPE. Les CGI semblent contenir des promoteurs « faibles », initiant la transcription sur une centaine de paires de bases contrairement aux promoteurs « forts » (ex : TATA ou DPE) qui le font sur une dizaine de bases. Il est intéressant de noter la proportion relativement importante, au sein des CGI, de sites de fixation possible pour le facteur de transcription Sp1. Ceci pourrait expliquer le maintien hypométhylé des CGI (Macleod et al., 1994), contrairement à la tendance hyperméthylé du couple de dinucléotides CG ailleurs dans le génome (Bird, 1986). Du point de vue transcriptionnel, les sites multiples pour Sp1, en association avec un Inr, seraient capables d'initier la transcription sans TATA ou DPE et créant ainsi des promoteurs CGI^+/Inr^+ (Smale et al., 1990).

A) Représentation linéaire du promoteur proximal d'un gène de classe II



B) Représentation, tel qu'*in vivo*, du promoteur proximal d'un gène de classe II

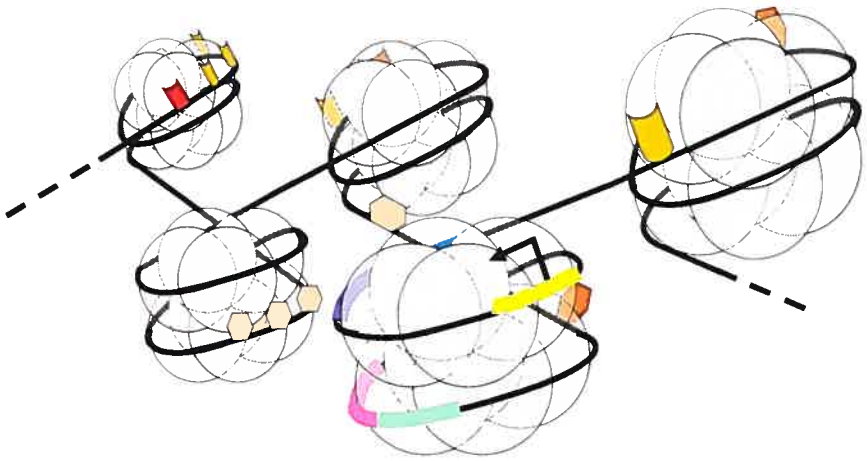


Figure 3. Représentation schématique de l'organisation du promoteur proximal d'un gène de classe II.

(A) Représentation linéaire, sans nucléosome, d'un promoteur d'un gène de classe II. (B) Représentation du même promoteur mais en incorporant les nucléosomes (plus proche de la réalité in vivo). Les différents éléments orangés représentent schématiquement les sites de fixation pour des facteurs de transcription. Les nucléosomes sont représentée par des octamères gris. Pour le promoteur basal, le code de couleur a été conservé par rapport à la figure 2.

I.2.1.2. Le promoteur proximal

Englobant le promoteur basal, le promoteur proximal s'étend de –1000/–250 à +100 pb par rapport au SIT (variable dépendamment des études ou des espèces, levures voir Pokholok et al., 2002; Ren et al., 2000 et humains voir Ren et al., 2002; Odom et al., 2004) (figure 3). C'est lui qui est analysé lorsque l'on parle de « promoteur » au sens large du terme. C'est dans cette portion du promoteur que se retrouve la majorité des sites de liaisons des facteurs de transcription (TF) dits spécifiques au gène (*gene specific*). Dans l'organisation de cette présentation, nous avons opté pour présenter le principe d'action des TFs plutôt que de présenter une liste des TFs. Pour de plus amples renseignements, il existe sur Internet plusieurs banques de données[†] qui compilent de nombreuses informations concernant les TFs (ex : la matrice d'ADN fixée, leurs séquences protéiques, ses domaines fonctionnels, sa structure tridimensionnelle). Brièvement, les TFs sont des modulateurs de la transcription qui possèdent au moins deux domaines caractéristiques et distincts. Le premier domaine sert à positionner le TF sur l'ADN, c'est le domaine de fixation à l'ADN. Le second contacte la machinerie transcriptionnelle pour en moduler l'activité, c'est le domaine d'activation (ou de répression). Ces domaines sont complètement dissociables. Le TF chimérique Gal4-VP16, qui possède le domaine de fixation de Gal4 fusionné à l'activateur de VP16 est un exemple parfait de cette dissociation (Sadowski et al., 1988).

Dans le cas où un TF ne posséderait pas de domaine de fixation à l'ADN, il influence donc la transcription via des interactions uniquement protéines/protéines et

---

[†] : TRANSFAC, www.gene-regulation.com

devient alors ce que l'on appelle un co-activateur (ou co-répresseur).

I.2.1.3. Les modules de régulation "longue distance" : enhancers, silencers et LCR

Que ce soit dans le promoteur où ailleurs dans le génome, les sites de fixation pour les facteurs de transcription sont souvent groupés (figure 3). L'hypothèse actuelle serait que le regroupement permettrait une synergie dans l'efficacité d'action des TFs s'y fixant. Les *enhancers* et les *silencers* (respectivement activant et réprimant la transcription) n'échappent pas à cette règle. Ces deux éléments ont la particularité d'agir de manière indépendante de l'orientation du gène, c'est-à-dire aussi bien en amont qu'en aval du gène régulé. Mais le plus surprenant c'est qu'ils sont capables d'agir à des dizaines voir des centaines de kilobases (kb) du gène régulé (Bejerano et al., 2004). L'incroyable conservation de ces éléments au cours de l'évolution pourrait être expliquée par un rôle régulateur au cours du développement. Compte tenu de la distance pouvant les séparer de leur cible, la recherche et la caractérisation des *enhancers* constituent un défi en soi. L'analyse comparative de génomes se révèle déjà un outil précieux dans cette voie (ex : comparaison du génome du fugu au notre, voir Bejerano et al., 2004).

Il existe aussi des régions dans le génome qui ont la propriété de pouvoir moduler la transcription en influençant la structure de large portion de chromatine : les LCR (*Locus Control Region*). Ces régions semblent capables d'activer ou de réprimer la transcription (Grosveld, 1999; Fraser and Grosveld, 1998). Leurs mécanismes d'action ne sont pas encore clairs et leurs interconnections avec les *enhancers* et les *silencers* sont encore méconnues.

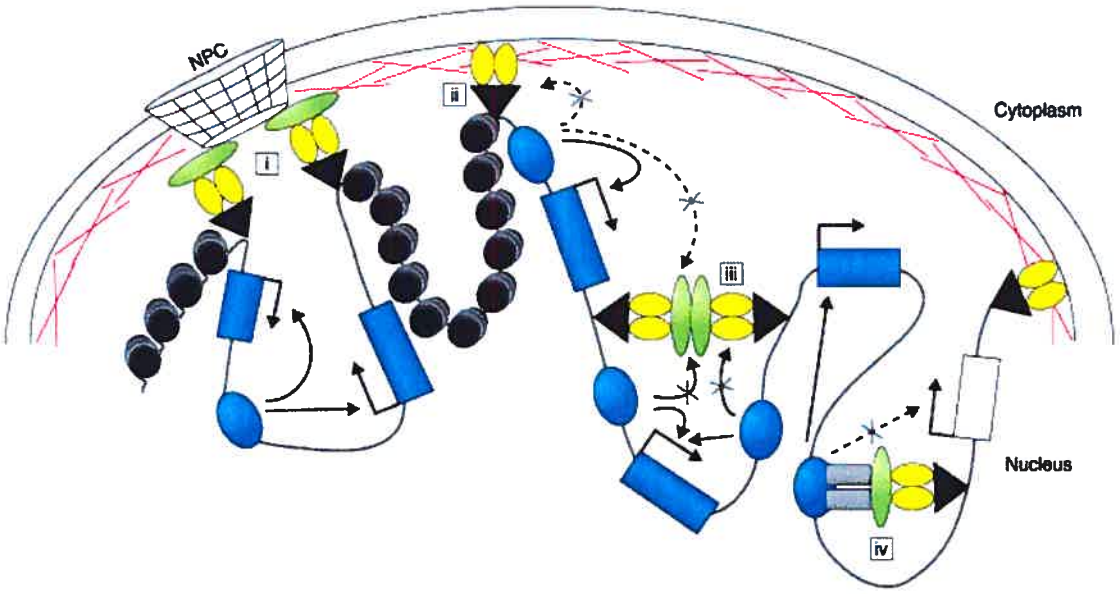


Figure 4. Schématisation de la topologie de la chromatine au sein du noyau d'une cellule eucaryote.

Représentation des liaisons possibles entre les isolateurs et la topologie de la chromatine dans le noyau des cellules eucaryotes. Différents domaines chromatiniens, en forme de boucle sont représentés et peuvent former par exemple une connection entre MAR/Isolateur et les pores nucléaires (NPC) [i]. Fixation de la chromatine sur la lamina nucléaire [ii] (par des MAR) ou à elle-même via des isolateurs entre-eux [iii] ou avec d'autres facteurs liant la chromatine [iv]. Dans tous les cas la formation de ces domaines architecturaux limite dans l'espace l'action des facteurs régulant la transcription ainsi que la propagation des *silencers* de la chromatine.

Les gènes actifs sont en bleu ; les triangles noirs représentent soit des MARs soit des isolateurs ; les ovales jaunes et verts symbolisent des complexes protéiques ; les rectangles gris, de possibles *enhancers* ; la chromatine silencieuse est représentée par un collier de perles noires (tiré de Kuhn and Geyer, 2003).

I.2.1.4. Compartimentation de la chromatine : les isolateurs et les MARs

Finalement, à l'intérieur du noyau des cellules eucaryotes, chaque gène se retrouve dans une mer de chromatine, foisonnante de régions régulatrices. Comment un gène peut-il suivre son profil d'expression spécifique sans subir les contraintes de son voisinage (Chambeyron and Bickmore, 2004) ? Il existe dans l'ADN des régions spécifiques qui servent à « isoler » certaines portions de la chromatine (Zhou and Berger, 2004 voir aussi la figure 4). Les mécanismes employés pour démarquer l'ADN ne sont pas encore connus, toutefois ils semblent agir principalement à deux niveaux (revu par Bell et al., 2001). Le premier serait au niveau transcriptionnel, où les isolateurs (*insulators*) formeraient une véritable frontière empêchant les TFs/enhancers/silencers d'agir au-delà. Indirectement ils protègent aussi le(s) gène(s) des effets de la chromatine qui l'(es) entoure(nt) (West et al., 2004). En second lieu, les MARs (*Matrice Attachment Region*) seraient capables d'organiser la chromatine dans le noyau, de lui donner une topologie particulière (en boucle, figure 4), ce qui aurait pour effet de compartimenter certaines régions du génome (eu- et hétérochromatine, Kuhn and Geyer, 2003). Des études supplémentaires sont nécessaires pour établir les relations précises entre enhancers, silencers, LCR, insulators et MARs.

I.2.2. Régions transcrites par l'ARN Pol II

À proprement parlé, c'est tout ce que transcrit l'ARN Pol II du site d'initiation au site de terminaison d'un gène. Dans un souci de clarté cette séquence a été nommée TDS (*Transcribed DNA Sequence*) par opposition au CDS (*Coding DNA Sequence*) trouvé dans les bases des données (type NCBI). Le TDS est de taille extrêmement variable chez l'humain et peut se subdiviser en 3 sous-parties : le 5'UTR, le cadre de lecture (ORF,

Open Reading Frame) et le 3'UTR du gène. Il est important de préciser qu'à l'origine la séquence d'un gène était obtenue par PCR inverse (*reverse transcriptase PCR*) sur l'ARN mature (épissé), ce qui avait pour conséquence de ne donner que les séquences strictement codantes d'un gène (sous forme d'un ADN complémentaire, cDNA). Depuis, des méthodes ont été développées afin de rechercher spécifiquement le premier nucléotide transcrit (ex : oligocapping, 5'SAGE, voir respectivement Maruyama and Sugano, 1994; Maruyama and Sugano, 1994; Wei et al., 2004).

I.2.2.1.1. Le 5'UTR

Le 5' *UnTranslated Region* est la première partie transcrite du gène et comme son nom l'indique, elle n'est pas traduite. Elle marque la différence entre le +1 de transcription (le SIT) et le +1 de traduction (codon *start*, ATG codant pour une méthionine). Cette région est, bien entendu, de taille extrêmement variable puisqu'elle dépend de la distance SIT/ATG qui peut varier en fonction de la présence d'un ou plusieurs SITs pour un même gène (SITs alternatifs, revu par Black, 2003). C'est aussi dans cette portion du gène que se fait une partie de la régulation de l'initiation de la transcription.

I.2.2.1.2. Le cadre de lecture (ORF)

Il est possible de définir cette région comme étant celle qui englobe la séquence codante du gène, délimitée par les codons « *start* » et « stop ». Par exemple, chez l'humain les gènes possèdent de nombreux exons entrecoupés d'introns alors que ces derniers sont rares chez la levure *S. cerevisae*. Lors de la transcription, les introns sont transcrits en ARN mais ne codent pas pour des protéines (épissé par la suite). En

terme de chiffres, un gène de classe II humain fait en moyenne 27,000 pb dont seulement 1500 sont réellement codantes (Lander et al., 2001; Venter et al., 2001). Donc dans le génome humain, il est actuellement estimé que 25% serait transcrit par l'ARN Pol II et que seulement 1.5% coderait pour des protéines. Toutefois, de plus en plus d'études tendent à indiquer qu'il y a une masse considérable d'ARN transcrit mais non codant (Bertone et al., 2004).

I.2.2.1.3. Le 3'UTR

Cette région est la moins bien caractérisée et caractérisable des trois car elle ne possède qu'un seul point fixe, le codon « stop ». En effet, celui-ci marque la fin de la partie codante du gène, mais l'ARN Pol II continue sa transcription ce qui permet, entre autre, la polyadénylation (poly(A),Colgan and Manley, 1997). Même basée sur des études à l'échelle du gène, la taille de cette région reste floue. Des recherches à grande échelle nous permettront sûrement de mieux caractériser cette région et peut-être de mettre en évidence de nouveaux mécanismes de régulation (ex : interférence transcriptionnelle chez la levure, Martens et al., 2004).

I.3. L'expression d'un gène, un processus dynamique hautement régulé

Le remodelage de la chromatine, la reconnaissance du promoteur, l'assemblage de la machinerie transcriptionnelle, l'initiation, l'élongation, la terminaison, le recyclage sont autant d'étapes liées entre elles, hautement régulées et impliquées directement dans la transcription. Toutes ensemble, elles composent le cycle transcriptionnel (Svejstrup, 2004). Seul une coordination précise de toutes ces étapes, incluant les machineries de maturation des ARNs, pourra conduire à la transcription efficace d'un gène en ARN

messager (revu par Lee and Young, 2000).

I.3.1. Création d'un environnement favorable à la transcription

Chez les eucaryotes, l'organisation de la chromatine se révèle la première barrière à franchir dans l'expression génique car elle empêche la reconnaissance du promoteur par la machinerie transcriptionnelle.

À l'aube de l'étude de la transcription, lorsque la chromatine était largement ignorée, un activateur reconnaissait le promoteur et activait l'initiation de la transcription en recrutant des protéines au SIT. La vision contemporaine voit l'(es) activateur(s) pionnier(s) recruter d'abord des complexes de remodelage de la chromatine afin de créer un environnement propice à la formation du complexe de pré-initiation (PIC). Comme il a été présenté jusqu'à maintenant, la chromatine peut être altérée par 1) remplacement des histones principales par des variants (voir II.1.1.1.1) 2) modifications covalentes des queues N-terminales des histones (code des histones, II.1.1.1.3) et enfin (et surtout) 3) remodelage ATP-dépendant des nucléosomes. Des complexes spécifiques dédiés à cette tâche sillonnent la chromatine et déplacent les nucléosomes sur de courtes distances. Les trois familles de complexes de remodelage de la chromatine ATP-dépendant sont : SWI/SNF (majoritaires), ISWI et Mi-2 (revu par Lusser and Kadonaga, 2003). Les mécanismes avec lesquels ils déstabilisent les nucléosomes ainsi que leurs cofacteurs sont différents, ce qui sous-entend une régulation spécifique au niveau génomique.

Plusieurs exemples peuvent être cités, comme le cas du gène de la levure *HO* (Fry and Peterson, 2001; Krebs et al., 1999) sur lequel l'activateur Swi5p recrute SWI/SNF dont l'activité est indispensable au recrutement de Gcn5p (HAT). S'ensuivent

le recrutement de SBF puis de l'ARN Pol II et ses GTFs. Chez l'homme, *IFN-β* (Agalioti et al., 2000) implique en premier un enhanceosome (Merika and Thanos, 2001; Thanos and Maniatis, 1995) qui recrute Gcn5p (HAT), ce qui permet l'acétylation des histones permettant à SWI/SNF de venir (probablement) stabiliser le tout. TBP reconnaît et fixe la boite TATA, ce qui conduirait à la formation du PIC. Enfin, en analysant l'expression de α_1-AT, il a été démontré que HNF-1, TBP, l'ARN Pol II et ses GTFs sont déjà présents sur le promoteur, en présence de nucléosomes, mais en pause transcriptionnelle (Soutoglou and Talianidis, 2002). La transcription n'est effective que lorsque des enzymes de remodelage (1 SWI/SNF-like et 2 HATs) sont recrutées.

Aujourd'hui, il est clair que les enzymes de remodelage de la chromatine peuvent réguler toutes les étapes du cycle transcriptionnel (Svejstrup, 2004). Finalement, savoir ce qui détermine la localisation de ces enzymes sur les gènes reste une question à élucider.

I.3.2. La machinerie transcriptionnelle lors de l'initiation de la transcription

L'ARN Pol II eucaryote seule est capable, in vitro, de synthétiser de l'ARN à partir d'ADN, mais elle est incapable de reconnaître un promoteur (Chambon, 1975). Il lui faut pour cela l'aide d'autres facteurs. Lors de l'initiation, la machinerie transcriptionnelle est composée des 12 sous-unités de l'ARN Pol II, des facteurs généraux de transcription et de complexe(s) multiprotéique(s) (coactivateurs et/ou médiateurs). L'holoenzyme, formée de la plupart de ces éléments, a été purifiée aussi bien chez la levure que chez l'humain (Koleske and Young, 1994; Kim et al., 1994). Ce complexe « tout en un » suggère la possibilité de son recrutement au promoteur en une seule étape.

Tableau III. Les douze sous-unités de l'ARN polymérase II humaine.

Sous-Unité	Masse (kDa)	Caractéristiques
RPB1	220	Contient le CTD ; orthologue β' de l'ARN polymérase procaryote ; fixe l'ADN ; impliquée dans la sélection du SIT.
RPB2	140	Contient une partie du site actif; orthologue β de l'ARN polymérase procaryote ; impliquée dans la sélection du SIT et dans l'élongation.
RPB3	33	Orthologue, avec RPB11, du dimère α de l'ARN polymérase procaryote.
RPB4	16	Sous-complexe avec RPB7; impliquée dans la réponse au stress.
RPB5	23	Commune aux ARN polymérases I, II et III ; cible des activateurs transcriptionnels.
RPB6	14.1	Commune aux ARN polymérases I, II et III ; impliquée dans l'assemblage et la stabilité.
RPB7	19	Sous-complexe avec RPB4 fixé préférentiellement durant la phase stationnaire.
RPB8	17	Commune aux ARN polymérases I, II et III ; possède un domaine de fixation pour les oligonucléotides/oligosaccharides.
RPB9	14.5	Pourrait être impliquée dans l'élongation ; impliquée dans la sélection du SIT.
RPB10	7.6	Commune aux ARN polymérases I, II et III
RPB11	13.3	Orthologue, avec RPB3, du dimère α de l'ARN polymérase procaryote
RPB12	10	Commune aux ARN polymérases I, II et III
ARN Pol II	497.5	Transcrit les gènes codant pour les protéines

I.3.2.1. L'ARN Pol II et son CTD

L'ARN Pol II, souvent considérée comme une entité unique, est composée de 10 à 12 sous-unités (tableau III) qui sont remarquablement bien conservées de la levure à l'humain, au point pour certaines d'être encore fonctionnelles si elles sont interchangées. L'ARN Pol II eucaryote partage avec son homologue bactérien de nombreuses caractéristiques (Zhang et al., 1999; Cramer, 2002). L'ARN polymérase bactérienne est pour sa part composée de quatre sous-unités formant l'enzyme (2α, β et β') et d'un

facteur de spécificité (σ) (Zhang et al., 1999; Cramer, 2002). Les deux plus grosses sous-unités chez les eucaryotes, Rpb1 et Rpb2, sont les homologues de β et β' chez les bactéries et Rpb3/Rpb11 partagent une certaine homologie avec la sous-unité α (Larkin and Guilfoyle, 1997). L'obtention de la structure cristallographique, à haute résolution (2.8Å et 3.3Å, voir respectivement Cramer et al., 2001; Gnatt et al., 2001), de l'ARN Pol II eucaryote a mis à jour de nombreuses caractéristiques pertinentes. Ainsi les comparaisons structurales entre les ARN polymérases eucaryotes et procaryotes ont révélé une très forte similarité dans l'organisation des sous-unités. Les cristallographies ont permis aussi de proposer de nouvelles implications fonctionnelles pour des sous-unités (Cramer, 2002), ce qui a donné lieu à des études de caractérisation fonctionnelle afin de mieux connaître ces domaines (Jeronimo et al., 2004). Parmi les sous-unités de l'ARN Pol II, Rpb4 et Rpb7 forment un sous-complexe dissociable du reste de l'enzyme (Choder, 2004). Ce sous-complexe est impliqué plus spécifiquement dans la réponse au stress et lors de l'initiation de la transcription (Choder and Young, 1993). Ceci explique l'existence des deux formes d'ARN Pol II lors des purifications (10 et 12 sous-unités).

Une caractéristique unique de l'ARN Pol II eucaryote est la présence d'un heptapeptide répété en tandem ($Y_1S_2P_3T_4S_5P_6S_7$) se trouvant en C-terminal de la plus grosse sous-unité Rpb1 : le CTD (*Carboxy Terminus repeat Domain,* Prelich, 2002). Le nombre des répétitions est très variable, passant de 26 chez *S. cerevisiae* à 52 chez l'homme (Corden, 1990). La présence du CTD est vitale aussi bien pour les levures que pour les organismes pluricellulaires, bien qu'*in vitro* elle semble dispensable Prelich, 2002. Les fonctions du CTD sont intimement liées à son degré de phosphorylation : non phosphorylé (forme IIA de l'ARN Pol II, dite hypophosphorylé) dans les complexes

d'initiation (Lu et al., 1991) ou phosphorylé (forme IIO de l'ARN Pol II, dite hyperphosphorylé) dans ceux d'élongation (Komarnitsky et al., 2000). Le rôle du CTD lors de l'élongation sera expliqué plus en détails dans la section s'y reportant (section II.3.3).

I.3.2.2. Les facteurs généraux de la transcription

Les facteurs « généraux » requis pour la reconnaissance *in vitro* du promoteur par l'ARN Pol II sont TFIIA, IIB, IID, IIE, IIF et IIH (voir tableau III). Malgré leur appellation de « généraux », il n'est pas encore clair qu'ils soient vraiment tous et toujours requis *in vivo* (Holstege et al., 1998). Si le modèle d'assemblage « étape par étape » du complexe de préinitiation (PIC) tel qu'établit *in vitro* semble différent de celui *in vivo*, ces études auront permis la caractérisation fonctionnelle de ces facteurs (revu par Lee and Young, 2000).

(TBPs, TAFs). C'est sous forme de dimère que TBP fixe la boite TATA, induisant une courbure significative de l'ADN (Chasman et al., 1993). TBP seul est capable d'initier la transcription lors d'essai *in vitro*, si la matrice d'ADN possède une boite TATA. Lors de la purification de TBP, de nombreux peptides co-fractionnèrent avec lui et furent nommés les *TBP Associated Factors* (TAFs) (Tanese et al., 1991; Dynlacht et al., 1991). L'association de TBP et des TAFs constitue le complexe TFIID (revu par Matangkasombut et al., 2004). Comme il a été présenté plus haut, lors de la définition du promoteur basal, en plus de leur association avec TBP, les TAFs peuvent reconnaître et fixer activement le promoteur basal et ce avec ou sans TBP (revu par Muller and Tora, 2004). Dépendamment des éléments présents dans le promoteur basal,

les TAFs sont capables de moduler la fonction de TBP (activer ou réprimer et/ou de reconnaître sélectivement un promoteur (section II.2.1.1).

(TFIIA). TFIIA fonctionne en partie en fixant TBP et en stabilisant l'interaction TBP-ADN (Dion and Coulombe, 2003). Il semble aussi être capable d'empêcher la répression de certains répresseurs transcriptionnels associés au complexe TFIID (Bleichenbacher et al., 2003; Ozer et al., 1998).

(TFIIB). TFIIB est impliqué dans la sélection (voir section promoteur basal), la fixation et la courbure du promoteur (Bushnell et al., 2004). Des études ont démontré que lorsque TFIIB est muté, il y a décalage du SIT et une perte de l'interaction avec l'ARN Pol II (Fairley et al., 2002). Une étude très récente (Chen and Hahn, 2004) propose un modèle de PIC dans lequel TFIIB permettrait d'orienter l'ADN et de sélectionner le site d'ouverture de l'ADN. Cette étude montre aussi une étroite collaboration entre IIB et IIF lors de la formation du PIC.

(TFIIF). TFIIF ressemble en de nombreux points au facteur σ des bactéries. Étroitement lié à l'ARN Pol II, il supprime les interactions d'ADN non spécifiques et stabilise le PIC. Il semble aussi affecter la topologie de l'ADN contribuant à son enroulement autour du PIC (Chasman et al., 1993; Robert et al., 1998; Coulombe and Burton, 1999). TFIIF peut aussi stimuler l'élongation en supprimant les pauses lors de la transcription (Elmendorf et al., 2001).

(TFIIE). TFIIE est fonctionnellement lié à TFIIH. Dans le modèle d'assemblage « par étape », l'arrivée de TFIIE suit celle de l'ARN Pol II et précède celle de TFIIH. Il active les fonctions kinases et ATPases de TFIIH (respectivement sur le CTD et sur l'ouverture de l'ADN, voir Watanabe et al., 2003). TFIIE est reconnu comme étant

capable de fixer de l'ADN simple brin ce qui suggère qu'il pourrait avoir comme un rôle dans l'ouverture, ainsi que le maintien de cette ouverture du promoteur (Forget et al., 2004).

(TFIIH). TFIIH, est un complexe multiprotéique composé de 10 sous-unités; la $10^{ème}$ sous-unité a été récemment trouvée chez la levure et l'humain, respectivement (Ranish et al., 2004; Giglia-Mari et al., 2004). Plusieurs maladies humaines ont pour origine des mutations au sein de ce complexe (Egly, 2001). Il est impliqué dans la transcription et la réparation de l'ADN (Zurita and Merino, 2003). Le complexe IIH possède au moins trois activités enzymatiques : ATPase dépendante de l'ADN, hélicase dépendante de l'ATP et kinase dépendante du CTD (Coin and Egly, 1998; Douziech et al., 2000). TFIIH, via son hélicase XPB, est impliqué i) dans l'ouverture de l'ADN in vitro ii) lors du passage de l'initiation avortive à l'élongation iii) dans la prévention des pauses précoces lors de la sortie du promoteur proximal. Via son domaine kinase (Cdk7/cyclinH), TFIIH est capable de phosphoryler le CTD de l'ARN Pol II ainsi que des protéines impliquées dans la régulation du cycle cellulaire, jouant alors le rôle d'activateur de kinase CDK (Chen et al., 2003). Des études ont aussi montré un lien entre transcription et réparation de l'ADN. Ainsi, si l'hélicase XPB est impliquée dans la transcription, l'hélicase XPD joue un rôle dans la réparation de l'ADN du type *Nucléotide Excision Repair* (NER) (Coin and Egly, 1998; Winkler et al., 2000; Douziech et al., 2000).

Tableau IV. La composition de l'holoenzyme

Facteur	Nb. de sous-unités	Taille du complexe (kDa)	Fonctions
ARN Pol II	12	497.5	Transcrit les gènes codant pour les protéines en ARN.
TFIID	13 à 15	TBP (38) TAF (1014)	Incluant TBP, reconnaissance et fixation du promoteur; possède une activité HAT
TFIIA	3	69	Fixe TBP; stabilise le complexe TBP-ADN
TFIIB	1	35	Impliqué dans la sélection du SIT et dans la courbure de l'ADN
TFIIE	2	90	Impliqué dans le recrutement et la stimulation de TFIIH; impliqué dans l'ouverture du promoteur.
TFIIF	2	84	Dimère d'hétérodimère. Contribue à l'enroulement de l'ADN autour du PIC. Supprime les pauses précoces de la polymérase.
TFIIH	10	475	Formé de deux sous-complexes, il est impliqué dans l'ouverture du promoteur, la phosphorylation du CTD mais aussi dans la réparation de l'ADN.
Mediateur	Combinaisons parmi 30	De 500 (CRSP) à 2000 (ARC-L)	Module l'initiation de la transcription en fonction des activateurs/represseurs transcriptionnels.
Holoenzyme	40 à 54	2800 à 4300	Machinerie transcriptionnelle capable de répondre aux stimulations de son environnement

I.3.2.3. Le médiateur

En utilisant des GTFs et de l'ARN Pol II hautement purifiés, la transcription *in vitro* n'est pas à son niveau maximum par rapport aux essais utilisant des extraits nucléaires. Il manque le médiateur (Malik and Roeder, 2000). La fonction actuellement connue du médiateur est de centraliser et de transmettre à la machinerie transcriptionnelle (ARN Pol II et GTFs) les signaux provenant des facteurs de transcription. En somme, il agirait comme un neurone : recevant de multiples « input » et ne donnant qu'un seul « output » (Kuras et al., 2003). Le médiateur a été purifié aussi bien chez la levure que chez l'humain, mais la notion d'un complexe unique a été abandonnée au profit des observations relatant plusieurs sous-complexes (Malik and Roeder, 2000). Leurs compositions semblent varier dépendamment des gènes considérés, des étapes du cycle

cellulaire, des conditions de l'environnement. Récemment, la nomenclature des sous-unités a été standardisée (Bourbon et al., 2004).

I.3.3. La machinerie transcriptionnelle lors de l'élongation

De nombreuses évidences indiquent que le passage de l'initiation à l'élongation implique i) la phosphorylation du CTD de l'ARN Pol II (Komarnitsky et al., 2000) ii) une stabilisation du complexe transcriptionnel (Ujvari et al., 2002) et iii) l'échange des facteurs et cofacteurs associés à la polymérase (Komarnitsky et al., 2000; Proudfoot et al., 2002; Pokholok et al., 2002).

I.3.3.1. Transition de l'initiation à l'élongation

La transition pour l'ARN Pol II de l'initiation à l'élongation processive se passe en plusieurs étapes critiques (Dvir, 2002). Une fois assemblé sur le promoteur du gène, le PIC est capable d'ouvrir l'ADN (12 à 15pb) créant ainsi une « bulle de transcription » (Coulombe and Burton, 1999). S'ensuit alors un cycle d'initiation dit « avortif » au cours duquel l'ARN Pol II synthétise un court transcrit d'ARN (<10pb), puis s'arrête, recule et recommence la transcription (Goodrich and Tjian, 1994). Lorsque le transcrit dépasse la longueur de 10 pb, la polymérase quitte le promoteur et entre alors en élongation (Spitalny and Thomm, 2003). Passé cette première difficulté, l'ARN Pol II en élongation connaît une deuxième période d'instabilité qui dure tant que le transcrit n'atteint pas environ 50 nucléotides. Cette étape peut aboutir à une pause et au décrochage de l'ARN Pol II (Ujvari et al., 2002). Selon cette étude, cette pause impliquerait la stabilisation du complexe transcriptionnel plutôt que la proximité du promoteur.

Le CTD, de part sa longueur, peut être positionné à une distance considérable du

cœur enzymatique de l'ARN Pol II (Cramer et al., 2001). Il sert d'échafaudage à de nombreuses protéines impliquées aussi bien dans la transcription que dans la maturation des ARNs (Maniatis and Reed, 2002). Avant l'initiation, hypophosphorylé, il sert de support au médiateur. Après l'initiation, rapidement hyperphosphorylé, le médiateur se détache et de nouvelles protéines sont recrutées (incluant la machinerie de maturation des ARNs).

La phosphorylation sur le CTD est localisée sur certains acides aminés et reflète la position relative de la polymérase sur le gène (Komarnitsky et al., 2000). Ainsi les sérines 5 sont phosphorylées en 5'du gène (par Cdk7 membre du complexe TFIIH, Lu et al., 1992) alors que les sérines 2 phosphorylées sont retrouvées en 3' du gène (exemple : Cdk9 faisant partie de P-TEFb, Shim et al., 2002). La phosphatase la mieux caractérisée actuellement est FCP1 (Kamenski et al., 2004). La forme humaine de cette protéine est capable de déphosphoryler à la fois les sérines 5 et 2 alors que la forme fongique (chez *S. pombe*) a une préférence pour les sérines 2 (Lin et al., 2002). Récemment découvert chez l'homme, SCP1 (Small CTD Phosphatase 1) aurait quant à elle une préférence pour la déphosphorylation des sérines 5 (Yeo et al., 2003).

L'émergence du CTD au cours de l'évolution reflète sûrement un besoin pour l'ARN Pol II d'un système de régulation conjoint entre transcription et maturation des ARNs Proudfoot et al., 2002. Par exemple, le spliceosome (machinerie d'épissage des ARNm) interagit avec le CTD et stimule la transcription *in vitro* (Fong and Zhou, 2001; Jurica and Moore, 2003). La mutation chez la levure du facteur Spt5 (complexe Spt4-5), fixant normalement l'enzyme de coiffage des ARN, entraîne l'accumulation d'ARN non épissés (pré-ARNm) (Lindstrom et al., 2003).

I.3.3.2. Les facteurs d'élongation

Un très grand nombre (et en constante augmentation) de facteurs possède la capacité d'influencer l'élongation transcriptionnelle. Parmi les facteurs influençant le départ du promoteur, peuvent être cités ceux qui ont un effet négatif tels que « Factor 2 » qui favorise un décrochage de l'ARN, (Xie and Price, 1996), DSIF et NELF qui confèrent une sensibilité à DRB, inhibant la synthèse d'ARNm et la phosphorylation du CTD (Xie and Price, 1996; Dubois et al., 1994). Ceux qui ont un effet positif regroupent TFIIF, TFIIH, P-TEFb. P-TEFb a pour fonction de contrer l'effet négatif de DSIF et NELF (Yamaguchi et al., 1998). L'homologue de Cdk9 (complexe P-TEFb) chez l'humain s'avère être un cofacteur de la protéine Tat du VIH qui stimule l'élongation en recrutant le facteur Tat-SF1 (Zhou and Sharp, 1996).

Une autre classe de facteurs influençant l'élongation son ceux qui augmentent sa processivité, citons par exemple TFIIS, Elongin, ELL, CSB, Tat-SF1 et l'élongateur (*elongator*). Découvert chez la levure à la fin du siècle dernier (Otero et al., 1999), l'élongateur a vu son rôle dans la transcription mis en doute (Pokholok et al., 2002) puis finalement ré-accepté par une étude présentant une variante intéressante de la technique d'immunoprécipitation de la chromatine (ChIP) transformée en immunoprécipitation de l'ARN (RIP) (Gilbert et al., 2004a).

De plus, la machinerie transcriptionnelle progressant sur la chromatine (à travers les nucléosomes), cette dernière doit nécessairement être dans un état permissif. Il convient alors d'ajouter les complexes de remodelage de la chromatine, qu'ils soient ATP-dépendants ou HAT (Bejerano et al., 2004; Ng et al., 2003; Robert et al., 2004).

