Analyse et modélisation de la stochasticité de l'expression génique dans des cellules eucaryotes

Gaël Kaneko

Numéro d'ordre 2013ISAL0099
Année 2013

Analyse et modélisation de la stochasticité de l'expression génique dans des cellules eucaryotes

Thèse présentée par
Gaël Kaneko

Devant
L'Institut National des Sciences Appliquées de Lyon

Pour obtenir
Le grade de docteur

Formation doctorale
Mathématiques et Informatique (InfoMaths)

Soutenance prévue le 26 Septembre 2013 devant le jury composé de :

Gilles Bernot	Professeur, Université de Nice Sophia Antipolis (examinateur)
Guillaume Beslon	Professeur, INSA de Lyon (directeur de thèse)
Olivier Gandrillon	Directeur de Recherche, CNRS, UCBL Lyon 1 (directeur de thèse)
Christophe Lavelle	Chargé de recherche, CNRS, MNHN (examinateur)
Andras Paldi	Professeur, École Pratique des Hautes Études (rapporteur)
Jean-Daniel Zucker	Directeur de recherche, IRD, UPMC Paris 6 (rapporteur)

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **http://www.edchimie-lyon.fr** Insa : R. GOURDON	M. Jean Marc LANCELIN Université de Lyon – Collège Doctoral Bât ESCPE 43 bd du 11 novembre 1918 69622 VILLEURBANNE Cedex Tél : 04.72.43 13 95 directeur@edchimie-lyon.fr
E.E.A.	**ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE** **http://edeea.ec-lyon.fr** Secrétariat : M.C. HAVGOUDOUKIAN eea@ec-lyon.fr	M. Gérard SCORLETTI Ecole Centrale de Lyon 36 avenue Guy de Collongue 69134 ECULLY Tél : 04.72.18 65 55 Fax : 04 78 43 37 17 Gerard.scorletti@ec-lyon.fr
E2M2	**EVOLUTION, ECOSYSTEME, MICROBIOLOGIE, MODELISATION** **http://e2m2.universite-lyon.fr** Insa : H. CHARLES	Mme Gudrun BORNETTE CNRS UMR 5023 LEHNA Université Claude Bernard Lyon 1 Bât Forel 43 bd du 11 novembre 1918 69622 VILLEURBANNE Cédex Tél : 06.07.53.89.13 e2m2@ univ-lyon1.fr
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTE** **http://www.ediss-lyon.fr** Sec : Samia VUILLERMOZ Insa : M. LAGARDE	M. Didier REVEL Hôpital Louis Pradel Bâtiment Central 28 Avenue Doyen Lépine 69677 BRON Tél : 04.72.68.49.09 Fax :04 72 68 49 16 Didier.revel@creatis.uni-lyon1.fr
INFOMATHS	**INFORMATIQUE ET MATHEMATIQUES** **http://infomaths.univ-lyon1.fr** Sec :Renée EL MELHEM	Mme Sylvie CALABRETTO Université Claude Bernard Lyon 1 INFOMATHS Bâtiment Braconnier 43 bd du 11 novembre 1918 69622 VILLEURBANNE Cedex Tél : 04.72. 44.82.94 Fax 04 72 43 16 87 infomaths@univ-lyon1.fr
Matériaux	**MATERIAUX DE LYON** **http://ed34.universite-lyon.fr** Secrétariat : M. LABOUNE PM : 71.70 –Fax : 87.12 Bat. Saint Exupéry Ed.materiaux@insa-lyon.fr	M. Jean-Yves BUFFIERE INSA de Lyon MATEIS Bâtiment Saint Exupéry 7 avenue Jean Capelle 69621 VILLEURBANNE Cedex Tél : 04.72.43 83 18 Fax 04 72 43 85 28 Jean-yves.buffiere@insa-lyon.fr
MEGA	**MECANIQUE, ENERGETIQUE, GENIE CIVIL, ACOUSTIQUE** **http://mega.ec-lyon.fr** Secrétariat : M. LABOUNE PM : 71.70 –Fax : 87.12 Bat. Saint Exupéry mega@insa-lyon.fr	M. Philippe BOISSE INSA de Lyon Laboratoire LAMCOS Bâtiment Jacquard 25 bis avenue Jean Capelle 69621 VILLEURBANNE Cedex Tél :04.72 .43.71.70 Fax : 04 72 43 72 37 Philippe.boisse@insa-lyon.fr
ScSo	**ScSo*** **http://recherche.univ-lyon2.fr/scso/** Sec : Viviane POLSINELLI Brigitte DUBOIS Insa : J.Y. TOUSSAINT	M. OBADIA Lionel Université Lyon 2 86 rue Pasteur 69365 LYON Cedex 07 Tél : 04.78.77.23.86 Fax : 04.37.28.04.48 Lionel.Obadia@univ-lyon2.fr

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Remerciements

En premier lieu, je souhaite remercier Andras Paldi et Jean-Daniel Zucker pour avoir accepté d'évaluer mon manuscrit de thèse et de me consacrer une partie de leur temps. Je remercie aussi Gilles Bernot et Christophe Lavelle pour avoir accepté de participer au jury de cette thèse et de nous avoir fait part du plaisir qu'ils auraient à venir.

Cette thèse n'aurait pas eu lieu sans la ténacité d'Olivier Gandrillon à porter des projets visant à étudier la stochasticité et à défendre des financements qui auraient pu, administrativement, ne pas nous être accordés. Par nos échanges, il m'a beaucoup appris en biologie et en méthodologie expérimentale mais aussi à ne jamais se cloitrer dans une vision obtuse des choses. Ses relectures de mon manuscrit ont été précieuses. J'ai apprécié le fait de travailler sous sa direction et l'en remercie.

Guillaume Beslon, mon autre directeur de thèse a toujours été là pour moi et ses conseils ont été déterminants. Si cette thèse a pu voir le jour, c'est en très grande partie grâce à lui. Ses multiples relectures de mon manuscrit m'impressionnent toujours. Je le remercie pour le temps qu'il m'a consacré à me former, à m'aider et pour les moments partagés, que ce soit en science ou autour d'une corde. Bien du chemin a été fait depuis qu'il m'a accueilli en stage de première année de master et c'est principalement grâce à lui.

Durant cette thèse, j'ai partagé ma vie avec les membres de plusieurs équipes, dans plusieurs Laboratoires.

Je remercie l'équipe BM2A, particulièrement Sandrine Gonin-Giraud toujours de bon conseil et Elodie Valin pour l'excellente qualité de son travail et sa bonne humeur. Je remercie aussi Anh Thu Lefebvre (anciennement dans l'équipe) pour ses discussions enflammées et ses macarons, ainsi qu'Ophélie Arnaud et Mimoun Maache avec qui il est toujours agréable de travailler. J'ai eu l'opportunité de co-encadrer le stage de Charles Rocabert ; je lui souhaite une thèse aussi riche en enseignements que l'a été la mienne.

Du côté informatique, j'ai eu l'opportunité de travailler au sein de Prisma puis Turing puis Combining et enfin Beagle. Je remercie leurs membres (anciens ou actuels) pour les discussions professionnelles ou personnelles, tout particulièrement Carole Knibbe, Hugues Berry, Jean-Francois Boulicaut, Christophe Rigotti, Serge Fenet, Céline Robardet et Fabrice Cêtre pour les permanents. Merci aussi à Bérénice Batut, Stephan Fischer, Jules Lallouette, Magali Vangkeosay, Bertrand Caré, Virginie Lefort et Yolanda Sanchez-Dehesa pour leur soutient et avec qui j'ai passé de très bons moments. En plus de son amitié, l'expertise d'Antoine Coulon m'a aussi été très précieuse.

Les bio-mathématiques et la bio-informatique allant souvent de paire ; l'équipe Dracula et l'équipe Beagle aussi. Merci à Fabien Crauste et Thomas Lepoutre pour leur soutien, particulièrement dans les moments d'intense fatigue.

Plus globalement, je remercie sincèrement le CGPhiMC pour avoir accepté en son sein un bio-informaticien pour qui la paillasse n'est que théories et le LIRIS pour permettre à des informaticiens de s'exprimer aussi dans des revue de biologie.

La vie entre deux laboratoires est parfois administrativement compliquée mais heureusement, Caroline Suter, Mabrouka Gheraissa, Caroline Ferri et Christine Bonnin m'ont facilité les choses. Merci à elles.

Durant ma thèse, j'ai pu enseigner au sein de différentes équipes pédagogiques. Elles étaient dynamiques et agréables, particulièrement grâce à Nadia Bennani, Nicolas Stouls, Mickaël Dardaillon et Sylvain Mousset.

Ma thèse s'inscrit directement dans un projet de plus grande envergure : le projet Stochagène. Il fait intervenir d'autres personnes que je souhaiterais aussi remercier pour leur enthousiasme et leur dynamisme sur ce sujet qui nous passionne. Je pense particulièrement à Guillaume Corre, Daniel Stockholm et Jean-Jacques Kupiec.

Je souhaite remercier la Région Rhône Alpes, l'ANR, l'INRIA et l'INSA de Lyon qui ont financé mes travaux ainsi que l'IN2P3 et Pascal Calvat pour leur aide précieuse.

José Viñuelas a été un pilier dans mes travaux. Je ne le remercierai jamais assez pour son amitié sans concession et le savoir faire qu'il m'a transmis. Que ce soit en science ou en dehors, sa rigueur est sans pareille. Toujours présent même dans les moments difficiles, il fait partie de ces gens moteurs (banc de puissance à l'appui) qui n'ont pas froid aux yeux et qui me sont chers.

David Parsons n'a pas travaillé directement avec moi mais ses remarques ont toujours été pertinentes. Toujours partant pour partager son temps ou pour aller faire des bulles, c'est un collègue mais surtout un grand ami. Je le remercie pour son aide et pour ses petites remarques piquantes toujours agréables. Nous nous reverrons dans le pédiluve.

J'aimerais aussi remercier mes amis qui m'ont soutenu et qui ont fait preuve de patience durant ma thèse, principalement Jeff, Emilie, Cédric, les Tulettiens, le gang des chauves souris des cascades, Chiara, Pablo, Audrey et Lolo (qui a en plus partagé le temps libre de sa moitié avec moi durant ces dernières années). Catherine, Pierre Marie, Soph, Vincent, Hélène et Aymeric, je vous remercie aussi pour les bons moments passés grâce à vous.

Je remercie ma famille pour son soutien inconditionnel, particulièrement mes parents sans qui je n'aurais pas pu arriver jusque là et ma soeur avec qui j'ai vécu une bonne partie de ma thèse et qui a réussi à me supporter.

Enfin, je remercie Aline, qui a vécu cette thèse avec moi au quotidien, l'a corrigée et qui a accepté de me partager avec mon travail. Elle a été mon bol d'air sans jamais que je ne lui en fasse la demande. Merci pour tous ces moments où même avec beaucoup de travail, j'ai réussi à penser à autre chose.

Résumés

Analyse et modélisation de la stochasticité de l'expression génique dans des cellules eucaryotes

Dans ce travail de thèse, nous avons étudié la variabilité (ou stochasticité) de l'expression des gènes en considérant que le signal stochastique que produit cette expression est porteur d'information quant au processus d'expression lui-même.

Cette stochasticité de l'expression génique peut être caractérisée par la variation observée du nombre de protéines produites soit entre différentes cellules isogéniques (portant le même génome) à un instant donné, soit au sein d'une même cellule au cours du temps.

Dans un premier temps, nous avons montré expérimentalement que le niveau de stochasticité de l'expression d'un gène change suivant son locus (sa position sur le génome). De plus, nous avons montré que, à locus constant, le niveau de stochasticité peut être influencé par des agents modificateurs globaux de l'état chromatinien.

Ensuite, nous avons analysé comment la dynamique chromatinienne peut influencer la stochasticité de l'expression génique d'un gène. Pour ce faire, nous avons utilisé une approche de modélisation et simulation que nous avons ensuite confrontée à des données biologiques. L'utilisation d'un modèle à deux états nous a permis de montrer que l'activité du promoteur est caractérisée par de longues périodes durant lesquelles la chromatine empêche la transcription, entrecoupées par de brèves périodes où la transcription est à nouveau rendue possible sous forme de « bursts » de forte intensité.

Pour finir, nous avons identifié, par des approches statistiques et par l'utilisation de bases de données génomiques, des éléments caractéristiques de la séquence génomique qui, lorsqu'ils sont présents dans le voisinage d'un gène, peuvent influer sur la stochasticité de celui-ci. Nous avons en particulier montré que, lorsque le gène rapporteur est inséré à proximité d'un autre gène, sa stochasticité augmente de manière significative.

Ce travail nous a permis de mettre en évidence un lien entre la dynamique chromatinienne, l'environnement génomique et la stochasticité de l'expression génique. Ce lien offre à la cellule des perspectives évolutives en lui permettant de réguler cette stochasticité, ouvrant ainsi la porte à la sélection d'un niveau approprié de variabilité.

Mots clefs : Modélisation, stochasticité, expression génique, chromatine, locus, environnement génomique.