Enfin, avant que l'ARNm ne soit transporté hors du noyau pour être traduit, il lui

faut subir les étapes de maturation. L'acquisition d'une coiffe en 5', d'une queue polyadénylée (poly(A)) en 3' et l'épissage des introns constituent trois réactions majeures, biochimiquement distinctes, mais toutes liées intimement à la transcription. Aujourd'hui il est accepté que la maturation des ARNm s'effectue non plus de manière post-transcriptionnelle mais co-transcriptionnelle (revu par Proudfoot et al., 2002).

I.3.4. Terminaison et recyclage

La terminaison de la transcription par l'ARN Pol II diffère fondamentalement de l'initiation dans le sens où elle n'est pas faite sur un site précis (à la base près) mais sur une zone plutôt vague en aval du site de polyadénylation (revu par Proudfoot et al., 2002). La terminaison de l'ARN Pol II est la moins bien caractérisée des trois polymérases (Wickens and Gonzalez, 2004). Néanmoins, plusieurs points clés dans ce processus peuvent quand même être cités. Premièrement, il semble que CPSF (*Cleavage Polyadenylation Specific Factor*), impliqué dans la reconnaissance du site de polyadénylation et la coupure du transcrit, s'associe à TFIID suggérant qu'il serait présent sur l'ARN Pol II dès le promoteur (Dantonel et al., 1997). Deuxièmement, une mauvaise phosphorylation du CTD conduit à la terminaison prématurée de la transcription, ce qui souligne une fois de plus l'importance du passage « initiation/départ du promoteur » (Morillon et al., 2003). Et finalement, même si la déstabilisation du complexe PolII/ADN/ARN reste obscure, il semble que la chromatine joue un rôle important (via Chd1, Isw1 ou 2 voir Alen et al., 2002).

Suite au décrochage de la matrice d'ADN, l'ARN Pol II est rapidement recyclée (déphosphorylée, passage de la forme IIO à IIA) afin de participer à un nouveau cycle de

transcription. Il est présumé que la phosphatase Fcp1 est impliquée dans cette étape car sa mutation chez la levure a démontré une accumulation de Pol II hyperphosphorylée (Cho et al., 2001). Les détails du mécanisme du recyclage restent encore inconnus.

I.4. La transcription à l'ère « omique »

Génomique, transcriptomique, protéomique (et d'autre encore[‡]), les mondes « -omiques » polarisent actuellement la communauté biologique. Ces domaines de recherche s'attachent à étudier les entités desquelles ils tirent leur nom (génome, transcriptome et protéome). Le suffixe « -ome » illustre très bien la volonté des biologistes d'étudier désormais les systèmes dans leur globalité (terme en anglais « *wide* »). Certains de ces « nouveaux » domaines, les techniques qui permettent de les étudier ainsi que les questions qu'ils soulèvent, seront brièvement présentés plus bas.

I.4.1. Le séquençage du génome humain, un tournant de la biologie moderne

Le séquençage du génome humain a donné lieu à un véritable duel opposant la recherche publique (Internationnal Human Genome Sequencing Consortium[§]) et la recherche privée (Celera Genomics[**]). La course s'est terminée par une double publication, début 2001, d'une version « brouillon » de notre génome (Lander et al., 2001; Venter et al., 2001). L'accessibilité de ces séquences a révolutionné la biologie moléculaire en permettant largement aux scientifiques de récupérer sur Internet, dans une banque de données, une séquence d'ADN (ex : un gène) plutôt que de tenter de la cloner

---

[‡] : voir le poster du mois d'octobre 2004 de Nature Genetic Review

[§] : IHGSC, http://www.genome.gov/

et de la séquencer. Toutefois, le qualificatif de « brouillon » était donné à juste titre puisque la séquence contenait des centaines de milliers de trous, représentant jusqu'à environ 10% de l'euchromatine et 30% de la séquence complète du génome (incluant l'hétérochromatine). Ces trous ont conduit à des artéfacts lors de l'annotation (présence de pseudogènes erronés) ou lors des comparaisons entre génomes (analyse phylogénétique). Depuis cette ébauche, un travail de finition a été entrepris par l'IHGSC aboutissant au séquençage complet de plusieurs chromosomes (voir tableau II) ainsi qu'a une version finie à 99.9% de l'euchromatine (International Human Genome Sequencing Consortium, 2004). Désormais les études phylogéniques devraient apporter encore plus de précisions quand à l'évolution subit par les espèces en se basant sur l'analyse de leur génome (exemple entre le tetraodon et l'Homme Jaillon et al., 2004).

I.4.2. La génomique fonctionnelle

Parmi les autres « omes », le protéome symbolise l'ensemble des protéines présentes dans les organelles, les cellules, les tissus (Kahn, 1995). La protéomique s'attache à caractériser les fonctions locales et globales des protéines (Dziembowski and Seraphin, 2004; Gavin et al., 2002). Nous pouvons aussi citer le transcriptome, qui représente l'ensemble des éléments transcrits provenant du génome (Kampa et al., 2004). Pour les ARNs, les « codants » font parti de l'ORFome (Harrison et al., 2002) alors que les « non codants » du RNome. La transcriptomique propose ainsi une annotation du génome par l'étude de ces produits (Saha et al., 2002). L'interactome, aux frontières de ces trois mondes, propose d'étudier leurs liens et interactions (Sanchez et al., 1999;

** : Celera Genomics, http://www.celera.com/

Coulombe et al., 2004). Les recherches « -omiques » sont rendues possible grâce à la robotisation et au perfectionnement des techniques de biologie moléculaire ainsi qu'au développement des outils informatiques, désormais omniprésents dans la science.

Une fois la séquence d'un génome obtenue, il reste à l'analyser, fonctionnellement parlant : Qu'est ce qui est codant ? Qu'est ce qui est régulateur ? Qu'est ce qui est actif d'une autre manière ? Ces questions constituent ce que l'on appelle désormais la « génomique fonctionnelle ». L'importance d'étudier de manière « fonctionnelle » le génome s'illustre parfaitement par les chiffres suivant : 1.5, 25 et 75%. Si nous savons que 1.5% de notre génome est codant et que nous estimons que 25% est transcrit, à quoi correspondent les 75 autres pourcent ? Et pourquoi ont-il été conservés lors de l'évolution ? Peut-être que l'analyse de cette matière noire (Bejerano et al., 2004) du génome permettra d'expliquer le *C-value paradox* (voir tableau I)?

I.4.3. Les perspectives de recherche

La régulation du dogme [ADN➔ARN➔protéine] nous apparaît de plus en plus complexe. Outre les régulations « classiques » (ex : de la transcription, de la traduction), il faut désormais ajouter les régulations « modernes » du type: transcription sens/antisens (Martens et al., 2004; Yelin et al., 2003), interférence d'ARN (revu par Hannon and Rossi, 2004), localisation intra-nucléaire (Misteli, 2004). Aux vues d'une telle complexification des systèmes, il est normal de se demander ce qui, parmi ce qui a été établi *in vitro* (à l'échelle d'un gène), est encore valable *in vivo*, à l'échelle du génome. Plusieurs questions émergent, par exemple 1) les composantes de la machinerie transcriptionnelle de base sont toujours toutes requises pour la transcription des gènes de

type II dans le génome ; 2) les facteurs « généraux » de transcription sont vraiment généraux ou seulement limités à certains groupes de gènes (Holstege et al., 1998) ; 3) S'ils ne sont pas tous toujours requis, où existe-t-il des combinatoires particulières en fonction des gènes (Harbison et al., 2004; Odom et al., 2004) ; 4) Plus directement vis à vis de l'ARN Pol II, où se situe-t-elle dans le génome ? Est-elle localisée seulement au niveau des gènes de classe II ? Et jusqu'où poursuit-elle sa course lors de la transcription d'un gène (Callen et al., 2004) ? Transcrit-elle l'ADN en d'autre ARN que l'ARNm (Kampa et al., 2004) ?

I.5. Contribution apportée par le présent projet

Au sein de l'armada de protéines impliquées dans la transcription, l'ARN Pol II reste la cible ultime : l'enzyme qui transcrit les gènes en ARNm. Bien que la plupart des gènes transcrits par l'ARN Pol II soit maintenant connu (International Human Genome Sequencing Consortium, 2004), de plus en plus d'évidences tendent à démontrer qu'elle transcrirait bien plus l'ARN que les ARNm « sens » (Callen et al., 2004; Cawley et al., 2004; Dermitzakis et al., 2002; Kampa et al., 2004; Lipman, 1997; Martens et al., 2004; Yelin et al., 2003). Jusqu'à maintenant la plupart des études génomiques cherchant à étudier l'activité de la machinerie transcriptionnelle utilisaient une approche orientée « coté gène », regardant et analysant les ARNm et/ou une (des) région(s) candidate(s) (généralement le promoteur). Mais pour comprendre la transcription par l'ARN Pol II telle qu'elle se dessine sous nos yeux, de moins en moins restreinte sur le génome, il était nécessaire de développer des approches orientées « coté génome ». Nous décrivons ici une méthode génomique d'analyses permettant d'identifier et d'étudier la localisation des

éléments composant la machinerie transcriptionnelle de base.

S'intégrant dans un vaste projet de génomique et de protéomique, la contribution de cette thèse a porté principalement sur 1) la recherche des sites d'initiation de la transcription d'un maximum de gènes humains, 2) la mise au point d'une technique d'immunoprécipitation de la chromatine (ChIP) utilisant la double affinité du système TAP (*Tandem Affinity Purification*), et 3) la validation de cette technique en localisant l'ARN Pol II à travers le génome.

Dans un premier temps, tous les SITs humains connus (février 2003) ont été rassemblés, comparés et classés. Cette étape a aboutit à la création d'une base de données de SIT disponible au laboratoire. En parallèle, l'essai TAP-xChIP en double affinité a été développé afin de permettre la localisation de l'ARN Pol II grâce à sa sous-unité RPB11 TAP-taggée. Validé à l'échelle du gène sur différentes régions transcrites ou non (set de gènes pilotes), le clonage a été utilisé comme approche « orienté génome ». Le clonage de séquences associé à l'ARN Pol II révèle que 17.3% des séquences correspondent à des promoteurs, 34.5% à des régions transcrites et 9.1% à des 3'*end*. En comparaison, ces régions ne représentent théoriquement que 1.5, 23 et 1.5% du génome humain (données selon NCBI, datant de juillet 2004).

A ce jour, aucune équipe n'a encore démontré une utilisation complète du TAP-tag lors d'une ChIP ainsi que son efficacité en clonage. La mise au point du TAP-xChIP en double affinité n'a constitué que le premier pas dans la cartographie des réseaux de régulateurs de la transcription, à l'échelle du génome, entrepris par le laboratoire du Dr Coulombe.

II) Matériels et méthodes

II.1. Construction de la banque de données de SITs

Deux sources différentes ont été principalement utilisées lors de la recherche de SITs humains connus. Dans un premier temps, c'est dans la littérature qu'on été recherchés les gènes dont le promoteur aurait été expérimentalement caractérisé. En second lieu, les bases de données publiques centralisant les ARN «pleine longueur» (*fl-RNA*) ont été utilisées, citons *Mammalian Gene Collection* (MGC, Strausberg et al., 1999; Strausberg et al., 2002) et *Database of Transcription Start Site* (DBTSS, Suzuki et al., 2002). « Pleine longueur » signifie ici que sa séquence en 5' comprend au moins l'ATG et souvent plus. Une fois récupérées, les extrémités 5' de ces gènes ont été comparés dans le but d'obtenir un SIT consensus. Afin de déterminer la fiabilité du SIT, les extrémités 5' ont été soumises à plusieurs filtres de sélection (voir figure 5) et une cote de confiance leur à été attribuée. C'est à partir de cette base de données qu'un ensemble de gènes pilotes a été choisi afin de tester, par PCR, le TAP-xChIP.

II.2. Design des amorces et choix des régions « négatives »

Les amorces de PCR utilisées dans cette étude ont été conçues avec le logiciel Primer 3 (Rozen and Skaletsky, 2000) en suivant les recommandations de Qiagen pour la PCR quantitative (QPCR, voir ci-dessous). Les paramètres du logiciel avaient les modifications suivantes : filtre : *human* ; taille des produits attendus : entre 150 et 250 nucléotides ; pourcentage de GC : entre 40 et 60% ; température d'hybridation : autour des 60°C et avec au maximum de 1°C de différence entre les deux amorces d'une même

région ; longueur de l'amorce : de 19 à 27 nucléotides, ne contenant pas de répétition de plus de trois nucléotides identiques ; concentration en sel (milieu réactionnel) : 100mM ; concentration des amorces (milieu réactionnel) : 300mM. Les contrôles négatifs choisis, théoriquement non transcrits, correspondent à 2 régions localisées sur les chromosomes 17 et 13. Le contrôle Desert_Island-ch13 a été recherché suite à la publication d'une étude qui a démontré l'absence de toute transcription dans cette région (Nobrega et al., 2003). Le Control-region_ch17 provient d'une recherche manuelle d'une région conservée, non répétée et non transcrite. Les amorces présentées dans cette étude sont rassemblées dans le tableau V et sont basées sur la version *build 33* (avril 2003) de l'assemblage du génome humain (voir UCSC hg17 basée sur NCBI *build* 33). ENO1 code pour l'enolase α, un gène fortement exprimé et spécifique au rein. FTL code pour la « Ferritin light polypeptide », un gène qui possède un très haut niveau d'expression et qui est exprimé de manière ubiquitaire. GTF2F2, codant pour RAP30 du complexe TFIIF, est exprimé de manière ubiquitaire et à un niveau modéré. Pour chaque gène, plusieurs paires d'amorces ont été créées afin d'amplifier spécifiquement le SIT et au moins une région transcrite en aval. La spécificité d'amplification a été vérifiée lors de la conception en utilisant UCSC in silico PCR[††].

II.3. Les lignées cellulaires et plasmides.

La création de la lignée cellulaire utilisée, ainsi que les conditions de cultures et d'expressions du peptide-taggé ont été mises au point avant le début du présent projet (description voir Jeronimo et al., 2004). Brièvement, l'ADNc codant pour Rpb11

[††] : UCSC in silico PCR, http://genome.ucsc.edu/cgi-bin/hgPcr

(Invitrogen ; accession number AA8141184) a été cloné dans le vecteur d'expression mammifère pMZI (No et al., 1996) de manière à porter en C-terminal le TAP-tag (Rigaut et al., 1999). Ce vecteur a été transfecté de manière stable dans des cellules EcR-293 (Invitrogen) par la méthode du phosphate de calcium. L'induction du peptide étiqueté est faite, pendant 24h avec 1 à 3µM de ponastérone A (un analogue de l'ecdysone) de manière à exprimer le peptide-taggé à un niveau proche de celui physiologique (Jeronimo et al., 2004). Les cellules sont ensuite récoltées, concentrées, puis congelées (-80°C) jusqu'à utilisation.

II.4. Immunoprécipitation de la chromatine (TAP-xChIP)

L'immunoprécipitation de la chromatine grâce au TAP-tag est déjà utilisée dans sa forme « simple affinité » (Krogan et al., 2002; Jeronimo et al., 2004). Elle est présentée ici dans sa version complète (double affinité). Les étapes de préparation de la chromatine se basent principalement sur un protocole déjà établit (Jeronimo et al., 2004) et celles d'immunoprécipitation suivent sensiblement le même protocole que lors d'une purification par TAP-tag (Rigaut et al., 1999). Tout comme lors d'essais protéomiques, des cellules induites ($^+$) et non induites ($^-$) pour le peptide « taggé » sont traitées en parallèle. Avec $2x10^7$ cellules, il est possible de voir un enrichissement spécifique en QPCR. Mais de manière routinière, environ $2x10^8$ cellules sont utilisées afin de pouvoir procéder au clonage. Brièvement, les cellules sont fixées au formaldéhyde (1% final à température ambiante, pendant 1min), puis lysées. Les noyaux sont isolés, nettoyés et re-suspendus dans un tampon de sonication (tampon RIPA, 10mM Tris-HCl pH 8.0, 1mM EDTA, 0.5mM EGTA, 140mM NaCl, 1% Triton X-100, 0.1% NP-40, 0.1% SDS,

1mMPMSF et 5μg/ml de leupeptin, pepstatin A et aproptinin). Des fragments de chromatine variant de 0.1kb à 0.7kb sont obtenus par sonication (5 fois pendant 20s espacé d'au moins 2 minutes à 50% de la puissance maximum de l'appareil, Fisher Sonic Dismembrator). Lors de la première affinité, les billes d'IgG-sepharose sont lavées avec un excès de tampon IgG-wash (50mM HEPES pH8.0, 1mM EDTA, 0.7% DOC, 1% NP-40, 500mM LiCl) avec un rapport de 20 volumes d'IgG-wash pour 1 volume de billes. Lors de la deuxième affinité, les billes de calmoduline sont lavées avec le même tampon mais comprenant 2mM de $CaCl_2$ en plus (20 volumes de tampon Calmo-wash pour 1 volume de billes). L'élution est faite par destruction des pontages au formaldéhyde (suite à une incubation toute la nuit à 65°C). Les fragments d'ADN sont ensuite purifiés (Qiaquick® PCR purification kit ; élution dans 100µl de EB) et leur taille analysée sur gel d'agarose 1.5%. L'ADN purifié est conservé à -20°C en attendant son utilisation.

II.5. Validation par PCR quantitative (TAP-xChIP-QPCR)

Après avoir renversé les pontages, une PCR quantitative (Sybrgreen II®, Qiagen) est réalisée sur l'ADN immunoprécipité induit (IP^+) et non-induit (IP^-) sur les régions d'intérêt (vis-à-vis de la transcription) afin de vérifier la spécificité de l'enrichissement. La quantification de l'ADN cible IP est calculée en fonction d'une courbe standard, résultant de dilutions en série d'ADN WCE (non immunoprécipité, à concentration connue) en respectant les recommandations du manufacturier (Qiagen). La courbe, dont les dilutions sont respectivement de $1/25^e$, $1/125^e$, $1/625^e$ et $1/3125^e$, est faite pour chaque couple d'amorce et à chaque expérience. Les expériences ont été faites en utilisant le « MX4000 Multiplex Quantitative PCR system » (Stratagene). Tout les fragments

d'ADN amplifiés sont uniques, tel que déterminé par BLAT (Kent et al., 2002), par courbe de dissociation (QPCR) et par analyse sur gel d'agarose. En remplacement de la QPCR (ou en la précédant), un test par PCR standard est souvent utilisé. Les amorces utilisées sont les mêmes que lors de la QPCR. Les conditions de réactions suivent les recommandations d'utilisation de la TAQ *polymerase* (Amersham). L'amplification est réalisée pendant 40 cycles (94°C pendant 60s, 60°C pendant 45s, puis 72°C pendant 30s) et les résultats sont ensuite visualisés sur gel d'agarose (1.5%).

II.6. Validation par clonage (TAP-xChIP-cloning)

Quelques précautions sont à prendre en considération, pour une efficacité maximale lors de cette étape, due à la très faible quantité d'ADN manipulée (entre 0.02 et 0.2 ng/μL, estimation par QPCR). À moins d'une indication contraire, toutes les étapes de cette technique sont réalisées selon les recommandations des manufacturiers. Brièvement, les extrémités des fragments de chromatine IP^+ et WCE^+ sont réparés en utilisant la T4 DNA polymérase (Invitrogen). Parallèlement à cette étape, le plasmide pBluescript II® SK^+ (Stratagene) est digéré par l'enzyme de restriction SmaI (Amersham) et déphosphorylé par la phosphatase alcaline de veau (CIAP, Invitrogen). La ligation est ensuite réalisée avec un excès de vecteurs (ratio de 1 insert pour 3 vecteurs). Pour cette étape, la T4 DNA ligase HC (Invitrogen ; *High Concentration* à 5U/µL) a été utilisée en suivant les recommandations valables pour le clonage de fragments *blunt*. Des bactéries ultracompétentes, XL10-Gold® ultracompetent cells (Stratagene), sont alors transformées avec une partie de la ligation (5µL de ligation par 100µL de cellules). Il est possible de congeler la ligation à condition de faire une congélation rapide (-80°C pendant 1min puis

-20°C). Les bactéries sont cultivées sur milieu solide permettant la sélection bleu/blanc (LB agar avec ampicilline, tétracycline, IPTG 80mM et Xgal 2%). Après vérification de la présence d'insert dans le vecteur par « criblage par PCR », les clones positifs sont cultivés en milieu liquide (au moins 18h) puis les plasmides sont purifiés par kit Qiagen (Plasmid miniprep purification kit). Les vecteurs purifiés sont ensuite séquencés par l'amorce universelle T7 en utilisant un séquenceur automatique. Les séquences sont disponibles en annexes.

II.7. Analyse des séquences clonées par BLAT

Une fois les inserts séquencés, ils sont soumis à une analyse par BLAT (Kent et al., 2002) sur le génome humain (version datant de juillet 2003 ;UCSC hg16 basé sur NCBI build.34). Les restrictions quant à l'acceptation ou au rejet d'une séquence sont les suivantes : 1) avoir une homologie maximale (≥ 99%) entre la séquence clonée et la séquence du génome, 2) posséder au minimum une séquence unique de 20 nucléotides qui permettent une localisation précise à travers le génome sinon elles sont considérées comme des séquences répétées (*repeat*), et 3) en cas de duplicatas de séquences (dus à la culture bactérienne) seule une séquence est considérée (les autres entrent dans la catégorie *duplicate*). Les données permettant de caractériser la séquence analysée (taille, localisation chromosomique, position intrachromosomique, brin) ont été compilées dans une base de données. Les séquences ont été majoritairement analysées de façon génomique (par rapport à leur environnement) plutôt que par rapport à leur composition nucléotidique (par rapport à leur séquence). Le calcul des distances séparant la séquence

analysée d'un élément d'intérêt (ex : une CGI, un SIT, un *3'end* et un gène connu[‡‡]) a été réalisé manuellement (Δ ±5%). Une distance négative indique que l'ADN cloné est situé en amont de l'élément analysé et à l'inverse une positive le situe en aval. L'« amont » et l'« aval » tiennent compte de l'orientation du gène, ce qui ne s'applique donc pas pour les CGI. Concernant les CGI, c'est le premier CGI « fort » (longueur ≥300pb) qui a été considéré. Les domaines géniques fonctionnels analysés lors de cette étude ne correspondent qu'à des éléments appartenant à des Refseq, ce qui veut dire qu'ont été volontairement exclus les gènes prédits par Ensembl (Birney et al., 2004; Hubbard et al., 2002) ou Genscan (Burge and Karlin, 1997). Enfin, aux vues de la taille du génome humain (2.8 Gb accessibles) la résolution de 50,000 bases (50kb) a été utilisée pour la figure 9. La figure 10A représentant une analyse à plus haute résolution des figures 9B, 9C et 9D, elle bénéficie d'un grossissement de l'échelle, qui passe de 50 à 10kb, voir 2kb pour les régions très spécifiques (ex : le SIT et de le 3'UTR).

---

‡‡ : Le terme « gènes connus » désigne tous gènes définis selon SWISS-PROT, TrEMBL, mRNA ou RefSeq. À des fins de simplification, nous l'appellerons désormais Refseq

III) Résultats

J'ai réalisé l'ensemble des expériences présentées dans ce mémoire. Les données ajoutées en compléments ou à titre explicatif sont suivies de la source d'où elles proviennent. Le protocole initial de TAP-tag utilisé dans la mise au point du TAP-xChIP provient de Célia Jéronimo (Jeronimo et al., 2004). Les cellules utilisées, induites ou non, ont été produites par Annie Bouchard.

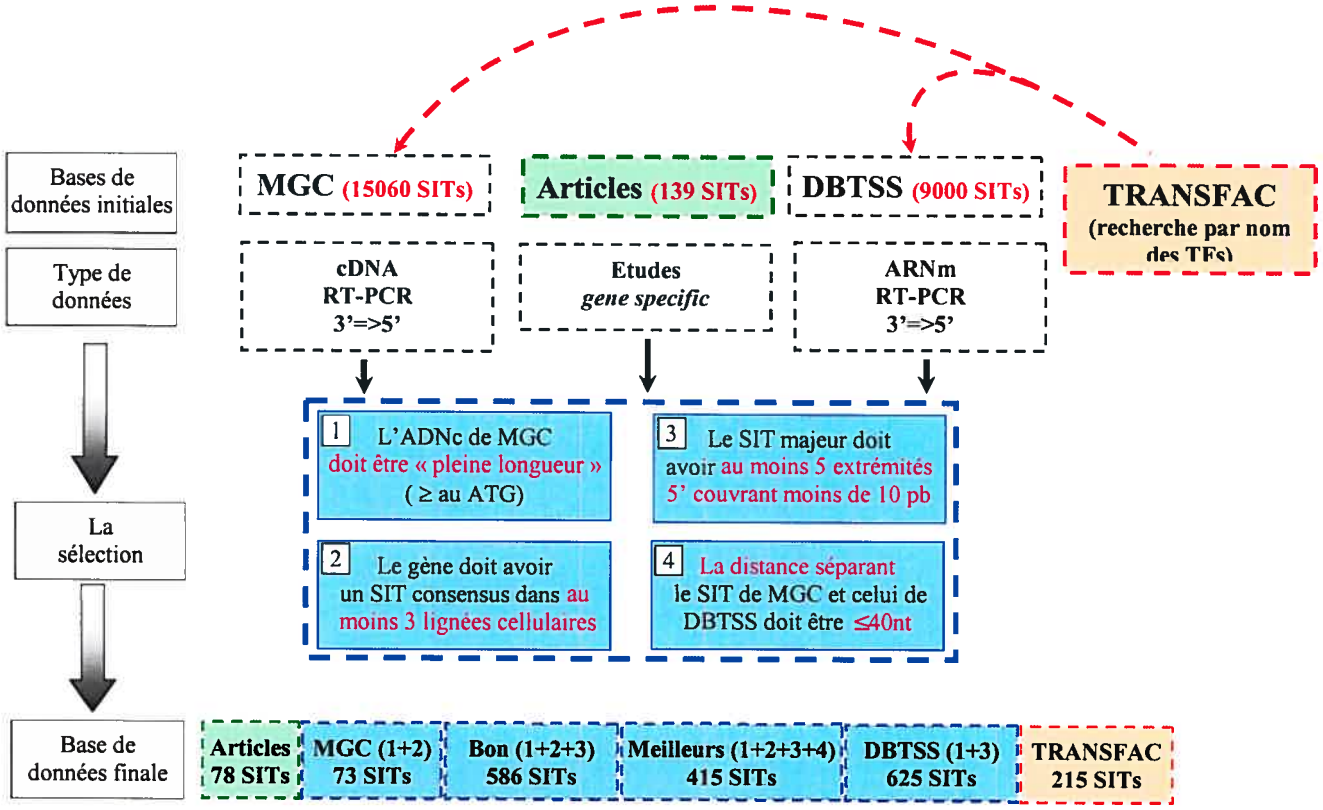
III.1. Construction de la base de données de SITs humains connus

L'approche principale visait à établir une procédure générale pour identifier le SIT d'un gène. Voulant analyser la transcription à grande échelle, un maximum de SITs a été recherché afin de s'approcher le plus possible de la réalité du génome. Pour cela, deux bases de données publiques proposant des séquences « pleine longueur » d'ARNm (fl-RNA) (figure 5A) ont été utilisées. La première, MGC (*Mammalian Gene Collection*) répertorie des cDNA (analyse 3'➔5'[§§]) provenant de nombreuses lignées cellulaires humaines (Strausberg et al., 1999; Strausberg et al., 2002). La deuxième, DBTSS (*DataBase of human Transcription Start Site,* Suzuki et al., 2002), référence le SIT de nombreux gènes définis par capture de la coiffe du transcrit, donc le premier nucléotide transcrit [***](*oligo-capping*, Maruyama and Sugano, 1994). Dans un premier temps, les cDNA de MGC ont été comparées aux clones disponibles dans DBTSS.

[§§] : Provenant de PCR inverse, c'est un cDNA qui est obtenu; le terme cDNA est utilisé.

[***] : Obtenu par séquençage direct, de 5'en 3', le terme de « clone » est utilisé.

A) Création de la base de SIT humain



B) Analyse de SITs provenant de MGC par le filtre de la base de données

Distances séparant les SITs MGC et DBTSS	**∞**	**± 100pb**	**± 50pb**	**≤40pb avec ≥5clones DBTSS regroupés sur 10pb**
MGC « pleine longueur »	100	X	X	X
Au moins un clone DBTSS	60	55	54	X
Au moins cinq clones DBTSS	39	37	36	26
Au moins dix clones DBTSS	28	26	25	24

Figure 5. Construction de la base de données de SITs humains.

A) Organigramme schématisant les sources utilisées, le nombre de SITs analysés, les critères de sélection et les différentes cotations possibles dans la base de données. Le système de classification (Articles, MGC, Bon, Meilleurs, DBTSS et TRANSFAC) rend les différentes classes mutuellement exclusives. B) Evaluation, sur un set de 100 fl-RNA provenant de MGC, des critères de sélection utilisés lors de l'élaboration de la banque de SITs. La case bleue représente l'ensemble des séquences répondant à l'ensemble des critères 1+2+3+4 décrit ci-dessus. ∞ aucune limite de distance.

Observant des différences dans la localisation précise du SIT, plusieurs critères de sélection ont été établis afin de sélectionner les SITs les plus fiables. Les critères utilisés sont : sur MGC, 1) être « pleine longueur », et 2) d'avoir une extrémité 5' « consensus » trouvée dans au moins 3 lignées cellulaires différentes; Pour DBTSS, 3) un SIT est considéré s'il a au moins 5 clones dispersés sur un maximum de 10 bases. Enfin, 4) les extrémités 5' trouvées sur MGC et DBTSS, pour un gène donné, ne doivent pas être séparées de plus de 40pb. Appliqués ensemble, ces critères de sélection composent la matrice de sélection à travers laquelle sont passés tous les cDNAs de MGC et les clones de DBTSS. D'après une estimation, ce criblage permettrait de récupérer autour de 25% des SITs, disponibles dans l'une et/ou l'autre de ces deux bases de données (figure 5B). Afin d'avoir une base de données plus complète, les critères de sélection ont été diminués et les sources diversifiées. Les gènes trouvés dans la littérature (voir liste en annexes) ont eux aussi été passés au crible, tout en bénéficiant d'un poids supplémentaire dû à la caractérisation fonctionnelle du SIT. Enfin, notre laboratoire s'intéressant de près à la transcription par l'ARN Pol II, les gènes codant pour les facteurs impliqués dans ce mécanisme représentent un intérêt tout particulier. Utilisant TRANSFAC (*transcription factor*), une base de données publique rassemblant toutes les informations disponibles sur les protéines impliquées dans la transcription, les gènes qui avaient été écartés de la base de données ont été insérés et ponctués d'une remarque les identifiant. L'établissement d'une classification entre les différents SITs permet de choisir avec plus de confiance les SITs pilotes nécessaires à la mise au point du TAP-xChIP. Au final, malgré les divergences des types de données analysées, une banque de SITs humains rassemblant près de 17,000 sites dont ~2,000 fiables a été créée. C'est à partir de cette banque qu'ont

été choisi les SITs utilisés dans la suite de l'étude.

Tableau V. Liste des amorces utilisées dans cette étude

Toutes les régions génomiques, sauf DI_ch13, sont basées sur la version d'avril 2003 du génome humain (build. 33). DI_ch13 est basée sur la version de juillet 2003 (build 34).

Nom du gène (ID) [a]	**Expression (niveau)** [b]	**Localisation génique**	**Nom de l'amorce**	**Séquence des amorces** [c]
Ferritin, light polypeptide (2312)	Ubiquitaire (+++)	Promoteur basal	FTL-pr	Fwd : gctgagactcctatgtgct rev: acactgttgaagcaagagac
		aval	FTL-av	Fwd : tatagaagccagctgaagat Rev : gtgaaatgaggctctgaa
GTF II F2 (2963)	Ubiquitaire (+)	Promoteur basal	GTF2F2-pr	Fwd : tttcttcagttatgctgacc Rev : ttacctgccagaacactg
		aval	GTF2F2-av	Fwd : ctaagaggtctttcttgtcg Rev : actattctgggtatgacagg
Enolase α (2023)	Spécifique au rein (+++)	promoteur proximal	ENO1-sp1	Fwd : agaaagggacagggtcac Rev : cgaccttgctgacaactt
		Promoteur basal	ENO1-pr	Fwd : ggtgagggaatgagtgac Rev : accgaggtgaacgtaaag
		dernier intron	ENO1-LI	Fwd : tcaagatccaaacagctc Rev : tagttctcctcatcccaac
		dernier exon	ENO1-LILE	Fwd : gcacaagtttagagggttta Rev : cagctcctcttcaattctt
Controle region ch17	Non transcrit	Balayage sur UCSC, build 33[d]	Ctrl_ch17	Fwd : agagaattgtctggatcttg Rev : tgaagctatatgacactactgc
Desert Island ch13	Non transcrit	Nobrega *et al.*, 2003 [e]	DI_ch13	Fwd : gagtatcactcagaaagaga Rev : attctcctcaggtatagaag

Légende :

[a] : Nom et numéro d'identification (ID) provenant de NCBI.

[b] : Spécificité cellulaire et niveaux d'expression provenant de CGAP.

[c] : Séquences des amorces (fwd = sens et rev : anti-sens) obtenues par Primer3.

[d] : Balayage du génome humain (built 33.)

[e] : La confirmation par Nobrega *et al.* de l'absence de transcrit à conduit à la sélection de ce contrôle.

III.2. Analyse de la localisation de l'ARN Pol II à l'échelle du gène

La localisation de l'ARN Pol II au niveau du gène a pour objectifs de valider la méthode de purification en double affinité ainsi que de déterminer son efficacité. Bien que plusieurs protocoles utilisant le système de purification par TAP-tag lors d'une ChIP aient déjà été publiés, aucun n'a encore présenté une méthode qui utiliserait successivement les deux affinités de l'étiquette (Jeronimo et al., 2004; Kim et al., 2004; Krogan et al., 2003; Kuras et al., 2003; Verdel et al., 2004; Westermann et al., 2003). Or, Rigaut *et al.* ont démontré que la majeure partie de la très haute spécificité de la purification par TAP-tag résulte en grande partie de la deuxième affinité, avec la calmoduline (Rigaut et al., 1999).

Brièvement, l'immunoprécipitation est effectuée sur des cellules humaines EcR293 exprimant une version TAP-taggée de la sous-unité Rpb11 de l'ARN Pol II (obtention des lignées cellulaires et culture voir Jeronimo et al., 2004). Les cellules sont induites à la ponastérone A afin d'exprimer le polypeptide-taggé. Traitées en parallèle avec des cellules « non induites » (-), les cellules « induites » (+) sont fixées au formaldéhyde. La chromatine est ensuite isolée, puis fractionnée par sonication de manière à obtenir des fragments de petite taille permettant une bonne résolution lors des analyses (de 100 à 700pb, figure 6C). Les fragments d'ADN portant des ARN Pol II possédant un Rpb11-TAP sont co-précipités par les IgG (fixées sur des billes d'agarose), élués par clivage enzymatique (coupure par la protéase TEV), puis re-précipités, en présence de calcium, en utilisant la calmoduline (fixée sur des billes de sépharose) (Figure 6B).