Analysis and modeling of gene expression stochasticity in eukaryotic cells

During my PhD, we have studied the variability (or stochasticity) of gene expression assuming that the stochastic signal it produces carries information about the process of gene expression itself. The stochasticity of gene expression can be characterized by the observed variation in the number of proteins produced either by different isogenic cells (cells that have the same genome) at a given time or within a single cell over time.
First, we showed experimentally that the level of stochasticity of a gene changes according to its locus (its position on the genome). We have also shown that, for a given locus, the level of stochasticity could be influenced by global chromatin-state modifier agents.
Then, we analyzed how the chromatin dynamics can influence the stochasticity of gene expression. This analysis was conducted by using a modeling and simulation approach, the results of which being in turn compared to biological data. Using a two-states model allowed me to show that the activity of a promoter is characterized by long periods during which the chromatin prevents transcription, interspersed by brief periods when transcription can occur in the form of intense bursts.
Finally, we identified characteristic genomic elements that, when in the neighbourhood of a gene, may influence its level of stochasticity. In particular, we have shown that when the reporter gene is integrated close to a neighbour gene, its stochasticity is significantly increased.
This work allowed me to unravel a link between the chromatin dynamics, the genomic environment and the stochasticity of gene expression. This link confers evolutionary perspectives to the cell by allowing it to regulate stochasticity, which allows for the selection of an appropriate level of stochasticity.

Key words : Modelisation, stochasticity, gene expression, chromatine, locus, genomic environment.

Table des matières

Table des figures

Chapitre I

Introduction

Le mécanisme de production des protéines à partir des gènes est un processus central chez tous les êtres vivants. Selon le "dogme de la biologie moléculaire", c'est en effet par ce mécanisme que les cellules biologiques produisent le matériel biologique nécessaire à leur fonctionnement à partir de l'information présente dans la molécule d'ADN. Cependant ce dogme n'est que l'idéalisation d'un processus complexe dans lequel interviennent de nombreux composés moléculaires – pour la plupart des protéines produites par le même mécanisme – et qui met en jeu de nombreux phénomènes biologiques. Cette complexité entraîne de nombreuses possibilités de régulation de la chaîne de transcription-traduction permettant aux cellules de maîtriser très finement la concentration des protéines. On sait depuis les années soixantes que cette régulation permet en particulier aux cellules de réguler le processus de production de protéines en l'adaptant aux conditions environnementales. Ainsi, cette production n'est pas identique pour tous les gènes ni, pour un même gène, stable au cours du temps (Figure I.1). Pourtant, dans sa description géné-

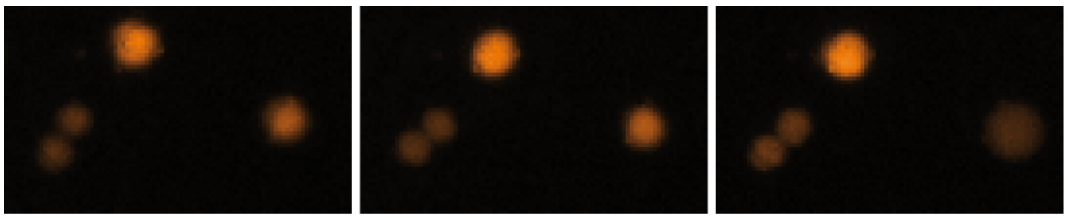

FIGURE I.1 – Variabilité du nombre de protéines fluorescentes dans le temps. Ici des cellules progénitrices aviaires de poulet ont été transfectées avec un rapporteur fluorescent hKO. En observant un même champ au microscope à trois instants différents (ici à $t = 0$h, 2h et 7h) on constate que la fluorescence des cellules varie au cours du temps.

rale, le dogme est longtemps resté essentiellement déterministe : pour un même gène d'une même cellule placée dans un même environnement, la production de protéines serait essentiellement constante. C'est cette idée même de déterminisme qui, depuis quelques années, est battue en brêche, d'abord dans le champ des idées (Kupiec, 1997) puis, au tournant des années 2000, expérimentalement (Levsky *et al.*, 2002; Elowitz *et al.*, 2002; De Krom *et al.*, 2002).
Depuis sa mise en évidence, cette "variabilité de l'expression des gènes" connaît un intérêt croissant même si elle a d'abord été considérée négativement par une grande partie de

la communauté (le terme le plus souvent employé et qui reste encore d'actualité pour caractériser cette variabilité est le "bruit", terme clairement connoté négativement et qui suggère une part d'incontrôlable, de méconnaissance, d'inqualifiable, d'inquantifiable ou encore de perturbateur dans un processus qu'il parasiterait). En effet, l'existence d'un "bruit" dans l'expression des gènes remet directement en cause l'idée d'un déterminisme génétique du phénotype de l'individu. Alors que l'idée même d'un fonctionnement non-déterministe a été énoncée très tôt dans l'histoire de la biologie cellulaire (Novick et Weiner, 1957; Spudich et Koshland, 1976; Rigney et Schieve, 1977; Berg, 1978; Kupiec, 1983), c'est probablement dans cette remise en cause du génotype d'un individu comme son essence, biologiquement/mécanistiquement parlant, qu'il faut chercher la résistance à la diffusion de cette vision non-déterministe de la production des protéines.

Une grande partie de la biologie moléculaire moderne a ainsi été guidée par l'idée qu'en étudiant le génome on parviendrait à percer le mystère du fonctionnement des cellules, quitte à ignorer à quel point cette vision rendait plusieurs processus difficiles à expliquer : comment en effet expliquer les différences entre individus génétiquement identiques si le "programme" de développement est totalement déterministe ?

Suite aux premières évidences expérimentales (Berg, 1978; Levsky *et al.*, 2002; Elowitz *et al.*, 2002; De Krom *et al.*, 2002), il est progressivement devenu de plus en plus clair qu'une partie des différences inter-individuelles ne pourrait pas être expliquée à l'échelle génomique mais trouverait peut être une part de ses explications dans les mécanismes non-déterministes de transcription-traduction (Figure I.2). Le "bruit" a alors progressivement cessé d'être un simple élément perturbateur dans le sens péjoratif du terme pour devenir un processus digne d'intérêt. La "stochasticité de l'expression génique" (et non plus le "bruit dans l'expression des gènes"), ses causes et ses conséquences, ont alors gagné leurs lettres de noblesse et ont commencé à être étudiés pour eux-mêmes.

Dans ce travail de thèse, nous étudierons donc la variabilité de l'expression des gènes en elle-même mais aussi en considérant que le signal stochastique qu'elle produit est porteur d'information quant au processus d'expression. Ainsi, non seulement nous proposerons une quantification de la variabilité à partir de mesures expérimentales mais nous essayerons de coupler cette quantification à une démarche de modélisation pour permettre une meilleur compréhension des mécanismes moléculaires de l'expression génique. Plutôt que d'observer directement le processus de transcription pour comprendre l'origine de sa variabilité, nous allons procéder suivant une approche inverse en observant la variabilité, témoin du fonctionnement du système, pour en inférer l'origine et en déduire des propriétés du mécanisme de transcription lui-même. Cette méthode d'observation indirecte du système nous permet de palier les limites des technologies d'observation directe de l'objet biologique étudié. En effet, il est aujourd'hui impossible d'observer directement l'ensemble des processus moléculaires impliqués dans la transcription des gènes. Cependant, cette approche inverse a elle aussi ses contraintes. En particulier, pour pouvoir espérer passer des conséquences de la variabilité aux causes de la variabilité, il est nécessaire de pouvoir observer un même système dans différents régimes de fonctionnement. Pour cela, une approche classique sera de perturber le système et d'observer les conséquences de ces perturbations sur la production de protéines (ou, plus exactement, ici, sur la variabilité de la production de protéines). La relation entre le type de perturbation et l'effet de la perturbation pourra alors nous permettre d'identifier la source principale de variabilité puis, d'en comprendre les modes de fonctionnement. Nous utiliserons pour cela des "outils moléculaires" (des

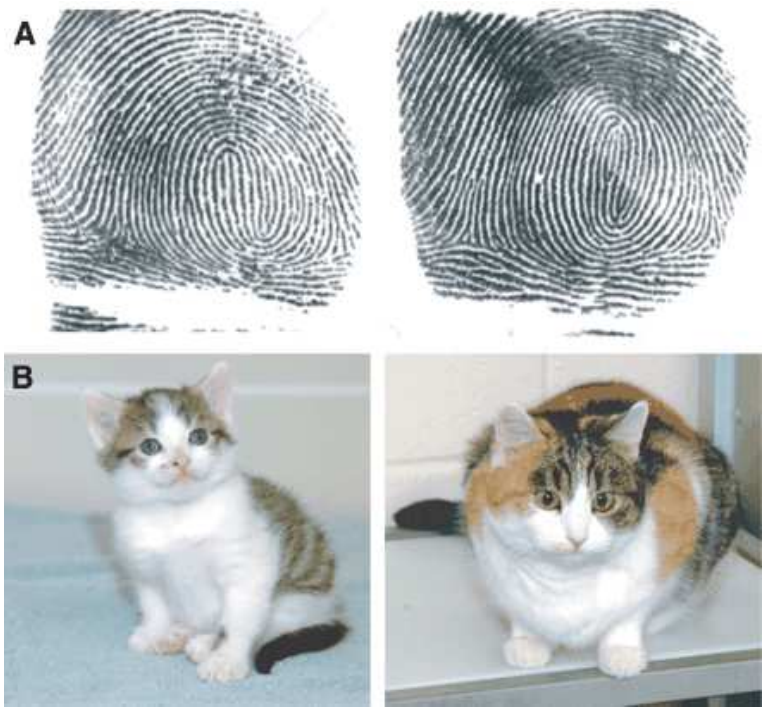

Figure I.2 – Différences phénotypiques entre individus génétiquement identiques. (A) Les empreintes digitales de deux vrais jumeaux présentent de nombreuses différences. (B) La robe d'un chaton et de son clone biologique est différente alors que le matériel génétique est strictement identique. Images à partir de (Raser et O'Shea, 2005)

drogues) permettant de modifier l'expression des gènes.
Cette rapide présentation méthodologique résume dans les grandes lignes notre travail de thèse. Dans un premier temps (chapitre II), nous avons étudié l'influence de modificateurs chromatiniens sur la variabilité de l'expression des gènes (ou, plus exactement, sur la variabilité de l'expression d'un gène rapporteur inséré dans le génome). Cela nous a permis de mettre en évidence deux propriétés importantes de cette variabilité : d'une part elle dépend fortement de la position du gène rapporteur sur le chromosome, d'autre part elle est fortement modifiée par l'action des modificateurs chromatiniens. Dans un second temps, ce constat nous a conduit à mettre en place deux analyses complémentaires :

- rechercher les propriétés dynamiques de la chromatine susceptibles d'expliquer ces observations (chapitre III),

- rechercher les propriétés statiques de la séquence chromosomique susceptibles d'expliquer ces observations (chapitre IV).

Ces deux analyses ont été conduites en combinant étroitement des approches de biologie moléculaire[1], de modélisation déterministe et stochastique, et d'analyses bioinformatiques.

[1]Toutes les expériences de biologie moléculaire ont été conduites par José Viñuélas, alors post-doctorant dans l'équipe BM2A du CGϕMC, Elodie Valin, technicienne, Valérie Morin, ingénieur d'études et Olivier Gandrillon, directeur de recherche.

Dans la suite de ce chapitre introductif, nous allons présenter les grandes lignes des différents concepts et outils utilisés pour étudier la stochasticité en insistant particulièrement sur ceux mis en œuvre dans le cadre de ce travail.

1 L'expression génique et ses mécanismes

Une cellule est principalement constituée de protéines qui forment sa structure[1] et lui permettent d'avoir une activité métabolique – donc de "vivre". Dans la plupart des cas, la cellule produit elle-même ses protéines. Ces dernières composent son ossature (le cytosquelette), produisent l'énergie nécessaire à son fonctionnement (au sein des mitochondries), organisent l'information génétique permettant de construire ces mêmes protéines (structure de la chromatine), réparent cette même information si elle présente des lésions ou la traduisent pour produire d'autres protéines ... Même si notre étude questionne le dogme de la biologie moléculaire, il n'en reste pas moins que, pour comprendre comment la cellule produit ses protéines, il est nécessaire de commencer par comprendre comment l'information permettant de construire ces protéines est stockée et traduite.

1.1 Le support de l'information génétique

À l'échelle cellulaire les chromosomes sont le support de l'information génétique. Ils sont rassemblés dans le noyau de la cellule (dans le cas des cellules eucaryotes. Pour les cellules procaryotes le chromosome, généralement unique, est organisé au sein d'une structure appelée nucléoïde). La substance de base des chromosomes est appelée chromatine. Cette dernière est formée du ruban d'ADN (Acide DésoxyriboNucléique[2]) enroulé autour des histones. Cette structure d'enroulement forme ainsi les nucléosomes, eux-même enroulés pour former la fibre de chromatine (figure I.3). Le niveau d'enroulement de l'ADN, et donc

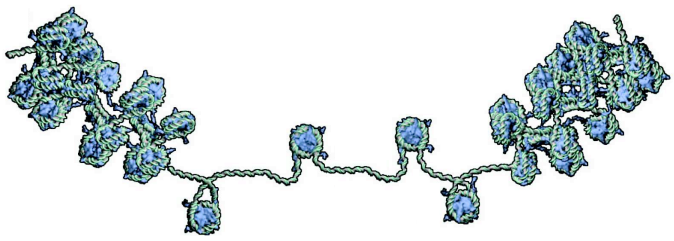

Figure I.3 – États de la chromatine. La chromatine (nucléosomes en bleu et ADN en vert peut changer d'état. Sur les extrémités droite et gauche de la figure, la chromatine est sous forme compactée. Une forme plus décompactée est visible au centre de la figure. Il existe différents niveaux de compaction (Source du schéma original : (Lavelle, 2009)).

[1] Les protéines ne sont pas les seuls constituants de la structure cellulaire. Les lipides et les glucides y jouent aussi un rôle important.

[2] Nous considérons ici connue la structure en double brins complémentaires structurée en double hélice (Crick et Watson, 1953). De même, nous ne reviendrons pas sur le principe de codage de l'information génique par l'ordre des nucléotides le long de la séquence d'ADN, considéré comme le support de l'information génétique à l'échelle moléculaire.

la structure de la chromatine, varient au cours du temps entre deux principaux états :

- Compacté : l'ADN est enroulé autour des nucléosomes qui sont eux-même compactés les uns contre les autres formant ainsi une structure de 30 nanomètres environ. Cet état de la chromatine est représenté sur la figure I.3.
- Décompacté : l'ADN est enroulé autour des histones qui ne sont plus les uns contre les autres et laissent ainsi plus d'ADN "nu" entre eux. Dans cet état l'ADN est moins contraint et les protéines de la machinerie transcriptionnelle ont plus de latitude pour accéder aux gènes.

1.2 L'expression génique

Comme nous l'avons vu en introduction, l'expression génique est le processus qui permet à un gène d'être traduit en protéines. Ce processus est basé sur un ensemble de mécanismes moléculaires extrêmement complexes (bien que souvent succinctement décrit sous la forme d'une chaîne de "transcription-traduction", voir Figure I.4) parmi lesquels on

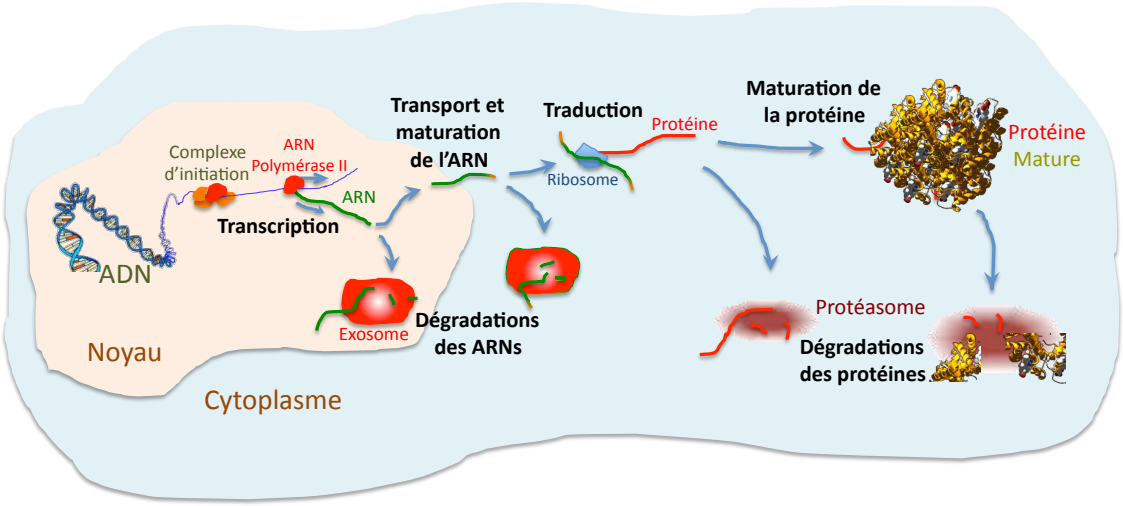

FIGURE I.4 – De l'ADN aux protéines, les principales étapes de l'expression génique et de la dégradation des ARNs et protéines.

peut distinguer cinq étapes principales[1] :

Initiation de la transcription Le processus d'initiation correspond au recrutement du complexe d'initiation de la transcription sur le promoteur du gène à transcrire. Ce complexe d'initiation est composé de protéines appelées facteurs de transcription (FT) et de l'ARN polymérase II. Le promoteur, quant à lui, comprend des zones appelées TFBS ("transcription factor binding sites") qui sont chacune spécifique de certains FT. En premier lieu, des facteurs de transcription vont se fixer sur les TFBS du promoteur du gène à transcrire. Cependant, l'évènement principal du processus de recrutement du complexe d'initiation est la fixation de l'ARN polymérase II

[1]Nous considérons ici le processus d'expression des gènes dans les cellules eucaryotes.

sur le promoteur. Les FT fixés facilitent cet évènement. Il est important de noter que tout ce processus de recrutement dépend de la concentration de chacune des protéines à recruter dans le noyau. Modifier ces concentrations est un des moyens de réguler l'expression génique. Une fois le complexe d'initiation en place, la polymérase commence la séparation des brins de la double hélice d'ADN et la transcription peut commencer.

Transcription de l'ARN messager L'ARN polymérase II va alors se déplacer le long de la double hélice d'ADN en la déroulant pour parcourir la séquence à transcrire (élongation). L'enzyme va ainsi générer un brin d'ARN pré-messager complémentaire de la séquence génique parcourue. Pour cette étape aussi, certaines concentrations vont être limitantes : celles des nucléotides qui vont composer le brin d'ARN en formation. La transcription se termine quand la polymérase rencontre une séquence terminatrice marquant la fin de la sequence transcrite. Elle arrête alors la transcription de l'ARN pré-messager qu'elle était en train de produire.

Export et maturation de l'ARN messager L'ARN pré-messager ainsi formé va migrer du noyau vers le cytoplasme et subir une maturation. Durant cette maturation, certaines séquences "non-codantes" (les introns – séquences qui sont présentes sur l'ARN mais qui ne seront pas traduites en protéines) sont supprimées. De plus, les extrémités de l'ARN vont être modifiées par des enzymes. Ces modifications vont faciliter ensuite la traduction, protéger l'ARN contre les enzymes de dégradation et faciliter l'export de l'ARN vers le cytoplasme. Sans ces changements, l'expression génique serait moins efficace. C'est ensuite dans le cytoplasme que le processus de traduction a lieu.

Traduction de l'ARN messager en protéine Pour cette étape, un ribosome va parcourir l'ARN messager et le traduire en une séquence d'acides aminés qui après maturation, formera une protéine fonctionnelle. Pour se fixer facilement sur l'ARN, le ribosome a besoin de facteurs d'initiation qui vont, avec lui, former un complexe d'initiation de la traduction. Une fois ce complexe recruté et fixé sur l'ARN, la traduction commence et les acides aminés sont recrutés selon les besoins, au fur et à mesure que le ribosome parcours l'ARN messager.

Maturation et repliement de la protéine Une fois la séquence d'acides aminés générée, elle doit subir différentes modifications pour devenir une protéine fonctionnelle. Elle doit principalement prendre une conformation en trois dimensions qui lui est spécifique. Pour ce faire, elle doit subir des modifications sur certains de ses acides aminés qui, dans leur conformation initiale ne permettent pas à la protéine de prendre la bonne conformation 3D. L'établissement de ponts disulfures entre certaines de ses sous-séquences d'acides aminés est aussi nécessaire pour contraindre la séquence totale à rester dans la bonne conformation.

Outre ces cinq processus participant à la production de nouvelles protéines fonctionnelles dans la cellule, il est nécessaire, pour spécifier totalement la dynamique du mécanisme d'expression des gènes, de rajouter les deux processus de dégradation des ARNm et des protéines. En effet, la quantité de protéines présentes à un instant donné dans la cellule

est le résultat d'un équilibre dynamique entre les processus de production (transcription puis traduction) et les processus de dégradation :

Dégradation des ARNm Les ribonucléases sont des enzymes qui dégradent l'ARN. Un complexe connu de ribonucléases est l'exosome. Il est présent en plusieurs copies dans le cytoplasme mais aussi dans le noyau. Il dégrade principalement l'ARN en partant de ses extrémités mais peut aussi cliver un ARN. Un ARN mature sera plus difficilement dégradable par ses extrémités qu'un ARN non mature car ces dernières ont été modifiées ce qui limite ce processus de dégradation. Néanmoins, la forte concentration de ribonucléases présentes dans les cellules induit une espérance de "vie" très faible pour les ARNs. Ainsi, un ARN peut être traduit plusieurs fois mais peut aussi être dégradé avant même d'avoir atteint le cytoplasme.

Dégradation des protéines À l'image des ribonucléases et des exosomes pour les ARNs, les protéines sont dégradées par les protéases et les protéasomes. Ces derniers sont des complexes enzymatiques qui dégradent les protéines mal repliées ou dénaturées en les clivant. Ce sont d'autres enzymes qui sont chargées de marquer les protéines à dégrader pour que les protéasomes les reconnaissent. Malgré ces complexes enzymatiques, l'espérance de vie des protéines est en général plus longue que celle des ARNs.

Enfin, même si la plupart des études de "transcriptomique[1]" considère l'expression génique comme un processus quasi-instantané, chacune des étapes présentées ci-dessus possède sa propre constante de temps. Qui plus est, celle-ci peut être extrêmement variable d'un type cellulaire à un autre, voire d'un type moléculaire à un autre. Ainsi, la durée de vie des protéines peut varier de plusieurs ordres de grandeur dans un même type cellulaire. A titre d'exemple, pour les rapporteurs fluorescents utilisés dans notre études, la demi-vie est d'environ 65h pour la protéine mCherry contre seulement 1h45 pour la protéine YFP déstabilisée[2] (dans les cellules 6C2[3]). Dans un autre type cellulaire, les rétinoblastes humains, cette même protéine YFP déstabilisée a une demi-vie d'environ 6h30 (Corre, 2012) soit plus de quatre fois plus que dans les 6C2. Enfin, toujours dans les rétinoblastes humains, on peut observer que la durée de la demi-vie des ARNs YFP est d'environ 3h contre seulement 1h30 pour les ARNs déstabilisés[4] de la même protéine. On a ici des variations d'efficacité à la fois au niveau des systèmes de dégradation des ARNs et des protéines suivant le type cellulaire ou moléculaire observé.
Avec le développement, dans les années 90 puis 2000, des approches haut-débit permettant de mesurer la quantité moyenne d'ARNm dans des populations de cellules (puces à ADN, SAGE, RNAseq ; (Metzker, 2010)), l'étude du processus d'expression génique s'est très rapidement développée et a ouvert une nouvelle fenêtre sur le fonctionnement des cellules. Ainsi, des études telles que l'analyse des co-expressions d'ensembles de gènes (gènes exprimés dans les mêmes conditions ou au cours d'une même phase du cycle cellulaire) ont permis d'inférer les premiers "réseaux génétiques" et très vite des méthodes

[1]La transcriptomique est l'étude, par des méthodes haut-débit, du processus d'expression des gènes, essentiellement considéré sous l'angle de la transcription.

[2]déstabilisée par la sequence PEST ajouté en queue de la sequence du gène d'YFP

[3]cellules d'origine aviaire qui seront décrites dans le chapitre II

[4]déstabilisés par la séquence ARE flanquée à la fin des ARNs

d'inférence dédiées ont commencé à apparaître (Friedman *et al.*, 2000). Ces techniques ont été par exemple employées pour identifier les gènes cibles de certains processus pathologiques comme le cancer (Wang, 2013; Waghray *et al.*, 2001; Claverie, 2001). Cependant, l'immense majorité de ces études se sont basées sur l'analyse des moyennes d'expression, entre autres parce que les principaux outils de mesure de l'époque ne permettent de mesurer que des moyennes sur de très grandes populations de cellules. En accord avec ces observations moyennes, mais aussi probablement du fait du caractère très bruité des premiers outils de mesure, la transcriptomique s'est développée dans le cadre d'une vision très déterministe de l'expression génique, les fortes variations constatées dans les observations étant considérées comme du bruit de mesure à supprimer statistiquement. Ce n'est qu'après les années 2000 que de nouvelles méthodes expérimentales, en permettant de mesurer l'expression des gènes[1] dans des cellules uniques, ont permis de montrer que la variabilité observée jusqu'alors n'était pas due à un bruit de mesure mais qu'elle comportait une importante part de variabilité intrinsèque au système biologique : la stochasticité de l'expression génique.

2 La stochasticité de l'expression génique

2.1 Introduction

La stochasticité de l'expression génique vient s'opposer à l'idée que l'expression d'un gène est stable au cours du temps, même dans un environnement constant. On regroupe donc derrière ce terme l'ensemble des mécanismes susceptibles de faire varier l'expression d'un gène au cours du temps dans une cellule, qu'il s'agisse de facteurs dits "extrinsèques" (c'est-à-dire extérieur au gène proprement dit mais influençant plus ou moins directement sa transcription) ou "intrinsèques" (c'est-à-dire propre au gène lui-même[2]).
La méthode la plus simple pour observer la stochasticité de l'expression génique consiste à cultiver des cellules isogéniques dans un environnement homogène. L'observation simultanée d'une population de cellules à un même instant t dans le champ d'un microscope permet de mesurer les variations de concentration de protéines et donc la stochasticité de l'expression (voir figures I.5 et I.6). La stochasticité de l'expression génique peut aussi être mesurée dans une même cellule unique, à condition de pouvoir suivre cette cellule au cours du temps (Figure I.1).
Les deux premières études de cette dernière décennie qui ont largement contribué, ensuite, à la richesse des études sur le bruit de l'expression des gènes sont celle de Levsky *et al.*

[1] En réalité, dans un premier temps, ce sont des concentration de protéines qui ont été mesurées. Le lien avec la variabilité de l'expression des gènes n'a ainsi été établi qu'indirectement.