A) Schématisation de polypeptide taggé : Rpb11-TAP

B) Schématisation du TAP-xChIP séquentiel

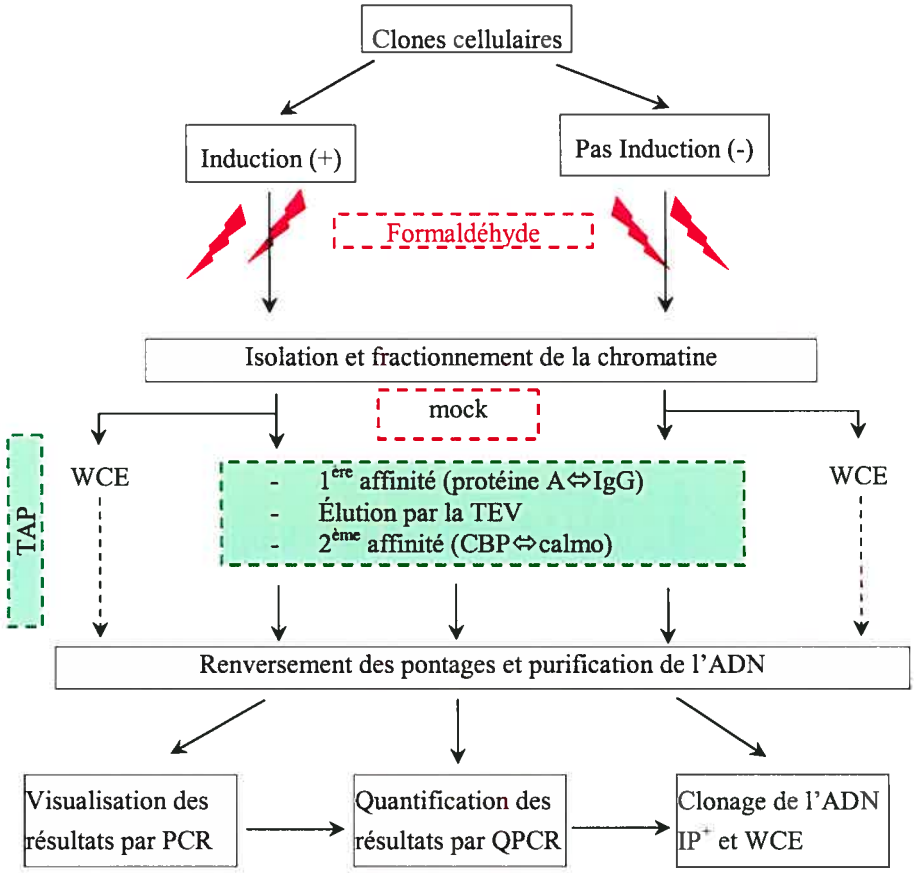
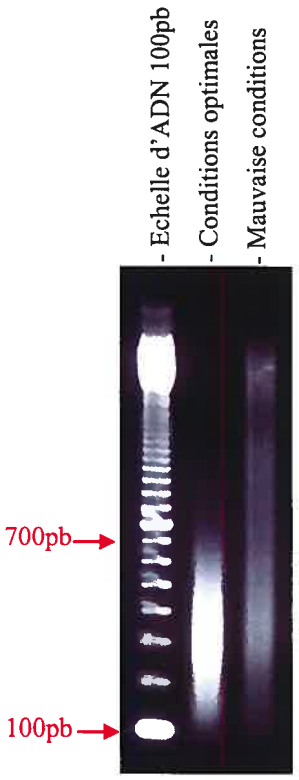


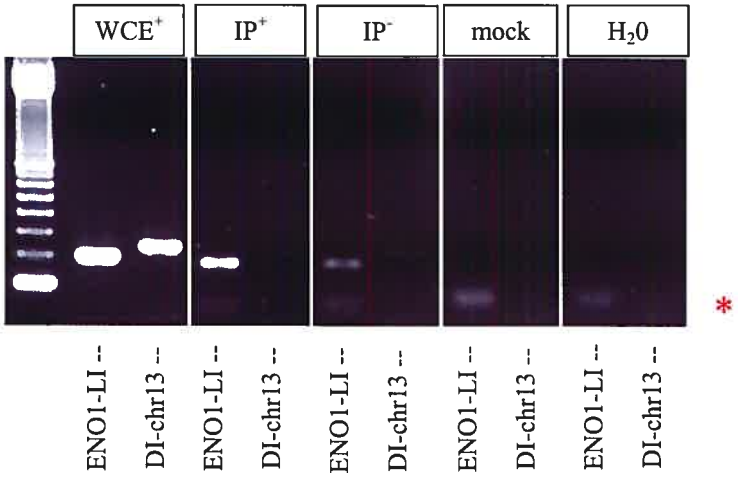
Figure 6. Le TAP-xChIP en double affinité.

A) Représentation schématique du polypeptide utilisé dans cette étude. B) Présentation de la méthode. TAP pour Tandem Affinity Peptide ; X pour « chimique » ; ChIP pour Chromatine immunoprecipitation ; QPCR, PCR quantitative. C) Fragmentation de la chromatine de manière à obtenir une taille variant de 100 à 700pb. D) Test d'amplification, par PCR, d'une région transcrite (ENO1-LI) par rapport à une région non transcrite (DI_ch13) lors du TAP-xChIP de RPB11. WCE+, *whole cell extract* (non immunoprécipité). IP+ et IP- correspondent à de l'ADN IP induit et non induit. Mock est un contrôle de la méthode. H20 représente le contrôle négatif de PCR. L'astérisque (*) correspond aux dimères d'amorces.

C) Fragmentation de la chromatine



D) Visualisation, par PCR, de l'efficacité de l'immunoprécipitation



Après plusieurs lavages des billes de calmoduline, les pontages sont renversés.

Avant d'être cloné (voir section suivante), l'ADN IP^+ purifié est d'abord testé qualitativement par PCR standard (figure 6D) et/ou quantifié par PCR quantitative (figure 7). L'efficacité est démontrée en comparant l'IP^+ et l'IP^- au niveau de l'amplification d'une région transcrite, par exemple le dernier intron de l'enolase-α (ENO1-LI), par rapport à une région contrôle (non transcrite), telle Desert_Island_ch13 (figure 6D).

Afin de quantifier l'enrichissement spécifique de l'ARN Pol II par le TAP-xChIP, en double affinité, une quantification par PCR (QPCR) de plusieurs régions d'intérêt a été réalisée (figure 7). L'enrichissement brut est déterminé par intrapolation sur une courbe de dilutions standard (à concentrations connues, voir section II.5) et résulte d'une moyenne entre deux duplicata. Pour chaque région cible, l'enrichissement réel est calculé par un ratio entre le nombre de copies obtenu dans l'IP^+ divisé par celui obtenu dans l'IP^- (figure 7B). Nous avons proposé qu'un enrichissement de 2 fois constitue le seuil au-dessus duquel un enrichissement est considéré comme effectif. Pour les contrôles non transcrits (DI_Ch13 et Ctrl_ch17), l'enrichissement est nul ou négligeable (respectivement 1x et 1.8x). Dans le cas de GTF2F2, ubiquitaire et modérément exprimé, l'ARN Pol II est détectée au niveau du promoteur (12.9x) et très faiblement dans la région transcrite (2.8x). Dans le cas de FTL, ubiquitaire et fortement exprimé, le ratio au promoteur révèle une absence presque totale de l'ARN Pol II (1.2x) alors que la région transcrite se révèle fortement enrichie (107.6x). Pourtant, le promoteur de FTL est très loin d'être dépourvu de polymérase comme le montre les résultats bruts de la QPCR (figure 7A, FTL-pr). Mais étant donné l'enrichissement dans l'IP^-, le ratio des deux donne une valeur proche de 1. Enfin, dans le cas de ENO1, dont l'expression est

forte et spécifique aux reins, plusieurs régions ont été analysées : en amont du SIT, sur le SIT, dans le dernier intron et sur le dernier exon. L'enrichissement détecté dans ces trois dernières régions est respectivement de 49x, 15.7x et 27.6x et démontre une localisation effective de l'ARN Pol II tout le long du gène. Le site en amont, possédant un site de fixation pour SP1, est lui aussi enrichi car, en plus d'être relativement proche du SIT (500pb), SP1 a été localisé sur cette séquence (Cojocaru M., résultats non publiés) et une possible interaction avec l'ARN Pol II est donc envisageable.

En conclusion pour cette section, localiser l'ARN Pol II sur des régions génomiques ciblées en utilisant le TAP-xChIP en double affinité s'avère une méthode précise (résolution, voir figure 6C) et efficace (peu de bruit de fond, voir Figure 6D et 7).

A) Résultats bruts de QPCR

		Régions contrôles		FTL		GTF2F2		ENO1			
		DI-chr13	Ctrl-chr17	pr	av	pr	av	sp1	pr	LI	LILE
Exp1	IP^+	1.64	2.32	562.7	84.13	10.36	13.65	60.25	92.11	33.06	11.20
	IP^-	2.38	1.01	388.3	0.78	0.57	7.8	3.53	1.32	1.94	0.82
Exp2	IP^+	1.02	2	378.1	137.1	37.9	10.9	62.73	205.9	43.78	41.56
	IP^-	0.79	1.43	408.5	1.45	5	2.79	1.92	7.28	3.05	1

B) Enrichissement quantifié sur des régions d'intérêt

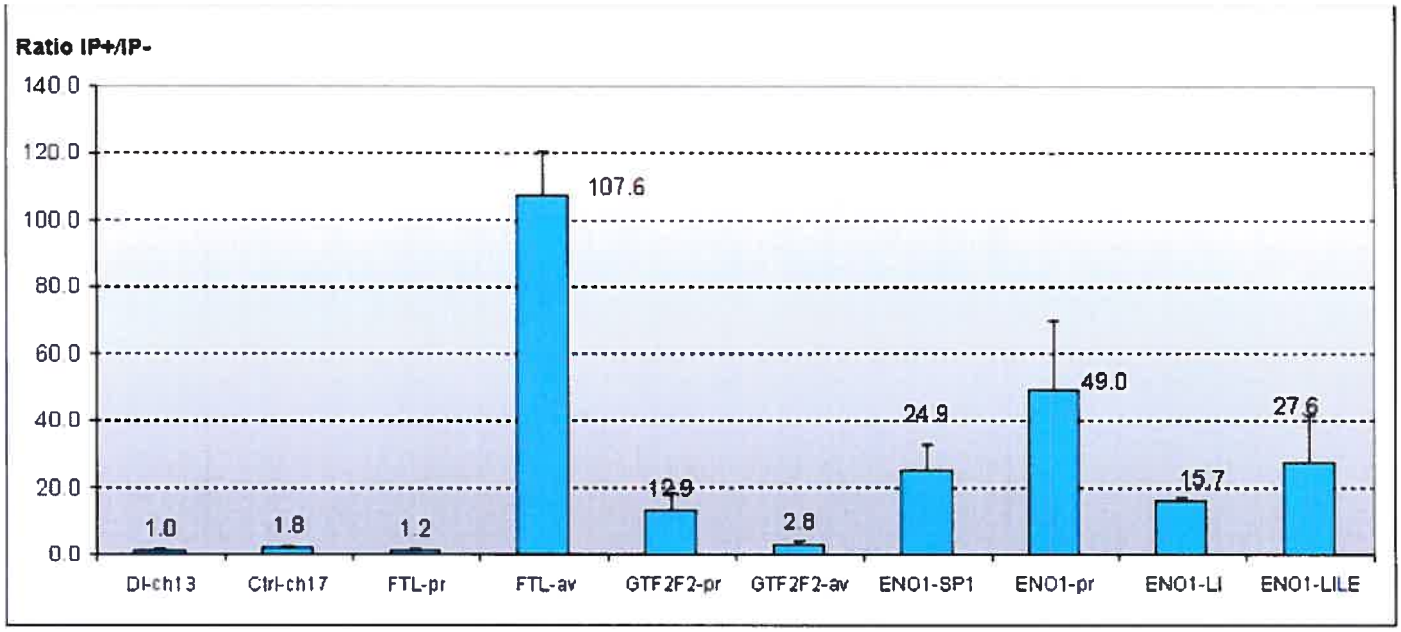


Figure 7. Localisation de l'ARN Pol II sur des cibles transcriptionnelles connues.

A) Nombre de copies de chaque région cible obtenue après IP^+ et IP^- et quantifiée par QPCR (pour plus de détails sur les amorces, voir le tableau V). av, aval ; pr, promoteur ; sp1, région en amont du SIT d'ENO1 contenant des sites de fixation pour Sp1 ; LI, dernier intron d'ENO1 ; LILE, limite dernier intron dernier exon d'ENO1. B) Représentation graphique des enrichissements obtenus après correction IP^+/IP^-.

III.3. Analyse de la localisation de l'ARN Pol II à l'échelle du génome

Afin de confirmer l'efficacité de la méthode de localisation par double affinité testée sur l'ARN Pol II, une analyse génomique de l'ADN IP a été réalisée par clonage. Bien que des protocoles de clonage à partir d'ADN immunoprécipité aient déjà été publiés (Laganiere et al., 2003; Weinmann et al., 2001; Weinmann and Farnham, 2002), aucune étude n'a encore proposé d'utiliser le TAP en double affinité. À partir de l'ADN IP^+ purifié, et après avoir vérifié la qualité de l'immunoprécipitation, les extrémités des fragments de chromatine sont réparées puis insérées dans des vecteurs préalablement coupés (*blunt)* et déphosphorylés. Malgré l'efficacité de l'immunoprécipitation séquentielle, la quantité d'ADN disponible est très faible (estimée par QPCR entre 0.02 et 0.20ng/μL). Ceci oblige l'analyse de tous les clones IP^+ obtenus lors des transformations bactériennes. Les ADNs IP^+ et WCE^+ (dilués au $1/125^{eme}$, quantitativement comparable à l'IP) sont traités en parallèle. Ne considérant que les clones possédant un vecteur recombinant (avec un insert), 113 clones provenant d'IP^+ de RPB11 et 129 provenant du WCE^+ ont été analysés (séquences en annexes).

Du point de vue de la longueur des fragments, aucune différence significative ne semble séparer les séquences IP^+ des séquences aléatoires WCE^+ (figure 8A), malgré une légère tendance à cloner, les fragments de chromatine de petite taille (<300pb, comparer les figures 6C et 8A). Leur localisation chromosomique est quant à elle légèrement différente vis à vis des chromosomes 3, 16, 17, 19 (figure 8B et 8C). D'après les estimations présentées dans le tableau II (voir I.1.1.2), les chromosomes 16, 17 et 19 font partie des chromosomes « riche en gènes ». Par contre, il est très surprenant de ne détecter aucune séquence IP^+ sur le chromosome 3 qui, pourtant, est de grande taille

(~199Mb) et pas particulièrement pauvre en gènes (6.23gènes/Mb, voir tableau II).

A) Classement, selon leur taille, des séquences provenant du clonage des IP^+ et WCE^+

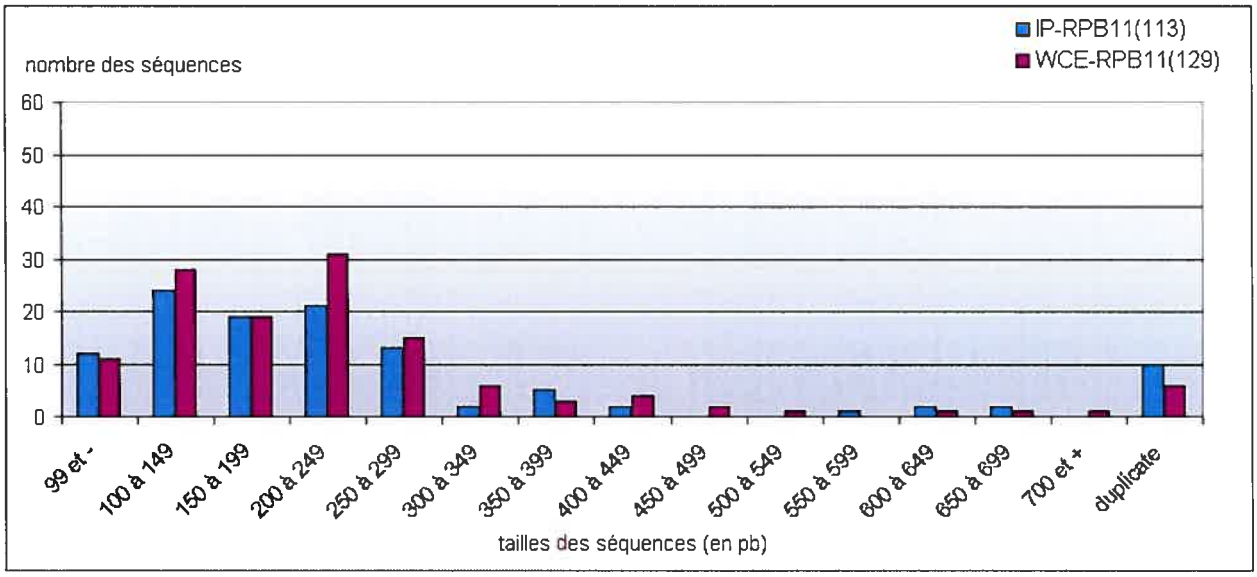
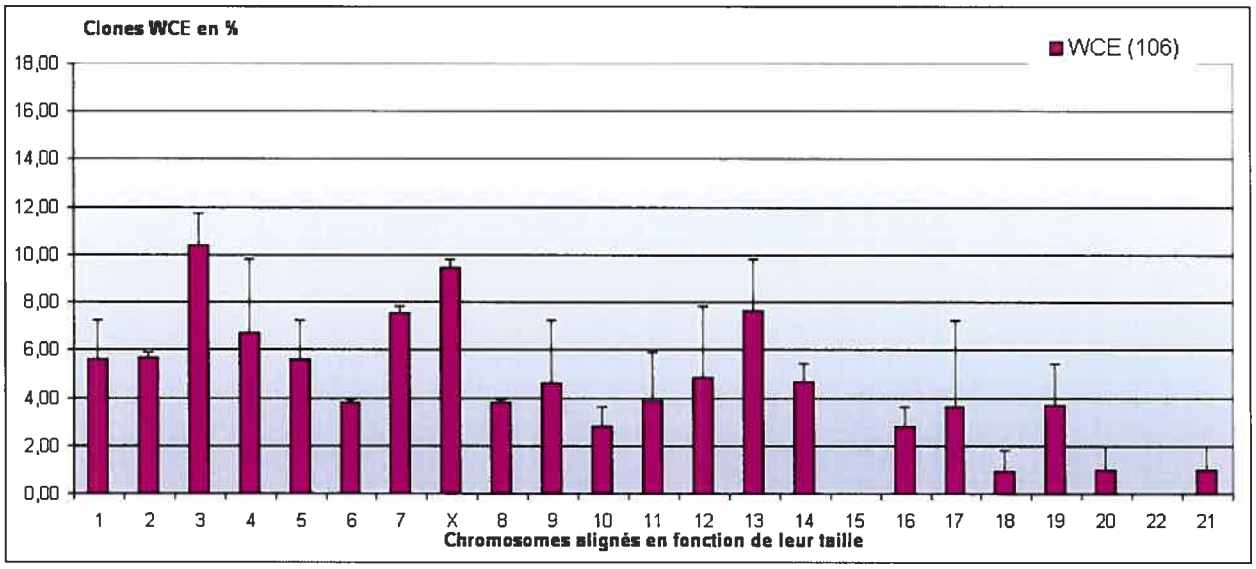


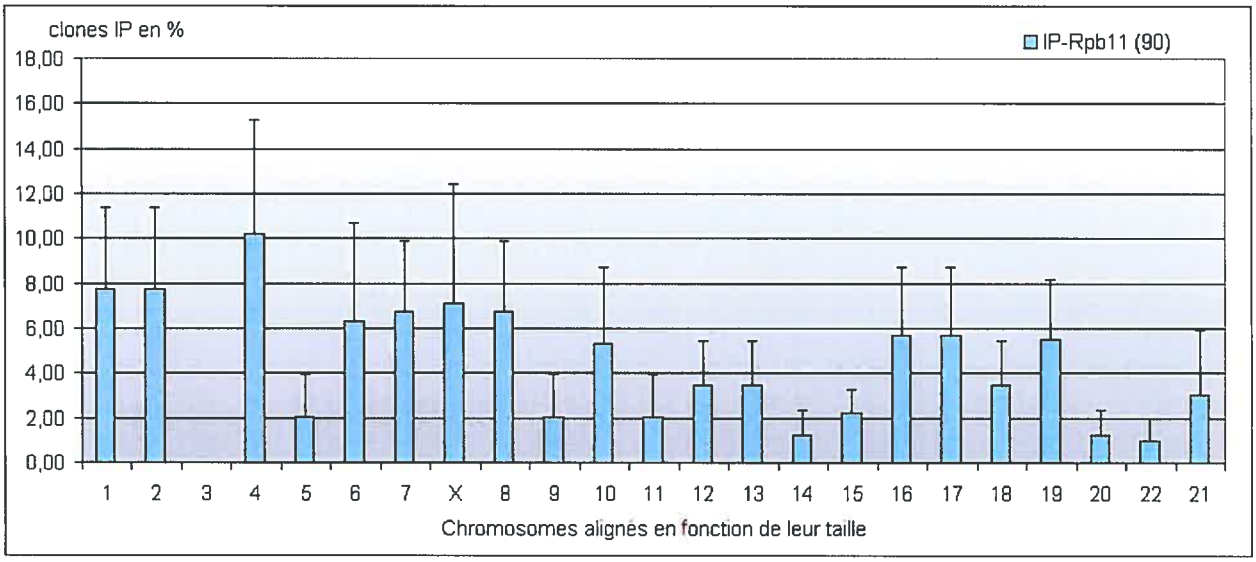
Figure 8. Analyse macroscopique des séquences provenant des clonages IP^+ et WCE^+.

A) Taille des fragments d'ADN séquencés. « Duplicate » indique le nombre de clones dupliqués lors de la croissance sur pétris. B) et C) correspondent à la répartition des séquences uniques et non répétées à travers le génome humain (respectivement pour WCE^+ et IP^+). Tenant compte du nombre limité de séquences obtenues, l'utilisation du pourcentage de séquences par chromosomes et par clonages s'avère plus représentative. Les séquences répétées et dupliquées ont été exclues de cette répartition (voir section II.3). X, chromosome X.

B) Répartition des séquences WCE^+ par chromosome

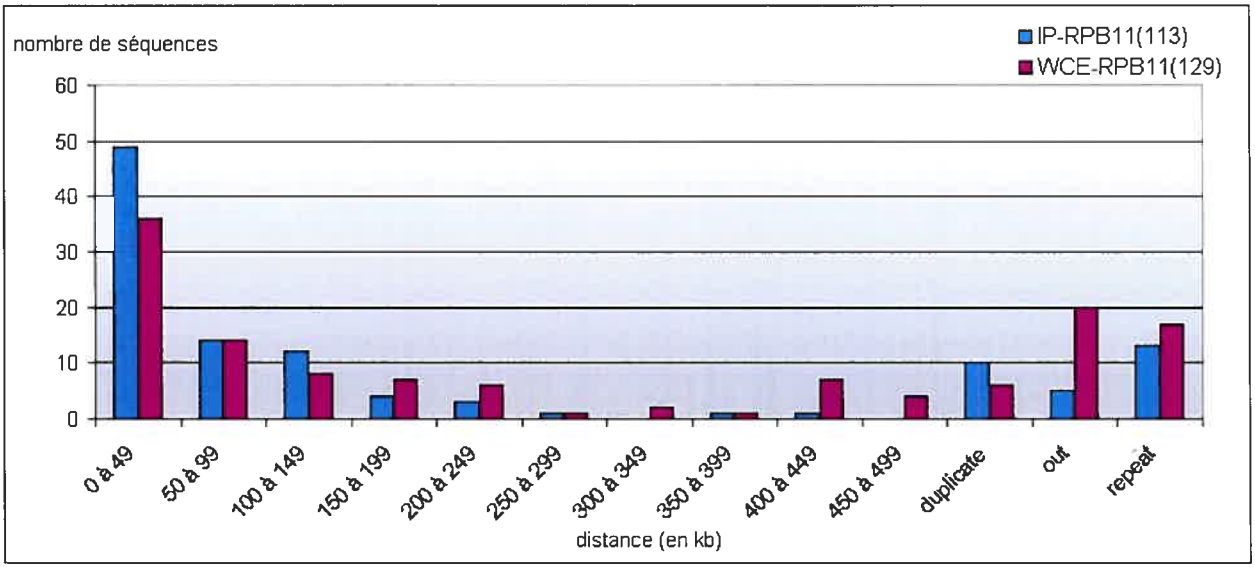


C) Répartition des séquences IP^+ par chromosome



Afin de déterminer le rôle potentiel des fragments d'ADN IP et WCE vis à vis de la transcription, différents éléments transcriptionnellement actifs tels que les CGIs, les SITs, les Refseq et les *3'end* ont été étudiés (figure 9). Suite à l'analyse des séquences par BLAT (Kent, 2002) et grâce à la visualisation graphique du génome humain disponible sur le site *Human Genome Browser Gateway* (aussi nommé UCSC, Karolchik et al., 2003), les distances en kb séparant le clone de l'élément d'intérêt ont pu être calculées. En ce qui concerne les éléments se rapportant au « Gène » (SIT, Refseq et *3'end*), l'orientation en amont (valeurs négatives) et en aval (valeurs positives) par rapport à l'élément a été précisée. Dans le cas des CGIs, aucune divergence significative n'est observée entre IP^+ et WCE^+, tout en notant que la majorité des clones est plutôt proche (<50kb) des GCIs. L'analyse de la distance séparant les clones d'un Refseq et compte tenu de l'immense diversité de taille des gènes humains, la valeur zéro (0) a été attribuée à toutes séquences tombant entre le SIT et le *3'end* d'un Refseq. Des valeurs négatives ou positives ont été attribuées pour tout clone tombant respectivement en amont ou en aval d'un Refseq. Suivant cette approche, un plus grand nombre de séquences IP^+ par rapport aux WCE^+ (respectivement 54 et 41) est retrouvé dans les gènes (figure 9B). Les différences apparaissent nettement lors de l'analyse de sites géniques précis. Ainsi, deux fois plus de clones IP^+ que de WCE^+ sont observés dans la région en aval lors de la recherche des SITs (figure 9C). La localisation des séquences clonées par rapport au *3'end* révèle un enrichissement IP/WCE très net en amont et en aval du site (respectivement 40 contre 24 clones et 19 contre 4 clones).

A) Localisation génomique des séquences clonées par rapport au CGI le plus proche



B) Localisation génomique des séquences clonées par rapport au Refseq le plus proche

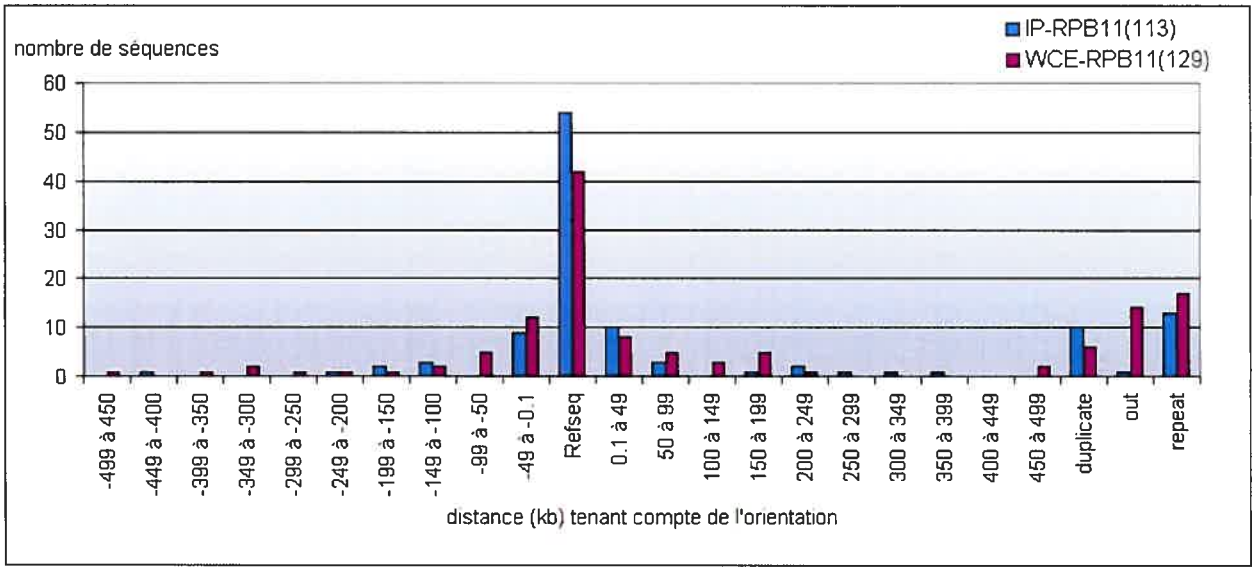
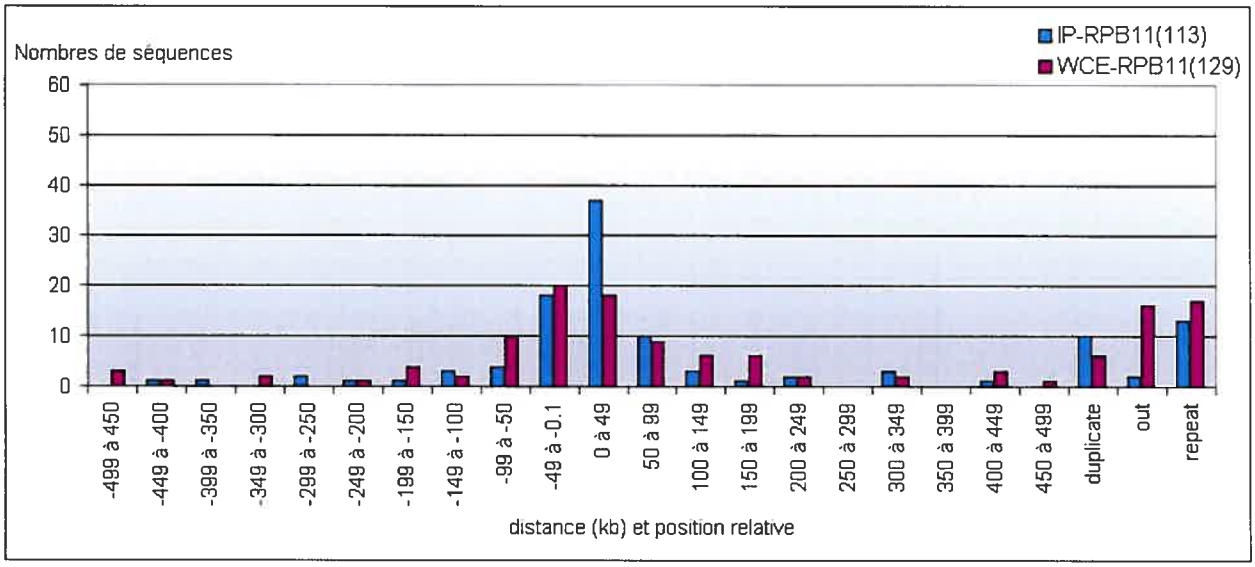


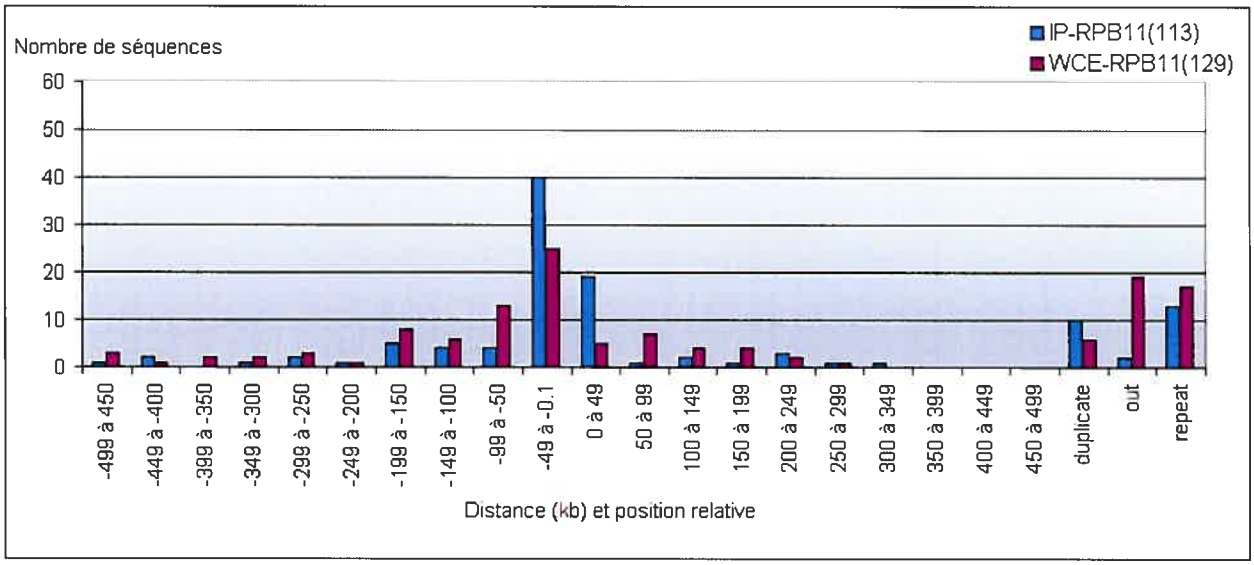
Figure 9. Localisation génomique de l'ARN Pol II.

A) Localisation par rapport au CGI fort (≥300pb, GC 0.6%, etc..) le plus proche. L'axe des ordonnées représente le nombre de clones obtenus. L'axe des abscisses indique la distance (en kb) séparant la séquence clonée du CGI. B) Localisation par rapport au Refseq le plus proche. C) Localisation par rapport au SIT le plus proche. D) Localisation par rapport au 3'end d'un Refseq. (Suite page suivante)

C) Localisation génomique des séquences clonées par rapport au SIT le plus proche



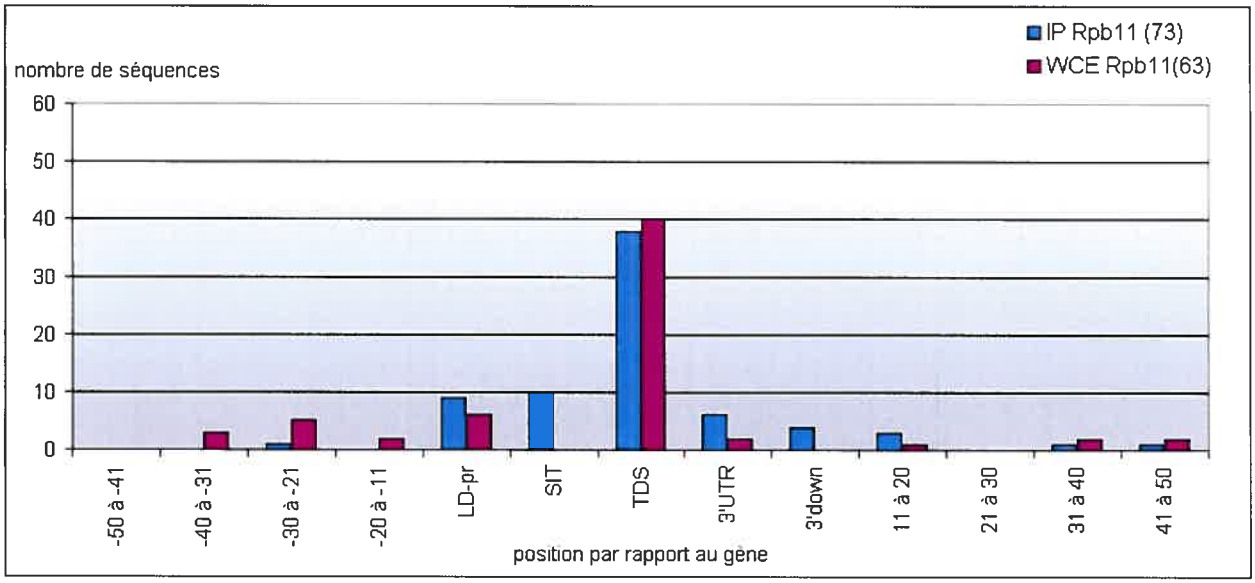
D) Localisation génomique des séquences clonées par rapport au *3'end* le plus proche



(Figure 9. suite)

Pour les graphiques B, C et D, la position relative de la séquence tient compte de l'orientation du gène : négatif en amont et positif en aval. *Duplicate* indique le nombre de clones dupliqués lors de la croissance sur pétris. *Out* indique le nombre de clones dépassant la distance limite de 500kb par rapport à l'élément considéré. *Repeat* indique le nombre de clones dont la séquence tombe dans une région répétée du génome.