[2] La distinction entre stochasticité extrinsèque et stochasticité intrinsèque a été proposée dès les premières observations (Elowitz *et al.*, 2002; Hilfinger et Paulsson, 2011) mais elle ne peut être comprise que par rapport à un système d'intérêt donné pour lequel les sources internes de stochasticité sont dites intrinsèques tandis que les sources externes sont dites extrinsèques. Dans la mesure ou les frontières d'un système biologique sont souvent floues (le facteur de transcription et/ou le promoteur font-ils partie du gène ?), cette distinction apporte peu d'information en pratique. De plus, les études utilisant cette distinction ((Elowitz *et al.*, 2002) par exemple) utilisent deux rapporteurs fluorescents différents dans une même cellule. Ils nomment alors "extrinsèques" les différences d'expression corrélées entre les deux rapporteurs et "intrinsèques", celles décorrélées. Ici, nous n'utiliserons qu'un rapporteur. C'est pourquoi cette distinction ne sera pas utilisée dans ce travail.

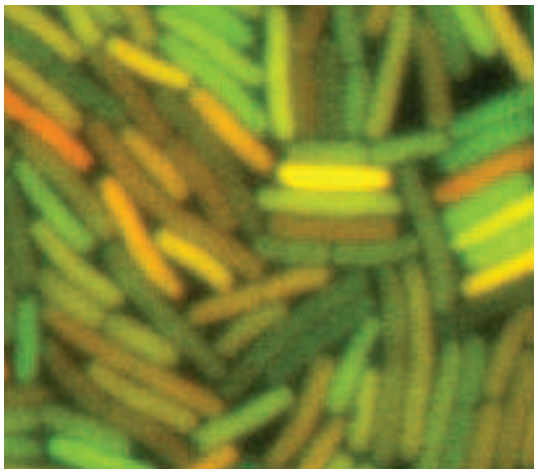

Figure I.5 – Variabilité du nombre de protéines dans des cellules clonales. Ici une population bactérienne a été transfectée par deux gènes rapporteurs fluorescents (CFP et YFP, réprésentés ici en fausses couleurs, respectivement en vert et en rouge). En observant la fluorescence au sein d'une population clonale, on observe que les deux protéines ne sont pas à quantité égale dans chacune des cellules observées : il y a donc une différence de production protéique entre des cellules identiques génétiquement. Image reproduite à partir de (Elowitz *et al.*, 2002).

(2002) et surtout celle d'Elowitz *et al.* (2002). L'équipe de Levsky a démontré la présence de stochasticité dans des fibroblastes humains, et ce en analysant 11 gènes différents et leurs transcrits. Ils ont démontré que la transcription y était soumise à variabilité. Elowitz et ses collaborateurs ont démontré qu'il existe bel et bien une variabilité de l'expression des gènes à génome et environnement constants. Pour ce faire, ils ont transfecté deux gènes dans des bactéries : l'un codant pour une protéine verte et l'autre pour une rouge. Cette souche de bactérie est donc capable d'exprimer chacune de ces deux protéines. Or, alors qu'elles sont isogéniques et qu'elles sont placées dans un même environnement, l'expression de chacune de ces protéines diffère d'une bactérie à l'autre (Figure I.5).

Ultérieurement, en étudiant le "bacteriophage lambda promoter" dans *Escherichia coli* à l'aide de protéines fluorescentes, Rosenfeld *et al.* (2005) ont observé que la protéine YFP était répartie selon une loi binomiale entre les deux cellules filles. Sachant cela et grâce à un modèle simple à un paramètre, ils arrivent à retrouver la distribution des protéines dans un groupe de cellules observées. Toujours chez *Escherichia coli* et à l'aide de protéines fluorescentes, Golding *et al.* (2005) déterminent le nombre d'ARN dans des cellules vivantes. Ils constatent aussi une répartition binomiale entre deux cellules filles mais cette fois-ci, pour les ARN.

La stochasticité de l'expression génique a ainsi été observée dans un grand nombre d'expériences et pour un très grand nombre de types cellulaires. Pourtant ce phénomène biologique est encore très mal compris. Ainsi, les sources de stochasticité sont mal explicitées. De même si, dans certains cas, des hypothèses plausibles ont pu être formulées, dans le cas général les conséquences de la stochasticité de l'expression des gènes sur la vie cellulaire ou, dans le cas d'organismes multicellulaires, sur le devenir des organismes sont encore très mal connues. Au vu du caractère apparemment universel de la stochasticité de

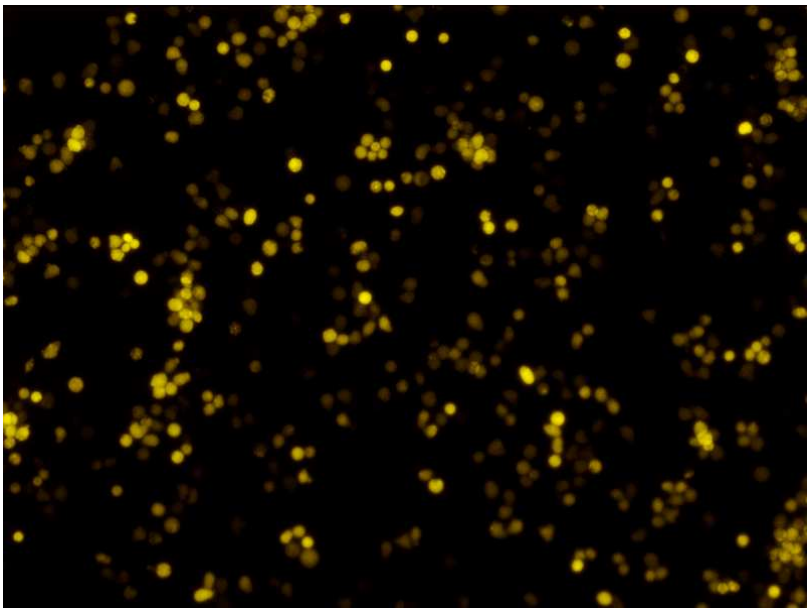

Figure I.6 – Stochasticité intercellulaire. Des cellules progénitrices aviaires isogéniques de type T2EC (cellules de type erythroblastes aviaires ; (Gandrillon *et al.*, 1999)), transfectées par un rapporteur fluorescent hKO et placées dans un environnement homogène, ne présentent pas le même niveau de fluorescence. Le niveau de fluorescence caractérise la concentration intracellulaire du rapporteur fluorescent et donc, indirectement, l'activité transcriptionnelle antérieure à la mesure.

l'expression des gènes, on peut se demander si cette variabilité confère un avantage sélectif aux organismes mais là encore, en dehors de quelques situations simples (généralement dans le cas d'organismes procaryotes), un tel avantage n'a jamais été clairement mis en évidence.

En dehors de ces questions purement biologiques, l'analyse de la stochasticité de l'expression génique vue comme un signal est elle-même balbutiante, entre autre du fait des difficultés expérimentales pour mesurer la variabilité d'une même cellule sur une longue période. Ainsi, on ne sait pas aujourd'hui si le processus est markovien ou non (du moins dans ses grandes lignes), ni s'il est ergodique[1]. À vrai dire, la simple question de la quantification précise de la stochasticité de l'expression génique (quelles sont, par exemple, les constantes de temps de la variabilité de l'expression génique intracellulaire ou la taille minimale des populations à mesurer pour estimer correctement la variabilité intercellulaire) est aujourd'hui encore très largement ouverte.

Dans la suite de cette section, nous aborderons successivement ces différentes questions en présentant les principales causes de la stochasticité de l'expression génique. Nous aborderons ensuite la question du ou des rôle(s) potentiel(s) de la stochasticité de l'expression

[1]On pourrait définir basiquement l'ergodicité de l'expression génique par le fait que la distribution de probabilité des niveaux d'expression d'un gène, pris au cours du temps dans une cellule C, est identique à la distribution de probabilité des niveaux d'expression de ce même gène observé à un instant t unique dans toute une population de cellules génétiquement identiques à C.

génique avant de présenter les principaux outils d'observation et les indicateurs de stochasticité les plus couramment utilisés. Dans un deuxième temps (section 3) nous présenterons les grandes approches de la modélisation de l'expression génique.

2.2 Causes biologiques de la stochasticité de l'expression des gènes

Dans une population de cellules, la variabilité phénotypique observée peut être due à de nombreux facteurs. Elle peut en effet être due à des variations environnementales, à des modifications de la séquence génétique au sein d'une sous-population, à des phénomènes d'auto-organisation liés à des mécanismes de communication intercellulaire ou même à de simples mécanismes de saturation spatiale ou de compétition pour les ressources. Même si, dans toutes ces situations, elles peuvent jouer un rôle important, ces variations phénotypiques, qui peuvent se traduire par d'importantes différences de concentrations protéïques, ne sont pas directement dues à la stochasticité de l'expression génique. Nous considérerons ici comme source potentielle de stochasticité de l'expression, tout phénomène susceptible d'induire une variabilité de la concentration protéïque au sein d'une population de cellules isogéniques placées dans un même environnement et supposées isolées les unes des autres[1]. Même avec cette définition stricte et restrictive, les sources potentielles de variabilité de l'expression génique sont nombreuses et peuvent se manifester à toutes les étapes du processus d'expression des gènes.

2.2.1 Stochasticité des réactions biochimiques

La cause la plus évidente de stochasticité de l'expression est la nature aléatoire des réactions chimiques (Raser et O'Shea, 2005). Ce "bruit" est principalement dû au caractère aléatoire de la présence/absence des réactants chimiques en même temps à un endroit donné. En effet, les réactions chimiques sont liées à la présence simultanée des différents composants de la réaction. Or, si certaines molécules chimiques sont présentes en grand nombre dans la cellule et peuvent donc, en première approximation, être considérées comme réparties de façon homogène dans l'espace cellulaire (ou dans l'espace nucléaire), cela n'est pas le cas pour un très grand nombre d'espèces moléculaires. En première approximation, la présence ou l'absence d'une molécule donnée à un endroit donné correspond donc bien à un processus stochastique. Cette source de stochasticité est probablement une des plus importante mais son mode d'action est relativement simple : plus faible est le nombre d'éléments appartenant à une espèce moléculaire donnée, plus importante est la stochasticité du processus auquel elle participe.
Les choses se compliquent notablement si on commence à prendre en compte les propriétés de déplacement des molécules. En effet, même si le transport moléculaire à l'intérieur du cytoplasme et du noyau cellulaire apparaît comme étonnamment efficace, il n'est reste pas moins qu'une même molécule ne peut pas s'éloigner infiniment vite d'un point donné, ni

[1]Cette dernière condition pose plusieurs problèmes cruciaux. En effet, des cellules isogéniques – donc issues d'un même clone – ne peuvent jamais être totalement indépendantes puisque, à minima, elles partagent une histoire commune. En outre, des différences intercellulaires peuvent être induites par le processus de division cellulaire lui-même (par exemple au cours de l'établissement de la colonie) qui n'est jamais totalement symétrique (Huh et Paulsson, 2011a). De même, si les cellules sont indépendantes, elles ne sont donc pas synchronisées et la présence de cellules identiques mais à des étapes différentes de leur cycle de vie peut être interprétée à tort comme une variation stochastique.

atteindre un autre point instantanément. En conséquence, une fois deux réactants proches, ils peuvent, suivant la nature de la réaction, leur affinité et la nature de leur déplacement, déclancher une bouffée (un "burst") d'activité par un phénomène de capture-recapture (Van Zon *et al.*, 2006). Or, ces phénomènes de capture-recapture sont très dépendants des propriétés de la diffusion (diffusion 1D/2D/3D, diffusion contrainte ou non, ...). En outre, l'encombrement cellulaire et les collisions entre molécules sont susceptibles de perturber fortement la diffusion (celles-ci étant d'autant plus contraintes qu'elles sont grosses), ce qui peut accentuer la variabilité liée à la diffusion de molécules présentes en faible concentration (Shahrezaei *et al.*, 2008).
La contribution de la stochasticité intrinsèque des réactions biochimiques à la stochasticité de l'expression génique peut se manifester à toutes les étapes de la chaîne de transcription-traduction. En effet, de très nombreuses réactions sont nécessaires à l'aboutissement du mécanisme de production de protéines, depuis le recrutement des facteurs de transcription par le promoteur du gène jusqu'à la présence de protéines chaperonnes impliquées dans le processus de repliement de la protéine en passant par le recrutement de l'ARN polymérase, des ARNt ou des ribosomes. En outre, ces différentes étapes vont induire une stochasticité plus ou moins importante suivant leur organisation spatiale et/ou la concentration des réactants. Cependant, il est aujourd'hui communément admis que la principale source de stochasticité de l'expression des gènes provient de la phase de transcription – plus précisément de la phase d'initiation de la transcription (Becskei *et al.*, 2005) – et que celle-ci aurait plus d'impact que les autres processus prenant part à l'expression génique.

2.2.2 Contribution des réseaux génétiques

Dans une cellule donnée, la trancription d'un gène n'est pas indépendante de la transcription des autres gènes puisque ceux-ci sont susceptibles, par leur produits finaux (protéines), ou intermédiaires (ARNm), d'en perturber la transcription. Or, suivant que cette perturbation corresponde à une activation ou à une inhibition, la contribution à la stochasticité de l'expression du gène cible peut être positive (augmentation de la stochasticité) ou négative (diminution de la stochasticité). Ces contributions peuvent prendre de multiples aspects, depuis l'auto-régulation d'un gène par son propre produit, ce qui peut contribuer à fortement réguler son activité stochastique (Lim *et al.*, 2013; Voliotis et Bowsher, 2012), jusqu'à l'inactivation de la traduction par RNA-interférence ou la présence de motifs locaux dans les réseaux de gènes (Alon, 2007). La contribution des réseaux génétiques à la stochasticité de l'expression des gènes est d'une telle complexité qu'elle ne peut pas être décrite en quelques lignes. Nous n'insisterons cependant pas plus sur ce point. En effet, il est relativement simple de s'en affranchir expérimentalement. Il suffit pour cela d'utiliser, pour mesurer l'expression des gènes, un rapporteur dont la protéine n'a pas de connection(s) au réseau génétique de la cellule considérée (voir section 2.4.1 ci-dessous). Comme nous utiliserons systématiquement cette approche pour toutes les expériences décrites dans ce travail, nous nous permettrons de rester très superficiels sur cet aspect de la stochasticité.

2.2.3 Maturation des ARNm et des protéines

Comme nous l'avons vu section 1.2, l'expression des gènes met en jeu de nombreux mécanismes biochimiques – tous susceptibles de contribuer à la stochasticité de l'expression

génique – mais aussi des mécanismes de transport et de maturation. Là encore, chacun de ces mécanismes est potentiellement source de stochasticité. Certains, comme la maturation des protéines, nécessitent même de la variabilité pour fonctionner (Figueirêdo *et al.*, 2010). Singh et Bokes (2012) montrent que le transport des ARNs du noyau jusque dans le cytoplasme peut augmenter ou diminuer la stochasticité de l'expression génique. De même, la dynamique de maturation des ARNm ou des protéines va induire un filtrage (filtre passe bas) plus ou moins important suivant les temps caractéristiques des processus moléculaires (De Jong *et al.*, 2010). Enfin, il est important de ne pas négliger la contribution de la dégradation des ARN ou des protéines. En effet, en première approximation, celle-ci correspond à un processus Poissonnien dont le temps caractéristique peut fortement contribuer à filtrer la stochasticité de l'expression produite en amont.

2.2.4 Contribution de la dynamique chromatinienne

Nous avons vu section 1.1 que l'information génique portée par l'ADN pouvait être plus ou moins accessible suivant l'état de la chromatine. Par ailleurs, nous venons de voir qu'il est communément admis que la phase d'initiation de la transcription est considéré comme un des principaux contributeurs de la stochasticité de l'expression génique (Becskei *et al.*, 2005). Il est donc naturel de supposer que la chromatine joue un rôle majeur dans ce processus (Raser, 2004; Schwabe *et al.*, 2012; Stavreva *et al.*, 2012; Field *et al.*, 2008).
La dynamique chromatinienne est un phénomène complexe qui n'a été étudié que récemment. Comme expliqué section 1.1, l'ADN est couplé à un ensemble d'histones autour desquels il est plus ou moins enroulé. L'acétylation/désacétylation ou la méthylation/déméthylation de ces histones sont les principales causes de l'ouverture de la chromatine même si de nombreux autres processus sont au moins partiellement impliqués (Smith et Denu, 2009). L'acétylation des histones est assurée par les enzymes "Histones Acétyl Transferase" (HAT) qui sont chargées d'acétyler la lysine des histones. À l'opposé, les enzymes HDAC ("Histone DesACetylase") permettent la désacétylation des histones, et donc la fermeture de la chromatine. L'inhibition des HDAC permet donc à la chromatine de rester ouverte.
Les méthylation/dé-méthylation des lysines ou des arginines des histones, sont regulées par d'autres enzymes telles que :

- les "Histone Lysine Methyltransferases",
- les "Protein Arginine Methyltransferases",
- les "Lysine Specific Histone Demethylases",
- les "Histone Arginine Demethylases".

Là encore, l'inhibition des méthyltransférases ou le remplacement des acides aminés des histones par d'autres non méthylables permet de favoriser l'état ouvert de la chromatine. En outre, l'état d'un histone peut influencer celui de son voisin. En effet, un histone méthylé peut indirectement faciliter le recrutement d'enzymes telles que certaines méthyltransferases. Ces dernières ont tendance à méthyler les voisines de cette histone,

propageant ainsi l'ouverture de la chromatine le long de la fibre d'ADN. Ainsi, on observe des phénomènes de décompaction par zone du génome. Ce processus de compaction/décompaction de l'ADN permet aux gènes de passer alernativement d'un état plus accessible à la machinerie de transcription à un état moins accessible (Figure I.7).

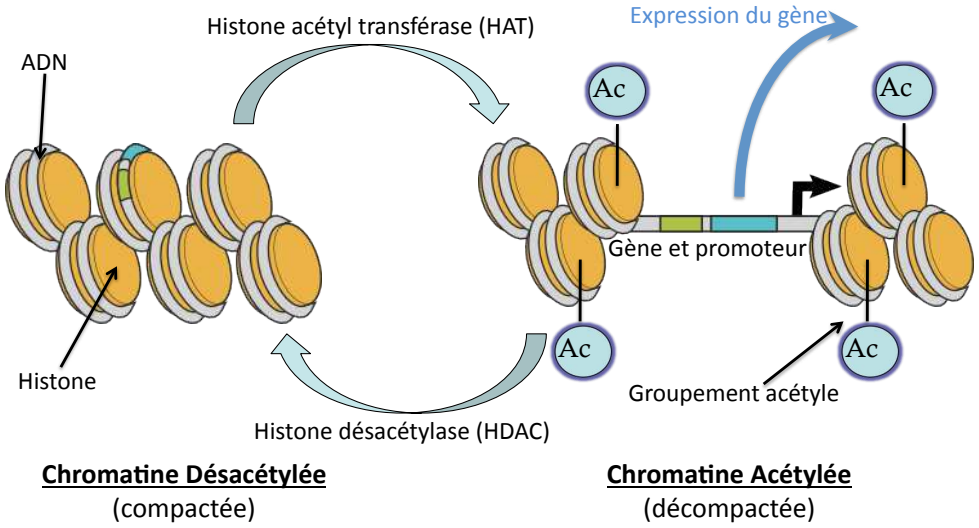

Figure I.7 – Dynamique chromatinienne. La compaction/décompaction de l'ADN autour des histones se fait principalement par l'action des HDAC/HAT. Ces derniers désacétylent/acétylent les histones et plus particulièrement leur lysine (acide aminé), ce qui a pour conséquence de fermer/ouvrir la chromatine. Quand la chromatine est décompactée, la machinerie cellulaire peut plus facilement accéder aux gènes et les transcrire en ARN.

Ainsi, cette dynamique chromatinienne peut influencer l'expression génique d'un ou de plusieurs gènes en leur permettant de passer d'un état actif (i.e. susceptible d'être transcrit) à un état inactif (silencieux). Dans ce cas le gène peut être considéré comme un système à deux états (actif/inactif) ce qui peut fortement augmenter la stochasticité de son expression. L'ensemble des histones, leur position et leur capacité à se méthyler, s'acétyler,... forment un système complexe tout au long du génome qui contribue à la stochasticité de l'expression génique (Verdone et Caserta, 2005; Mellor, 2006; Li *et al.*, 2007). On ignore cependant à peu près tout de son fonctionnement et de son influence vis à vis de la stochasticité de l'expression génique, de même que, en dehors de certains modèles simples (comme par exemple la levure), on ne connait pas les caractéristiques temporelles de la dynamique chromatinienne, ni globalement, ni localement. Il n'existe en effet pas, à l'heure actuelle, de technique expérimentale permettant d'observer en temps réel le niveau de compaction de la chromatine. Par contre, il est clair que cette dynamique peut être une forte source de variabilité dans l'expression génique puisqu'elle influe directement sur l'expression elle-même.

2.3 Fonctions biologiques de la stochasticité de l'expression génique

Le changement de statut épistémologique de la stochasticité de l'expression génique – du statut de "bruit" à celui de "processus biologique" – fait émerger de nouvelles questions. En effet, si la stochasticité n'est pas un bruit qu'il convient de réduire ou de négliger, alors elle devient susceptible d'apporter un bénéfice aux cellules en participant directement ou indirectement à certaines fonctions biologiques. Ainsi, dès les années quarante, Bigger (1944) a constaté qu'au sein d'une même souche de bactéries, certains individus était résistants à la Pénicilline alors que d'autres non. Ultérieurement, il a été montré que cette résistance pouvait être perdue par la bactérie et qu'elle n'était pas systématiquement transmise à sa descendance (Moyed et Broderick, 1986), même en l'absence de mutation génique. Ce résultat à priori surprenant montre que l'expression des protéines liées aux traits de résistance ne dépend pas uniquement de la séquence génétique mais que d'autres causes de variation se superposent à l'action de la séquence. Ce n'est cependant que près de vingt ans plus tard que les fonctions biologiques de la stochasticité ont été systématiquement étudiées.

La variabilité est un processus clé du vivant. En provoquant l'apparition de variants au sein d'une population, elle permet à la selection naturelle de "filtrer" ces variants, constituant ainsi le moteur de l'évolution. Dans ce contexte, la stochasticité de l'expression génique pose question puisqu'elle constitue une variabilité non-héritable[1] et donc sur laquelle la sélection n'a pas directement prise. La contribution de la stochasticité aux fonctions cellulaires n'est donc a priori pas directement liée à l'évolution. En revanche, il a été rapidement montré que, à l'échelle d'une population, la stochasticité de l'expression génique permet de générer des sous-populations adaptées à des environnements différents et donc d'implémenter simplement un mécanisme de "bet-hedging[2]" (Ito *et al.*, 2009; Beaumont *et al.*, 2009; Veening *et al.*, 2008). Ainsi, la stochasticité semble pouvoir être sélectionnée et participer à l'activité "normale" de la cellule.

L'adaptation des cellules à des changements environnementaux est un des processus biologique où la stochasticité joue un rôle. Certaines souches bactériennes ou certains types cellulaires utilisent la stochasticité de l'expression génique pour maintenir des sous-populations capables de résister à des traitements antibiotiques ou à des environnements stressants (Stern *et al.*, 2007; Thattai et Van Oudenaarden, 2004; Balaban *et al.*, 2004; Ackermann *et al.*, 2008; Beaumont *et al.*, 2009; Eldar *et al.*, 2009; Ito *et al.*, 2009; Zhang *et al.*, 2009; Süel *et al.*, 2007; Mettetal et Van Oudenaarden, 2007; Veening *et al.*, 2008; Çağatay *et al.*, 2009; Locke *et al.*, 2011). Dans un tel environnement, la stochasticité permet à certaines colonies variantes de disposer d'un phénotype résistant (Thattai et Van Oudenaarden, 2004) tout en permettant une réversibilité rapide lorsque le stress dis-

[1] Notons cependant que Rosenfeld et al (Rosenfeld *et al.*, 2005) ainsi que Golding et al (Golding *et al.*, 2005) ont montré que la stochasticité peut être conservée sur plusieurs générations cellulaires, même si elle n'est pas strictement héritable génétiquement.

[2] Le bet-hedging correspond à un mécanisme de pari dans lequel une population se diversifie stochastiquement pour faire face à des événements aléatoires. Ce mécanisme est surtout connu chez les plantes pour lesquelles la plupart des graines germent dans l'année mais certaines ne germent qu'au bout d'une ou plusieurs années, garantissant ainsi la survie de l'espèce même en cas de conditions environnementales temporairement très défavorables. En termes triviaux, le bet-hedging correspond à "ne pas mettre tous ses œufs dans le même panier".

paraît. En outre, la stochasticité peut être filtrée par le réseau génétique de la cellule de façon à réguler les proportions de variants ou la fréquence de variation (Çağatay *et al.*, 2009). Dans ce contexte, une des contributions les mieux connues de la stochasticité à la prise de décision cellulaire est la compétence de *Bacillus subtilis*. Maamar *et al.* (2007) ont démontré que, chez cette bactérie, réduire la stochasticité du gène ComK du circuit génétique de décision dans une population cellulaire fait diminuer le nombre de cellules qui entrent en compétence. De plus, Süel *et al.* (2007) démontrent qu'en modifiant plus ou moins la stochasticité d'expression de ce gène, on peut influer sur la "décision" de la bactérie.

La stochasticité de l'expression génique peut aussi participer directement à la dynamique cellulaire. Ainsi, Nachman *et al.* (2007) et Di Talia *et al.* (2007) ont montré que la stochasticité joue un rôle dans la régulation du cycle cellulaire. Di Talia *et al.* (2007) ont mesuré, chez la levure, la stochasticité d'un gène contribuant au cycle cellulaire ainsi que la durée des différentes étapes du cycle cellulaire. Ils constatent, entre autres, que le cycle cellulaire dépend du niveau de variabilité de l'expression génique. Toujours chez la levure, Nachman *et al.* (2007) ont étudié la variabilité du temps de méiose. Ils constatent que cette variabilité nécessite que lme1 (un des principaux régulateurs de la méiose chez *Saccharomyces cerevisiae*) soit stochastique et fortement exprimé. Le bon déroulement de la méiose et donc du cycle cellulaire dépend donc de la stochasticité. Par ailleurs, il a été montré que la stochasticité de l'expression génique est impliquée dans d'autres rythmes biologiques tels que les rythmes circadiens. Ullner *et al.* (2009) utilisent un modèle de réseaux de gènes impliqués dans les rythmes circadiens et étudient le lien entre stochasticité et perturbations du cycle classique jour/nuit. Ils démontrent que le cycle de ce réseau, intrinsèquement stochastique, s'adapte aux conditions lumineuses et que les cellules se synchronisent globalement entre elles.