A) Localisation fonctionnelle des séquences clonées par rapport à Refseq



B) Liste des gènes classés en fonction de la région transcriptionnelle

	LD-PR	SIT	TDS	3'UTR	3'down
WCE	DSCR1 BAF53A COLQ C9orf27 CITED1 TNFRSF13B		TNKS; NP220; TPR; NAV2; PCDH16; SMOC2; FLJ22405;AK122624; GPC6; TCF7L2; C7orf23; PCDH11X; CCDC9; FLJ32685; AB020672; GRID2;NFATC2; NRXN3; HGNT-IV-H; SLIT2; PTPN11; GHR; STAG2; SPTLC2; SGCZ; PPARG; RHOBTB3; DND; FLJ22679; BC063432; CACNG3; KHDRBS2; AKR1D1; VRK2; NT5M.	KIAA0828 C10orf46	
IP	CTPS FLJ20527 UPK3A HLA-G LTC4S TUFM DUFD1 ZNF567 PITPNM2	BC036762 IMP-1 STMN1 TRRAP PPP3CA ZNF43 HMGB2 COX7B DAFT1 C6orf119	ZNF515; RGS7; FAF1; LOC84524; ARID1B; C21orf42; M22406; TFCP2L1; NUP214; TCTEL1; KIF5C; ARMC4; COG1; WRN; LOC285513; BC057245; DD5; ZNF585B; MLR2; ZNF533; CTBP2; WBP11; FLJ32743; AK000210; KIN; FLJ34870; HDAC8; AFAP; HNRPC; SPATA5; C1orf24; PSPC1; NRK; CACNB4; PLEC1; TG; ZCCHC14; ATP6V0A1.	PDXK SULT1A3 FLJ11126 RNF135 LOC115294 BAZ2A	KCNJ6 BC041587 RANBP2L1 MBP

Figure 10. Analyse fonctionnelle des séquences associées à l'ARN Pol II humaine.

A) Répartition des séquences clonées IP^+ et WCE^+ provenant du TAP-xChIP de RPB11. Une échelle de 10kb a été utilisée en amont et en aval du gène. Les régions SIT et 3'UTR bénéficient d'une échelle plus précise de 2kb. LD-pr, *long-distance promoter* s'étend de -10 à -1kb. 3'down, *3'downstream* s'étend de 0 à +10kb. La région « TDS » dépend de la taille du gène, elle s'étend de 1kb après le SIT jusqu'au codon STOP de chaque gène. B) Liste des gènes auxquels les séquences clonées sont associées.

L'exploration du génome nécessite une échelle de distance adaptée : 50kb nous a semblé être une taille raisonnable comparée au 2.9Gb du génome entier. Toutefois, la taille moyenne des gènes humains étant estimé à 27kb (Lander et al., 2001; Venter et al., 2001), l'échelle de 50kb englobe l'amont, l'aval et le gène en un tout peu représentatif. Afin de mieux départager les différences entre l'enrichissement spécifique et le bruit de fond, il nous faut augmenter la résolution de notre échelle. Tenant compte de ce paramètre, une analyse à plus haute résolution (2 ou 10kb) de la localisation des clones à proximité d'un gène a été pratiquée (figure 10A). Différentes échelles ont été utilisées afin de mettre en évidence les régions d'un gène telles qu'elles ont été décrites au cours de l'introduction (promoteur, SIT, TDS, 3'UTR et 3'down). C'est dans le TDS, représentant tout ce qui est transcrit de +1kb par rapport au SIT au codon STOP, que nous trouvons le plus de séquences (38 et 40 respectivement IP^+ et WCE^+). Toutefois, en tenant compte du nombre total de séquences analysées (38 sur 113 pour IP^+ et 40 sur 129 pour WCE^+), nous nous rendons compte que ces valeurs sont assez proches de l'estimation du pourcentage du génome couvert par les Refseq (25% comparé à 33% et 31% respectivement IP^+ et WCE^+). Donc par rapport aux TDS, l'IP^+ ressemble au WCE^+, et serait plus l'effet d'un bruit de fond que d'un réel enrichissement spécifique. Les régions SIT et 3'UTR, respectivement englobant le SIT (±1kb) et la fin non codante du gène (entre 1 et 2kb), bénéficient d'une résolution d'environ 2kb. Sur le SIT, seuls des clones IP^+ (10 clones) ont été retrouvés. En 3' du gène (3'UTR et 3'*down*), une proportion largement supérieure de clones IP^+ par rapport aux WCE^+ est détectée (respectivement 6 et 4 contre 2 et 0). Les régions SIT et 3'UTR, représentant statistiquement 1.8% du génome, sont trouvées respectivement à 8.8% et 5.3% dans

l'ADN IP^+ et 0% et 1.5% dans l'ADN WCE^+. Bien que le nombre de clones soit faible, de manière générale, les séquences IP^+ révèlent bien la présence de l'ARN Pol II sur un gène transcrit. Une analyse par ontologie, utilisant FATIGO[†††] (Al Shahrour et al., 2004), des gènes obtenus lors des clonages (figure 10B) n'a pas révélé d'appartenance particulière à une classe fonctionnelle (données non montrées).

Pour résumer, l'approche génomique par clonage après TAP-xChIP a permit la localisation de l'ARN Pol II sur des régions transcriptionnelle très petites et précises (SIT et *3'end*), contrairement à des régions plus générales (TDS) qui, compte tenu de leurs grandeurs, se perdent dans le bruit de fond. Associée aux résultats à l'échelle du gène, en terme de quantification de l'enrichissement, l'approche génomique présentée ici démontre la faisabilité et l'efficacité de la localisation par TAP-xChIP en double affinité.

[†††] : FATIGO, http://fatigo.bioinfo.cnio.es/

IV) Discussion

Séparément, les techniques de ChIP (revues par Orlando, 2000) et de TAP-tag (Rigaut et al., 1999) ont prouvé leurs efficacités et se révèlent de précieux et puissants outils en génomique et en protéomique (respectivement Harbison et al., 2004 et Gavin et al., 2002). La combinaison des deux, en utilisant la première affinité du TAP-tag (par IgG) a été publiée par notre laboratoire (Jeronimo et al., 2004) ainsi que plusieurs autres (pour exemple Kim et al., 2004; Krogan et al., 2004; Verdel et al., 2004; Westermann et al., 2003). Actuellement cette technique reste limitée à une approche par « cibles présumées » (ex : amplification par PCR ou puce à ADN). Afin de profiter au maximum du potentiel qu'offre le TAP-tag, nous avons développé un essai ChIP utilisant les deux affinités de l'étiquette TAP et avons confirmé son efficacité par clonage, une approche par « cibles inconnues ».

IV.1. Mise au point de la méthode

La difficulté majeure rencontrée avec le TAP-xChIP réside dans la structure de l'étiquette elle-même. Pour être optimale, l'étiquette TAP doit avoir une conformation permettant à la fois 1) l'accessibilité des billes d'IgG à la protéine A, 2) l'accessibilité et la reconnaissance par la TEV de son site de coupure, et 3) l'accessibilité des billes de calmoduline au CBP. Étant de nature protéique et de grande taille (20kDa), le TAP est sujet aux pontages fait par le formaldéhyde (Metz et al., 2004) et cela s'avère critique lorsque l'on passe en double affinité. Pour permettre une conservation conformationnelle la plus native possible, tout en permettant la fixation des protéines à l'ADN, le temps d'action du formaldéhyde a été diminué au minimum (voir matériel et méthodes, inspiré

de Saunders et al., 2003). De plus, il est connu depuis longtemps que les pontages à la formaldéhyde rendent la chromatine plus ou moins bien fractionnable. Or, l'une des conditions critiques pour une efficacité optimale de l'immunoprécipitation reste la taille des fragments de chromatine qui, lorsqu'ils sont trop grands, nuisent considérablement à la résolution de la technique (Orlando et al., 1997).

IV.2. La méthode TAP-xCHIP : analyse au niveau du gène

Dans un premier temps, la méthode a été validée par des essais classiques de ChIP (PCR et/ou QPCR), en ciblant des régions candidates (tableau I). Étudiant la localisation de l'ARN Pol II (via sa sous-unité taggée RPB11), le choix des contrôles négatifs (non transcrits) est très important. De nombreuses études proposent comme contrôle lors d'une ChIP d'utiliser une région en amont ou en aval de la région candidate (Laganiere et al., 2003; Liang et al., 2004; Parekh and Maniatis, 1999). Si cette approche est valable lors de la localisation de facteurs de transcription, fixes sur l'ADN, elle s'applique plus difficilement à l'ARN Pol II, extrêmement mobile sur l'ADN. Suivant l'exemple des études faites chez la levure (par exemple Wood et al., 2003), l'utilisation de régions non transcrites situés à de plus grande distance des gènes, principalement dans les *gene desert island* (Nobrega et al., 2003), représente la meilleure solution (tableau V).

Ainsi, à l'échelle du gène, nous avons montré que la méthode fonctionne et que le bruit de fond est faible (figure 6D). L'amplification par QPCR indique que l'ARN Pol II est retrouvée sur les gènes transcrits, avec une proportion plus importante sur le promoteur (figure 7B). Ceci peut s'expliquer par une probabilité plus importante de trouver l'ARN Pol II sur les promoteurs (initiation de la transcription) comparativement à

une région transcrite quelconque le long du gène. Toutefois, le cas du promoteur de FTL est à noter. FTL est un gène de très petite taille (~1500pb), ubiquitaire et très fortement exprimé, il possède au niveau du promoteur une CGI particulièrement grande. Or, la méthode TAP-tag, chez l'humain, immunoprécipite de manière non spécifique TFII-I (retrouvé dans les éluats provenant des cellules non-induites, communication personnelle de Jeronimo C.). TFII-I est un TGF retrouvé au promoteur de certains gènes $TATA^-$ /Inr^+/CGI^+ et il interagit directement avec la machinerie transcriptionnelle (Roy et al., 1993). FTL, un gène de maintenance, possède un promoteur du type Inr^+/CGI^+ et aucune boite TATA n'a été formellement identifiée (utilisation de TFSEARCH, Dragon PF, Promoter 2.0, résultats non montrés). Ainsi, la possible présence de TFII-I au niveau du promoteur de FTL pourrait expliquer l'enrichissement observé dans la lignée non-induite (enrichie au promoteur et non dans la région transcrite, figure 7.A).

IV.3. Le TAP-xChIP enrichit les régions géniques

À ce jour et à notre connaissance, c'est la première fois que la possibilité de faire une double immunoprécipitation de la chromatine en utilisant le système TAP-tag et de valider la méthode par clonage est démontrée. Toutes les études traitant actuellement de clonage suite à une ChIP n'analysent que l'ADN IP. Pour palier à un possible biais lors du clonage, nous avons décidé de traiter en parallèle l'ADN IP^+ et de l'ADN non immunoprécipité (WCE^+).

Si aucune différence notable n'est observée entre les tailles des séquences IP^+ et WCE^+ (Figure 8A), une légère tendance à être localisée sur les chromosomes riches en gènes (ch16, 17, 19) est observée pour les séquences IP^+ (figure 8B). L'absence de tout

clone IP^+ sur le chromosome 3 est plus surprenante et reste inexpliquée. Le chromosome 3 n'est ni un chromosome pauvre en gène ni de petite taille (tableau III).

Du point de vue génomique, le patron de distribution pratiquement identique entre IP^+/WCE^+ par rapport aux CGIs est assez surprenant (figure 9A) sachant qu'il est prouvé que les CGIs ne sont pas distribuées aléatoirement dans le génome Craig and Bickmore, 1994). Le fait est encore plus troublant lorsque l'on regarde la distribution des séquences par rapport aux gènes connus, aux SITs et aux *3'end*. Si l'on s'attend à voir un enrichissement particulièrement fort pour l'ADN IP^+ provenant de RPB11, la même chose n'est vraiment pas attendue pour WCE^+ (figure 9B, C et D). Une hypothèse pouvant expliquer ce biais implique à la fois le formaldéhyde (et ses pontages) et une destruction totale ou partielle des régions intergéniques. Baignant encore dans un milieu nucléaire lors du renversement des pontages, contrairement à l'ADN IP, la chromatine « WCE » pourrait être la cible de nucléases s'attaquant aux régions moins protégées en protéines (les régions intergéniques). Une étude réalisée chez la levure démontre que les pontages au formaldéhyde, puis leur renversement, enrichissent les régions chromatiniennes contenant des gènes par rapport à celles intergéniques (Nagy et al., 2003).

Par conséquent, malgré l'efficacité du TAP-xChIP observée au niveau des gènes (voir figure 7), il est indispensable de traiter en même temps IP/WCE lors des clonages.

IV.4. Localisation de l'ARN Pol II sur des régions fonctionnelles très précises.

C'est en considérant l'enrichissement de fragments près d'un SIT que l'on se rend le mieux compte de l'efficacité du TAP-xChIP. Cette région, considérée ici de ±1kb autour du SIT de chaque gène de classe II, ne devrait théoriquement pas représenter plus de ~1.8% du génome lors d'une analyse aléatoire ; elle est ici de l'ordre de 8.8% (figure 10A). La même observation est aussi valable pour la région 3'UTR (du codon « stop » au *3'end*), avec les valeurs de 5.5% pour l'IP^+ et 1.5% pour WCE^+. Conscient de la possibilité d'obtenir des résultats faussés par un contaminant de la méthode (du type TFII-I, voir section précédente) l'ADN IP^- a été analysé et traité dans les mêmes conditions (résultats non montrés). Outre le fait que cet ADN était en quantité encore plus faible que celui IP^+ (tel qu'attendu si la méthode fonctionne), un nombre limité de séquences a été obtenu (16 clones dont 13 uniques). Parmi celles-ci, 6 tombent dans un ORF, 1 dans le *3'downstream* (de 0 à +10kb en aval du gène) et le reste au-delà de ±10kb autour d'un gène.

Ainsi, malgré l'absence d'enrichissement sur les régions TDS, la fraction des séquences clonées correspondant aux SIT ainsi qu'aux 3'UTR semble bel et bien résulté d'un enrichissement spécifique (figure 10). Si par ces résultats, la méthode démontre bien son efficacité au niveau génomique, l'analyse à plus grande échelle sur un ensemble de régions précises permettrait de diminuer le bruit de fond observé lors des clonages. Une étude par *location array* possédant des SIT, des 3'UTR en plus des TDS devrait permettre cela.

IV.5. Le TAP-xChIP efficace en clonage.

Malgré le nombre limité de séquences analysées, l'ADN IP^+ indique la localisation de l'ARN Pol II de -10 à 10/20kb autour des gènes de classe II (limité au Refseq, voir figure 10). La détection de l'ARN Pol II entre -10 et +10/20kb autour des Refseq peut être expliquée par plusieurs hypothèses. Les pontages indirectes, lors d'interactions entre l'ARN Pol II et des protéines localisées loin des bornes des gènes, sont une possibilité. Toutefois, gardant à l'esprit la topologie de la chromatine in vivo (voir figure 4) et compte tenu des résultats obtenus en PCR, l'hypothèse d'une interaction directe entre l'ARN Pol II et la région clonée semble la meilleure. Allant dans ce sens, un étude à très haute résolution du transcriptome humain a démontré, entre autre, l'existence de nombreux exons transcrits au-delà des limites actuellement établies des gènes (Bertone et al., 2004).

Mis ensemble, ces résultats démontrent : 1) qu'il est possible d'utiliser les deux affinités de l'étiquette TAP en ChIP, 2) l'efficacité du TAP-xChIP dans la localisation génomique d'un complexe aussi mobile que l'ARN polymérase II humaine, 3) que le bruit de fond est assez réduit pour permettre la détection de régions très spécifiques dans le génome, et 4) qu'en analysant uniquement les gènes connus, la machinerie transcriptionnelle a été détectée bien en dehors des bornes limitantes du gène.

V) Conclusions et perspectives

Durant les dernières pages de ce mémoire, je résumerai brièvement la contribution apportée par le présent travail au sein du projet global entrepris par notre laboratoire, ainsi que ses perspectives. Enfin, je terminerai par une projection personnelle des challenges génomiques à venir.

V.1. « Du génome à l'organisme »

S'inscrivant dans un vaste projet de génomique et de protéomique, notre laboratoire a entrepris de caractériser, chez l'humain, les réseaux de régulateurs de l'expression génique. Dans ce projet, les volets de génomique et de protéomique avancent conjointement, étant connectés l'un à l'autre (Coulombe et al., 2004). Utilisant la technologie de TAP-tag, plusieurs facteurs généraux de transcription (incluant l'ARN Pol II) ont été purifiés (Jeronimo et al., 2004). Etiquetés à leur tour, ils servent à purifier d'autres complexes impliqués dans la transcription.

Le but initial de mon projet était de développer une méthode permettant l'analyse et la caractérisation des réseaux régulateurs, à l'échelle du génome et *in vivo*, des régulateurs de la transcription. Démarré en septembre 2002, l'approche choisie consistait à développer et à utiliser une puce à ADN (type *location array*) et une technique d'immunoprécipitation de la chromatine associée au système de purification par double affinité du TAP-tag.

Le premier volet du projet consistait à construire et à utiliser une puce à ADN du type *location array* possédant le maximum de promoteurs humains de -900 à +100 autour du SIT (exemple de design Odom et al., 2004) afin d'y détecter la présence de la

machinerie transcriptionnelle. Si l'étape de construction a été atteinte sous la forme de notre base de SITs, son utilisation par contre n'a pas été possible (celle-ci a été rendue opérationnelle à la fin de la rédaction de ce mémoire). Il serait donc intéressant, dans un premier temps, d'analyser, via cette *location array* de promoteurs humains, la localisation de l'ARN Pol II (via sa sous-unité RPB11).

Le deuxième volet du projet a été atteint en déterminant la localisation de RPB11 à l'échelle du gène (QPCR), puis du génome (par clonage), démontrant l'efficacité du TAP-xChIP séquentiel. L'utilisation de cette technique sur d'autres membres de la machinerie transcriptionnelle de base constitue l'étape suivante de cette partie du projet. L'accessibilité à une *location array* de promoteurs humains devrait permettre de détecter, pour les facteurs plus difficiles à immunoprécipiter, les promoteurs auxquels ils sont associés. Une étude pilote réalisée avec TFIIB a démontré que la technique, sans aucun changement par rapport à celle utilisée avec RPB11, fonctionnait mais que l'enrichissement était plus faible (~5 fois moins fort que celui de RPB11 sur le promoteur de l'énolase α). Un enrichissement trop faible nuit aussi, par la trop faible quantité d'ADN récupérée, à la localisation par clonage. Pour pallier à ce manque d'ADN, une étape d'amplification par PCR (telle la LM-PCR, Ren et al., 2000) pourrait être ajoutée.

V.2. Corréler « localisation génomique » avec « expression génomique »

Réussir à localiser l'ARN Pol II sur le génome ne répond qu'à l'une des deux questions qu'il faut se poser lorsque l'on cherche à étudier la transcription. La seconde étant « Quel est le profil d'expression génomique de la cellule étudiée ?». Autrement dit, l'expression corrèle-t-elle avec la localisation ? Répondre à cette question constitue la

suite logique d'un projet de localisation.

La technologie des puces à ADN du type *expression array* de gènes de classe II représente bien sûr un excellent moyen d'y parvenir (Schena et al., 1995 et revue par Young, 2000). Dans l'hypothèse d'une analyse plus exhaustive, l'utilisation des *tiling arrays* (Selinger et al., 2000 qu'utilise aussi Bertone et al., 2004) présenterait un très bon parallèle avec des analyses par clonage. À plus petite échelle, il serait possible de faire, en plus d'une analyse de localisation complémentaire (le long du gène), une analyse par PCR inverse (RT-PCR) afin de mesurer l'expression des gènes en question. Ainsi, dans le cas des gènes dont le SIT a été localisé, il serait possible de savoir si l'ARN Pol II est en pause (gène non transcrit) ou active (gène transcrit). En étudiant l'expression des gènes, dont une région transcrite (nommé Refseq dans cette étude) a été analysée lors des clonages, il serait facile de faire la différence entre l'immunoprécipitation spécifique et le biais qui semble être imputable au formaldéhyde (Nagy et al., 2003). Enfin, l'analyse par RT-PCR des régions 3'UTR obtenues lors des clonages pourrait se révéler triplement intéressante. Premièrement, l'expression d'un ARNm est détectée et les essais de localisation confirment la présence de l'ARN Pol II le long du gène. Dans ce cas, l'hypothèse d'une densité d'ARN Pol II équivalente à celle du promoteur serait envisagée (présentée dans l'Introduction). Deuxièmement, l'ARN Pol II n'est pas détectée sur le SIT et un ARN est détecté non pas sur le brin sens mais sur celui antisens. De plus en plus d'évidences convergent vers l'existence d'une transcription antisens, qui pourrait jouer un rôle dans la régulation de la transcription (Yelin et al., 2003; Bertone et al., 2004; Kampa et al., 2004; Lipman, 1997; initialement revue par Inouye, 1988). Enfin, aucun ARN n'est détecté dans aucun des deux sens et l'ARN Pol II n'est pas détectée

dans l'unité de transcription. Dans ce cas, plusieurs hypothèses peuvent être émises dont par exemple le cas d'une pause sur la fin du gène en attendant peut-être le signal d'une transcription antisens. L'hypothèse d'une boucle régulatrice, entre le SIT et la fin du gène (démontrée chez la levure O'Sullivan et al., 2004), est aussi à considérer auquel cas le test de localisation au niveau du SIT devrait révéler la présence de l'ARN Pol II.

Si ces modes de régulation de la transcription sont désormais connus, la composition de la machinerie transcriptionnelle nécessaire au fonctionnement de ces mécanismes reste en revanche inconnue.

V.3. Au-delà des bornes du gène : la zone de transcription s'étend.

Comme l'ont démontré les nombreuses études le caractérisant manuellement ou à grande échelle (MGC et DBTSS, respectivement Strausberg et al., 1999; Suzuki et al., 2002), le site de l'initiation de la transcription chez l'humain possède une organisation particulière. Supportés par de récentes évidences (Bertone et al., 2004), les gènes humains de classe II semblent posséder en plus de leur SIT majeur un ou plusieurs SIT(s) mineur(s) pouvant être utilisé(s), dépendamment du type ou de l'environnement cellulaire. Selon les résultats publiés par Bertone *et al.*, la transcription pourrait commencer jusqu'à 10kb en amont des gènes actuellement connus. Cette possibilité met alors l'emphase sur les séquences clonées qualifiées de « promoteur longue distance » (LD-pr, figure 10). Avec des évidences telles que 1) la transcription inverse (Yelin et al., 2003), 2) les sites d'épissage alternatifs, ou encore 3) la transcription intergénique (Martens et al., 2004), l'analyse de l'expression des gènes associée aux séquences localisées en aval des gènes (0 à +10/20kb) revêt-elle aussi un intérêt d'étude

supplémentaire. Ainsi la question de savoir si les sites de fixation éloignés en « amont » ne seraient pas tout simplement en fait proches d'un site d'initiation alternatif peut être posée.

V.4. Quand le transcriptome parle, l'obscurité s'éclaircit.

Depuis la publication de la séquence du génome humain, les études à grandes échelles ont explosé et on ne compte plus les articles étudiant le génome humain, ainsi que plus récemment le transcriptome. Qu'ils utilisent des techniques du type SAGE (5' ou 3') ou une approche par puce à ADN (d'*expression* ou de *location*, *tiling* ou pas), leurs résultats démontrent qu'une proportion énorme du génome, jusqu'à maintenant considérée comme du *junk-DNA* (dans certains cas appelé *DNA dark matter*, voir Pennisi, 2003), serait transcrite mais pas forcément codante. Passant outre le fait qu'actuellement toutes les études, même les plus récentes (Bertone et al., 2004), se basent sur la version du génome criblée d'erreurs et de trous, l'ampleur du phénomène est de taille (pour la nouvelle version du génome humain voir International Human Genome Sequencing Consortium, 2004, et comparer avec Lander et al., 2001; Venter et al., 2001). La séquence euchromatienne du génome humain étant maintenant de très bonne qualité, les analyses transcriptomiques futures devraient permettre d'évaluer précisément la portion du génome transcriptionnellement active. De telles études pourraient permettre entre autre de porter un regard différent sur le contenu d'un génome et ainsi peut-être permettre de mieux comprendre ce qui est actuellement appelé le *C-value Paradox*. L'exploration du *dark matter* de notre génome, qui semble être le prochain défi au niveau du transcriptome humain, révélera peut-être un rôle crucial dans l'intégrité cellulaire.

VI) Références

VI.1. Références électroniques

CGAP, Cancer Genome Anatomy Project
http://cgap.nci.nih.gov/

Celera Genomics
http://www.celera.com/

DBTSS, Database of human transcription start site
http://dbtss.hgc.jp/samp_home.html

Dragon PF, Dragon promoter finder
http://research.i2r.a-star.edu.sg/promoter/promoter1_5/DPF.htm

ENSembl
http://www.ensembl.org/Homo_sapiens/34dbuild.html

FATIGO, gene ontology
http://fatigo.bioinfo.cnio.es/

IHGSC, internationnal human genome sequencing consortium
http://www.genome.gov/

Human genome gateway
http://www.nature.com/nature/focus/humangenome/

MGC, mammalian gene collection
http://mgc.nci.nih.gov/

NCBI, nationnal center for bioechnology information
http://www.ncbi.nlm.nih.gov/

Primer3, Webware, design PCR amplification primer
http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi

Projet général du Dr Coulombe
http://www.ircm.qc.ca/microsites/expressiongenetique/fr/

Promoter 2.0, promoter prediction program (PPP)
http://www.cbs.dtu.dk/services/Promoter/

Refseq, NCBI reference sequence
http://www.ncbi.nlm.nih.gov/RefSeq/

SOURCE
http://source.stanford.edu

TFSEARCH, transcription factor search binding site
http://molsun1.cbrc.aist.go.jp/research/db/TFSEARCH.html

TRANSFAC
http://www.gene-regulation.com/

UCSC Genome Browser
http://genome.ucsc.edu/

UCSC in silico PCR
http://genome.ucsc.edu/cgi-bin/hgPcr

VI.2. Bibliographie

Adachi,N. and Lieber,M.R. (2002). Bidirectional gene organization: a common architectural feature of the human genome. Cell *109*, 807-809.

Agalioti,T., Lomvardas,S., Parekh,B., Yie,J., Maniatis,T., and Thanos,D. (2000). Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. Cell *103*, 667-678.

Ahmad,K. and Henikoff,S. (2002). Histone H3 variants specify modes of chromatin assembly. Proc. Natl. Acad. Sci. U. S. A *99 Suppl 4*, 16477-16484.

Al Shahrour,F., Diaz-Uriarte,R., and Dopazo,J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics *20*, 578-580.

Alen,C., Kent,N.A., Jones,H.S., O'Sullivan,J., Aranda,A., and Proudfoot,N.J. (2002). A role for chromatin remodeling in transcriptional termination by RNA polymerase II. Mol Cell *10*, 1441-1452.

Bejerano,G., Pheasant,M., Makunin,I., Stephen,S., Kent,W.J., Mattick,J.S., and Haussler,D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321-1325.

Bell,A.C., West,A.G., and Felsenfeld,G. (2001). Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. Science *291*, 447-450.

Bernardi,G. (2000). Isochores and the evolutionary genomics of vertebrates. Gene *241*, 3-17.

Bertone,P., Stolc,V., Royce,T.E., Rozowsky,J.S., Urban,A.E., Zhu,X., Rinn,J.L., Tongprasit,W., Samanta,M., Weissman,S., Gerstein,M., and Snyder,M. (2004). Global Identification of Human Transcribed Sequences with Genome Tiling Arrays. Science.

Bickmore,W.A. and Sumner,A.T. (1989). Mammalian chromosome banding--an expression of genome organization. Trends Genet. *5*, 144-148.

Bird,A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6-21.

Bird,A.P. (1986). CpG-rich islands and the function of DNA methylation. Nature *321*, 209-213.

Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T., Down,T., Eyras,E., Fernandez-Suarez,X.M., Gane,P., Gibbins,B., Gilbert,J., Hammond,M., Hotz,H.R., Iyer,V., Jekosch,K., Kahari,A., Kasprzyk,A., Keefe,D., Keenan,S., Lehvaslaiho,H., McVicker,G., Melsopp,C., Meidl,P., Mongin,E., Pettett,R., Potter,S., Proctor,G., Rae,M., Searle,S., Slater,G., Smedley,D., Smith,J.,

Spooner,W., Stabenau,A., Stalker,J., Storey,R., Ureta-Vidal,A., Woodwark,K.C., Cameron,G., Durbin,R., Cox,A., Hubbard,T., and Clamp,M. (2004). An overview of Ensembl. Genome Res. *14*, 925-928.

Black,D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annu. Rev. Biochem. *72*, 291-336.

Bleichenbacher,M., Tan,S., and Richmond,T.J. (2003). Novel interactions between the components of human and yeast TFIIA/TBP/DNA complexes. J. Mol Biol *332*, 783-793.

Bourbon,H.M., Aguilera,A., Ansari,A.Z., Asturias,F.J., Berk,A.J., Bjorklund,S., Blackwell,T.K., Borggrefe,T., Carey,M., Carlson,M., Conaway,J.W., Conaway,R.C., Emmons,S.W., Fondell,J.D., Freedman,L.P., Fukasawa,T., Gustafsson,C.M., Han,M., He,X., Herman,P.K., Hinnebusch,A.G., Holmberg,S., Holstege,F.C., Jaehning,J.A., Kim,Y.J., Kuras,L., Leutz,A., Lis,J.T., Meisterernest,M., Naar,A.M., Nasmyth,K., Parvin,J.D., Ptashne,M., Reinberg,D., Ronne,H., Sadowski,I., Sakurai,H., Sipiczki,M., Sternberg,P.W., Stillman,D.J., Strich,R., Struhl,K., Svejstrup,J.Q., Tuck,S., Winston,F., Roeder,R.G., and Kornberg,R.D. (2004). A unified nomenclature for protein subunits of mediator complexes linking transcriptional regulators to RNA polymerase II. Mol Cell *14*, 553-557.

Breathnach,R. and Chambon,P. (1981). Organization and expression of eucaryotic split genes coding for proteins. Annu. Rev. Biochem. *50*, 349-383.

Burge,C. and Karlin,S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol Biol *268*, 78-94.

Bushnell,D.A., Westover,K.D., Davis,R.E., and Kornberg,R.D. (2004). Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms. Science *303*, 983-988.

Callen,B.P., Shearwin,K.E., and Egan,J.B. (2004). Transcriptional interference between convergent promoters caused by elongation over the promoter. Mol. Cell *14*, 647-656.

Carruthers,L.M. and Hansen,J.C. (2000). The core histone N termini function independently of linker histones during chromatin condensation. J. Biol Chem. *275*, 37285-37290.

Cawley,S., Bekiranov,S., Ng,H.H., Kapranov,P., Sekinger,E.A., Kampa,D., Piccolboni,A., Sementchenko,V., Cheng,J., Williams,A.J., Wheeler,R., Wong,B., Drenkow,J., Yamanaka,M., Patel,S., Brubaker,S., Tammana,H., Helt,G., Struhl,K., and Gingeras,T.R. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell *116*, 499-509.

Chalkley,G.E. and Verrijzer,C.P. (1999). DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator. EMBO J. *18*, 4835-4845.

Chambeyron,S. and Bickmore,W.A. (2004). Does looping and clustering in the nucleus regulate gene expression? Curr. Opin. Cell Biol *16*, 256-262.

Chambon,P. (1975). Eukaryotic nuclear RNA polymerases. Annu. Rev. Biochem. *44*, 613-638.

Chasman,D.I., Flaherty,K.M., Sharp,P.A., and Kornberg,R.D. (1993). Crystal structure of yeast TATA-binding protein and model for interaction with DNA. Proc. Natl. Acad. Sci. U. S. A *90*, 8174-8178.

Chen,H.T. and Hahn,S. (2004). Mapping the Location of TFIIB within the RNA Polymerase II Transcription Preinitiation Complex; A Model for the Structure of the PIC. Cell *119*, 169-180.

Chen,J., Larochelle,S., Li,X., and Suter,B. (2003). Xpd/Ercc2 regulates CAK activity and mitotic progression. Nature *424*, 228-232.

Cho,E.J., Kobor,M.S., Kim,M., Greenblatt,J., and Buratowski,S. (2001). Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain. Genes Dev. *15*, 3319-3329.

Choder,M. (2004). Rpb4 and Rpb7: subunits of RNA polymerase II and beyond. Trends Biochem. Sci. *29*, 674-681.

Choder,M. and Young,R.A. (1993). A portion of RNA polymerase II molecules has a component essential for stress responses and stress survival. Mol. Cell Biol. *13*, 6984-6991.

Coin,F. and Egly,J.M. (1998). Ten years of TFIIH. Cold Spring Harb. Symp. Quant. Biol *63*, 105-110.

Colgan,D.F. and Manley,J.L. (1997). Mechanism and regulation of mRNA polyadenylation. Genes Dev. *11*, 2755-2766.

Corden,J.L. (1990). Tails of RNA polymerase II. Trends Biochem. Sci. *15*, 383-387.

Coulombe,B. and Burton,Z.F. (1999). DNA bending and wrapping around RNA polymerase: a "revolutionary" model describing transcriptional mechanisms. Microbiol. Mol Biol Rev. *63*, 457-478.

Coulombe,B., Jeronimo,C., Langelier,M.F., Cojocaru,M., and Bergeron,D. (2004). Interaction networks of the molecular machines that decode, replicate, and maintain the integrity of the human genome. Mol Cell Proteomics *3*, 851-856.

Craig,J.M. and Bickmore,W.A. (1994). The distribution of CpG islands in mammalian chromosomes. Nat. Genet. *7*, 376-382.

Cramer,P. (2002). Multisubunit RNA polymerases. Curr. Opin. Struct. Biol. *12*, 89-97.

Cramer,P., Bushnell,D.A., and Kornberg,R.D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science *292*, 1863-1876.

Dantonel,J.C., Murthy,K.G., Manley,J.L., and Tora,L. (1997). Transcription factor TFIID recruits factor CPSF for formation of 3' end of mRNA. Nature *389*, 399-402.

Dantonel,J.C., Quintin,S., Lakatos,L., Labouesse,M., and Tora,L. (2000). TBP-like factor is required for embryonic RNA polymerase II transcription in C. elegans. Mol. Cell *6*, 715-722.

Dermitzakis,E.T., Reymond,A., Lyle,R., Scamuffa,N., Ucla,C., Deutsch,S., Stevenson,B.J., Flegel,V., Bucher,P., Jongeneel,C.V., and Antonarakis,S.E. (2002). Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature *420*, 578-582.

Dion,V. and Coulombe,B. (2003). Interactions of a DNA-bound transcriptional activator with the TBP-TFIIA-TFIIB-promoter quaternary complex. J. Biol Chem. *278*, 11495-11501.

Douziech,M., Coin,F., Chipoulet,J.M., Arai,Y., Ohkuma,Y., Egly,J.M., and Coulombe,B. (2000). Mechanism of promoter melting by the xeroderma pigmentosum complementation group B helicase of transcription factor IIH revealed by protein-DNA photo-cross-linking. Mol Cell Biol *20*, 8168-8177.