Chez les organismes multi-cellulaires, les processus de morphogénèse et de différentiation cellulaire pourrait paraître assez déterministes. Or, alors qu'on pourrait imaginer que, dans ce contexte, la stochasticité de l'expression génique soit réduite à son plus petit niveau, il a été montré qu'au contraire, la stochasticité contribue à la différenciation cellulaire, en particulier chez les cellules souches (Wernet *et al.*, 2006; Hume, 2000; Kupiec, 1997; Paldi, 2003; Chang *et al.*, 2008; Halley *et al.*, 2008; Hoffmann *et al.*, 2008; Kalmar *et al.*, 2009; Wu *et al.*, 2009; Wu et Tzanakakis, 2012; Stockholm *et al.*, 2010; Rao *et al.*, 2002; Ungrin *et al.*, 2008; Discher *et al.*, 2009). Ainsi, Wu et Tzanakakis (2012) montrent, par une approche couplant modélisation et expérimentations biologiques, que les cellules souches embryonnaires humaines peuvent comporter des gènes à l'expression stochastique tels que Nanog, confirmant ainsi que la stochasticité est présente chez les cellules souches. Stockholm *et al.* (2010) utilisent une approche couplée similaire sur des myoblastes primaires humains. Ces cellules peuvent passer à un état formant des myotubes et exprimant fortement une protéine de surface CD56 ou revenir à leur état "normal". Ils montrent que la stochasticité permet de passer d'un état à l'autre de cet équilibre bistable. Enfin, Wu *et al.* (2009) montrent un processus similaire dans le cas de la differentiation des cellules souches spermatogonial chez le rat. La stochasticité permet ici, pour une même lignée, d'obtenir des cellules différenciées et non différenciées.

Le rôle de la stochasticité dans l'embryogénèse et le développement a été historiquement évoqué très tôt (Kupiec, 1983). Kupiec évoque en effet dès 1983 la possibilité, en théorie, que des événements aléatoires dans la cellule, touchant en particulier les gènes et des

molécules de régulation soient nécessaires à la différenciation cellulaire. Ces travaux théoriques ont par la suite été partiellement confirmés (Arias et Hayward, 2006; Rossant et Tam, 2009; Boettiger et Levine, 2009; Weinberger *et al.*, 2005). Rossant et Tam (2009) montrent que pendant l'embryogénèse de la souris, la differentiation d'un types de cellule souche précis en l'un ou l'autre des 3 types de blastocyste possibles nécessite que certains gènes régulant cette différentiation soient stochastiques (Cdx2, Nanog, Gata6). Par ailleurs, en étudiant le profil d'expression de 14 gènes de contrôle du développement dans une centaine d'embryons de drosophile, Boettiger et Levine (2009) démontrent que les gènes peuvent être répartis en deux catégories : les gènes synchrones et les gènes stochastiques. Les gènes synchrones sont exprimés "uniformément" dans toutes les cellules d'un tissu contrairement aux gènes stochastiques. Boettiger et Levine montrent que ces deux types de gènes sont nécessaires au développement des embryons.

À l'opposé, Bahar *et al.* (2006) et Spencer *et al.* (2009) ont montré que la stochasticité peut déterminer l'entrée dans un processus de mort cellulaire (apoptose). En quantifiant les transcrits dans ces cellules musculaires cardiaques, Bahar *et al.* (2006) ont montré que, chez la souris agée, la stochasticité de certains gènes augmente. Ils concluent à un lien entre stochasticité et vieillissement. Spencer *et al.* (2009) observent, dans des cellules Hela, un réseau de gènes pouvant causer l'apoptose et induit par un facteur d'induction de cette dernière. Ils montrent que la stochasticité interne du réseau est une des sources de basculement de la cellule vers une décision d'apoptose.

L'utilisation de la stochasticité de l'expression génique dans les processus de décision cellulaire peut bien évidemment s'appliquer aux organismes pathogènes. Ainsi, il a été montré que la stochasticité de l'expression des gènes est utilisée pour certaines prises de décisions chez le virus HIV (Weinberger *et al.*, 2005, 2008; Miller-Jensen *et al.*, 2011; Singh *et al.*, 2010a; Skupsky *et al.*, 2010). Skupsky *et al.* (2010) ont montré que la vitesse de progression de HIV dépend fortement de la stochasticité de ses gènes et de leur régulation par des facteurs cellulaires. La stochasticité serait ainsi impliquée dans le "choix" du virus de rester latent ou non. Dans un article de synthèse, Singh et Weinberger (2009) suggèrent qu'en modulant la stochasticité de HIV-1, il serait possible de bloquer le virus en phase latente et ainsi, d'endiguer la progression du virus.

Enfin, de très nombreux auteurs ont suggéré que la stochasticité de l'expression génique puisse jouer un rôle dans l'établissement de certaines formes de cancer ou dans l'acquisition de résistances aux traitements (Capp, 2005; Laforge *et al.*, 2005; Cohen *et al.*, 2008; Gascoigne et Taylor, 2008; Brock *et al.*, 2009; Mayburd, 2009; Singh *et al.*, 2010b; Kumar *et al.*, 1998; Hoek *et al.*, 2008; Roesch *et al.*, 2010). Capp détaille dans ses travaux (Capp, 2005) un mécanisme possible à l'origine du cancer. Selon lui, un dérèglement de l'expression de certains gènes pourrait entraîner des interactions cellulaires anormales. La stochasticité de l'expression, couplée aux interactions entre gènes produirait alors un dérèglement global de la cellule qui basculerait alors en une forme cancéreuse. Laforge *et al.* (2005) appuient cette théorie par une démarche de modélisation. Ils montrent en effet que la stabilité des cellules dépend de l'auto-stabilisation de la stochasticité d'expression de leurs gènes et des interactions intercellulaires.

2.4 Observation et quantification de la stochasticité de l'expression génique

L'observation expérimentale et la quantification de la stochasticité de l'expression génique demandent d'utiliser des outils différents de ceux de la transcriptomique "classique". En effet, il devient nécessaire d'observer non seulement la moyenne de l'expression génique mais aussi sa variabilité intracellulaire ou intercellulaire. Pour cela, il est nécessaire de mesurer l'expression d'un gène au cours du temps dans une cellule unique ou à un instant t donné mais dans toutes les cellules d'une population.

2.4.1 Méthodes d'observation

Les méthodes d'observation de l'expression génique passent généralement par la mesure du produit de la transcription, c'est à dire de la concentration moléculaire d'ARNm ou, plus couramment, de la concentration de protéines. Ces concentrations peuvent être mesurées par différentes techniques comme la rtq-PCR (pour l'ARNm) ou par Western blot (pour les protéines). Cependant, ces méthodes sont soit rapides, mais alors elles utilisent une quantité de matériel biologique équivalent à plusieurs cellules (ce qui interdit la mesure de la stochasticité inter- ou intra-cellulaire), soit adaptées à la cellule unique, mais sans autoriser la mesure au cours du temps (pas de mesure de la stochasticité intra-cellulaire). De plus, dans ce dernier cas, mesurer plusieurs milliers de cellules (mesure de la stochasticité inter-cellulaire) se révèle souvent extrêmement fastidieux. De fait, les premières études quantitatives de la stochasticité de l'expression génique connues ont plutôt utilisé des techniques indirectes liées à l'utilisation de rapporteurs fluorescents (Elowitz *et al.*, 2002).

Introduction d'un rapporteur fluorescent ou luminescent La principale technique permettant de mesurer la concentration de protéines dans une cellule unique au cours du temps – ou dans une population de cellules à un moment donné – est l'utilisation d'un rapporteur fluorescent (Komorowski *et al.*, 2010). Cette technique combine des approches de biologie moléculaire, permettant d'insérer une nouvelle séquence dans un génome. Cette séquence comporte un gène codant pour une protéine fluorescente (le rapporteur). Une fois le gène inséré (stablement ou non) dans l'organisme, sa transcription en ARNm puis sa traduction en protéines vont, *in fine*, produire une fluorescence dont l'intensité sera directement proportionnelle à la concentration de protéines.
Il existe aujourd'hui une grande variété de séquences fluorescentes disponibles, chacune avec ses spécificités (longueur d'onde de fluorescence, délai de maturation, demi-vie, ...). Cela permet de contruire des dispositifs expérimentaux différents, par exemple en insérant plusieurs gènes rapporteurs différents au sein d'une même cellule pour mesurer simultanément leurs variations respectives. Les dispositifs expérimentaux peuvent par ailleurs différer par les techniques moléculaires employées pour insérer le rapporteur dans la cellule cible. On distingue essentiellement trois approches :

- Une des méthodes consiste à introduire dans les cellules des plasmides (molécules d'ADN courtes autonomes et dispensables à la cellule) contenant un gène codant pour un rapporteur fluorescent. Cette technique ne permet cependant pas l'établissement de lignées stables chez les eucaryotes car les plasmides vont être progressive-

ment éliminés par les cellules. En outre, on ne maîtrise pas la quantité de plasmides présents et celle-ci sera donc variable d'une cellule à l'autre. On ne pourra donc pas observer des cellules dans le temps ni utiliser la différence d'expression de plusieurs cellules pour étudier la stochasticité.

- Une deuxième technique consiste à utiliser des anticorps fluorescents capables de se fixer sur une cible moléculaire ; par exemple, une protéine précise. Cette approche a l'avantage de permettre des mesures à d'autres niveaux que les protéines (en utilisant des anticorps capables de se fixer sur d'autres molécules). En revanche, elle ne permet pas une estimation très précise de la stochasticité car l'efficacité des anticorps à se fixer sur chacune des protéines ciblées influence le résultat et est difficile à maîtriser. De plus, cette méthode ne permet pas d'observer l'expression génique d'une cellule au cours du temps.

- Enfin, l'approche la plus directe consiste à transfecter (insérer) stablement le gène rapporteur dans le génome d'une ou plusieurs cellule(s). Ainsi, ces cellules expriment durablement la protéine fluorescente introduite et transmettent le rapporteur à leur descendance. Cette technique permet de générer des populations clonales, génétiquement identiques. Il suffit pour cela de sélectionner une cellule transfectée et de la laisser s'expendre en une population. Celle-ci sera alors isogénique. On peut grâce à cette méthode, mesurer la stochasticité intercellulaire dans un grand nombre de cellules mais aussi mesurer la stochasticité intracellulaire dans une cellule unique au cours du temps sans variation du nombre de copies du rapporteur. C'est une technique puissante même si l'insertion du rapporteur dans le génome n'est pas très facile à maîtriser chez les eucaryotes supérieurs.

Outre ces trois techniques relativement classiques, il existe une approche alternative permettant d'utiliser un rapporteur fluorescent pour mesurer les concentrations en ARNs : le système MS2 qui permet de mesurer des ARNs spécifiques en les couplant à une protéine fluorescente et ce dans une cellule vivante (Querido et Chartrand, 2008). Pour ce faire, on insère dans le génome ou dans des plasmides, un gène contenant une protéine fluorescente qui a la propriété de se lier à un ARN spécifique dont on veut évaluer l'expression. Cet ARN contient un site de fixation pour la protéine fluorescente (une structure particulière contenant plusieurs tige-boucles – le gène codant pour cet ARN devant bien sûr avoir été préalablement inséré dans le génome). On obtient ainsi un système de deux types d'éléments complémentaires qui permet de mesurer la stochasticité inter mais aussi intra-cellulaire. L'avantage de cette technique est que l'on mesure la présence des ARNs. L'inconvenient est que ces ARNs doivent rencontrer et se fixer avec la protéine fluorescente pour être observables. Ces deux processus ainsi que l'expression de la protéine fluorescente peuvent faire varier les mesures car ils sont tous deux source de stochasticité.

La plupart des rapporteurs demandent, pour émettre une fluorescence, d'être préalablement stimulés par une brève exposition à une lumière d'une longueur d'onde spécifique. Malheureusement, au cours de ce processus, des résidus toxiques sont produits. Qui plus est, lorsqu'elle est trop souvent stimulée, la protéine va perdre de son pouvoir fluorescent (phénomène de "bleaching"). Ces deux propriétés interdisent des mesures trop fréquentes et l'utilisation d'un rapporteur pour des mesures de stochasticité au cours du temps devra

prendre cela en compte pour optimiser la qualité du signal mesuré sans pour autant entrainer une dégradation de ce signal du fait des effets toxiques des résidus ou de l'extinction progressive de la protéine fluorescente.
Une solution pour palier ces difficultés est d'introduire stablement dans le génome des cellules un gène codant pour la luciférase (Waidmann *et al.*, 2011). Il s'agit d'une protéine qui réagit avec de la luciférine en émettant de la lumière. Elle n'a pas besoin d'être stimulée par une longueur d'onde d'excitation mais par de l'ATP pour être observée. La quantité de lumière est alors plus précise, plus ponctuelle (la durée de vie de cette protéine est très courte) mais malheureusement beaucoup moins intense que pour une protéine fluorescente. C'est un excellent moyen pour observer la stochasticité intracellulaire au cours du temps dans une cellule unique (Suter *et al.*, 2011) mais il n'existe malheureusement pas pour l'instant d'outil efficace pour mesurer ce type de lumière indépendamment dans une très grande quantité de cellules. Dans l'état actuel des techniques de mesure, la luciférase ne permet donc pas d'observer efficacement la stochasticité intercellulaire, contrairement à la cytométrie (voir ci-dessous).
Les rapporteurs fluorescents ou luminescents ont plusieurs avantages pour étudier la stochasticité de l'expression génique. Ils permettent en particulier d'utiliser un gène fluorescent ou luminescent exogène n'ayant pas (ou peu) d'interactions avec le réseau génétique de la cellule (et donc, a priori, sans auto-régulation). Cela permet de s'affranchir des modifications que les rétro-contrôles engendreraient sur la stochasticité de l'expression du gène observé (cf section 2.2.2). De plus, on peut parfaitement maîtriser la séquence insérée (promoteur, leader, gène) ce qui permet de s'affranchir des variations de stochasticité potentiellement dues aux interactions ou aux variations de ces divers éléments (tout en permettant d'étudier ces variations au besoin).

Une fois que la cellule produit durablement la molécule fluorescente, il faut mesurer aussi précisément que possible la quantité de lumière émise de façon à pouvoir estimer la concentration de protéines dans la cellule. Pour cela plusieurs techniques expérimentales existent dont trois sont classiquement utilisées :

La cytométrie en flux Pour mesurer la fluorescence de chacune des cellules d'une population cellulaire, la méthode la plus classique est la "cytométrie en flux". Un cytomètre[1] va permettre de trier les cellules pour en mesurer leur fluorescence une par une en émettant une lumière correspondant à la longueur d'onde d'excitation de la protéine fluorescente et en enregistrant la lumière émise par les protéines de chacune des cellules dans leur longueur d'onde de réponse (celle-ci dépendant du type de rapporteur utilisé). Grâce à cette technique, on peut donc mesurer la distribution de fluorescence (Figure I.8) d'une grande population mais il n'est en revanche pas possible de suivre chacune des cellules dans le temps.

La cytométrie par microscopie Une alternative est la cytométrie par microscopie dans laquelle les fluorescences sont mesurées au microscope pour toutes les cellules d'un champ. À partir de l'image du champ, les cellules sont détectées plus ou moins automati-

[1]Souvent improprement appelé "FACS" pour Fluorescence Activated Cell Sorter". Le FACS est en réalité un cytomètre qui a en plus la capacité de trier les cellules suivant, par exemple, leur fluorescence.

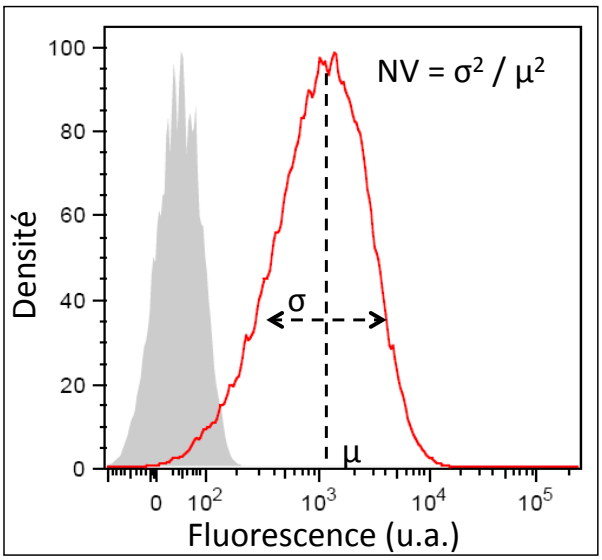

Figure I.8 – Distribution de l'expression : la cytométrie en flux. La densité de cellules est ici représentée pour chaque valeur de fluorescence. Il en résulte une distribution de fluorescence d'une population cellulaire. Cette fluorescence étant due à des protéines fluorescentes, cette distribution peut être considérée comme une représentation de la distribution de protéines par cellule. La fluorescence est ici exprimée en "unité arbitraire" (u.a.) car la calibration des appareils pour la mesurer permet de régler la sensibilité des capteurs de ces derniers pour éviter que les niveaux de fluorescence d'une expérience ne saturent ou ne soit pas détectable. Cela décale artificiellement la distribution sur l'axe des abscisses. Il n'y a pas alors d'unité fixe entre deux appareils réglés différemment. La densité est ici le nombre de cellules multiplié par cent et divisé par le nombre de cellules maximum rencontré pour un niveau de fluorescence donné.

quement (par traitement d'images) et la quantité de fluorescence totale de chaque cellule est mesurée. Cette technique souffre d'un manque d'automatisation qui rend son usage à grande échelle fastidieux pour mesurer plusieurs milliers de cellules comme en cytométrie en flux. Son avantage est qu'elle peut utiliser le même type de microscope qu'en vidéo-microscopie (voir ci-dessous). Les quantités mesurées avec ces deux procédés (cytométrie par microscopie et vidéo-microscopie) sont donc comparables. On peut aussi utiliser ce système pour la luciférase contrairement aux appareillages de cytométrie en flux existant.

La vidéo-microscopie Pour mesurer la stochasticité intracellulaire, on utilise une technique alternative : la vidéo-microscopie. Dans ce cas, on mesure la fluorescence par microscopie time-lapse en acquérant une image d'un objet d'intérêt (une ou plusieurs cellules) par microscopie et ce à une fréquence choisie. On peut ainsi capturer la stochasticité intracellulaire en mesurant la concentration de protéines dans une cellule au cours du temps (Figure I.1).

2.4.2 Indicateurs

Une fois mis en place un dispositif expérimental permettant d'acquérir les mesures de fluorescence, soit pour une population de cellules, soit pour une même cellule au cours du temps, plusieurs indicateurs statistiques permettent de quantifier la stochasticité de l'expression génique.

La distribution de fluorescence : Dans le cas d'une mesure par cytométrie, on a directement accès à la distribution de fluorescence d'une grande population de cellules (plusieurs dizaines de milliers de cellules pour la cytométrie en flux ; plusieurs centaines pour la cytométrie par microscopie). La distribution est un indicateur très complet de la stochasticité mais il est peu lisible : comparer la stochasticité de deux distributions différentes n'est pas facile. On utilise donc souvent des indicateurs agrégés tels que la variance ou encore la variance normalisée :

La variance : elle permet de caractériser la dispersion d'un signal x autour de la moyenne $\langle x \rangle$. C'est une mesure de la somme du carré de l'écart à la moyenne de chaque mesure d'un échantillon. Elle est notée σ^2. La variance n'est pas un très bon indicateur de stochasticité car elle est très fortement influencée pas la moyenne : à stochasticité égale, un ensemble de mesures ayant une moyenne forte aura une variance plus forte. Compte tenu du fait que nous désirons étudier la différence de stochasticité entre plusieurs populations de moyenne potentiellement différente, nous n'utiliserons pas cet indicateur.

Le facteur de Fano : Il est égal à $\sigma^2(x) \,/\, \langle x \rangle$ avec $\sigma^2(x)$ la variance et $\langle x \rangle$ la moyenne de la distribution. Il a la même unité que la moyenne et permet de diminuer l'influence de cette dernière sur la variance. Pour une distribution suivant une loi de Poisson, on a $\sigma^2(x) \,/\, \langle x \rangle = 1$. La différence entre le facteur de Fano d'une distribution donnée et la valeur 1 permet donc de quantifier à quel point cette distribution suit une loi de Poisson ou non. C'est une des caractéristiques intéressante de ce facteur.

La variance normalisée : La variance normalisée (NV) est égale à $NV(x) = \sigma^2(x) \,/\, \langle x \rangle^2$. Comme son nom l'indique, la NV est normalisée. Cet indicateur sans unité a l'avantage de s'affranchir totalement de l'influence de la moyenne et donc de permettre une meilleure caractérisation de la stochasticité de l'expression génique (Figure I.9). C'est l'indicateur que nous utiliserons préférentiellement dans ce travail (avec les histogrammes lorsque plus d'information sera nécessaire).

Indicateurs de la dynamique de données temporelles Dans le cas d'une mesure par vidéomicroscopie (time-lapse), le signal mesuré ne correspond plus directement à une distribution mais à un signal temporel direct. Dans ce cas, **Le spectre de puissance** est particulièrement utile pour étudier les séries de mesures temporelles obtenues. Le spectre de puissance permet en effet d'obtenir plus d'informations que la seule distribution des valeurs mesurées. Il représente la mesure des fréquences d'une série temporelle. Dans cette thèse, nous avons abordé mais pas utilisé ce type d'analyse temporelle, c'est pourquoi nous ne détaillerons pas plus cet indicateur.

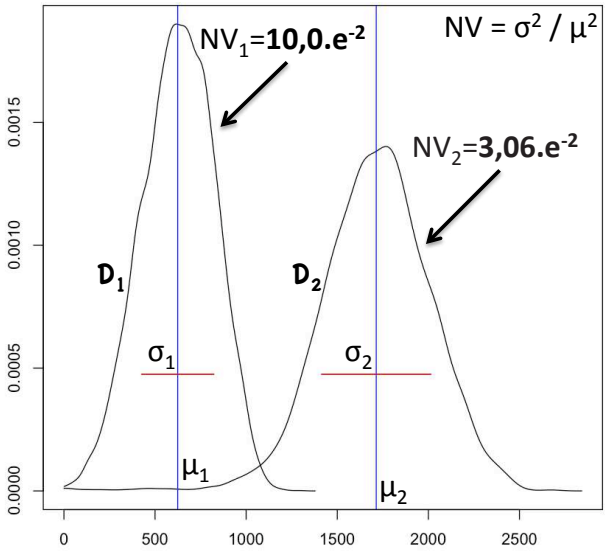

Figure I.9 – Comparaison de deux distributions et de leurs indicateurs. Ici, D_1 possède une moyenne μ_1 et un écart type σ_1 inférieurs à ceux de D_2. Par contre, la variance normalisée de D_1 est supérieure à celle de D_2. C'est un bon exemple montrant que la variance n'est pas un bon indicateur de la stochasticité. Elle est très influencée par la moyenne. Ici, c'est en effet D_2 qui est la moins stochastique des deux distributions.

3 La modélisation de la stochasticité de l'expression génique

3.1 Introduction

Depuis son apparition, l'informatique a évolué en symbiose avec l'électronique, la physique mais aussi les mathématiques. Elle s'est progressivement imposée comme un formidable outil de modélisation au service des sciences et des techniques. Ainsi, grâce au développement de modèles "performatifs", l'informatique est naturellement devenue un outil incontournable pour l'industrie. On peut maintenant concevoir un prototype de voiture ou encore visualiser un composé chimiques avant qu'il soit produit. Cet usage de la modélisation n'est cependant pas le seul : grâce aux modèles "explicatifs", la modélisation et la simulation accompagnent la recherche et permettent à cette dernière d'avancer. Ce sont les "sciences computationnelles".

Dans ce domaine, la biologie computationnelle, l'alliance de la biologie et de l'informatique, est apparue très tôt. Dès l'apparition des premiers ordinateurs, la biologie a inspiré l'informatique (réseaux de neurones artificiels, algorithme génétique, ...). Réciproquement, l'informatique assiste la biologie pour lui permettre, par exemple, d'obtenir le séquençage complet de génomes de diverses d'espèces, de les comparer, de les aligner, ... Ces vingt dernières années, la quantité de données biologiques disponibles a explosée. L'informatique, là encore, permet de stocker et de manipuler ces grands volumes de données. La bioinformatique permet d'obtenir de manière "descendante" (c'est-à-dire par des systèmes de requêtes) des séquençages d'ADN, des protéomes, la localisation ou la fonction des gènes annotés, etc. De manière "ascendante" (par inférence), on peut aussi, par recherche de

similarités entre génomes, prédire la localisation d'autres gènes ou encore leurs fonctions. Si on a pu avoir l'illusion qu'obtenir ces masses de données permettrait de comprendre tous les objets biologiques étudiés, il a rapidement fallu déchanter. En effet, un système biologique (tel qu'une cellule) est un ensemble de sous-systèmes en interaction. Acquérir des connaissances sur les constituants ne suffit alors pas à comprendre comment fonctionne le système dans son ensemble. Pour cela, une approche par modélisation systémique est nécessaire afin de modéliser l'organisation du système en plus de ces constituants (l'un comme l'autre étant souvent non observables directement). Par confrontation des résultats de la modélisation avec des expériences biologiques ciblées, on peut supporter ou infirmer les hypothèses à l'origine du modèle.
Cette approche se heurte cependant à une difficulté majeure : modéliser un système aussi complexe qu'une cellule dans son ensemble est impossible. Trop de paramètres, de mécanismes, de composants sont mal connus, voir totalement inconnus. Il est donc nécessaire d'isoler des sous-systèmes et de créer des modèles pour chacun d'eux. Ce processus de simulations/confrontation peut alors permettre de mettre en évidence des inexactitudes quant aux hypothèses initiales. Selon les résultats, on pourra alors modifier ces hypothèses et recommencer un cycle de modélisation/simulation. Ce principe est illustré figure I.10. On peut ainsi imaginer que :

- La partie gauche de la figure part de connaissances acquises telles que des bases de données de séquences génomiques et de leurs annotations ("O_1" dans ce cas). Ces connaissances permettent de mieux comprendre le système a étudié, de le modéliser et de le simuler. Les prédictions obtenues *in silico* permettent de faire des hypothèses que l'on peut alors vérifier en les confrontant avec "O_1" ou avec d'autres informations génomiques et ainsi de suite.

- La partie de droite de la figure fonctionne sur le même principe que la partie de gauche mais se base sur un autre ensemble d'observables que "O_1" : "O_2" ici. On passe, par exemple, d'informations génomiques et leurs annotations à des résultats d'expériences bio-chimiques révélant des caractéristiques *in vitro/in vivo* sur les capacités métaboliques du système a. On a alors deux processus permettant d'acquérir des connaissances complémentaires sur a.

Ces deux boucles, portant sur le même sujet d'étude, sont bien évidement liées et peuvent être utilisées ensembles. Les choix de modélisation peuvent changer suivant la boucle et il peut y avoir plusieurs boucles.