Dubois,M.F., Nguyen,V.T., Bellier,S., and Bensaude,O. (1994). Inhibitors of transcription such as 5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole and isoquinoline sulfonamide derivatives (H-8 and H-7) promote dephosphorylation of the carboxyl-terminal domain of RNA polymerase II largest subunit. J. Biol Chem. *269*, 13331-13336.

Dvir,A. (2002). Promoter escape by RNA polymerase II. Biochim. Biophys. Acta *1577*, 208-223.

Dynlacht,B.D., Hoey,T., and Tjian,R. (1991). Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. Cell *66*, 563-576.

Dziembowski,A. and Seraphin,B. (2004). Recent developments in the analysis of protein complexes. FEBS Lett. *556*, 1-6.

Egly,J.M. (2001). The 14th Datta Lecture. TFIIH: from transcription to clinic. FEBS Lett. *498*, 124-128.

Elmendorf,B.J., Shilatifard,A., Yan,Q., Conaway,J.W., and Conaway,R.C. (2001). Transcription factors TFIIF, ELL, and Elongin negatively regulate SII-induced nascent transcript cleavage by non-arrested RNA polymerase II elongation intermediates. J. Biol Chem. *276*, 23109-23114.

Evans,R., Fairley,J.A., and Roberts,S.G. (2001). Activator-mediated disruption of sequence-specific DNA contacts by the general transcription factor TFIIB. Genes Dev.

15, 2945-2949.

Fairley,J.A., Evans,R., Hawkes,N.A., and Roberts,S.G. (2002). Core promoter-dependent TFIIB conformation and a role for TFIIB conformation in transcription start site selection. Mol Cell Biol *22*, 6697-6705.

Fazzari,M.J. and Greally,J.M. (2004). Epigenomics: beyond CpG islands. Nat. Rev. Genet. *5*, 446-455.

Felsenfeld,G. and Groudine,M. (2003). Controlling the double helix. Nature *421*, 448-453.

Fong,Y.W. and Zhou,Q. (2001). Stimulatory effect of splicing factors on transcriptional elongation. Nature *414*, 929-933.

Forget,D., Langelier,M.F., Therien,C., Trinh,V., and Coulombe,B. (2004). Photo-cross-linking of a purified preinitiation complex reveals central roles for the RNA polymerase II mobile clamp and TFIIE in initiation mechanisms. Mol Cell Biol *24*, 1122-1131.

Fraser,P. and Grosveld,F. (1998). Locus control regions, chromatin activation and transcription. Curr. Opin. Cell Biol. *10*, 361-365.

Fry,C.J. and Peterson,C.L. (2001). Chromatin remodeling enzymes: who's on first? Curr. Biol *11*, R185-R197.

Fuks,F., Hurd,P.J., Wolf,D., Nan,X., Bird,A.P., and Kouzarides,T. (2003). The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. J. Biol Chem. *278*, 4035-4040.

Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M., Remor,M., Hofert,C., Schelder,M., Brajenovic,M., Ruffner,H., Merino,A., Klein,K., Hudak,M., Dickson,D., Rudi,T., Gnau,V., Bauch,A., Bastuck,S., Huhse,B., Leutwein,C., Heurtier,M.A., Copley,R.R., Edelmann,A., Querfurth,E., Rybin,V., Drewes,G., Raida,M., Bouwmeester,T., Bork,P., Seraphin,B., Kuster,B., Neubauer,G., and Superti-Furga,G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141-147.

Giglia-Mari,G., Coin,F., Ranish,J.A., Hoogstraten,D., Theil,A., Wijgers,N., Jaspers,N.G., Raams,A., Argentini,M., van der Spek,P.J., Botta,E., Stefanini,M., Egly,J.M., Aebersold,R., Hoeijmakers,J.H., and Vermeulen,W. (2004). A new, tenth subunit of TFIIH is responsible for the DNA repair syndrome trichothiodystrophy group A. Nat. Genet. *36*, 714-719.

Gilbert,C., Kristjuhan,A., Winkler,G.S., and Svejstrup,J.Q. (2004a). Elongator interactions with nascent mRNA revealed by RNA immunoprecipitation. Mol Cell *14*, 457-464.

Gilbert,N., Boyle,S., Fiegler,H., Woodfine,K., Carter,N.P., and Bickmore,W.A. (2004b). Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. Cell *118*, 555-566.

Gnatt,A.L., Cramer,P., Fu,J., Bushnell,D.A., and Kornberg,R.D. (2001). Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. Science *292*, 1876-1882.

Goodrich,J.A. and Tjian,R. (1994). Transcription factors IIE and IIH and ATP hydrolysis direct promoter clearance by RNA polymerase II. Cell *77*, 145-156.

Govin,J., Caron,C., Lestrat,C., Rousseaux,S., and Khochbin,S. (2004). The role of histones in chromatin remodelling during mammalian spermiogenesis. Eur. J. Biochem. *271*, 3459-3469.

Grosveld,F. (1999). Activation by locus control regions? Curr. Opin. Genet. Dev. *9*, 152-157.

Grueneberg,D.A., Henry,R.W., Brauer,A., Novina,C.D., Cheriyath,V., Roy,A.L., and Gilman,M. (1997). A multifunctional DNA-binding protein that promotes the formation of serum response factor/homeodomain complexes: identity to TFII-I. Genes Dev. *11*, 2482-2493.

Hannon,G.J. and Rossi,J.J. (2004). Unlocking the potential of the human genome with RNA interference. Nature *431*, 371-378.

Hansen,S.K., Takada,S., Jacobson,R.H., Lis,J.T., and Tjian,R. (1997). Transcription properties of a cell type-specific TATA-binding protein, TRF. Cell *91*, 71-83.

Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J., Jennings,E.G., Zeitlinger,J., Pokholok,D.K., Kellis,M., Rolfe,P.A., Takusagawa,K.T., Lander,E.S., Gifford,D.K., Fraenkel,E., and Young,R.A. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431*, 99-104.

Harrison,P.M., Kumar,A., Lang,N., Snyder,M., and Gerstein,M. (2002). A question of size: the eukaryotic proteome and the problems in defining it. Nucleic Acids Res. *30*, 1083-1090.

Hartl,D.L. (2000). Molecular melodies in high and low C. Nat. Rev. Genet. *1*, 145-149.

Hayes,J.J. and Hansen,J.C. (2001). Nucleosomes and the chromatin fiber. Curr. Opin. Genet. Dev. *11*, 124-129.

Holstege,F.C., Jennings,E.G., Wyrick,J.J., Lee,T.I., Hengartner,C.J., Green,M.R., Golub,T.R., Lander,E.S., and Young,R.A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. Cell *95*, 717-728.

Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T., Durbin,R., Eyras,E., Gilbert,J., Hammond,M., Huminiecki,L., Kasprzyk,A., Lehvaslaiho,H., Lijnzaad,P., Melsopp,C., Mongin,E., Pettett,R., Pocock,M., Potter,S., Rust,A., Schmidt,E., Searle,S., Slater,G., Smith,J., Spooner,W., Stabenau,A., Stalker,J., Stupka,E., Ureta-Vidal,A., Vastrik,I., and Clamp,M. (2002). The Ensembl genome database project. Nucleic Acids Res. *30*, 38-41.

Inouye,M. (1988). Antisense RNA: its functions and applications in gene regulation--a review. Gene *72*, 25-34.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931-945.

Jaillon,O., Aury,J.M., Brunet,F., Petit,J.L., Stange-Thomann,N., Mauceli,E., Bouneau,L., Fischer,C., Ozouf-Costaz,C., Bernot,A., Nicaud,S., Jaffe,D., Fisher,S., Lutfalla,G., Dossat,C., Segurens,B., Dasilva,C., Salanoubat,M., Levy,M., Boudet,N., Castellano,S., Anthouard,V., Jubin,C., Castelli,V., Katinka,M., Vacherie,B., Biemont,C., Skalli,Z., Cattolico,L., Poulain,J., De,B., V, Cruaud,C., Duprat,S., Brottier,P., Coutanceau,J.P., Gouzy,J., Parra,G., Lardier,G., Chapple,C., McKernan,K.J., McEwan,P., Bosak,S., Kellis,M., Volff,J.N., Guigo,R., Zody,M.C., Mesirov,J., Lindblad-Toh,K., Birren,B., Nusbaum,C., Kahn,D., Robinson-Rechavi,M., Laudet,V., Schachter,V., Quetier,F., Saurin,W., Scarpelli,C., Wincker,P., Lander,E.S., Weissenbach,J., and Roest,C.H. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature *431*, 946-957.

Jeronimo,C., Langelier,M.F., Zeghouf,M., Cojocaru,M., Bergeron,D., Baali,D., Forget,D., Mnaimneh,S., Davierwala,A.P., Pootoolal,J., Chandy,M., Canadien,V., Beattie,B.K., Richards,D.P., Workman,J.L., Hughes,T.R., Greenblatt,J., and Coulombe,B. (2004). RPAP1, a novel human RNA polymerase II-associated protein affinity purified with recombinant wild-type and mutated polymerase subunits. Mol Cell Biol *24*, 7043-7058.

Jurica,M.S. and Moore,M.J. (2003). Pre-mRNA splicing: awash in a sea of proteins. Mol Cell *12*, 5-14.

Kahn,P. (1995). From genome to proteome: looking at a cell's proteins. Science *270*, 369-370.

Kamenski,T., Heilmeier,S., Meinhart,A., and Cramer,P. (2004). Structure and mechanism of RNA polymerase II CTD phosphatases. Mol. Cell *15*, 399-407.

Kampa,D., Cheng,J., Kapranov,P., Yamanaka,M., Brubaker,S., Cawley,S., Drenkow,J., Piccolboni,A., Bekiranov,S., Helt,G., Tammana,H., and Gingeras,T.R. (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res. *14*, 331-342.

Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J., Weber,R.J., Haussler,D., and Kent,W.J. (2003).

The UCSC Genome Browser Database. Nucleic Acids Res. *31*, 51-54.

Kent,W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res. *12*, 656-664.

Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M., and Haussler,D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996-1006.

Khorasanizadeh,S. (2004). The nucleosome: from genomic organization to genomic regulation. Cell *116*, 259-272.

Kim,M., Ahn,S.H., Krogan,N.J., Greenblatt,J.F., and Buratowski,S. (2004). Transitions in RNA polymerase II elongation complexes at the 3' ends of genes. EMBO J. *23*, 354-364.

Kim,Y.J., Bjorklund,S., Li,Y., Sayre,M.H., and Kornberg,R.D. (1994). A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. Cell *77*, 599-608.

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

Komarnitsky,P., Cho,E.J., and Buratowski,S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. Genes Dev. *14*, 2452-2460.

Krebs,J.E., Kuo,M.H., Allis,C.D., and Peterson,C.L. (1999). Cell cycle-regulated histone acetylation required for expression of the yeast HO gene. Genes Dev. *13*, 1412-1421.

Krogan,N.J., Dover,J., Wood,A., Schneider,J., Heidt,J., Boateng,M.A., Dean,K., Ryan,O.W., Golshani,A., Johnston,M., Greenblatt,J.F., and Shilatifard,A. (2003). The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation. Mol. Cell *11*, 721-729.

Krogan,N.J., Kim,M., Ahn,S.H., Zhong,G., Kobor,M.S., Cagney,G., Emili,A., Shilatifard,A., Buratowski,S., and Greenblatt,J.F. (2002). RNA polymerase II elongation factors of Saccharomyces cerevisiae: a targeted proteomics approach. Mol Cell Biol *22*, 6979-6992.

Krogan,N.J., Peng,W.T., Cagney,G., Robinson,M.D., Haw,R., Zhong,G., Guo,X., Zhang,X., Canadien,V., Richards,D.P., Beattie,B.K., Lalev,A., Zhang,W., Davierwala,A.P., Mnaimneh,S., Starostine,A., Tikuisis,A.P., Grigull,J., Datta,N., Bray,J.E., Hughes,T.R., Emili,A., and Greenblatt,J.F. (2004). High-definition macromolecular composition of yeast RNA-processing complexes. Mol. Cell *13*, 225-239.

Kuhn,E.J. and Geyer,P.K. (2003). Genomic insulators: connecting properties to mechanism. Curr. Opin. Cell Biol *15*, 259-265.

Kuras,L., Borggrefe,T., and Kornberg,R.D. (2003). Association of the Mediator complex

with enhancers of active genes. Proc. Natl. Acad. Sci. U. S. A *100*, 13887-13891.

Kurdistani,S.K. and Grunstein,M. (2003). Histone acetylation and deacetylation in yeast. Nat. Rev. Mol Cell Biol *4*, 276-284.

Kutach,A.K. and Kadonaga,J.T. (2000). The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. Mol. Cell Biol. *20*, 4754-4764.

Lachner,M., O'Sullivan,R.J., and Jenuwein,T. (2003). An epigenetic road map for histone lysine methylation. J. Cell Sci. *116*, 2117-2124.

Laganiere,J., Deblois,G., and Giguere,V. (2003). Nuclear receptor target gene discovery using high-throughput chromatin immunoprecipitation. Methods Enzymol. *364*, 339-350.

Lagrange,T., Kapanidis,A.N., Tang,H., Reinberg,D., and Ebright,R.H. (1998). New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. Genes Dev. *12*, 34-44.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., Funke,R., Gage,D., Harris,K., Heaford,A., Howland,J., Kann,L., Lehoczky,J., LeVine,R., McEwan,P., McKernan,K., Meldrim,J., Mesirov,J.P., Miranda,C., Morris,W., Naylor,J., Raymond,C., Rosetti,M., Santos,R., Sheridan,A., Sougnez,C., Stange-Thomann,N., Stojanovic,N., Subramanian,A., Wyman,D., Rogers,J., Sulston,J., Ainscough,R., Beck,S., Bentley,D., Burton,J., Clee,C., Carter,N., Coulson,A., Deadman,R., Deloukas,P., Dunham,A., Dunham,I., Durbin,R., French,L., Grafham,D., Gregory,S., Hubbard,T., Humphray,S., Hunt,A., Jones,M., Lloyd,C., McMurray,A., Matthews,L., Mercer,S., Milne,S., Mullikin,J.C., Mungall,A., Plumb,R., Ross,M., Shownkeen,R., Sims,S., Waterston,R.H., Wilson,R.K., Hillier,L.W., McPherson,J.D., Marra,M.A., Mardis,E.R., Fulton,L.A., Chinwalla,A.T., Pepin,K.H., Gish,W.R., Chissoe,S.L., Wendl,M.C., Delehaunty,K.D., Miner,T.L., Delehaunty,A., Kramer,J.B., Cook,L.L., Fulton,R.S., Johnson,D.L., Minx,P.J., Clifton,S.W., Hawkins,T., Branscomb,E., Predki,P., Richardson,P., Wenning,S., Slezak,T., Doggett,N., Cheng,J.F., Olsen,A., Lucas,S., Elkin,C., Uberbacher,E., Frazier,M., Gibbs,R.A., Muzny,D.M., Scherer,S.E., Bouck,J.B., Sodergren,E.J., Worley,K.C., Rives,C.M., Gorrell,J.H., Metzker,M.L., Naylor,S.L., Kucherlapati,R.S., Nelson,D.L., Weinstock,G.M., Sakaki,Y., Fujiyama,A., Hattori,M., Yada,T., Toyoda,A., Itoh,T., Kawagoe,C., Watanabe,H., Totoki,Y., Taylor,T., Weissenbach,J., Heilig,R., Saurin,W., Artiguenave,F., Brottier,P., Bruls,T., Pelletier,E., Robert,C., Wincker,P., Smith,D.R., Doucette-Stamm,L., Rubenfield,M., Weinstock,K., Lee,H.M., Dubois,J., Rosenthal,A., Platzer,M., Nyakatura,G., Taudien,S., Rump,A., Yang,H., Yu,J., Wang,J., Huang,G., Gu,J., Hood,L., Rowen,L., Madan,A., Qin,S., Davis,R.W., Federspiel,N.A., Abola,A.P., Proctor,M.J., Myers,R.M., Schmutz,J., Dickson,M., Grimwood,J., Cox,D.R., Olson,M.V., Kaul,R., Raymond,C., Shimizu,N., Kawasaki,K., Minoshima,S., Evans,G.A., Athanasiou,M., Schultz,R., Roe,B.A., Chen,F., Pan,H., Ramser,J., Lehrach,H., Reinhardt,R., McCombie,W.R., de la,B.M., Dedhia,N., Blocker,H., Hornischer,K., Nordsiek,G., Agarwala,R., Aravind,L., Bailey,J.A., Bateman,A., Batzoglou,S., Birney,E., Bork,P.,

Brown,D.G., Burge,C.B., Cerutti,L., Chen,H.C., Church,D., Clamp,M., Copley,R.R., Doerks,T., Eddy,S.R., Eichler,E.E., Furey,T.S., Galagan,J., Gilbert,J.G., Harmon,C., Hayashizaki,Y., Haussler,D., Hermjakob,H., Hokamp,K., Jang,W., Johnson,L.S., Jones,T.A., Kasif,S., Kaspryzk,A., Kennedy,S., Kent,W.J., Kitts,P., Koonin,E.V., Korf,I., Kulp,D., Lancet,D., Lowe,T.M., McLysaght,A., Mikkelsen,T., Moran,J.V., Mulder,N., Pollara,V.J., Ponting,C.P., Schuler,G., Schultz,J., Slater,G., Smit,A.F., Stupka,E., Szustakowski,J., Thierry-Mieg,D., Thierry-Mieg,J., Wagner,L., Wallis,J., Wheeler,R., Williams,A., Wolf,Y.I., Wolfe,K.H., Yang,S.P., Yeh,R.F., Collins,F., Guyer,M.S., Peterson,J., Felsenfeld,A., Wetterstrand,K.A., Patrinos,A., Morgan,M.J., Szustakowki,J., de Jong,P., Catanese,J.J., Osoegawa,K., Shizuya,H., and Choi,S. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Larkin,R.M. and Guilfoyle,T.J. (1997). Reconstitution of yeast and Arabidopsis RNA polymerase alpha-like subunit heterodimers. J. Biol. Chem. *272*, 12824-12830.

Lee,T.I. and Young,R.A. (2000). Transcription of eukaryotic protein-coding genes. Annu. Rev. Genet. *34*, 77-137.

Lewis,J.D., Saperas,N., Song,Y., Zamora,M.J., Chiva,M., and Ausio,J. (2004). Histone H1 and the origin of protamines. Proc. Natl. Acad. Sci. U. S. A *101*, 4148-4152.

Liang,G., Lin,J.C., Wei,V., Yoo,C., Cheng,J.C., Nguyen,C.T., Weisenberger,D.J., Egger,G., Takai,D., Gonzales,F.A., and Jones,P.A. (2004). Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. Proc. Natl. Acad. Sci. U. S. A *101*, 7357-7362.

Lim,C.Y., Santoso,B., Boulay,T., Dong,E., Ohler,U., and Kadonaga,J.T. (2004). The MTE, a new core promoter element for transcription by RNA polymerase II. Genes Dev. *18*, 1606-1617.

Lin,P.S., Dubois,M.F., and Dahmus,M.E. (2002). TFIIF-associating carboxyl-terminal domain phosphatase dephosphorylates phosphoserines 2 and 5 of RNA polymerase II. J. Biol Chem. *277*, 45949-45956.

Lindstrom,D.L., Squazzo,S.L., Muster,N., Burckin,T.A., Wachter,K.C., Emigh,C.A., McCleery,J.A., Yates,J.R., III, and Hartzog,G.A. (2003). Dual roles for Spt5 in pre-mRNA processing and transcription elongation revealed by identification of Spt5-associated proteins. Mol Cell Biol *23*, 1368-1378.

Lipman,D.J. (1997). Making (anti)sense of non-coding sequence conservation. Nucleic Acids Res. *25*, 3580-3583.

Lu,H., Flores,O., Weinmann,R., and Reinberg,D. (1991). The nonphosphorylated form of RNA polymerase II preferentially associates with the preinitiation complex. Proc. Natl. Acad. Sci. U. S. A *88*, 10004-10008.

Lu,H., Zawel,L., Fisher,L., Egly,J.M., and Reinberg,D. (1992). Human general transcription factor IIH phosphorylates the C-terminal domain of RNA polymerase II.

Nature *358*, 641-645.

Luger,K., Mader,A.W., Richmond,R.K., Sargent,D.F., and Richmond,T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251-260.

Lusser,A. and Kadonaga,J.T. (2003). Chromatin remodeling by ATP-dependent molecular machines. Bioessays *25*, 1192-1200.

Macleod,D., Charlton,J., Mullins,J., and Bird,A.P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev. *8*, 2282-2292.

Malik,S. and Roeder,R.G. (2000). Transcriptional regulation through Mediator-like coactivators in yeast and metazoan cells. Trends Biochem. Sci. *25*, 277-283.

Maniatis,T. and Reed,R. (2002). An extensive network of coupling among gene expression machines. Nature *416*, 499-506.

Martens,J.A., Laprade,L., and Winston,F. (2004). Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. Nature *429*, 571-574.

Maruyama,K. and Sugano,S. (1994). Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene *138*, 171-174.

Matangkasombut,O., Auty,R., and Buratowski,S. (2004). Structure and function of the TFIID complex. Adv. Protein Chem. *67*, 67-92.

Merika,M. and Thanos,D. (2001). Enhanceosomes. Curr. Opin. Genet. Dev. *11*, 205-208.

Metz,B., Kersten,G.F., Hoogerhout,P., Brugghe,H.F., Timmermans,H.A., de Jong,A., Meiring,H., ten Hove,J., Hennink,W.E., Crommelin,D.J., and Jiskoot,W. (2004). Identification of formaldehyde-induced modifications in proteins: reactions with model peptides. J. Biol. Chem. *279*, 6235-6243.

Miklos,G.L. and John,B. (1979). Heterochromatin and satellite DNA in man: properties and prospects. Am. J. Hum. Genet. *31*, 264-280.

Misteli,T. (2004). Spatial positioning; a new dimension in genome function. Cell *119*, 153-156.

Morillon,A., O'Sullivan,J., Azad,A., Proudfoot,N., and Mellor,J. (2003). Regulation of elongating RNA polymerase II by forkhead transcription factors in yeast. Science *300*, 492-495.

Muller,F. and Tora,L. (2004). The multicoloured world of promoter recognition complexes. EMBO J. *23*, 2-8.

Nagy,P.L., Cleary,M.L., Brown,P.O., and Lieb,J.D. (2003). Genomewide demarcation of

RNA polymerase II transcription units revealed by physical fractionation of chromatin. Proc. Natl. Acad. Sci. U. S. A *100*, 6364-6369.

Ng,H.H., Robert,F., Young,R.A., and Struhl,K. (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Mol Cell *11*, 709-719.

No,D., Yao,T.P., and Evans,R.M. (1996). Ecdysone-inducible gene expression in mammalian cells and transgenic mice. Proc. Natl. Acad. Sci. U. S. A *93*, 3346-3351.

Nobrega,M.A., Ovcharenko,I., Afzal,V., and Rubin,E.M. (2003). Scanning human gene deserts for long-range enhancers. Science *302*, 413.

O'Sullivan,J.M., Tan-Wong,S.M., Morillon,A., Lee,B., Coles,J., Mellor,J., and Proudfoot,N.J. (2004). Gene loops juxtapose promoters and terminators in yeast. Nat. Genet. *36*, 1014-1018.

Odom,D.T., Zizlsperger,N., Gordon,D.B., Bell,G.W., Rinaldi,N.J., Murray,H.L., Volkert,T.L., Schreiber,J., Rolfe,P.A., Gifford,D.K., Fraenkel,E., Bell,G.I., and Young,R.A. (2004). Control of pancreas and liver gene expression by HNF transcription factors. Science *303*, 1378-1381.

Olins,D.E. and Olins,A.L. (2003). Chromatin history: our view from the bridge. Nat. Rev. Mol Cell Biol *4*, 809-814.

Orlando,V. (2000). Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. Trends Biochem. Sci. *25*, 99-104.

Orlando,V., Strutt,H., and Paro,R. (1997). Analysis of chromatin structure by in vivo formaldehyde cross-linking. Methods *11*, 205-214.

Otero,G., Fellows,J., Li,Y., de Bizemont,T., Dirac,A.M., Gustafsson,C.M., Erdjument-Bromage,H., Tempst,P., and Svejstrup,J.Q. (1999). Elongator, a multisubunit component of a novel RNA polymerase II holoenzyme for transcriptional elongation. Mol Cell *3*, 109-118.

Oudet,P., Gross-Bellard,M., and Chambon,P. (1975). Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. Cell *4*, 281-300.

Ozer,J., Mitsouras,K., Zerby,D., Carey,M., and Lieberman,P.M. (1998). Transcription factor IIA derepresses TATA-binding protein (TBP)-associated factor inhibition of TBP-DNA binding. J. Biol Chem. *273*, 14293-14300.

Parekh,B.S. and Maniatis,T. (1999). Virus infection leads to localized hyperacetylation of histones H3 and H4 at the IFN-beta promoter. Mol. Cell *3*, 125-129.

Pennisi,E. (2003). Bioinformatics. Gene counters struggle to get the right answer. Science *301*, 1040-1041.

Pokholok,D.K., Hannett,N.M., and Young,R.A. (2002). Exchange of RNA polymerase II initiation and elongation factors during gene expression in vivo. Mol Cell *9*, 799-809.

Prelich,G. (2002). RNA polymerase II carboxy-terminal domain kinases: emerging clues to their function. Eukaryot. Cell *1*, 153-162.

Proudfoot,N.J., Furger,A., and Dye,M.J. (2002). Integrating mRNA processing with transcription. Cell *108*, 501-512.

Ptashne,M. (1988). How eukaryotic transcriptional activators work. Nature *335*, 683-689.

Ranish,J.A., Hahn,S., Lu,Y., Yi,E.C., Li,X.J., Eng,J., and Aebersold,R. (2004). Identification of TFB5, a new component of general transcription and DNA repair factor IIH. Nat. Genet. *36*, 707-713.

Redon,C., Pilch,D., Rogakou,E., Sedelnikova,O., Newrock,K., and Bonner,W. (2002). Histone H2A variants H2AX and H2AZ. Curr. Opin. Genet. Dev. *12*, 162-169.

Ren,B., Cam,H., Takahashi,Y., Volkert,T., Terragni,J., Young,R.A., and Dynlacht,B.D. (2002). E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. Genes Dev. *16*, 245-256.

Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E., Volkert,T.L., Wilson,C.J., Bell,S.P., and Young,R.A. (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306-2309.

Richmond,T.J. and Davey,C.A. (2003). The structure of DNA in the nucleosome core. Nature *423*, 145-150.

Rigaut,G., Shevchenko,A., Rutz,B., Wilm,M., Mann,M., and Seraphin,B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. Nat. Biotechnol. *17*, 1030-1032.

Robert,F., Douziech,M., Forget,D., Egly,J.M., Greenblatt,J., Burton,Z.F., and Coulombe,B. (1998). Wrapping of promoter DNA around the RNA polymerase II initiation complex induced by TFIIF. Mol Cell *2*, 341-351.

Robert,F., Pokholok,D.K., Hannett,N.M., Rinaldi,N.J., Chandy,M., Rolfe,A., Workman,J.L., Gifford,D.K., and Young,R.A. (2004). Global Position and Recruitment of HATs and HDACs in the Yeast Genome. Mol Cell *16*, 199-209.

Roy,A.L., Malik,S., Meisterernst,M., and Roeder,R.G. (1993). An alternative pathway for transcription initiation involving TFII-I. Nature *365*, 355-359.

Rozen,S. and Skaletsky,H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. *132*, 365-386.

Sadowski,I., Ma,J., Triezenberg,S., and Ptashne,M. (1988). GAL4-VP16 is an unusually potent transcriptional activator. Nature *335*, 563-564.

Saha,S., Sparks,A.B., Rago,C., Akmaev,V., Wang,C.J., Vogelstein,B., Kinzler,K.W., and Velculescu,V.E. (2002). Using the transcriptome to annotate the genome. Nat. Biotechnol. *20*, 508-512.

Sanchez,C., Lachaize,C., Janody,F., Bellon,B., Roder,L., Euzenat,J., Rechenmann,F., and Jacq,B. (1999). Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database. Nucleic Acids Res. *27*, 89-94.

Saunders,A., Werner,J., Andrulis,E.D., Nakayama,T., Hirose,S., Reinberg,D., and Lis,J.T. (2003). Tracking FACT and the RNA polymerase II elongation complex through chromatin in vivo. Science *301*, 1094-1096.

Schena,M., Shalon,D., Davis,R.W., and Brown,P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467-470.

Schumacher,M.A., Lau,A.O., and Johnson,P.J. (2003). Structural basis of core promoter recognition in a primitive eukaryote. Cell *115*, 413-424.

Selinger,D.W., Cheung,K.J., Mei,R., Johansson,E.M., Richmond,C.S., Blattner,F.R., Lockhart,D.J., and Church,G.M. (2000). RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. Nat. Biotechnol. *18*, 1262-1268.

Shi,Y., Lan,F., Matson,C., Mulligan,P., Whetstine,J.R., Cole,P.A., Casero,R.A., and Shi,Y. (2004). Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell *119*, 941-953.

Shim,E.Y., Walker,A.K., Shi,Y., and Blackwell,T.K. (2002). CDK-9/cyclin T (P-TEFb) is required in two postinitiation pathways for transcription in the C. elegans embryo. Genes Dev. *16*, 2135-2146.

Smale,S.T. and Baltimore,D. (1989). The "initiator" as a transcription control element. Cell *57*, 103-113.

Smale,S.T., Schmidt,M.C., Berk,A.J., and Baltimore,D. (1990). Transcriptional activation by Sp1 as directed through TATA or initiator: specific requirement for mammalian transcription factor IID. Proc. Natl. Acad. Sci. U. S. A *87*, 4509-4513.

Soutoglou,E. and Talianidis,I. (2002). Coordination of PIC assembly and chromatin remodeling during differentiation-induced gene activation. Science *295*, 1901-1904.

Spitalny,P. and Thomm,M. (2003). Analysis of the open region and of DNA-protein contacts of archaeal RNA polymerase transcription complexes during transition from initiation to elongation. J. Biol Chem. *278*, 30497-30505.

Strahl,B.D. and Allis,C.D. (2000). The language of covalent histone modifications.

Nature *403*, 41-45.

Strausberg,R.L., Feingold,E.A., Grouse,L.H., Derge,J.G., Klausner,R.D., Collins,F.S., Wagner,L., Shenmen,C.M., Schuler,G.D., Altschul,S.F., Zeeberg,B., Buetow,K.H., Schaefer,C.F., Bhat,N.K., Hopkins,R.F., Jordan,H., Moore,T., Max,S.I., Wang,J., Hsieh,F., Diatchenko,L., Marusina,K., Farmer,A.A., Rubin,G.M., Hong,L., Stapleton,M., Soares,M.B., Bonaldo,M.F., Casavant,T.L., Scheetz,T.E., Brownstein,M.J., Usdin,T.B., Toshiyuki,S., Carninci,P., Prange,C., Raha,S.S., Loquellano,N.A., Peters,G.J., Abramson,R.D., Mullahy,S.J., Bosak,S.A., McEwan,P.J., McKernan,K.J., Malek,J.A., Gunaratne,P.H., Richards,S., Worley,K.C., Hale,S., Garcia,A.M., Gay,L.J., Hulyk,S.W., Villalon,D.K., Muzny,D.M., Sodergren,E.J., Lu,X., Gibbs,R.A., Fahey,J., Helton,E., Ketteman,M., Madan,A., Rodrigues,S., Sanchez,A., Whiting,M., Madan,A., Young,A.C., Shevchenko,Y., Bouffard,G.G., Blakesley,R.W., Touchman,J.W., Green,E.D., Dickson,M.C., Rodriguez,A.C., Grimwood,J., Schmutz,J., Myers,R.M., Butterfield,Y.S., Krzywinski,M.I., Skalska,U., Smailus,D.E., Schnerch,A., Schein,J.E., Jones,S.J., and Marra,M.A. (2002). Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc. Natl. Acad. Sci. U. S. A *99*, 16899-16903.

Strausberg,R.L., Feingold,E.A., Klausner,R.D., and Collins,F.S. (1999). The mammalian gene collection. Science *286*, 455-457.

Suzuki,Y., Tsunoda,T., Sese,J., Taira,H., Mizushima-Sugano,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Nakamura,Y., Suyama,A., Sakaki,Y., Morishita,S., Okubo,K., and Sugano,S. (2001). Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res. *11*, 677-684.

Suzuki,Y., Yamashita,R., Nakai,K., and Sugano,S. (2002). DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. Nucleic Acids Res. *30*, 328-331.

Svejstrup,J.Q. (2004). The RNA polymerase II transcription cycle: cycling through chromatin. Biochim. Biophys. Acta *1677*, 64-73.

Tanese,N., Pugh,B.F., and Tjian,R. (1991). Coactivators for a proline-rich activator purified from the multisubunit human TFIID complex. Genes Dev. *5*, 2212-2224.

Thanos,D. and Maniatis,T. (1995). Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. Cell *83*, 1091-1100.

Ujvari,A., Pal,M., and Luse,D.S. (2002). RNA polymerase II transcription complexes may become arrested if the nascent RNA is shortened to less than 50 nucleotides. J. Biol Chem. *277*, 32527-32537.

Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A., Gocayne,J.D., Amanatides,P., Ballew,R.M., Huson,D.H., Wortman,J.R., Zhang,Q., Kodira,C.D., Zheng,X.H., Chen,L., Skupski,M., Subramanian,G., Thomas,P.D., Zhang,J., Gabor Miklos,G.L., Nelson,C., Broder,S., Clark,A.G., Nadeau,J., McKusick,V.A., Zinder,N., Levine,A.J., Roberts,R.J., Simon,M., Slayman,C., Hunkapiller,M., Bolanos,R., Delcher,A., Dew,I., Fasulo,D., Flanigan,M.,

Florea,L., Halpern,A., Hannenhalli,S., Kravitz,S., Levy,S., Mobarry,C., Reinert,K., Remington,K., Abu-Threideh,J., Beasley,E., Biddick,K., Bonazzi,V., Brandon,R., Cargill,M., Chandramouliswaran,I., Charlab,R., Chaturvedi,K., Deng,Z., Di,F., V, Dunn,P., Eilbeck,K., Evangelista,C., Gabrielian,A.E., Gan,W., Ge,W., Gong,F., Gu,Z., Guan,P., Heiman,T.J., Higgins,M.E., Ji,R.R., Ke,Z., Ketchum,K.A., Lai,Z., Lei,Y., Li,Z., Li,J., Liang,Y., Lin,X., Lu,F., Merkulov,G.V., Milshina,N., Moore,H.M., Naik,A.K., Narayan,V.A., Neelam,B., Nusskern,D., Rusch,D.B., Salzberg,S., Shao,W., Shue,B., Sun,J., Wang,Z., Wang,A., Wang,X., Wang,J., Wei,M., Wides,R., Xiao,C., Yan,C., Yao,A., Ye,J., Zhan,M., Zhang,W., Zhang,H., Zhao,Q., Zheng,L., Zhong,F., Zhong,W., Zhu,S., Zhao,S., Gilbert,D., Baumhueter,S., Spier,G., Carter,C., Cravchik,A., Woodage,T., Ali,F., An,H., Awe,A., Baldwin,D., Baden,H., Barnstead,M., Barrow,I., Beeson,K., Busam,D., Carver,A., Center,A., Cheng,M.L., Curry,L., Danaher,S., Davenport,L., Desilets,R., Dietz,S., Dodson,K., Doup,L., Ferriera,S., Garg,N., Gluecksmann,A., Hart,B., Haynes,J., Haynes,C., Heiner,C., Hladun,S., Hostin,D., Houck,J., Howland,T., Ibegwam,C., Johnson,J., Kalush,F., Kline,L., Koduru,S., Love,A., Mann,F., May,D., McCawley,S., McIntosh,T., McMullen,I., Moy,M., Moy,L., Murphy,B., Nelson,K., Pfannkoch,C., Pratts,E., Puri,V., Qureshi,H., Reardon,M., Rodriguez,R., Rogers,Y.H., Romblad,D., Ruhfel,B., Scott,R., Sitter,C., Smallwood,M., Stewart,E., Strong,R., Suh,E., Thomas,R., Tint,N.N., Tse,S., Vech,C., Wang,G., Wetter,J., Williams,S., Williams,M., Windsor,S., Winn-Deen,E., Wolfe,K., Zaveri,J., Zaveri,K., Abril,J.F., Guigo,R., Campbell,M.J., Sjolander,K.V., Karlak,B., Kejariwal,A., Mi,H., Lazareva,B., Hatton,T., Narechania,A., Diemer,K., Muruganujan,A., Guo,N., Sato,S., Bafna,V., Istrail,S., Lippert,R., Schwartz,R., Walenz,B., Yooseph,S., Allen,D., Basu,A., Baxendale,J., Blick,L., Caminha,M., Carnes-Stine,J., Caulk,P., Chiang,Y.H., Coyne,M., Dahlke,C., Mays,A., Dombroski,M., Donnelly,M., Ely,D., Esparham,S., Fosler,C., Gire,H., Glanowski,S., Glasser,K., Glodek,A., Gorokhov,M., Graham,K., Gropman,B., Harris,M., Heil,J., Henderson,S., Hoover,J., Jennings,D., Jordan,C., Jordan,J., Kasha,J., Kagan,L., Kraft,C., Levitsky,A., Lewis,M., Liu,X., Lopez,J., Ma,D., Majoros,W., McDaniel,J., Murphy,S., Newman,M., Nguyen,T., Nguyen,N., and Nodell,M. (2001). The sequence of the human genome. Science *291*, 1304-1351.

Verdel,A., Jia,S., Gerber,S., Sugiyama,T., Gygi,S., Grewal,S.I., and Moazed,D. (2004). RNAi-mediated targeting of heterochromatin by the RITS complex. Science *303*, 672-676.

Watanabe,T., Hayashi,K., Tanaka,A., Furumoto,T., Hanaoka,F., and Ohkuma,Y. (2003). The carboxy terminus of the small subunit of TFIIE regulates the transition from transcription initiation to elongation by RNA polymerase II. Mol Cell Biol *23*, 2914-2926.

Wei,C.L., Ng,P., Chiu,K.P., Wong,C.H., Ang,C.C., Lipovich,L., Liu,E.T., and Ruan,Y. (2004). 5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation. Proc. Natl. Acad. Sci. U. S. A *101*, 11701-11706.

Weinmann,A.S., Bartley,S.M., Zhang,T., Zhang,M.Q., and Farnham,P.J. (2001). Use of chromatin immunoprecipitation to clone novel E2F target promoters. Mol. Cell Biol. *21*,

6820-6832.

Weinmann,A.S. and Farnham,P.J. (2002). Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. Methods *26*, 37-47.

Weinmann,A.S., Yan,P.S., Oberley,M.J., Huang,T.H., and Farnham,P.J. (2002). Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev. *16*, 235-244.

Weis,L. and Reinberg,D. (1997). Accurate positioning of RNA polymerase II on a natural TATA-less promoter is independent of TATA-binding-protein-associated factors and initiator-binding proteins. Mol. Cell Biol. *17*, 2973-2984.

West,A.G., Huang,S., Gaszner,M., Litt,M.D., and Felsenfeld,G. (2004). Recruitment of Histone Modifications by USF Proteins at a Vertebrate Barrier Element. Mol Cell *16*, 453-463.

Westermann,S., Cheeseman,I.M., Anderson,S., Yates,J.R., III, Drubin,D.G., and Barnes,G. (2003). Architecture of the budding yeast kinetochore reveals a conserved molecular core. J. Cell Biol. *163*, 215-222.

Wickens,M. and Gonzalez,T.N. (2004). Molecular biology. Knives, accomplices, and RNA. Science *306*, 1299-1300.

Winkler,G.S., Araujo,S.J., Fiedler,U., Vermeulen,W., Coin,F., Egly,J.M., Hoeijmakers,J.H., Wood,R.D., Timmers,H.T., and Weeda,G. (2000). TFIIH with inactive XPD helicase functions in transcription initiation but is defective in DNA repair. J. Biol Chem. *275*, 4258-4266.

Wood,A., Krogan,N.J., Dover,J., Schneider,J., Heidt,J., Boateng,M.A., Dean,K., Golshani,A., Zhang,Y., Greenblatt,J.F., Johnston,M., and Shilatifard,A. (2003). Bre1, an E3 ubiquitin ligase required for recruitment and substrate selection of Rad6 at a promoter. Mol. Cell *11*, 267-274.

Xie,Z. and Price,D.H. (1996). Purification of an RNA polymerase II transcript release factor from Drosophila. J. Biol Chem. *271*, 11043-11046.

Yamaguchi,Y., Wada,T., and Handa,H. (1998). Interplay between positive and negative elongation factors: drawing a new view of DRB. Genes Cells *3*, 9-15.

Yelin,R., Dahary,D., Sorek,R., Levanon,E.Y., Goldstein,O., Shoshan,A., Diber,A., Biton,S., Tamir,Y., Khosravi,R., Nemzer,S., Pinner,E., Walach,S., Bernstein,J., Savitsky,K., and Rotman,G. (2003). Widespread occurrence of antisense transcription in the human genome. Nat. Biotechnol. *21*, 379-386.

Yeo,M., Lin,P.S., Dahmus,M.E., and Gill,G.N. (2003). A novel RNA polymerase II C-terminal domain phosphatase that preferentially dephosphorylates serine 5. J. Biol Chem. *278*, 26078-26085.

Young,R.A. (2000). Biomedical discovery with DNA arrays. Cell *102*, 9-15.

Zhang,G., Campbell,E.A., Minakhin,L., Richter,C., Severinov,K., and Darst,S.A. (1999). Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 A resolution. Cell *98*, 811-824.

Zhou,J. and Berger,S.L. (2004). Good Fences Make Good Neighbors; Barrier Elements and Genomic Regulation. Mol Cell *16*, 500-502.

Zhou,Q. and Sharp,P.A. (1996). Tat-SF1: cofactor for stimulation of transcriptional elongation by HIV-1 Tat. Science *274*, 605-610.

Zurita,M. and Merino,C. (2003). The transcriptional complexity of the TFIIH complex. Trends Genet. *19*, 578-584.

VII) Annexes

>IP_14-07-04_07
tttcattataaagagtaggaaatggacatgatggagaaga
ggggcttgccctgaaatcacacggccagtgatgaaagag
ctgcttcccaactccaaaagtaacat

>IP_14-07-04_09
tttcattataaagagtaggaaatggacatgatggagaaga
ggggcttgccctgaaatcacacggccagtgatgaaagag
ctgcttcccaactccaaaagtaacat

>IP_14-07-04_10
atactaactttacttgcaacaacttaatggatatcatcatccc
ccttttttatatatggacactaaggcacaaagaggttaaatac
attgctcaaggaacttacctgagttagaaaatggctgagcc
agggtttgcagtcagattctagaacctgcagttgtaacgact
gtgctatgcatgaaacttctcttgcagaatgtacagtctagtg
actccagcataccagcaatagtaatagtc

>IP_14-07-04_11
taactttctgtcttgttgatctgtctaatgttgacagtggggtgtt
aaagtctcccattattattgtgtgggagtctaagtctctttgtag
gtcactcaggacttgctttatgaatctgggtgctcctgtattgg
gtgcatatatatttaatatagttagttcttcttgttgaattgatccc
tttaccattatgtaatggccttctttgtctcttttgatgtttgttggttt
aaagtctgttttatcagagactaggattgcaacccctgccttttt
tttgttttccatttgcttggtagatcttcctccatccttttattttgag
cctatgtgtgtctctgcacgtgagatgggtttcctgaatacag
cacactgaagggtcttgag

>IP_14-07-04_12
ctccctgtcttggtgagggccaggggacaggtcggggcag
tatttgggggaagcgagaggcaagcagggcgggctggat
acggagcgccctatgactgaggttagggaggacgcccga
gacccccggcatcctagtcctgggccgtgggctccggacg
ccgggcca

>IP_14-07-04_13
ctgtcctttaaatattgcctacagaggatatagagtcaaaat
actgtactatactgtgtttttaagtaattttaattcagatttttgatt
atgtcatcaactcgatacacggatacattcacttcaatttgttt
acaattcctagtgattgccgtttaaatttaatttcattttggagta
tgtaaacacatcaatgtaaaatagaagtgtatagtaatata

>IP_14-07-04_14
ccaagaaaaattcaaatgcttaagaatggaatattactaaa
ctaatatggttctagttactcttttgtgttcttcttgtatacactagc
accattgtttgcttacacacagcacaaagcataggggccat
agtgatgaatgcagtaatttctaacctattaacatttccagatt
atttgaaaacctacatattgtaatctgtaatggga

>IP_14-07-04_15
ctctcgttgtggcttttccgggaggtggctgtgttcttaccgga
gaaagtttcgatggccttcatcgcccagtgctcgtccgggc
agtccacgaaggcgtagccggatttgaccaagaactggc
cgctgtaggagatcttgtgctccgcaaacactttctccaagt
ccgcgggggtcacgctct

>IP_14-07-04_16
cagcactttgggaggccgaggcaggtggatcacctgaggt
caggagttcgagaccagcctggccaacatggtgaaagcc
cgtctctactaaaaatacaaaattagccgggtgtggtgg

>IP_14-07-04_17
gatttttttaattgtgttgttttcttattgttgaattgtaagagttcctt
gtatattgtggataccagtgctttatcagatacgttttgcaaag
attttcccccagtctgtagtttgtctttcagtgtcttaatggtgtatt
taaaagaatcaaagttttaaatttttaataaattccaacttaca
atttttttctcatagattgtgcttttgatgttgtgt
>IP_14-07-04_18
ctatctcaaaaaaaaaaaaaaaaaagtttgggttgtctttgaa
ggtactctacagccacatcttctttcaagaagcatttgatgttt
atcaaagtgcctggaacataacagatgctgagaacatact
tgctgaggtaatgaagatgctatgggagatggaaaaatga
gccattcacccaaccactcctctgaatgaccacttgacattg
tcttcttcctcaaagctcttccatctcctcccccatattctcctca
gctggtggtcttg

>IP_14-07-04_19
catgcttcattctttttttattctttttttcttctgtctcctctgactgtattt
ttaagcaggttgtcttcacactcactaatactttctgtctgatca
attctgctactaaaagacttaaaatgcattcttgagtagttca
attgcattttcaactccagagtttctgcttgattctttttttaattaat
tcagtctctctgatagacttctgaattccttttccatattatcttga
atttctttgagtttccctcaaagcagctctttggaattctctgtct
gaaaaaggtcacatatctgtttctccaggattggtccctggtg
cctcatttaactcatttggtgaagtcacgttttcctggataatctt
gatgcttatggatgtttgctggtgtctgggcatat

>IP_14-07-04_20
ggtcagcaggttaaaagaggccatcagggctctgggtgg
gccctgatccaagctgcttggtgtccttgtaagaagaagag
atgaggacacagacacacacag

>IP_14-07-04_21
ctatctcaaaaaaaaaaaaaaaaagtttgggttgtctttgaa
ggtactctacagccacatcttctttcaagaagcatttgatgttt
atcaaagtgcctggaacataacagatgctgagaacatact
tgctgaggtaatgaagatgctatgggagatggaaaaatga
gccattcacccaaccactcctctgaatgaccacttgacattg

tcttcttcctcaaagctcttccatctcctcccccatattctcctca
gctggtggtcttg

>IP_14-07-04_22
ttttcatttttgtatgtcctcttcaatttctttcatcattgttttctagtttt
cattgtagagatcttgcacatttttggttaaattcctaggtgttct
attatttttttgttacaattacaaatgcgattgatttctttttcagcta
gttcgttttggtgtatagaaacactactgatttttatatgttgatttt
gtatcctacaatgttaccaaattggtttatcaattttaagagtta
tttaatagagtttttttgtctttttagtttttctatatataagactatgttt
ctgcaaaca

>WCE_14-07-04_24
cgatcctcccacctgagccccacgaatagctgggaccac
aggtgcgtgccatcacgcctggctcatttttttacttttttgaaaa
gtccaggtctccctatgttgcccaggctggtcttaaactccta
ggcttaagcaattcttccacctcagcctcccaaagtgctggg
aatataggcatgagacaccatgctcagcttgtgttatcagta
gttttttcctctgttg

>WCE_14-07-04_25
caggaacagtaccatattttccaaatatttccaggctacagtt
gcattcggtgcctttccagaacaactgaacaatgaaatgga
tagtaatctgccaaatgcatacagctgtatgcctgtttcctg

>WCE_14-07-04_26
aatggcctctcatttaaagtggctattacacagggggtctc
aaaagtgctaggcatacagaaccaacttcattgtacctgga
aaccacaatcctcaggaaaggcagagaaaacaaatgat
attatcaaaagacacctcagaaaatgttccaaagtaactca
cctgccttgggttacccttcttcctatagcttcactaaggataa
ggaacacctgttcaaagttagttttaagcccattagaaattat
gtatgggtggcctttcctcaccaataacctttt

>WCE_14-07-04_27
cctgcttttagagcaggaagtttttctccccacccaccacctc
ccccccccctttttttttttttttttttttgtcgagatggagcctggctc
tgttgtccaggctggagtgcagtggcgtgacctcggctcact
gcagcctccacctgccgagttca

>WCE_14-07-04_29
caatttgtggctgacagagaaggtattaaagagccggaaa
ggtaagtgattttcccagattgcagctagttagcagcaaag
ctacagctggcacaagaggcttttcttttctttgtaattgatga
agctggtttccttttaaacctggaccctggcttgagttagaag
ttcaactcgactgtttcctggatgtgtgaacacaggcaaact
atctaaactgcctgaagctcaattcaaagaggaagttcagc
atcgagatgaaaaagtctgggggcttccagcaccacatca
tgtctggattcttgatctgccactttcttgccattgagacctgca
aacattatttacccttttaaggttcagttttccactctccaaaat
gggtgtgagtactcagactcattgcatagaggtggcaaga
gaattaaatgcttaatacacataaagcacttaattcctgaca
tttaagagtcaaataataagtgttagctgtttttttatgatgatac
tgttttcttatcagtgaaaatgggataataatgtaaaactcact
gaatcaccaagaagaggaatccaggaactaagaatagt
accacgcacacagcagacatatcatgaacactagactgtt
agtagcattatagattttgaagtctgctaaacctaacatttcttt
tttcaataaacctctggaa

>WCE_14-07-04_31
catactcctcttcaccgttctcgactactcatacacgccctgct
cttgtttacactgnnccggtttacactgtttttccaagccatca
cagctgatatctcctggtgctatccccaaactgccactcttaa
ctcttgaag

>WCE_14-07-04_33
aatggcctctcatttaaagtggctattacacaggggtctca
aaagtgctaggcatacagaaccaacttcattgtacctggaa
accacaatcctcaggaaaggcagagaaaacaaatgata
ttatcaaaagacacctcagaaaatgttccaaagtaactcac
ctgccttgggttacccttcttcctatagcttcactaaggataag
gaacacctgttcaaagttagttttaagcccattagaaattatg
tatgggtggcctttcctcaccaataacctttt

>WCE_14-07-04_34
tatttttatttttttttattatactttaagttttagggtacatgtgcaca
acgtgcaggtttgttacatatgtatacatgtgccatgttggtgt
gctgcacccattaactgatcatttacattaggtatatctcctaa
tgctat

>WCE_14-07-04_35
tgtgtcaaaagaaaggttcaactctgttagttgaggacaca
catcacaaataagtttctgagaatgcttctgtctagttcttattt
gaagacatttcctttctcaccttaggcctgaaaacgctcgaa
atatccacttccagatacgacagaaacagtgattcaaacct
gctctatgaaagggaatgttcaactaggtgacttgaatgca
aacatcacaaagcagtttctgagaatgctgctgtctactttct
atttgtaatcccgtttccaacgaaatcctcagaactatcgaa
atttccaattgcagattccacaaaaagcgtgtttca

>WCE_14-07-04_37
acactataatttttttggttgtttactttgtttttttcttgtaattaagt
aataataaaaagttcaataacaacagctcactttttttttttttttt
ttgagacagagtctcgctctgtcccccaggctggagtgcac
tggcgagatctcagctcactgcaagctccaccttccaggtt
cacaccattctcctgcctcagcctcctgagtagctgggacta
caggcatccgccaccatgcccggctaatttttttgtatttttttta
gtcg

>WCE_14-07-04_38
caccaaaatctcacacatcatcactaaagaacttactcatct
gttccccaaaaaactgatgggaatttttttaaaaagaaaca
tatagagtggcagatgttcccaacctctgtatgtgagagaat
gtctttgttaaattcacggcagaaagatattcttataaatacat
cataatgcagtcacaaggtggcatcttcctcctatctccactg
cctgccgctgtgattcactgacactgtgttatagttttgttttaa
aacaaaacatg

>WCE_14-07-04_39
aacaatttatttaaaactatacaaaaccagattagtcatcatt
ttttaatatacacaattctaattagcaatgagttataagaacat
agctttgaaaatggagagtatttgaaaggtgaagactgaa
ctatctagtatttaccaaaatttgtgagtttataatgcattttcag
acatattaaagaactaatcttaaataaagatatcatatatgct
tacacgatttcggtatctaaaagctagtaatattcaaaattac
aaagaaataagatttcataacattcatcctagttttttttgtaa
gataagtttcttcaaaataccttatcctgttgtgctttaagttcttc
atattgagtcttgtacctacgtccaattttcttaacttgagtaat
agttttgacttttttcttggat

>WCE_14-07-04_40
ctgccaatggaatctgctctggatataagataaggtacaaa
gatctgaatcccttgctaaagggaaccaattaatacaaaa
atcaggactaggtaggcaaaaggaagaactgaatggctt
agtggaatttttagcatttatgatcttccttacaaaaaaagaat
ataaccagtacttgctgagacatatggcctctgaagaactgt
agtctatgggctgagcctgtatactattaaatgaggcaaac
agtcatttaccaaagttccacttccctacttccgaaagtaag
gtcactttcagactagctttgagagtatatgtgaagaccagg
agaagttggagagggtagaggtagaagaagagaatcag
gatgggaaggttttggtgatcatgcgtggtctggtgttatcttg
agtcaccttggtcttgagactttgaggcagaccagcatagg
aaaaaagtgggaaacaaacaaacctgagagtgttaaaa
aggatcggtaagggga

>IP_15-07-04_14
gaacttatttcaagaaatggggagtcctcacctgcgtcaaa
cactgagaggttgagtgatgagagcttagcatattgaccac
tggatttggagacctggaggtttcctgtcatcttaagagcaa
attaggggcatgatgtgggtaaaagcctaatgccagagga
ataaggagagaatgggag

>IP_15-07-04_15
cactaaggttggttaagttaacaaatcataaccctcaaaaa
tgttaacccaacttgaaagaagcttttgaaactgtttacag

>IP_15-07-04_16
caaaataaataaataaataaataacaagaggaggttagt
gggctgggcgtggtaactcacacctgtaatcccagcacttt
gggaggctgaggtgggaggattgcttgaggtcagaagttc
aagaccagcctgggcaacaaagtgagctcttgtctctaca
aaaatttaaaaata

>IP_15-07-04_17
ctcattctcttctactttctatccagttgtgtcctagaggaaaac
ttaccccattagtttctcgtttgatcatacttatgcaaagtgttgtt
tgtcaaaagtccaaagtaatattttattcacatttgtctcaaga
gctactatact

>IP_15-07-04_19
cagccccagcggaccactgcactgatgaagagaaattca
ggctggtcataaacaagaaacaaacaaaaaaacaaaa
aacaaaccctatgaat

>IP_15-07-04_20
tttctctttccgtcctgctctttgccctgtgtctgcacagcaataa
cttagtgcactaatatctgtaaaaccagcacgtatccaatgt
acaaggacatgatacatactattatgtgcttctaagtaacac
aaatggtgatacatctttctgcaatttcattaacaatatgtggg
atccaaaaccattgaaaaatactgatctaattcttttatctgc

>IP_15-07-04_21
gcctttcttaattctcattaccaggggaaaaaacagtaattt
tgtgaatactgcggcatttggaattaaaaataattacacaaa
cagttttccaacagcacagagacagtatctgaactgagac
ctttttttctgcatgcattacttccaacattactttgactttctctacc
attcagcaatataacacatacacacttttttttatgaagagaat
caaaacgttttctttttaagaaaagttatttttctgtttttagaaaa
gtttaaaaagcccacttgaaggggggaaattactaaattaaa
tttgatatgaacaaacatgccaaataatagcataattataatt
attttc

>IP_15-07-04_22
ctaccaacaagcactagtgaaataaacatgtaatcatcatc
gtttctaaaactctgatagaaatcacaaaacattcaagcaa
aagggcaaggtcttccaggagcctaggagcagacttttgg
ctacaaagagcctagtccaggctgaatgccttatggagtat
gtcagt

>IP_15-07-04_23
tgctacgtaactgttctttttcacagtggccctgctcctgtgagt
cggagtgtgtcatttctccacttaaaacactccagtgtctcca
cctcggtcttgtgaagcttctggagtgtcaggcacttgagcat
atgaggggatacctcgttcattgtagggactaagttaatttttg
ttgacttaatgaatgaaatatgagtgtattaaattgcatcaca
gataattataaactgtaaaacactgaaaaagttcagaaag

>IP_15-07-04_24
cctacccaggcattgatattcacctggtgatactaataattgt
ctcaggtgataaattgtctcagctgctctcccacatggccca
caggtccctggaggttccccggatcctttcagggaagacc
at

>IP_15-07-04_26
catcttgtggccttcgttggaaacgggatttcttcctattctgct
agacagaagaattctcagaatcttccttgtgttgtgtgtattca
actcacagagttgaaggatcctttacacagagcagacttga
aacactctttttgtggaatttgcaagtggagatttcagccgctt
tgaggtcaatggtagaaaaggaaatatcttcgtataaaaac
tagacagaatgattctcagaaacttctttgtgatgtgtgcgttc
aactcacagagtttaacctttctttttcatagagcagttaggaa
acactctgtttgtaaagtctgcaagtggatattcagacctctt
gaggccttcgttggaaacgggttttttcatataaggctagac
agaagaattccca

>WCE_15-07-04_27

ctttctcttttgtatgctttccctgaatcaactgtgactcatggtttt
gactttcacttccatgttgatagccctcctctttacagaacttctt
ctatttatgctttggttaggttccaaaaggttggtttaagtctgttt
gcaaagtgacacttaaaatgagtaactgctaacaccctctg
atggtaccaggcaaagttgattg

>WCE_15-07-04_28
tgctggagaagcaagcagattgtaacgcacatctaaatag
cagagccgagcttcctaagtgggaaatggatgttgtaatac
aaaggcagttagcagggtccttggaggtttcagcaggatg
taattatggttctttaaaaagaggctcgcatatgtttttgagca
aacttaacagatgcccggat

>WCE_15-07-04_29
caccacaccctgctaattttgtatttttttagtagaaatggggttt
catcctgttggccaggctggtctcaaactcctgacctcaagt
gatccacccgccttggcctcccaaaatgctgggattacagg
caagaacgaccatgccctgcctgttttcttgtatcttgtccaa
atttacttttttttttttttctttctcaatccaggacaaagctgaaca
aaatccactcttcttacaaggacgccaa

>WCE_15-07-04_30
aataaattccctaaaggaagataagtatgtaaagttccattg
gcacatttaatgtttatggataagtatacagatgcctctccac
cttagagttatgtccagataaacaataatgatattttcaattta
caatagaggtttatcaaaaaataaccccataataagtcaa
agagtatactgaatgcctatcacttttttcaccatagtaaaa

>WCE_15-07-04_31
catttggcacatggtaagtcttcaacaagtatttgttgaataa
cagatttatcaatgaatcacttcacatgtgtaccctacgacct
cctct

>WCE_15-07-04_32
gtgtgtattcaactcacagagttgaacgatcctttacacaga
gcagacttgaaacactctttttgtggaatttgcaagtggagat
ttcagccgctttgaggtcaatggtagaataggaaatatcttc
ctatagatactagacagaatgattctcagaa

>WCE_15-07-04_33
ctcacaaaggggactcaggtgaggcactagttccttaga
catggcaccaaaagcatagccacaaaagaaaagatag
gtaaactg

>WCE_15-07-04_34
ctgttggttcagcagcttggaaacccccttttctgagaatctgg
aaagggaaaatctttagtccattgaggtctatggaaaaaaa
actgaatattctcgaaaaaaaaaaaaaactggaaagaagc
tatctgtgaaactgctttatgatgtgtgggttcatcttagagtta
aacatttcttgtgattcagcatgttggaaacattatttttggaga
atcttcgaagggacatttgggagaccattgaggcctatgag
gaaaatatgaatatctccagttaaaaacttgaaaaaagcta
tctgtgaaacagatttgagatgtgtggattcatctcacagagt
taaatctttctttcaatttatcatgttggaaacactctttttggaa

>WCE_15-07-04_35
tttgatatgcacacattctaccaacagtctccacattttgctta
ctgatgcataaggaaggacagacagaaaggaaaaggat
ttataggtcgcatgtctttccctttccttctgtgtcatcattttctgc
attagtggttgactaacagggaggtaatgtgaggaagaaa
aaatagtatgttctgtatctgtt

>WCE_15-07-04_36
ttttttttttttgagaaagagtcttgctctgtcacccaggctaga
gtgcaatggtgtgatctctgctcactgcaacctctgcctcctg
ggtgagccaccacgcccagccaacaaaaaatttttaatag

>WCE_15-07-04_37
actcaggaggctgaggcaggagaattgcgtgaacccgg
gaggcagaggttgcagtgagccgagatcgtcccactgca
cttcaatctgggcgacagagcaagactccgtctcaaaaaa
aaaacaagaaagagaacgaaataggggataggattatg
aaaaaataggtgagatagaggttaaagttttagacagtga
gaaaggtaattatgacagaatg

>WCE_15-07-04_38
aattgcttagtggtgaatatttttcagtgaacccatatcttaca
accacatctgctgaaccataaggtttgtaaagctgtggtcatt
ctacaaaactgtcccatggcctacttaatttatacccaaagg
aaagcatgaata

>WCE_15-07-04_39
ccatcaatcaagctgctgtttgttctacatgatagcattgtgta
gaacaaacagcatgtttgattgatttttttggagcaaggtgac
tgaaattccacttcatagcaaacacagcctcaatctctagtg
ccaatgcttaagtgccagatagttcatctcccaac

>WCE_15-07-04_40
ctccaaaggggactcaggtgaggcactagttccttagac
atggcaccaaaagcatagccacaaaagaaaagataggt
aaactg

>WCE_15-07-04_41
ttgaaagtgataattagataattcgggaaagtatttggaagc
ttcaatgcagacagatgctacagatgtacaaagcaattctc
cttccacctagctgacttgcaaaacctgtggattttagg

>WCE_15-07-04_42
ctcactatgccctcaagtgagtgaagactgttctgaatgcca
gtactaaccacggagcctgacacctagtaggttctcaaag
atctatcaag

>WCE_15-07-04_43
gacttttttccccttttttttcttcttagagaattttggtctcaggcaat
gttgcaagtcaatacttctttccccaaatgtgagtaggtggat
acgtatagaagctgtaaaagttaatgatcgttctctgactcat
agctgatggtttcaggttgaggaaaag

>WCE_15-07-04_44
atatcctgggggtaaagaggaaagcttgacttcctcctttcct
tcaaagagctcttaattcccttgtgtttttacttactgtccccagt
gtgctagctactgcagggacacgaagacaaataggacat
ctgggatgagccagatggagcttaatttaggctttgacatgg
gcagatgacctttcttaattttagattcttcaagtcacaagag
ataaaattatgacctttcagattgttgagaacatatttaaaagt
ctcctgcaccttgtattggcatatggttaaccagtcgcttccttc
cttcctgtttctttt

>WCE_15-07-04_45
ccatcaatccagctgctgtttgttctacatgatagcattgtgta
gaacaaacagcatgtttgattgatttttttggagcaaggtgac
tgaaattccacttcatagcaaacacagcctcaatctctagtg
ccaatgcttaagtgccagatagttcatctcccaac

>WCE_15-07-04_46
tgaaccattcccaagttttgtgctatagtacttctctttcaggtg
aaaaatatggcacagataccaagatgataagaaatatctg
aagtttccaagcctgtgttatggcttagcattttctaaaagagt
gctgtccagtagaaatatgatgtgagcagccatatatgtaat
ttacatttcttagtagccatatgaaaaagtataaagaaacaa

>WCE_15-07-04_47
tgaccaaggttcccaacctctctgattctcagtttccttctgtat
cccctcataggattaaattggatgaaaaggagtgttcttgct
ggccagaggtgagaatcacatacttttaggttgatgatgac
atcctgtctatgaa

>IP_21-07-04_01
agccgtgatccctggcccaggagaagctgtgatcctgggc
ctaacgcctcactcaatggctttacccttgttcaagggctgc
cccttatgttcaggcctggaagttgctctctacaggacagga
ggtgcttgctgtttggcgtgagctgaggttgccctgcagagc
ctggggtcagtgctataactgggccctgtggcacccgtaaa
ctcacttaatgaccacctggacattgggcctcatgctaggg
ccgctgtcagcattctttatctgtgaaggcagaggacattgg
aaaaaccagttagggtgtgtgaggcatgttttatgttgatatc
aagggagtgttggactcttggttcaggagcctggaaatgtat
gcaatctggctttttcagaaagtcagccatgttgaataggga
agcaagacctttttaggcgtcacatctcctgaacaggcaag
tgggcaggatgagagtttggggtttgctcctcttccctctgatg
ggaaaggcgtccagcgcgtcaccatgagcagcactggg
gtttggcatggtcaagaagtgttcactggcatgtgaccaga
gaccttcctaatggctgagacggagtctgacccagtgggc
atggctgtgctcggccagtgtcatggccggtg

>IP_21-07-04_02
gaacccgagtttccagactcttagtgatctgtgcctctgtcgc
tcatttctgaagcaggaactttagtgttgctggcaatgtcagt
gtcatttgtagactcactcagtaaagagatgtctctttggggc
tctgttatagcaccattttcttcttagaacaaatat

>IP_21-07-04_03
agccgtgatccctggcccaggagaagctgtgatcctgggc
ctaacgcctcactcaatggctttacccttgttcaagggctgc
cccttatgttcaggcctggaagttgctctctacaggacagga
ggtgcttgctgtttggcgtgagctgaggttgccctgcagagc
ctggggtcagtgctataactgggccctgtggcacccgtaaa
ctcacttaatgaccacctggacattgggcctcatgctaggg
ccgctgtcagcattctttatctgtgaaggcagaggacattgg
aaaaaccagttagggtgtgtgaggcatgttttatgttgatatc
aagggagtgttggactcttggttcaggagcctggaaatgtat
gcaatctggctttttcagaaagtcagccatgttgaataggga
agcaagacctttttaggcgtcacatctcctgaacaggcaag
tgggcaggatgagagtttggggtttgctcctcttccctctgatg
ggaaaggcgtccagcgcgtcaccatgagcagcactggg
gtttggcatggtcaagaagtgttcactggcatgtgaccaga
gaccttcctaatggctgagacggagtctgacccagtgggc
atggctgtgctcggccagtgtcatggccggtg

>IP_21-07-04_04
ctcacttcttaaatgtttaatggcttatattagagaaaataatgt
gttttaaatactatcttggaattgtttcgactgagtctgtgtgtttt
ggtgtttcaaataacagtttttacttcagtaaatagctagataa
gagttgcttctatattatgtaaattacatattttaaatcagataa
ggagttttatgtacatagcattattactgcatatcaccaagatt
tgattctataatcttttaagcataaaagaatcctggtatacaa
ataaatgtctccagggttgaatgaagaaagagtgagtata
agaaaattgaacaatcaatttccagttattctggacctagtg
atccagtttggggcctatttgcttacttttcttacattcctaaattt
cacactttattgtgttgttgatcatgaagaggaagagagaca
gaagggatgtttgttactcattcagaaagtacacactgagg
gagcgtgcagggcatagacccttctggagtttacagctgag
gggcaggcagggcagacacatcaaacagtcctgctgga
aggtgtcagcctgcatttgctgtgtggtaagtgccactagt

>IP_21-07-04_05
gaatggaggactccctctcacccccctctggctggcatttcac
caaaccgagggttccctctcacccccctctggctggcatttca
ctgagcaggggggccccctctcaccctcaggcagacctgtc
attgcacccagcctttctctctcctgatgtgcaacaccaggct
gggccggcgctgaccaccct

>IP_21-07-04_06
atgaaatatttcacgtttgaaattaaaagtacccataaaattg
cagcaaaggcagactatggcatctaaatagaggactgca
accttgtagtggagaaaataacattattcattggatcattga
ggatcagaaccagaagcgttcttatgactaccggtttccca
atgaaatttacaaagcctagctgtcagtgttcttaaccaacat
aatgaaggt

>IP_21-07-04_07
cttcgcggccggacgcgcggacagcttgggtggcggcag
gttgggcatggggacggcgggaggcggtggcgagctca
ccg

>IP_21-07-04_08

gcaaactaaacaagttcttctctcctgataatatttgcagcta
ccatttattgtgtgtttcctatgcactaggcaaggtacaaaa

>IP_21-07-04_09
gcaaactaaacaagttcttctctcctgataatatttgcagcta
ccatttattgtgtgtttcctatgcactaggcaaggtacaaaa

>IP_21-07-04_11
ctttatgccatggttgtaattgctgatataagctgatgtttcact
ctgttgccatttttcttgctttttgtttttttaatttttgctttgggtgtga
gtcccaccagttaagtcttaaatcatcatttaaaattcaaaac
agaaaaaatgcacgaagtcagcttccaggataaggaga
aggaacatctgacgcggttgcaggatgctgaagaaatga
aggtgtgtgtggctgccccttccatcaagcccagtgtgcattt
gatgcatttggattggaaatgcttgttctaggaaaatctgtca
ccttcagatccgtatatggcaagggacaggggaggggctt
tgttgggcattcagagtcgatctcagagtatagcctggc

>IP_21-07-04_12
gcctgggaggtcgaggctgcagtgagctgtgaatgcacta
ccatactccagcctgaatgacagagcaagatcctttctcta
aacaaaaaacaaacaaaaaaaagatgatcactcttggag
gtgtatagggacaggatcaga

>IP_21-07-04_13
cgcctgggaggtcgaggctgcagtgagctgtgaatgcact
accatactccagcctgaatgacagagcaagatcctttctct
aaacaaaaaacaaacaaaaaaagatgatcactcttgga
ggtgtatagggacaggatcaga

>IP_21-07-04_16
ttgttccttcgtaacccaagctcccagcggctgacacctgga
tcgtggcaacagcctcctaagtgagctcttggcatccattct
agccccaaggggatctattgccacctaaaagccaaattc
atttatttcatttcagttgcgataaatgctaccaaataaaagta
caggatgcacgggcacacataacagagggtccgcacca
agcctggaaggtcccaggaggcttccattgttctttggataa
agtctaaattccctagcctcgct

>WCE_21-07-04_17
cacgggccaggcccagagtccctctgctatgtgcagtctag
ggacttggtgccctgtgtcccagccaccacagccatgacta
aaaggggcaaaggtacagctcaaggcatggcttcagag
ggtgcaagccctaagccttggcaccttccacatggtgttga
acctataggtgcacaaaagtcaagaattgagtcttgggaa
catctgcctaggtttcagaggatgt

>WCE_21-07-04_19
gtgttcatttcatgatatattatgggccttttctattcacagattta
gttctctggatatttttttctgttttatttattgtcttcctctgcatttatt
gtgagtctcttaagccttttttttacacaccattaattagatattt
aaaaactttttattataaaaaatttcaaatatatacagtagag
agactagtgttatggattctcatg

>WCE_21-07-04_20
ccagttcgaacttcctggtggctttgtttacactgtgagggga
aaaccgcctactcaagcctcagcaatggcagatgcccctc
cccccaccaaacccaagcatcccaggtcgagctcagact
gctgtgctggcagcaagaatttcaagccagtggatcttagct
tgctgggctccatgggggtgggacctgcca

>WCE_21-07-04_21
tctctccttagcttctggtaaacatgaatctgttctccatcctttg
ttttttttttccagactataatacaaatagaatcatactgggtgg
agctttttgggtctggattctttggcttacaaaataatgcattta
agattcaagtatgttaatgcatgaatcaatacacatccctttttt
gttgctg

>WCE_21-07-04_22
tctttcactcagcattacatctataagattgatcgatactgtttta
catagcaaaataagtttttaaaaccactgtgtagtatttcata
aattattcatctgttctatagttgatggatatttgtttgttgtagcat
ttggttattatgaatacaagtgctatgaacattcttgtctata
tgtgctggtggatggatataaagcactcatttatgtttattgtct
gttcaggtgtggaattactgggtcatagagtggacatatctg

>WCE_21-07-04_23
tttttcccgagccgtgggcggggccagagcggtttccagga
tgtcggggacgtggccctggcattcccagccggattgagg
ctggagctttgcacagaaaatctggggcgggcctgggtag
ttttgtccttgttggggcgggtggagtattcagtggggcaaaa
cgccagtt

>WCE_21-07-04_24
tttcatatgagaatataaatatgatttgttgctccatcttggctg
aaaaaataaattttataatctaatgtaaagccttttgcctttca
ggcattacagaaagaaat

>WCE_21-07-04_25
caatgaatgtgagaggcaaggaaatgaaaacagtgaatt
tgataatgcatttgagaaattattcgaggatgtagaatagag
agatggctgggaagaatccacaaccctcatctcccctatct
acgtacaatatttttaattgagtctagctgtggttcaggaagct
gtgaaatttgtgaaaattgagttcatgagcacaggttactttgt
gtgaagactacaaagaaggaagtaatttgtcttacatttttca
tatttcaaaaa

>WCE_21-07-04_26
tgaaaaggttagaatgatttgagatgtcaccatggtgtaatg
tgatgtgaaactatactgacagaccgaatgctttatgttccat
ttgtatttagagagaa

>WCE_21-07-04_27
gctgataaagacatactagagactgggaagaaaaagag
gtttaattggacttat

>WCE_21-07-04_28
ctggaatcccagctacttgggaggctgaggcaggaaaatc
gcctgaacccaggatgtgaaggttgcagtgaacagagat
gttgccattgcactccagcctgggcaacagagtgagacttc
atctccaaaaaaag

>WCE_21-07-04_29
gctgataaagacatactagagactgggaagaaaaagag
gtttaattggacttat

>WCE_21-07-04_30
cacctgaaagaaaaacaaaaccctaatagcatgattaaa
aaaaatacaaccaaggcaaaaatacatcaaaggacca
ataaatgagtggatc

>WCE_21-07-04_31
atgttcatgaatattgccaagatgtgcttggatatctttttttctc
ctcaattctctttggaactgtgagccttttgaatcttaagattaa
ttcttcgcttagcttatggatattttctttagtatttgtttaggaata
tggcatatctaatatttacacattaagtcccctggatctatgcc
tgtctcttttccctc

>IP_22-07-04_08
ctagatgacacctcagaaccaggagggaacaaaaaag
acctctgagcagtcaaaataatgctataagcatgtattccag
cttactttcggcatttctgtggagtgcaggaggccacattttgtt
catagatccagtttgacaagcttgg

>IP_22-07-04_09
caccttcaccgtggcgcagaatgagcgcttcgatgcggac
tatgcggagaagatggcaggctgcagcctcagcttccgct
ctgagctgtgagaggggctcctggagtcactgcagaggg
agtgtgcgaatctaccctgaccaatgggctcaagaataaa
gtatgatttttgagtcaggcacagt

>IP_22-07-04_11
nggtctttgatcatatggaatttgttcagaacctaaataaaaa
gtcattgtattatttgtaaaatcataatcatttctatatgataatg
gatttccatcttttcctcttactaggccccacctcccaacactg
ttgcactggggattaagtccaacacatgaacttttggggcat
gcattcagcagatttgctttgtctggatatttcatataaatgca
actatacaatatgtggtct

>IP_22-07-04_12
taaaaaatagtttgcgtcataagaagtaacgatatgagtag
aaatattttggtacggacggggtacgtgggttggttggtgaa
tatccgatcaccccggaatttgtcatgaaactgggt

>WCE_22-07-04_13
c
acattgtcacttctgccatttgcatcagttcctgggttcagttaa
ccctgcttcattataatagcacagcttcgtggactgact

>WCE_22-07-04_14
cattgaccaccaatcggctgaaatctccacttgcaaattcca
caaaaagagtgtttcaagtctgctctgtgtaaaggatcattc
aactctgtgagttgaataaacacaacacaaggaagttact
gagaattcttctgtctagcagaatatgaagaaatcccgtttcc
aacgaaggtctcaacgaggtctgaatatccacttgcagact
ttacaaacagagcgtttcctaactgctctatgaaaagaaag

>WCE_22-07-04_14_bis
gcctgtgaagatatagctactagtttagcaaatgaaacctta
agcatcatctgaaaactctgaatgttaaggttgtattctagta
aatgttctaagccagggaatagaatgccttcctgcttctcctg
tgct

>WCE_22-07-04_15
caacattgtcacttctgccatttgcatcagttcctgggttcagtt
aaccctgcttcattataatagcacagcttcgtggactgact

>WCE_22-07-04_16
ggataccctgggcatcagtgcatgggacgatgaaatattg
gggcagggctggtgaaagccacttgtgttcatctccatcttct
cagtaaaagaggagtctgagtgtgaggatgaggaaggtg
gcattgagtgttgaaggcaagagaagaaaatgtgaagtg
atcttctaggaatgtgggag

>WCE_22-07-04_17
atcagttatattgaactttatttacaaagggatgtgtattactctt
tttctttctttttctttttctttttttttgagacagaacctcactttgtca
cccaggctggagtgcagtggtgtgatcccagctaactgca
acctctgcctccagggttcaagcaattctcctgcctcagcct
ccccagtagctggggctacaggcgtgcaccaccacacct
ggctaatttttgtatttttagtacaaacagggttttgctatgttgc
caaggctggtcttgaactcctggcctcaagtgatccacccg
ccttggcctcccaaagtgctgggattacaggtgtgagccac
cgtgcccgacctcgctctcttctttctccctctgccttcattcctc
ctctgcagattgtgctcttatttgcatgtaagctattatggcactt
aggtcaccttccatgtccagcaccgataattaagtaactagt
ct

>WCE_22-07-04_18
aaataaattccacatttaagatctctgtcaaagttatgaagtc
cacctgaactccccaactatttttttaaaaggcagaaaatga
acaacaacaacaaacaaactgtcaggataagactttaaa
atgtcaaaaactgacagcccagaattctggaaacattgaa
ggtactctaggctagaggtgagcaact

>WCE_22-07-04_19
caaaatattgctgtgtagctaggcaccatagcaggcaccta
cagtcccagctatgctggaggctgaggcaggaagattgttt
gagcccaggaaatcgacactgcagtgcaccatgattaca
cctgtgaatagccactatgctccagcctgggcaacatagc
aagaccctatgtcttaaaaaaa

>WCE_22-07-04_20
ctacttacatcatatagttgaataagcgtcaactaaataatta
aaaatccaagactgcattttctcatggatgggtaatttcaata
tactggaaaaaaacttgtggctctgttaacaatgatggaaat
actgcggtaaataccctcttggcaggatgctggcatttgaag
gattaatggaaaaaggaagcacaaaggttatgtaaagctt
aaaaagccatgaaagaagaaaaatctaattagaaaatta
tatataagggtttaattcaaattatatgcctagaaatgagtt

>WCE_22-07-04_21
c
attgaccacaaagcggctgaaatctccacttgcaaattcca
caaaaagagtgtttcaagtctgctctgtgtaaaggatcattc
aactctgtgagttgaataaacacaacacaaggaagttact
gagaattcttctgtctagcagaatatgaagaaatcccgtttcc
aacgaaggtctcaacgaggtctgaatatccacttgcagact
ttacaaacagagcgtttcctaactgctctatgaaaagaaag
gcctgtgaagatatagctactagtttagcaaatgaaacctta
agcatcatctgaaaactctgaatgttaaggttgtattctagta
aatgttctaagccagggaatagaatgccttcctgcttctcctg
tgct

>WCE_22-07-04_22
ccgttttccccttgttctgccctcatttaaattatgcatgcattca
attcaacccatttgaaatatttacctgtgtaccttgcactgaaa
gaaatacagacagccttgcctctgaaaga

>WCE_22-07-04_23
acctaaaatctaaatgtcttatttttcttctgtcataccacttaac
cccaatacaaactcaacaatggttccaattagccagaaaa
tggcactttcaatctttccat

>IP_25-07-04_01
attatcaaaaaggatcttcacctagatccttttaaattaaaaa
tgaagttttaaatcaatc

>IP_25-07-04_02
naagtgcgggccccctcgaggtcgacggtatcgataagct
tgatatcgaattcctgcagcccccaccgatgcgcagcagat
ggaaaaagcgtctccaccctgagtaaaagtaaaagtgaa
gtagccgtttctgacgagtaggagccgcgcctcgcggcga
atggttcaggcaacctcgatcatgagctgaacccaccattc
gccccggggcggggctcctacagttcgtggacaa

>IP_25-07-04_03
tgaagatgtcaaatcaaaatgaatgaatggaagaaaaag
tcccaatactggttgcccaagtttagacagctcaaagggca
attctttatcaagaatatcatgtttccagacctggcaattgatt
atgacggtacttctatgcatgaattcagccaggagtcttttgg
caaagatttgctctctattatctaaagttaaacataattaagct
aagtgttgat

>IP_25-07-04_04
agactctactaaaaatacaaaaaattagcatggtggctgg
cacctctagtctcagctactcgggaggctgaggcaggaga
atcgcttgaacctgggaggcggaggttgcagtgagctgag
a

>IP_25-07-04_08
cccacatacctctggactccatatttgggaagttaccatcag
ggtcagatctttaagtgcgtcattttg

>IP_25-07-04_09
atgtttaagaagttttcttatatagacttctatagacttatgcaa
attttcttctatacacaactagaacttgttg

>IP_25-07-04_10
ccccgtgcccgcgcaggccgcacccacttgggggcggc
gcgccggggcctggtgctcggtcggcgggtgctgccgcttt
aagcgggggcgggactgcgcgcggccgagcggttgcga
cgagggctcg

>IP_25-07-04_11
attgataaaatgttgatggcaatagttgtttttatgtatttttttgttt
gtttgttttgttttgaga

>WCE_25-07-04_13
t
ctagtctccttttgtaaagaatcgttcactatgctgattaaaaa
tgaccaagaaaatgataccagtaacttcttattataaaaaa
gactttatccctttggatac

>WCE_25-07-04_13_bis
ggtgaggtatgatctcactgctgtattccagcctgcatgataa
agtgagacactctcaaaaaaaaaaaaaaaaaaaaaga
aagacattgtaatgattaaatatttagaaaaaaatcacatac
actgtacattctgtatggattacaattatttctagtattgaaaag
ataataaccaactctttaacatgataatataactgaacctga
tttaatttgtttttttctttg

>WCE_25-07-04_15
ggtcgaaaccgtgtgtttccctcccccatcgtggagcagcg
actcgggcatcgcgctgatgtggtcccctcccctgggagga
gtggaatgcaatgatgtcacagtgcccctaggaact

>WCE_25-07-04_16
ggtttctgtttcaggtgatacaaacattcttgaaatggataatg
gtgatggttgcattacattgggaatatacttagtgctgtggact
tgcacactttaaagtagttaaaatcgcaatttgtatgttatgtgt
tttttaccatgaagaataaacaaaggattagtttaatatcagtt
gccccttcataaatttttctccgtgtttacagaattggctgattc
cagacaacaacactttaagagagacagataacaggacc
gacatattttatgaggaatggatgttggactctcctgagtagg
gcaagatcacttcaatacatcaaaattctgtgatgacaggtc
tgaacagcaaccctca

>WCE_25-07-04_17
caaggaaaaattaccaacaggaacgcatgagtctatctgt
ttatctacgcatatgcctgtctacagcctgactcgggacactt
ctcgaaaacatgaaagaaacaaaaggactcactatgaa
aaggaaaacagtcaataagggccattttatgtttaaaaaaa
aaaaaaacacaggccaggcgtgatagctcacgcctgtaa
tcccagcacttcgggaggccaaggcaggcagatcacaa
agtcaggagatcaagaccatcctggctaacacggtgaatc
cctgtctctactaaa

>WCE_25-07-04_19
tgtaatatgtaccgaaatgttagcctaaggactaaaacctc
aacctgggattgtaaaagcttttcttagtataaaagtttaaattt
ttttttttttttttttttttgagacagagtctcactctg

>WCE_25-07-04_20
gactgtacttaacaaaaatcactgctaaagtgacatctgact
ctgatggacactatgtgacttgcccaagatcacatggagag
tcagggttagagttctgctctcccagattagtggccttttggtgt
ggcagatcccaggggacactgcctctttttcttggggaagc
gtttaggggcctctgacatcagccaggccagagagtgatg
cgggagtcacccagattaggctgga

>WCE_25-07-04_21
cactgtgcctggtctccttcactgttgtaagatacttgaattgg
gtcaatatttgtggagaagtctcttaaaagttcacttgattgtc
agtactagaactctacatttaatattgacatattcctgggagc
atttcagagcattctattagcttagaaaggtccaggataattt
gactttagaagttactgttaccatgaatctcaatgacttttgaa
atccatgaagaatatctttttttttttttttgagacggagtctcact
ctgtcgcccaggctggagtgcagt

>WCE_25-07-04_22
cttccttgggctgggcacagtggctcatacctataatgccaa
ctactctgcaggctgaagaaggattgtttgaagctagtaga
actgtttggggcctgcacaacatagcaagttcctatctctaa
aac

>WCE_25-07-04_23
cattccactccacccgcatttattcgattccattccattccattc
cattccattccattccagttgattccattggattccattccgacc
gaatccatttcattccattccattccattccattccattcc
a

>IP_28-07-04_01
attaaacctagaagaaaatcagaaatatactgtatttagatg
aatgcaaattcaactaccttaaatttcaaagtcaatgcaaa
gaaatgagacaaatgtaagaaaacatgacagctaaagat
tatccttccctagcatctgtgttgttaatgcctttctctccctatgg
gttttcagggttggaggtgattgtgtagaatatacctggcacg
aaatagatgttcactcaatatttgttaatcaattttctcccttgtc
atgtcaaagctgtgaaataatatatagcaagtaaggttatg
atagtgtcactataaataacgctgaaatttcttacatgactag
actcatttaaatatcctctgttcttagtttttggcatgaaagc

>IP_28-07-04_03
gccaccgcggcgctccgcttacctttcaccaccctgtcggtc
gtcgacaacttgggatcaattgccttgggctcggacatctcc
agctgccggaggacagcgacgcgctgctcgtccgtccga
ctgcacaccccgaccggaccggcgggccagacactcaa
cgccgccgccgccgccgccgcc

>IP_28-07-04_04
gccaccgcggcgctccgcttacctttcaccaccctgtcggtc
gtcgacaacttgggatcaattgccttgggctcggacatctcc
agctgccggaggacagcgacgcgctgctcgtccgtccga
ctgcacaccccgaccggaccggcgggccagacactcaa
cgccgccgccgccgccgccgcc

>IP_28-07-04_05
gcctcagcctcccgagcaggtgggtctataggcacacagc
accatgccagactaattgtttttcttttttttttttttgagacagag
gctcactctgtcacccagactggagtgcagtggtgcaatct
cagctccctgcaacctctgc

>IP_28-07-04_06
gtgcatggtggcacgtgcccataatcccagcttctcaggag
actgaggcaggagaatcgcttgaacccggaaggtggag
gctgcagtgcggtgagccaagatcacaccactgcactcca
gcctgggcaacagagcgagactccacctcaaaaaaaca
caaaaggtgacaacaataacaaaataaagtgtattcctatt
atcattaca

>IP_28-07-04_07
ctatctaaaaccaggaggttatggagagaatcaaagctaa
tgatagcatctcctgtcactgcctgaacttcttggcatgtgatg
catggagtaggggcagccaggttatccctaaataggaga
acaagaggaatcaggaaagcactaaacttcgctttgcttgc
ctctaagctaaaacttccatgtttctgctgccagtagtgacaa
ggtacctcaaaagttcttggcttcagcagcccccagatccta
agcttaaaact

>IP_28-07-04_09
caatattaactccctccctatttataattattttgaatgtgaata
aattaaactcctcaatcaaaaaacatactattccagaaaac
tcaaaaggagggaacacttctacacatattcaataagacc
agcattaccctgatccaaagccaggcaa

>IP_29-07-04_07
ctatatgattgaatttcttaaaattatagaaaaggcaaagat
atactgatacaaagcagattagtgcttgccaaggtggatgg
gaggaagaacttttgaggaaataaaaatgttctatgttatg
attgcaatgatgattacacagcagttcgcgtttggcaaaatc
attgaataatgtccccaaaactggtaaattttagttttataaatt
atat

>IP_29-07-04_08
gaaatgtaaattcgtacagcctttatggaaaatggtatgaaa

ctttctgagaaaattaaaaatagaactaccataagatccag
caattctccttctgggtatatccacaaaggaattgaaatcaa
cacctcagcgagatatctacatatgttcccatgttcattggag
cattactcatgatagccaagatatggaaacaacctaagtgt
ctgtcagtggatg

>IP_29-07-04_11
gtcacactgtgtggccggcctgggtgggttcctgctgagttg
ctccttcgctggaggggtacctccgagcagggcatctgca
ggatggtgggcacacttgcctctcactcttggcggggccag
gagtctctgttgccactgagtgacgttatgctgggaaatgga
ggcgtggtgtccgcttctct

>IP_29-07-04_12
taattatccctgctgatatttagaccagaggcaggaagtgg
aatcagttacaactgaactgggaaatatctgggatcctgaa
gcctgacccagaactcttagccaggctagctgcacagattc
tcccacattatttcagaaaccttgtaatagcctaatgccatga
ctgaatttttttttgtaccccatttaatgtatcattgaggagtttc
attccatctcttttacttactactctgtttttgggcaaa

>IP_19-08-04_01
acagaagtgaaagggagaagggatgagcaagaaagtc
tctggatttcaggagaaggaagacagccccaagggcag
gagaaacacttgtgaagggtccattgaagag

>IP_19-08-04_02
acagaagtgaaagggagaagggatgagcaagaaagtc
tctggatttcaggagaaggaagacagccccaagggcag
gagaaacacttgtgaagggtccattgaagag

>Non-IND_19-08-04_05
caaagaagtttattctaggcttaccactctaacccatgttcac
aagctcccccagtcatgcttaat

>Non-IND_19-08-04_06
caaagaagtttattctaggcttaccactctaacccatgttcac
aagctcccccagtcatgcttaat

>WCE_19-08-04_07
catatttatcgccaattttagatatgtttaatggcgaagttattg
cttaagatttgtctacaagcgctaataaccagcaaattaga
aaaatgcttcaaagagctttttttcaaagcacaaaaatcttga
gggtttaatttttcacattga

>WCE_19-08-04_08
agtaagcaggcacatcttacatggctggagaagaagaaa
gagagaggggagaggtattacacanctttgaacaatcac
aacttgtgataactcacaattacaagagtggcaccaatggt
gaaatcttcccccatgatctaatcacttcccaccaagcccc
acctccaacactggaaattacaacttgatatgagatttgggt
ggggatacagaactaaaccatatcactttgcagctggtcac
tggtccctttttgtttgtttggcgaagtcgtgttttcctggatggt
cttgctgctggtgaatgtttgtcagtatctgagcact

>WCE_19-08-04_09
gctttatatatctgggtttttgatgctgggtgcatatatagttag
aattgttatatcttcttgctgaattaatccctttatcattatataat
gaccttctttgcctttcttctaatatttgacttaaagcttgtgttatc
tgatataagtatagctattcttcctcatctttgacttctgtataca
agtaatatcttttcccatcccctttaccatcatactatatttgttttta
caggtgacaagagtttctggtaggcatcatatagt

>WCE_19-08-04_10
aataaaaatgggtttgagggtcaaatatgtttggggtaaaa
gcaaagtcaatgcatttctttactgccggacttctcagtgttttt
gatatgccaataaatgttgtgaatttccaacagtgacataaa
acaggcagggtttcctaatgtccatgaactcctctgcttttca
aaaaccatgcatctcagaatgagcacacccccagcata
gtatgaaaaacactgtattggttaagattactgtcgctagtgtt
gtgattgcactccagatggctcaaggccagcctactatctct
gccctgaaagtgaggaaatctaagctcctggccctggaga
t

>WCE_19-08-04_11
ctaatatagtacattcagctatgattatacatgtgattaggcc
aaatgtggccaagcgtgacattcccagaagctctatgagc
aaaaacgtgcaggaagaatgagtctggctttgccaggcc
gactccttctattactaatcatat

>WCE_19-08-04_12
ccaaatatcccagcctgggaaatatctgattttaccataaag
attcagggagaattacaaatatatttttcaaatgaactaaggt
taaatgaggctgtgattaataacaggggtttgtgttatttattat
attttgagttgcctctgaggtgagctgctctttaagtacttttttctt
tccaaaagagaaaatgggcaaaataataatgaatgctgc
ctattattaggtacatatca

>19-08-04_13
ttttagttatttcttgccttctgctagcttttgaatgtgtttgctctt
gcttttctagttcttttaattgtgatgttagggtgtcaattttggatc
tttcctgctttctcttgtgggcatttagtgctataaatttccctcta
cacactgctttgaatgcgtcccagagattctggtatgtggtgt
ctttgttctcgttggtttcaaagaacatctttaaatcatgctgcta
taaagacacat

>WCE_19-08-04_14
gtttagataccatacacttctccaggaggagtttgacaacat
ggtcacacagttgcaaagtgcacaaagctcaaggcacag
agtagaagaaagttgtgaactgatgggggaaagttctagc
aaaagcagagttagcattttcctttaacaagactttctaatgc
taaacaaagaccaactcttttaaaaggggttgttttggttgtg
ggtgaaaaatactgtactgtaatgatctgctt

>WCE_19-08-04_15
caatttctgcatttataaaatgaaagtaacattagtaccaatt

ccatcagcttgttttgagaatatattggaaaaatttgtaaagtg
tttaccatatgtccagcatctattacatactaaatacattctag
ctattagtacagctaatataggattctattcatttaaagtatcc
aatcatttaaacttgccattctttaatgagcctgcattgtttaaa
ttctttattacttgaatttgaagacaagaacaattgtattcaact
gatagctcaattgatttcttggagtagttgagacactgaatttt
gatatatatatatatatatacacacatacacacacacac
acacatacacatacacacttggaatttgtattatggcccaat
aaatacttttta

>WCE_19-08-04_16
ttaaaaattaataaaagaacaaatcaaatagacaatgaa
attgtacagataagagtaaattaaattgaaacatttactaca
aagattggtaaatccaaaagatggtttatttgaattaacgaa
taaaattgaaccagtatacgtctatttaattttatcaaagtgac
aaaggaggaaatgacagatgagaataatccatttatgag
atggaaacatattctgatgtgcattcaatcctagatttattgta
attgtgaaatagaatataacctgggtgtcagggaacataatt
atgtgtgaattatttgagaaataacactctacacatgtccca
gttgagtctatatgccaatctgaagtggttctaatatagtttga
cattgtaatcatattagag

>WCE_19-08-04_17
ctagtccagccctacaagaaatgcttaaaaacatcctgcat
ttgacagtaacaggatgatatctaccatcatgaaaacacac
ataagtattaccctggtagagcagtcacacaaatgagaaa
gagaaatgactccaattttaccactacaaaaaaaaaaaa
aaacactccaatgataaataataaaagagaaaggaaca
aaagataaactgcatctgagtgtctctttgatatttgactaaa
gtaggttataagatacctatacgttcatttatatttatatctatac
ctctatttatctatctatatccctataaacatctactcgaaatca
tgtggcaataaccattttctcttcaattttattgtacatcagatca
gaatggctttgagcctagagtccccaaaggcttaaactattt
attgctgtccctcaatagactttaagctccatgggggcaaaa
acttctctcttcaatgttgtaaacatggtgtctctaatagtgcct
gatatatagcaggctctcaagaaatatttgttgaatgatttaat
ttagtgctttatctggatactgatcaaaatattttatttaatttaaa
ttttttttgaggtgaggtctcactctgtcaccgaggctggagt
gatctcagttcactgcagcttctgcctcccagggctcaagtg
atcctcccacctcagcctacggagtagttgggcccacagg
cacacgccaccacacctggttcatttttgtatttttggtaaag
gcag

>WCE_19-08-04_18
gtcctgaaaactaacccatgtcaatgtagtctttttaaaattat
taactgaaaaaaattggtgacctcaaaaaaatttttttaagatt
aggaaacaaaaagtacttaaaaggagcctcaaacaaatt
tacaagaaaaagacaaacaaccccatcaaaaagtgggc
aaaggatatgaacagacacttctcaaaagaagacatttat
gcagccaaaagacacatgaaaaaatgctcatcatcactg
gcc

>WCE_19-08-04_19
ttacctgggtcccacacacggcagcaaagaaagtgtggtg
tgcatggctaaagtcaggctatccctgagtctatagaggag
gatacatcagcataacaaataggttcctttctattcacactac
ttaggaatcatggttttctaaatgctttcacaaggctgtaaaat
ggaaacttgatatcaccagtaaacccatgaagaagcctct
gttcatctctatggtgagagagccatttcttta

>WCE_19-08-04_20
tgcaattcggtgttcttgtaagttggtctgtcaccatcaggctc
atgaggaaaggacccaagagaaaacagaaaccagc
ctgagccctcgagggtcttatgtgtggcatagtttaggctagt
gttaccgggggggtgaaattgtttgctattgcccagactcag
atggcaaagacaagtgtggcccagattgaaaggagtgca
ggagggctgtgggctgtgcccaggctcatccctgacaatg
ctgcctgttgggtttatcatact

>WCE_19-08-04_21
gtcaatttattgtcagatatataaatagaatacagaagtctcc
atttatccacaggggatacattccaagaccaccagtggatg
cctgaaaactcagtaccaaaccctctatatgctgtgtttttttctt
a

>WCE_19-08-04_22
agttacaaaacttttttttcctgtgttgcaggggcttctgctgtgg
taaggcctgtgagcaggcaatgtttcattcatttttccgtcctat
gggtacagcacacagtataataaattgctataaggaacta
ataataacagttattgaataaacaaagaaatata

>WCE_19-08-04_23
cacataacagtctaaaattaaaaactaccttccccagactc
ccttgcagctaggtgggctcccatgaatctgttctggccagt
gggatgtaaatagaaatgatgagccctggcacagtggctc
atgcctgtaatcccagcactttgggaggccgaggcaggca
gatcatgag

>WCE_19-08-04_24
ttataatttatcttaagtcataggcattatactcatagtccattta
tttacgtccacccaatagattgtcaggctctgaagacaaca
cctggccttgcccaggactgcacgtacagtgctcacagtag
gtgcactggcttcttttaaataaaggcttttagtatctggatact
tggttttattagattaatcagaatttg

>WCE_19-08-04_25
gtcctgaaaactaacccatgtcaatggtagtctttttaaaatt
attaactgaaaaaaattggtgacctcaaaaaaatttttttaag
attaggaaacaaaaagtacttaaaaggagcctcaaacaa
atttacaagaaaaagacaaacaaccccatcaaaaagtg
ggcaaaggatatgaacagacacttctcaaaagaagacat
ttatgcagccaaaagacacatgaaaaaatgctcatcatca
ctggcc

>WCE_19-08-04_26
cagtcaacgctatcatccccattttaggaagaagggcatcg

aggcacagagaggggaaacaagcctaagatgacacaa
agggttaagtggcagagctgggattcacagccgggtctgg
ctcactcctgaggccaaac

>WCE_19-08-04_27
ctctggaatattcctgggggtgtgaagctcgtcctgccttcct
caggccctactccatttggcgaagaccatggaactagaga
ttctctcgcctgcctggatgtcaa

>WCE_19-08-04_28
gtaagtttaagagaaaaattgaggattctgttttggacatgtt
aaatttgggatggatgttagctatcctagcagagatgtcaag
caggagcaaggaagagagaagagtgtggaattcagaa
gacactcttaactgaagatattaaaaa

>WCE_19-08-04_29
ctgaatttttgaattcagaaagagcaaaatagtaggactaat
aaataaattcaaaatcttgtgctctcttataattagaaggtatt
cactctgtgatatttttctttctttagaataaatatgggctgggc
gcagtggctcatgcctgtaatcccagcactcttgggaggcc
gaggcgggaggatcccttgagctcaggagttcaagacca
gcctgggcaacatggcaaaaccccgtctttacaaaaa

>WCE_19-08-04_30
atccatttagctgaaaatgacaggatttcattcgtttttatggct
gaatactattctattgtgagtatattcccatttttctttatccattcat
ccattgattgacacttagattgattccatatctgggctattgta
aagagtgctgcagtaaatatgggggtacagatatcccgttg
atacactgatatctttttttggatatatacgcaggagtgggatt
gctggatcatatggtagatctgttcttagtttttgagaaatctct
gtacttttttcataatggctgtactaatttacattcccaccaac
aatatacgataattttctttttcttcacatgcttgccagcatttgttg
tgctttgtctttttaatacccattctaacaagtgtgagatgatat
ctcac

>WCE_19-08-04_31
attgataaaataatgcaacactaaatattaaatattatatactt
atatataattattatattgcccaaaatagcatattctctaatgc
gtattttgcagtaggtttgtaaatagcttctttctctcatgacaat
ctttcttctttttctaatgtctgatcaagcccatag

>WCE_19-08-04_32
gacattttgctaaattatcattcaataaacacacatcgaatttc
tatgctgattttatcactacttaac

>WCE_19-08-04_33
cattccagcatcagtgacagagggtgaccctgtctcaaaat
tagaaa

>IP_20-08-04_01
caggcgtgaggcaccatgcctggccaatatatgtatttttttta
aaaaaggccaggcgcggtggctcatgcctataatcccag
cattttgggaggctgaggcgggtggatcacaaggtcagga
gttcgagaccagcctggccaacatggtgaaactccatctct
actaaaaatacaaaaaaattacctgggcgtggtggcagg
cgcctggaatc

>IP_20-08-04_02
tctgaataaaggaaggttccactctgtgagttgaatacaca
caacacaaaggatttactgagaattcttctgtctagcagtaa
atgagaaatcccgcttccaacgaaggcctcaaaggggtct
aactaatcacttgcagactttacagacagagtctttccaaac
tgctctatgaagagaaaggtgaaactctgtgaactgaacg
cacagatgacaaagcagtttctgagaatgattctgtgtagttt
ttacacgaagatatttccatttcaaagattagcctcaaatcgc
ttgaaatctccacttgcaaactccacagaaagaatttttcaa
aactgctctgtctaaaggaaggttcaactctgtgacttgaat
acacacaacacaaagaagtgactgagaattcttctgtctag
cattatatgaagaaatcccgtttccaacgaaggcctcaatg
aagtccaaaaaagcacttgcaggctttacaaacagagtgt
ttccaaactgctctatgaaaagaaaggttaaactctgtgagt
tgaacgcacacatcacaaagtagttgttgagaatgattctgt
gtagtttttatacgaagatatttccttttctgccataggcctaga
agcgcttgaaatctgcacttgcaaattccaaaaaca

>IP_20-08-04_03
tctgaataaaggaaggttccactctgtgagttgaatacaca
caacacaaaggatttactgagaattcttctgtctagcagtaa
atgagaaatcccgcttccaacgaaggcctcaaaggggtct
aactaatcacttgcagactttacagacagagtctttccaaac
tgctctatgaagagaaaggtgaaactctgtgaactgaacg
cacagatgacaaagcagtttctgagaatgattctgtgtagttt
ttacacgaagatatttccatttcaaagattagcctcaaatcgc
ttgaaatctccacttgcaaactccacagaaagaatttttcaa
aactgctctgtctaaaggaaggttcaactctgtgacttgaat
acacacaacacaaagaagtgactgagaattcttctgtct
agcattatatgaagaaatcccgtttccaacgaaggcctcaa
tgaagtccaaaaaagcacttgcaggctttacaaacagagt
gtttccaaactgctctatgaaaagaaaggttaaactctgtga
gttgaacgcacacatcacaaagtagttgttgagaatgattct
gtgtagtttttatacgaagatatttccttttctgccataggccta
gaagcgcttgaaatctgcacttgcaaattccaaaaaca

>IP_20-08-04_04
ctcatcatcactggccatcagagaaatgcaaatcaaaacc
acaatgagataccatctcacaccagttagaatggcaatcat
tacaaagtcag

>IP_20-08-04_05
cacatggctaacttagtctcctcagcttgaccttgcaaggtc
aagctgatactgcatggcccaaggcccacaggtaaacaa
aaacactcttatcaggcaagatattctaatggcttagaggtt
atctcccaggagccaggcaggggcctattctttatttggaat
gtccagggttcaaacatcccaagttcatcctgtatcacac

>IP_20-08-04_06
tcaatatagtgagctaattcaatataatgagctaatacttgac
aactagatactttaaaaagtgttaacatgcagcttgtgaata

gcaagttttctacctggatcaaactgggggaggggaacac
aggaaaactccagaaagcaagc

>IP_20-08-04_07
gacctcaggtaatacacccacctcagcctcccaaagtgct
ggaattacagacatgagccactgcgcccggccaatttttta
ttttttgtggaaataaggtcccactatgttgcccaggctggtct
caaactcctggcctcaagtgatcaaagtgctgggattatag
gtgtgagccactgtgcccagccaaataacagcaattcttaa
tgtttacactcttagaagtatcagaagcactgaaaatataac
tgaggaattatagtaacattataaatatataatagctcttgtt
ctaagagctgtagctctagggagggagactatttagtatattt
ttgctttagaggtaaggaatgtcttggtttgtgaaactaaaag
catctgccaagcttctgattatataacactatagatcatctcc
agaggatcactcttgggtactcccatacaaatcttccctctgt
gtaagtgcacagtagtgggtagtctatcagggcacagact
cagagtatcctacgatggaaaaccacaatgaaaagggat
acagaggcaaccacagtaccagattgagtagatt

>IP_20-08-04_08
ttcgctgggtaaactggctggcgcgaaaaccgaagcgaa
aggcgcatatcctggctggcagccggacgctaattctccg
gtgatgcatctggtacgtgaaacctatcagcgcctgttcaac
aagacgccgaacatccagattatccacgcgggcctggaa
tg

>IP_20-08-04_09
tatgtgaaatatgcatgtggaatttctaggaaaagcaataca
agaatagaaacagaggacatagcttcaaatctagtagatg
gaaaaagaaaggaaaaagataaaagtaattaatctaaa
aaaagaaagtaagcaaaacaaaatgttag

>IP_20-08-04_10
tgaataggaagagctttggtccgtggcatttctagtccctctttt
ggtctattagaaacaggagagtcggccgggcgcggtggg
tcacgcctttaatcccagcactttgggaggtcgaggtgggtg
gatgacgaggtgaggagttcgagagcagcctggctaatat
ggtgaaccccgtctatactaaaaatacaaaaattagctgg
gcgtggtggcgcgcgcctgtaatcg

>IP_20-08-04_12
agatcagtagggtggagtggagagtggaatcattaggaat
caaatccagtattagtacccttcccctataggcagctaagg
aggaagag

>IP_20-08-04_13
gcctacatttgcttttgatctatctagtataagaaatattcaca
gaaattgtatgtagataccctttgttttgaaaatgct

>IP_20-08-04_14
cactgagtgaaagtagccagaccagaaaacaggacaca
tttatatttatagaaacttaatttatataaaattctagaaaaata
tgaactgaactatgacagaaagcatgtcagcagatgtcga
ggaagacagaggatgcaggagcagaaggtaggaggg

agggattagagagggaaatgagaggatctattcatcagct
agattgtagtgatggttttatggaagtaggcacatgttga

>IP_20-08-04_15
cctggctaatttttgtatttttattatttatttatttatttttattt
ttttgagatggagtttcactcttgttgcccaggctggagtgca
gtggcacaatctcggctcactgcaacctccacctcccaggt
tcaagcaattctcctgcctcagcctcctgagtagctgggact
acaggcacgt

>IP_20-08-04_16
ctgaagccacagtccacagatgggatttttctctgcctcagg
gaaacctcggttctgctcttaaagcctttcagttgattgggtat
agccctcctagattgt

>IP_20-08-04_17
ctgaagccacagtccacagatgggatttttctctgcctcagg
gaaacctcggttctgctcttaaagcctttcagttgattgggtat
agccctcctagattgt

>IP_20-08-04_18
gacggggtttcaccatatttgccaggctggtctcgagctcct
gaccttgtgatctacctgcctcggactcccaaactgctggga
ttacaggcatgagccaccatgcctg

>IP_20-08-04_19
ttattattctctgggcggcgcacgggagaccctgcagggca
acagagcagagcgactacagctcccaggagccaacgct
gcagggctgagccgacgcgggggacagacaggaccta
aatg

>IP_20-08-04_20
agcattaggagatatacctaatgtaaatgataagttaatgg
gtacagcacaccaatatggcatatgtatacatatgtaacaa
acctgcacactgtgcacatgtaccctagaacttaaagtata
ataaaaaaattttttttaa

>IP_20-08-04_21
catccatgttgctatgaatgccattatctcattccttttatggct
gagtagtattccatggtgtatatataaaacattttctttatccac
ttgttgattgatgggcatttgggctggttccatatttttgcatttgc
aaattgtgctgctataaacatgcgtgtgcaagtatctttttcgt
ataatgacttctttttttctgggtggatacctagtagtgggattg
ctaaatcaaatggtagatctacttttagttctttaaggagtctc
cacactgttttccatagtgattgtactagtttacattcccacca
gcagtgtaaaagtgttccctttttcaccacatccat

>IP_20-08-04_23
ggatttttttgggggggaaaattgggggtattacgataattgca
tataaaat

>IP_20-08-04_24
ttctctgtatttcctgaatctgaacgttggcctgccttgctagatt
ggggaagttctcctggataatatcctgcagagtgttttccaac

ttggttccattc

>IP_20-08-04_25
ccgtcctggccacgcaggaagaggatgaaacttcaccgc
gcgcgctggcgttgcgtgtgcttgacc

>IP_20-08-04_26
tgggaggccgaggcgggtagatcacgaggtcaggagat
cgagaccgtcctggctaacatggtgaaaccccatctctact
gaaaacacaaaaaattagccgggcatggtggagggcgc
ctgtagtcctggctactcgggaggctaaggcaggagaatg
gtgggaacctgggaggcagagcttgcagtgagccgagat
cgcaccactgcactccagcctggacaacagagcaagact
ccctctcaaaaaaaaaaaaaaaaaaaaaaaagacttca
tttccttgttatcttacagaaagtttcaattttatcagtgacctca
gactgagcatgagttgtaacaaagtatactttaagtctactct
ggtctacaatgcttgcagccaggtgtggtggctcacacctgt
actcccaacactttgggaggcagaggcgggcagataactt
gaggtcaggagttccagaccagcctggcctacacggtga
aaccctgtctccactaaaaatacaaaatctagccaagcgt
ggtagtggtgggtgcctgtaatcccagctacttgggaggat
gaggcaggagaactgcttgaacccgggaagtggaggttg
cagtgag

>IP_20-08-04_27
gtgtgagccgccaacggactaatgcatgcaggctgcaaa
gattcgtgggagaagtgtggtttcctaagagtagtcgcaca
ataactcactgcttcccttggct

>Non-IND_20-08-04_33
ctcctgcagtcaacttcaatgtagtgaatggcaataccatgc
tttcagttgcttagaccaaacatcttaacaatgatctccaact
ccttttttttcccactcatactctacattcaatctgtcaagcaatc
ctgttggttcttcatcaaaatacacccaaggtctcattatcatc
actacctccattgctactaccctagtacaagccactatcatc
aattatctagattactgctgcaattagcaaactatctgcttcca
atccacagtctattgtcaacactgcagccaaagtgatactg
aaaactagactgtgtaactacacggcataaaa

>Non-IND_20-08-04_34
ttcaatgaaaattattataaaccatcacactcagttttataag
gctgcttctttgttttcagcttttccaagttaatttcagtcattagg
gggccgtgataatatatgtaattttatacatgtatgtatagattt
gaaggtgcatatatttcataactttcattcttggcagacaggg
tcagaaa

>Non-IND_20-08-04_35
ttcaatgaaaattattataaaccatcacactcagttttataag
gctgcttctttgttttcagcttttccaagttaatttcagtcattagg
gggccgtgataatatatgtaattttatacatgtatgtatagattt
gaaggtgcatatatttcataactttcattcttggcagacaggg
tcagaaa

>Non-IND_20-08-04_36
cacttggttacgccggttagagtctcagatgctttgttacagg
ggcacaaactaccctacccctactgtgcgtgggtgcggga
gccttggtccataagggagccatgtgggcaatggcttagg
gtggtttagggttcctcagagacagagaaccagct

>WCE_20-08-04_39
acctctatatgtccaaatttttagttcccacatatcagttagag
catgtgatatttgtctttgcgtgcctgacttatgtcacttaagat
aataacttccagtgccattcatgttgttgcaaatgacatgattt
cattatgtatatcacattttctttatctattcatccattaaggaac
acttaggttgatttcttatctttactatcgtgaatagtactgcaat
aaacatgagtgcaggtattgctttgataatattaatttattttcct
tagatacccagtgtgaaggctgctggatca

>WCE_20-08-04_40
atatgagaattaagcatatgctcagattttcttaaggagtttcc
ttattattgaagttaatttgtttcttcgtaaagattgtaaatttata
gaatgtgaaatttgatagctagatttttttgcttgtcctttaggac
tatttattttctcttcctgcttcctaagttgtcacatatgttacaac
aag

>WCE_20-08-04_41
aatttttgctttagttgtggcaattattacatttagcaccataata
atttcaacagaaaagaaacaagctcttgaacgtgaagata
aacttttcca

>WCE_20-08-04_42
ccagtcaggcactttgatgccttcaagcagcctaggaccta
gaggtttcaatgcctctgctttcctatttataaatgggcagact
ggacctgaagtcctgaaggctcacccgcctggaagccac
ctctcctctggagttctgatgacgctgcagaagcctgtacgt
atctctgcccaataccacatgtacaacagaggactttaact
aaaacgagtgagagagagagattccttttcccaacttagg
atctaaaagtttcttgaacgtt

>WCE_20-08-04_43
aacttacggaccaatcaatacagccattacccttggacca
tggcaccgacacaggtgcaaggagccaggggagcagg
cctcgcccctgcacctctgccatccttgaggtccctgagctct
gcccagcttctgcagctccttctggtggctctcacaaggtctt
ctcattgttgccccctcttctgtctttcttgagtcctcttacac

>WCE_20-08-04_44
attaaggtatctacattaacttactgaggaggtagagtaaaa
tattttaactgccttaacgagcttagaaattgttgtgtggccac
ttgaagtaacttagaaaaatattatttagcatattattatggaa
ttttgacaggttagtaactatcaaacataaagaaatgtttaca
ttgttaataattctacatgtacaaaactatttacatctattctttttt
aaaaaaattttgaatgactgaactcaaatatgcagaataac
catgag

>WCE_20-08-04_45
atgatagctctctctgtgcatcatcttagaacttgagacagta
aaatatcgttttcaggtacaagtttatgctagtctttcgtgcatc

aaaagtatctgccaacgttccgtgtagtgacaacagaggct
ggaacccatcaaacaatcctaaaattcaactaataccaaa
gaatgttcatgttaaaaagttaagaatttttggtgagaaaat
gaaggatgccaacttaagaaatatactattt

>WCE_20-08-04_46
cttcaattctacgtgctttcaaaggattctggagggtgagga
caagggaagggagattggggtcttagagtccctgatgctgt
taagccttttcccggcagtggcttctcctcccctgggccgtgg
gcaactgttgatctcacgctgcagagagggctttgctccgta
atccattttcctctgcctttctttggagattagacaagtgcccc
gctggccaccaaagccagctaaagctctctggggcctgg
aggctctctgtttccagtcagggaaggattcttgagggtctg
gacagccacctctactgctggatgaaaaaggcaggggct
agggatcctgctctctagggccatggggtccagccaaggg
agccaagatatctggccaaagtaagtaaataaattgatcat
atttaacaaatctttaaagaaatcaaacaaggcctgtttggg
agtgatgtgagttaagtattggttgagccacccagacagag
ttgtttaatatttctgagcctctgttttgccatctgtaaaagggg
aataaagatacctcagcccccacttccccaggcagtttgga
ggcccttcgaggcccctgtgtgacctgagccttcttatttccct
ccctaatctcat

>WCE_20-08-04_47
gtagaggataaagaaacacaactttaatttttagagtggcc
tatatattttgtttggaagcttaaattctctgttatcacctggtagt
ggccatccctgctgctgtggcccctcttactgttactctgacct
tcacaggtctttgggaacttatcactcattgttgatggcatttttc
ttctgtatagcagtctttgcttttcttcctatcatcaaacagatta
agaatggacac

>IP_21-08-04_01
gatgcagtttcttcatagcgttgatggtctttcaatttggtatgttt
gtacagtggatatttagtgcttccttcaggag

>IP_21-08-04_02
tattcgtagggtacactgaaactaaggataaagccctctag
tatttttacaacatgatcttctttaaaagatagccacctaaag
tcaaaagttaaatggcaactgtcatatgacattgtacagtttt
gttgaaattatttaggcaaaaaaatctggacttgatttattctta
ttttgttctaataaatagggcagcattgcaactaaacttatttat
aaaataaaataagcatattttaataatattaagcccaaatta
aagact

>IP_21-08-04_03
ttctctcaaattcagattaaggtggatatngtgttttataaaatg
agctgaggcccaccagccccatggaaatatttgcccatgg
tgcagtgcctgagatttttgcatatgaggtttgacactttccctc
ctgagtgctgagggaaggtgagagaaagaaagttaaata
tggaaactacttttcttcaatattctgtctcttgactgcaataca
agtctt

>IP_21-08-04_04
aggaacatgtgaaggttccgcttcaaaatatgaatgaatgg
gctcctcgtgcccctattattttctggccacccccaggggtcc
tgatggaagggaggctggcctgcggggtgcaagaacca
cgattcactgctccccgctgaaggtgcttggtc

>IP_21-08-04_05
cctaacagctttcttttctttctttcttttgagacgtagtctcgctc
tgtcgcccaggctggagtccagtggcacgttctcggctcact
gcaacctccgcctcccgggttcaagtgattattctgcctcag
cctcctgagtagctgggatgacaggcgcgcgccaccacg
cc

>IP_21-08-04_06
ctgggacctcctgagtagctgggaccacacacatttaacct
gtatttataaaattactgtttagagaataacatttgatggaatc
atgcttttactttctgcttatgactcaattgtttgtactgacattaa
catcccaaatccttagcatggcctacaaggccctgagcaa
tgtggcacctgctgaagcctgctgcctcat

>IP_21-08-04_07
ctagctcccatgtcttgaccaactttgcattcccctccaggac
cttatgggcacatgggttcctatcatttcacgggtatgtagact
caggaatgtggcattgacttctgggtggagtacagatgact
gacacccccaccccgattcg

>IP_21-08-04_08
tattcgtagggtacactgaaactaaggataaagccctctag
tatttttacaacatgatcttctttaaaagatagccacctaaag
tcaaaagttaaatggcaactgtcatatgacattgtacagtttt
gttgaaattatttaggcaaaaaaatctggacttgatttattctta
ttttgttctaataaatagggcagcattgcaactaaacttatttat
aaaataaaataagcatattttaataatattaagcccaaatta
aagact

>IP_21-08-04_09
tagaaatctaagatggatttttcagttttatttggttttgcgttata
cacagtttaaggggggaagaaaacagaaataacttagta
cagtactttccaaatctagctgtacatcagaatcaccagata
agctttgaaaaaccaaggatttcctggccatgaaccag

>IP_21-08-04_11
tcttcttgttgaattgatccctttatcattatgtaatggtcttctttgt
ctcttttgttctttgttggtttaaagtctgttttatcagagactagg
attgcaacccctgccttttttttgttttccatttgcttggtagatctt
cctccatccttttattttgagcctatgtgtgtctctgcacatgag
atgggtttcc

>IP_21-08-04_14
tttctctgtttttctctaagtttctttctgtctttcctgtattggtctcta
actttcatgttaaatccttgcctccaagtctggtgattctggggt
atccatttacatttgagttagatgctaaaagtctgattgcatgt
ggataga

>IP_21-08-04_16
cctcggcgcgggcttgctgatcctgcc

>IP_21-08-04_18
agctatgtttggttgaaagcagaaaggccttggcagcactc
acggactcagcctcctcctgtttgcactgcggcccatggcg
ggggtgcctgtcctgctgctgagcactactccatctgaaccc
agctcactcccaccccagtgctgcgtgacg

>IP_21-08-04_19
tgcaggtttgttacgtaggtatacacgtgccatggtggttgat
gggtgcacccatcaaccctacattangtatttctcctaatgct
atttctcccctantccaccacccccaaacaggccccantgt
gtgatgttcccctttcctgtgtccatgtgttctcatcgttcagctcc
cacttat

>IP_21-08-04_20
tgatggagtgagactctgcctcaaaaaaaaagcaataga
aacaatatcattgttagttttatttgtatcgtttttctggattactaa
tgaggttaaacatattttcatatgtctattgaccactcagaatt
cctttc

>IP_21-08-04_21
ctaatttttttaatatttttagtagagatggagtttcaccatgttgg
ccaggctggtcttgaactcctgacctcaagtgatctgcccat
cttggtctcccaaagtgctgggattacaggtgtgacccactg
tgcccagccctctccagcatattttatacaaagttttcctatcc
acaatccaatcccaggcaaggcatgattcctgaggatact
ggatgccggagtcccaccgagacttgcgtggcactgcatc
atcaccctgcaccttttttctgtctggctgagtggagtct

>IP_21-08-04_22
aaaaaaatccaaattccaatacacttctagtcccaaacattt
tagataagggatactcaacacgtacattttgtgggtgtatatg
tgtgtgtgtattaccaataacctaatttaaaggaatgacttaa
cacagtactcaggtaaagctaaagtgtagtatgaccaccta
aacttggggatacactcactgaactaaataaaaactgcttt
aactgtgtcaactggaacaccatttaatgatcaatgacgag
aggtattcatctaacactttatcaatataa

>IP_21-08-04_23
ggtggctcatgactgtaatccgagcaccttgggaggccga
ggtgggtgatcatttgaggtcaggagttcgagaccagccta
gccaacatggtgaaaccccacccgtac

>IP_21-08-04_25
ggccgtcctgggctatgcctctatggaattcaccatcggttcc
agccagccaacggtcgcaacgtgccatcgtctatcc

>IP_21-08-04_26
gttaggggaggcctggggagccttgtctggctgcatctgga
gaggaagtacttttaccattgcctgcttcagctcctgattggg
aaccagtaggtttgggaagcagagagcagaaggaagct
ctcttagagcaagaggagagggactcaagggaggcaga
aggaggttatggatacccacagtcctgagtcccaggcccc
cagacctgcc

>IP_21-08-04_27
gtcctcacaagccctgagttcagtttgggaaggtgcctgaa
attcatgtgtctgcattgctccagataaggagcaaaagttgt
gcatttcccaccaacccatgacctaagctgctgcagcctgg
tgccatcttgaatctagaggcactgctgaaatacactatgttc
tgggggccagtaaacactgtgcctctccagtcttgcagctcc
actgtcatttcaccaagccccaactgatggctgaaatgctgt
gaccctgactgctcagatcttgggcccaggaaaggt

>IP_21-08-04_28
ngtttttagcatgaagggttgttgaattttgtcaaaggccttttct
gcatctattgagataatcatgtggtttttgtctttggctctgtttat
atgctggattacatttattgatttgcgtatattgaaccagccttg
catcccagggatgaagcccacttgatcatggtggataagct
ttttgatgtgctgctggattcggtttg

>IP_21-08-04_29
atcacttttttgtctttgggaaaccacaggaacaatttctctgg
agacaaggctgtgtctctctcctggtcatttttgttccagcctct
tcagactgtgcaatctttcagcaggaactccctctcttctcggt
agctttgaatcttaagcttctacgggagagtggtagaactgg
atcatttcctaatcccatttagttgtgcttttcttcatttacttcata
ccccaggacccctttccccagcagcagagaccctggagc
acaggagagtagggag

>Non-IND_21-08-04_31
cagcccggtcttgtaggcctgataagacgcgtgagcgtcg
catcaggcaaggcaaacagtgaggaatctatgtccaaact
cgatctaaacgccctgaacgaactgccgaaggtagatcg
cattctggcgctggcggaaactaacgcccaactggaaaa
actggacgctgaaggccgcgtagcctgggcgctggataat
ctgcc

>Non-IND_21-08-04_32
cagcccggtcttgtaggcctgataagacgcgtgagcgtcg
catcaggcaaggcaaacagtgaggaatctatgtccaaact
cgatctaaacgccctgaacgaactgccgaaggtagatcg
cattctggcgctggcggaaactaacgcccaactggaaaa
actggacgctgaaggccgcgtagcctgggcgctggataat
ctgcc

>Non-IND_21-08-04_33
ttaattgtgatgttaggatgtcaattttagatctttcctgcttcctc
ttgtgggcatttagtgctataaatttccctctacatactgctttaa
atgtgtcccagagattctggtacgttgtgtctttgttctcattggt
ttcaaagaacatctttatttctaccttcatttc

>Non-IND_21-08-04_34
ctctacaaaaaattagaaacttaccaggtatagtggcgcac
acnctgttgtcccagctactcggaaggctggacag

>WCE_21-08-04_36
atatattttatatatatcattatatatataccatatatatatatttt
tgaggcagagtctcactccattgcccaggttgaagtacagt
agtgtgaacacagctcactatataccaaagaaaattaaag
aaaaatcaaaacatattaccagaaagaaaaaaaatatga
atgaaacacaaaggaaacagtaggagaggaaaagatg
ggcaaaaaattatcagatagataaaattaattaacaaaac
agcaagacatttgctattaataattactttaaatat

>WCE_21-08-04_37
atctttaattatggattattgtaaagaagagatcatgataatttt
taattttatttctttgaaagatgtatctacacacacacacac
acacacacacacacacactgtatacaaagattggaagttg
ggaatggagttaaaagggcatggtctttgaaattaacctttta
ttcattcgttcaatcattcattcagtgaatcatccaatcagtcat
tcatt

>WCE_21-08-04_38
cattaaaaatacaaaaaaattagccaggctttgtggtacat
gcctgtagtcccagctacacaggaggctaaggcaggagg
atcgcttgaacc

>WCE_21-08-04_39
gtatgtataataataattattattatttcgagacagagtttcact
cttgttgcccaggctggagtgcaatagcgtgatctcggctca
ctgcaacctctgcctcccgggttcaaccgattctcctgcctca
gccacctgagtagctgggattaaaggcacgtgccaccatg
ccgagctaatttgtatttttagtagaggtggggtttctccatgtt
ggttaggttggt

>WCE_21-08-04_40
actaccattcaacccagcaattccattactgggttatataccc
aaaggaaaataagctattctaccaaaaagacacacgcac
ttgcaggttcatcacagcacggttcacaatagcaacaaca
tggaatcacctaggtacccatcgacagtggactgaataaa
gaaaatgtcactcatagtcaccatggaatactacacagcc
ataccaaaatgagatcatgtcctttgcagcaatgtggatg
cagctagaggccaccaccctaaatgaattaatgtaggaat
agaaaaccaaacactgcatgttcttgcttacaagtgaaagc
taaatatagggtacacttggaaataaagatggtaataataa
acaccagtgagtactagaaaggggagagagaagaggct
cagggcctgaaaaactaccaattgggtactatt

>WCE_21-08-04_41
actaccattcaacccagcaattccattactgggttatataccc
aaaggaaaataagctattctaccaaaaagacacacgcac
ttgcaggttcatcacagcacggttcacaatagcaacaacat
ggaatcacctaggtacccatcgacagtggactgaataaag
aaaatgtcactcatagtcaccatggaatactacacagccat
accaaaatgagatcatgtcctttgcagcaatgtggatgcag
ctagaggccaccaccctaaatgaattaatgtaggaataga
aaaccaaacactgcatgttcttgcttacaagtgaaagctaa
atatagggtacacttggaaataaagatggtaataataaac
accagtgagtactagaaaggggagagagaagaggctca
gggcctgaaaaactaccaattgggtactatt

>WCE_21-08-04_42
aacttacggaccaatcaatacagccattacccttggacca
tggcaccgacacaggtgcaaggagccaggggagcagg
cctcgcccctgcacctctgccatccttgaggtccctgagctc
tgcccagcttctgcagctccttctggtggctctcacaaggtctt
ctcattgttgccccctcttctgtctttcttgagtcctcttacac

>WCE_21-08-04_43
gagtatcccttggatgaggcccaaggtccacatggctttaa
atcaaacgctatttaattacataggaatgtaatatttcatattc
agtattttaaaaacagaagaataaaggtgggtataatttgc
ctgatatttttttctgcacatagattaaaaaaaattcatttccaat
agtttaaaatatgcagacaaagccttatgaaatgcaaaaa
ccaaaagctcagacaggataa

>WCE_21-08-04_44
tatctgttattttgtgggtttttttccttatgaaggacctcccaaat
tgtgtaagtttgaggccctacaaaatctgaaaccatccctta
aagagaccaaaaagaaggccaccatgacaaataacta
aatgactagaatgcgaaatccaatggagatgaagag

>WCE_21-08-04_45
gtattaaaaatctctgttgccagtacttttgtccctacttcaagg
caagactgaagaatcgaagactaaaatatgagaagaaat
cata

>WCE_21-08-04_46
cacacacaatggctgacaaaacaatgtggtgaggaaga
aatggatacttttctggacacataacccaagactaaaccag
gaagaaattaaatatctgaacagaccaataacaaactcc
aaacttgaatcagtaataaacagtcaggggacaaggcca
aaagaaccagcctgatgctagtatcactaggaactggccc
cactcctctaactctcaggcccacccataaggacaggcac
ctcttgtctttaaggtgagtttcacaaaaaggaaaaaaaaa
aatgaggaactgaagtagaaaagtcataaacacactgag
cttgggaaagacccttaggaggagtaggaaagtgcccta
attagactagaatcagtggttgtctcagtaaagggaaaatta
gccatctttagaactcttgggcttgagccagaggtaaccagt
gctggcccttcccactcaaggttttaagaa

>WCE_21-08-04_47
taattattctggtctataatagattattattactactattttgttctta
ttaatcataacaatgctgctaatttcaaatatacattttctgcca
tattattatttttaccttaaggagaattattgtcaagatatcagg
agggaaatatttgaatttcctttgagtcaagtttcttatttgttcgt
cagttcccacctctctgctgagggtcatatgaa

>IP_24-08-04_02
atctgtgtatgtgtttttagctcacaaacattttgttttggcccat
agtctttggtttcagaaagtaagcaggctaacatgtgcatag
cactgttgtgtttttaatggggatgatgggtttaggtggagattt
gcagtcactgactaccatcttgggcttgtcagcatttcatgaa

aggagaattagaagaaacagtataacctagtgaggtcttttt
ttctgagtcc

>IP_24-08-04_03
atcggccaggacgtagtggaattgcgaaccgatggtagc
aacatgtccgccgagcaggttaacaatccccatcaccgcg
caaccagccagcaaggccggaaggtgaccgggtaaac
gaatgcccagacgcggccaaaaaacaagaatacctagc

>IP_24-08-04_05
atcggccaggacgtagtggaattgcgaaccgatggtagc
aacatgtccgccgagcaggttaacaatccccatcaccgcg
caaccagccagcaaggccggaaggtgaccgggtaaac
gaatgcccagacgcggccaaaaaacaagaatacctagc

>IP_24-08-04_06
ctctgtccgtttgcccgcctgatggacagcagcagcggtgc
ctgcagccctcctcatgtgctgtggcagctgcccccagccg
ctcccctcggacttcctcccgcaagaccccagcttcttcccc
aggccctagtcctctccgctcaagcctcgcatccctcatccc
tccaggcctggcagcccagcacccccccttcagatccttgtg

>IP_24-08-04_07
cc
taacactttgggaggctgaggcgggtggatcacctgaggt
caggagttcgagaccagcctggccaacatgacaaaaacc
ccgtatctactaaaaatacaacaattagccagacgtggtgg
tgcgcacctgtaatcccaactactagggaggctgaggcag
gagaatcacttgaacctggtggggcggaggctggagtga
gctgagatcatgccattgcactccagcctgggcacactcgc
aaaaaaaaaaagtgcaatatagcttttcacaaaatatgga
actgtggtagtgtagaacaatgtctcaatatacctcctacact
aagt

>IP_24-08-04_09
aactgagaactcttcatgtgcaggcctccatagaccccagt
cacgcgctgcccatgtctgagtgacggtgtcttcctagcaca
ccacacttgccatagcccggtactgcttatagctctgaagttt
gcttgccacggtacagatgaaactgtctcccagg

>IP_24-08-04_10
gatgtctcaatgcaatgttgtctccgtccgatgacgacaggg
agacgtgagctgggctcttccttccagcaagtgccctgtgg
ccgtacaggcctccaaggcagcttccagtcacttctattctg
cttatcacaggga

>IP_24-08-04_13
attcccagataaaattggcgtctcttaacagtttctctggacc
ctcgatcgtgtcttggctcacgtggcacacgaagtaacatg
gccgaccagttccgtaaatgtgctgctgccggggccctgg
ggaccagtgtccaagcacctttctggctggtgggtggcact
gatctcatgtcacagggcatgactgctgtcgtcaaccttggt
ggccatgaatacagcattgagaggaagggcttgaagaag
cagtgg

>IP_24-08-04_14
ctcccttctccggtgattcccagaacttccgtctctacccgtca
cctgctcccaatcaaaggagcttaatagacggcgacgtttt
cgactgcccagccaccaggcaggctgaaagtgcatgga
gggcatcatccacctttagacagctacttgcccacacctcat
cgagctctgggcgttgtgccagcatttgaagaagcgatctgt
tgcaaagcgatccacacggagagccttagttatcgcgaga
cttgacaggcaccctcaactttagcccgccggaagcgga

>IP_24-08-04_15
gctatcagcaaatctgagttcttcttttgtgttttgaccaagtta
ctttaggacttgagacaagtatcacttcctgtctgaaatggtc
atgtaatatgtgttct

>IP_24-08-04_17
tatgcctcctgaccttgtcctccccaacccaatataccctac
aacacagccactgggggtatttcaatgcgccatggctgtgt
caccatcagcttcaccgctctctcttggatggctccttgttgtc
caaatgaggcaaaaactactatggctggtttctgccaacct
ctagtctcacctgtccccagttccccactcctgcctctgctcc
agccaagctgaatacttgctagtccccaacacgtgatgga
cctacaggtg

>Non-IND_24-08-04_18
agtccaatagtggcccagctaaagaaaggatttcagtaag
aacccaccctctgtccatttgtcagaaattcttcacagagga
aaaatctatccagtacccaagaggaaaaccttcaatgaga
ttggttcttcataagcaagtcagaa

>Non-IND_24-08-04_19
gatagcattaggagatatacctaacgctaaatgacgagtta
atgggtgcagcacaccagcatggcacatgtatacatatgta
actaacctgcacattgtgcacatgtaccctaaaacttaaaat
ataataataataaaataaaaaggagatctatt

>Non-IND_24-08-04_20
ttgtttagtgcttgttctattctatttttccatatgaagcaagagg
aatacagaagccaaagtcaaacctggctgctttccctaga
ggtcctaagcacacatgcacagcatgaactcg

>Non-IND_24-08-04_21
gacaataggcaaatgccaccacattcatctaatttattgtattt
tttgtagagatggagtctctataaaagttgcccaggctggtct
taaactcttgagctcaagcgatctacccaccttggcctccca
aagtgcagagattacaggtgtgagtgacttcgcccaacca
acaattggatacttttcagtactatctgaattattattatttgcca
tacacataagttacatctattaataaaatagtcaggccagg
cgcggtggctcatgcctgtaatccagctacttgggaggctg
aggcaggagaatcacttgaacccgggaggcagaggttat
agtgagccaagatcgcgccattgcactccagcctaggca
acaagagcgaaactccgtctcaaataaacaaataaatag
g

>Non-IND_24-08-04_22
agtccaatagtggcccagctaaagaaaggatttcagtaag
aacccaccctctgtccatttgtcagaaattcttcacagagga
aaaatctatccagtacccaagaggaaaaccttcaatgaga
ttggttcttcataagcaagtcagaa

>Non-IND_24-08-04_23
tattagcatatatagttttatttaaatattaagttcagtgtcgttac
atctagaaagccttttctaactttccagaaaggagtgaccatt
ccctcttgtaatctcctttgtaatctgtttatatttatgtcatttgttg
aactaattcacaaatatctccccgctgaagctctgagctctt
cgtttgtttgtttgttttgtttgtttgttgttgttgtttttcgagatg
gagtcttgctctgttacccagactggagtgcagtggtgtgatc
ttggctcactgcaacctctgcttcccgggttcaagagattttcc
tgcctcagcctcctgagtagttgggattacaggttccagcca
ccacacccggctaatttttgtatttttagtagaaacagggtttc
accatgt

>Non-IND_24-08-04_24
acagcagaaaaaaacatttttaattgataatgggaagagg
gtagcaagttaagggcacaagtaattaatattacctgctga
aaacaggcctggacaagattctcagggaacatattcctgtt

>Non-IND_24-08-04_28
ctcataggtgttcgcagtgagtaaaggaactcttgtctttcatt
aatcttgacaaatgtataattattacctcttcacatattgtatctt
ttctacttttaaaattatctttttctgtaactataggtagatacata
gtaatccattttaattactctgccatgactcttaaattgccatgt
ctcttttccttgagtttatatatg

>WCE_24-08-04_29
cctcagtttcacttctaagctctctctttgtgatttgtaacattttg
gagtatgagattctttcatagaagtgcccattcatgtatctttc
atgaagtttcatgagcatactgaacctcatccatggagctaa
atgacacagtattcgtacctaaactcccaacaccagctgcc
tgtggtgggtcacccaccgagtaaccgtgtcttacagagtct
actgtcaaacttatctatagtgttcctttctttttattgtcagtaga
aacactggctcccaaggatttcttctttt

>WCE_24-08-04_30
caataatgtacctaattcttgcatgaaatgaggaaaactgtc
ttctaccatcgcattactcatgatagctgaaatttttaacacct
caacatc

>WCE_24-08-04_31
gtcatcttaaatcgacatactcagaagataaaatgtatgtta
caaggaagcaacaattcgtcaatacatcccaacattcagg
aaaagcaatagataatggaaaaccattttgaagcaattac
ctcactgatgaggataatcataaaatgaactgatttgttataa
tgagtatcatgttctgagagaaatatcgcatactaagacaat
atcctggaagaaatattctatttctgaatttgctcacttg

>WCE_24-08-04_32
aattacatattatataggtaatatataaatatgaaaatgcac
acagagacacatttacatcttggatattagcttttaattaagc
agacttttaacctttgagctcttaaaaaatccttttaaatctcaa
taccatttttt

>WCE_24-08-04_33
atccacccaccttggcctcccaaactgctgggattacaggc
gtgcaccaaacgctcctggccagaaaaccatattctaagg
aaagcaaacagttatcacaattacacacttcagcaacctc
catctcctctttgctacttaagggatgaaaacatcaactgtgt
atgtaaaag

>WCE_24-08-04_34
cttactatgtgactttccaaacacacggacctcttaagaccc
ctgcacacactatgggccccgcctccacctgggctgccca
gcttcctcccaccatgtctgtgtggcatcctcccacccatcctt
accgggacgcttgtttaaatcacgtcttctgcacctgctggtt
ggtatcctgctctcgggaactgctgaccaat

>WCE_24-08-04_35
ccttttggaattggagtatttacccaatgcctaaatccccattg
tatcttggaagtaaataacttgtttctgatttcacaggctcatg
ggcaaaagggacttgccttgtctcagaagagacttgggac
tttggactttactggtaatgctggaaagaggtaatggactgtt
ggaaaaccattgttacattttgaaatgtgagaaatgtcaggc
ctctgagcccaagccaagccatcgcatcccctgtgac

>WCE_24-08-04_36
gctctgtcgcccaggttggggagcagtggtgtgattatggct
cactgcagcctcaacctccttggctcaagtgatcctcctccc
acctcagcctcacaagcagctgggactacaggcaagtgc
caacccttttgtatttttttgtagagatggggtttcaccatgttgc
ccaaactggtcttgaactcctagactcaagcgatcctcctgc
ctcag

>WCE_24-08-04_37
cctccttactgccccgtgcctttgccttgtgcttggaaaccagt
aaccccacacgtacagagcctacatatttgcacaaagaa
ctggaaaagctgctgcacttttcggagttgggggtgggcttg
gtttccacgttaatacactcccgggcggatatatggtgtcca
gactccccatgccgagcgctgcctataaatgaacactcgc
cacacactggagcctattcctgaggctgcagaaagtacca
gggcattttctccagtttaaagtttatcattaggcatccgctgta
gaagagctcagcacactgggctctgtgaaagtcaccttctg
aagtgcctggtgttgatggagacaaacggtacttggtttcctt
ttttaacacctgtccctcaacaaaccacattactgatgaggc
atctttctgatgtcagaagaaatgtcacttgttagagacagta
gtacagagaacaggactgtggagtggggagcactaactg
tgcttgggaggcagacagacttgggtttgagtcctgactga
gcccctgagcagcatgtgacttggggcaagtttcctcatctg
gaaaacagagataataatatctacccaacaaaagagag
gatgcagtgaacttgtggctaaaaaacgggctcttgtaac
ttccttctcctctttatcaaaccctccagcagcaaccctaggt

ggccaatcaagggatgccaaaaaacactcagcctctttaa
ttttgt

>WCE_24-08-04_39
tacttgggaggctgaggcaggagaatcgcttgaactcggg
gaggtagaggttgcagtgagttgagatagccccactgcac
tccagcctgggtgactctgtctcaaaaaaaaaaaaaaaat
agaatacagt