Parmi les différentes sources de la stochasticité de l'expression génique, nous avons vu que certaines proviennent de caractéristiques spatiales (diffusion des FT, accessibilité du promoteur,...) tandis que d'autres sont dues au(x) faible(s) nombre(s) d'élèments dans le système (ARN, Polymérase, Ribosomes, FT,...) ou à leurs interactions (réseaux de gènes, interactions entre facteurs de transcriptions,...). Malgré les progrès rapides des techniques d'observations et de mesures en biologie moléculaire et cellulaire (séquenceurs haut débit, multi PCR ciblés permettant de mesurer le niveau d'un grand nombre de gènes dans plusieurs cellules simultanément, ...), la plupart de ces sources sont encore inaccessibles à l'observation ou à la mesure directe. Nous en sommes donc réduits à des mesures indirectes à partir desquelles nous essayons d'inférer l'origine de la stochasticité mesurée. Dans ce

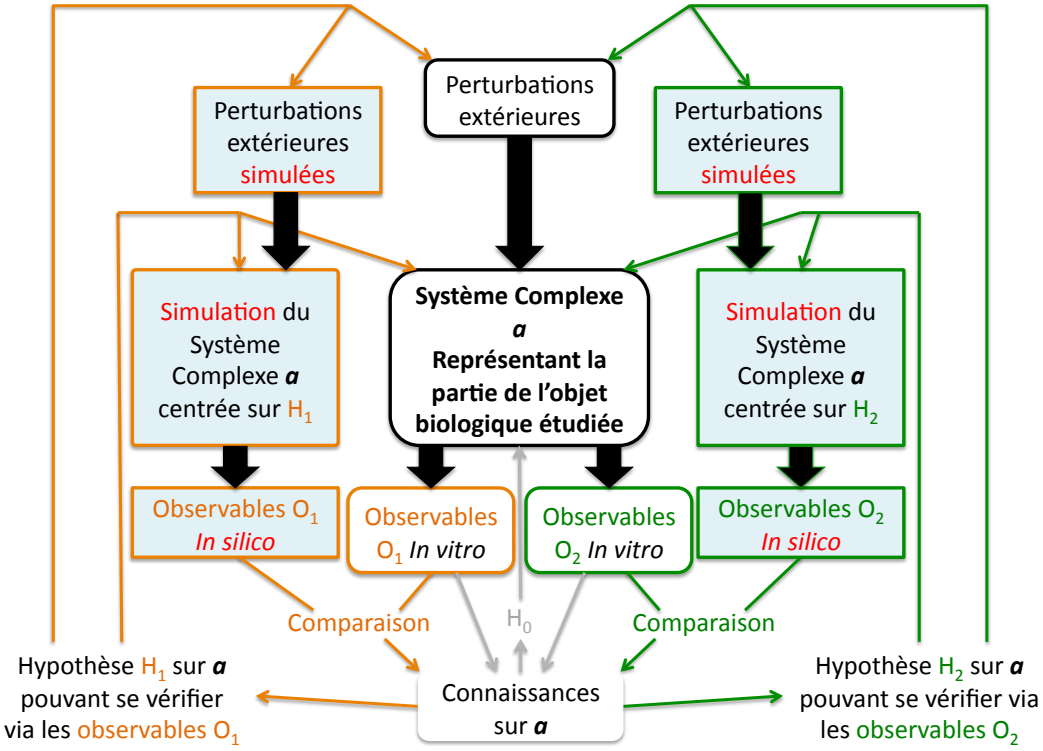

Figure I.10 – Cycle de développement des connaissances par modélisation de systèmes biologiques complexes : la connaissance peut principalement s'acquérir en validant ou infirmant des hypothèses, soit directement par l'observation directe *in vivo* ou *in vitro* ("O_1" ou "O_2"), soit indirectement par la simulation *in silico*. Cette connaissance fraîchement acquise permet alors de faire de nouvelles hypothèses "H_0", "H_1" ou "H_2" que l'on pourra à nouveau valider ou infirmer dans un nouveau cycle. Ce sont les caractéristiques de l'hypothèse qui dicteront les expériences (*in vivo*, *in vitro* ou *in silico*) qui permettront de la confirmer ou l'infirmer par un ensemble d'observables. En effet, le nouveau cycle ne passera pas forcément par le même processus de validation que le cycle précédent. On peut ainsi alterner entre O_1, O_2, ... et O_n. Il est aussi possible de perturber le système et d'analyser les variations induites sur les observables, que ce soit *in vivo*, *in vitro* ou *in silico*.

contexte, les modèles sont des outils essentiels afin de tester les hypothèses reliant les sources aux observables. De fait, la modélisation est couramment utilisée pour étudier la stochasticité et un très grand nombre de modèles ont été proposés dans la littérature afin de tester les contributions des différentes sources de stochasticité. Ainsi, certains modèles représentent explicitement les contraintes spatiales tandis que d'autres les négligent en ne représentant pas les composés moléculaires (TF, ribosomes,...) pour ne prendre en compte que la concentration (ou le nombre) de molécules présentes et leurs interactions entre elles et avec le promoteur.

Nous avons vu que les avis convergent pour considérer que l'initiation de la transcription est la principale source de stochasticité. La stochasticité observée (sur les protéines) serait essentiellement issue de l'initiation mais filtrée par l'ensemble du processus de l'expression génique. La plupart des modèles décrits ici se concentrent donc sur cette phase du processus de transcription-traduction. Néanmoins, comme nous n'avons accès expérimentalement qu'aux concentrations de protéines pour observer la stochasticité de l'expression génique, il est important de considérer aussi les autres sources de stochasticité qui interviennent entre l'initiation de la transcription et la production de protéines.
Nous allons passer en revue ces modèles en commençant par les modèles centrés sur la chaîne de transcription-traduction (réseaux de gènes, variabilité de la transcription due aux ARNt, variabilité de la traduction due aux ribosomes, maturation des protéines (De Jong *et al.*, 2010)) non spatialisés et spatialisés.
Ensuite, nous étudierons les modèles non spatiaux centrés sur le promoteur et l'initiation de la transcription et ce sous différents niveaux de complexité :

- le modèle à un état – ou modèle Poissonnien (McCullagh *et al.*, 2009; Golding *et al.*, 2005; Munsky *et al.*, 2012) – qui correspond à la description la plus simple possible du promoteur,
- le modèle à deux états – ou modèle de "random telegraph" (Paulsson, 2004; Raser, 2004; Paulsson, 2005a; Bar-Even *et al.*, 2006; Kaufmann et Van Oudenaarden, 2007; Raj et Van Oudenaarden, 2008; Dar *et al.*, 2012; Munsky *et al.*, 2012; Weinberger *et al.*, 2012) – qui intègre plus de réalité biologique en complexifiant l'initiation,
- les modèles à plus de deux états, qui permettent d'étudier des promoteurs de complexité plus élevée. Il s'agira, par exemple, de modèles incluant une période réfractaire (Suter *et al.*, 2011) ou de modèles prenant en compte l'ensemble des états traversés par le système au cours de l'assemblage du complexe d'initiation (Coulon *et al.*, 2010).

Enfin, pour conclure, nous aborderons le cas des modèles intégrant les aspects spatiaux en nous focalisant sur ceux qui sont centrés sur le promoteur et l'initiation.

3.2 Modèle de la stochasticité des réseaux de gènes et de la chaîne de transcription-traduction

3.2.1 Modèle de la chaîne de transcription-traduction

Nous avons vu, section 2.2 qu'il existe de nombreuses sources potentielles de stochasticité de l'expression génique. Certaines causes de la stochasticité de l'expression sont issues du transfert de l'information entre l'ADN et les protéines. Par exemple, l'ARNm peut être dégradé avant d'être exporté du noyau et traduit par un ribosome ; même en présence du ribosome, la traduction peut être bloquée par un manque d'acides aminés, ... Tous ces processus sont des causes potentielles de stochasticité qui ont été étudiées et modélisées (Aitken *et al.*, 2010; Garai *et al.*, 2009).
La contribution de l'ensemble de la chaîne post-transcriptionnelle à la stochasticité peut être modélisée de manière plus ou moins complexe, c'est à dire en incluant différents niveaux de détails suivant la précision souhaitée et les moyens disponibles (observables

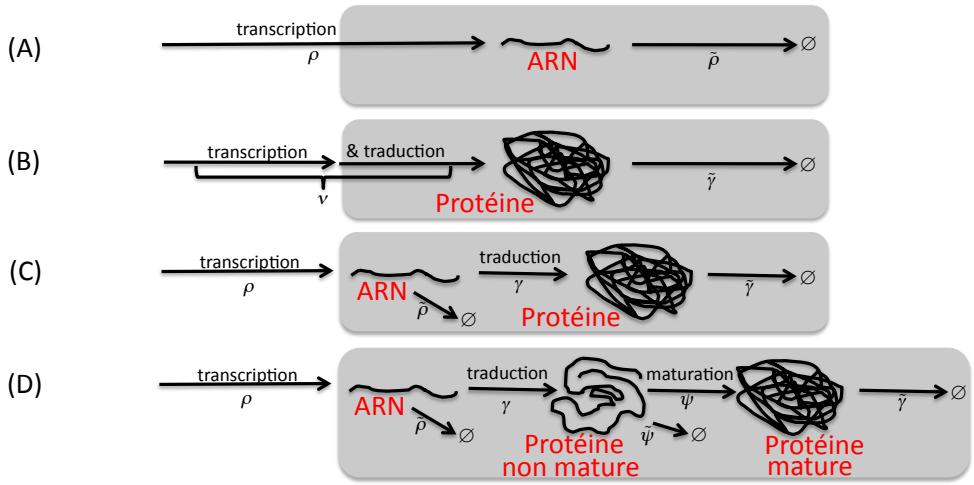

Figure I.11 – Exemples de modèles de l'expression post-transcriptionnelle de différentes complexités. (A) Dans ce modèle, seuls les ARNs sont représentés. Il peuvent être transcrits et dégradés. Chacun de ces deux processus à son propre taux (ρ pour l'un et $\tilde{\rho}$ pour l'autre). Les protéines ne sont pas représentées ici. (B) Ici, la transcription et la traduction sont agrégées en un seul et même processus et les ARNs ne sont pas représentés. L'étude de ce modèle est similaire au précédent sous condition de remplacer les ARNs par les protéines, ρ par ν et $\tilde{\rho}$ par $\tilde{\gamma}$. (C) Les ARNs et les protéines sont représentés par des entités séparées. Elles ont toutes deux leurs propres taux de production et de dégradation. (D) Une étape de maturation des protéines est ajoutée en supplément du modèle précédent.

biologiques, temps, puissance de calcul, ...). Les modélisations les plus courantes sont représentées sur la figure I.11, depuis la vision la plus simpliste résumant la chaîne en un seul processus (Fig. I.11.A), jusqu'à des modèles très complexes (Fig. I.11.D) prenant par exemple en compte la maturation des protéines pour estimer son impact sur la stochasticité. Ce dernier modèle a, par exemple, été utilisé par De Jong *et al.* (2010) et a servi de base pour le travail de (Komorowski *et al.*, 2010).

Une fois la structure du modèle choisie, il est nécessaire de spécifier une méthode d'implémentation. Celle-ci va dépendre de la complexité du modèle mais aussi des observables que l'on souhaite en extraire.

Les modèles les plus simples (Figures I.11.A, I.11.B et I.11.C) sont souvent décrits sous la forme d'un système d'équations différentielles. Il n'est cependant pas toujours possible de trouver une solution analytique à ce système d'équations. En outre, dans le cas de systèmes stochastiques, l'étude analytique du modèle ne permet pas toujours d'étudier la variabilité induite par les différentes étapes du modèle.

Pour calculer l'évolution d'un tel système stochastique, une approche, plus coûteuse en calculs mais qui permet d'obtenir une représentation de la solution quelle que soit la complexité du système, est d'utiliser la simulation numérique. Pour ce faire, l'algorithme le plus classique est l'algorithme de Gillespie (Gillespie, 1977). Son *Stochastic Simulation Algorithm* (SSA) permet en effet de simuler très précisément l'évolution temporelle de réactions chimiques stochastiques. Cependant, cet algorithme ne permet de calculer qu'une seule réalisation à la fois et il peut être nécessaire de l'exécuter un très grand nombre de

fois pour estimer la stochasticité du système, ce qui alourdit considérablement le calcul. Des solutions ont été développées pour rendre ces méthodes plus rapides comme, par exemple, le tau-leaping (Gillespie, 2001). Ce dernier permet de calculer plus rapidement mais avec moins de précision, ce que calculerait un système utilisant le SSA. Une autre de ces solutions est, par exemple, le Finite State Projection (Munsky *et al.*, 2005) qui combine tau-leaping et SSA pour plus de précision que le tau-leaping seul.

3.2.2 Modèle spatialisé de la stochasticité post-transcription

Comme nous l'avons vu, la stochasticité peut venir du faible nombre de molécules et de leurs interactions. Le déplacement de ces dernières étant aléatoire et contraint par leurs interactions, plus leur nombre est faible, plus leur action est aléatoire dans le temps (Kærn *et al.*, 2005). L'algorithme SSA peut partiellement rendre compte de ce phénomène mais il suppose que les concentrations moléculaires sont homogènes dans l'espace. Or, lorsque le nombre de molécules devient très faible, cette hypothèse devient peu réaliste. Il est alors nécessaire de modéliser le temps mais aussi l'espace. Pour ce faire, il faut prendre en compte les caractéristiques physiques des molécules : mode de diffusion et contraintes sur le déplacement, interactions avec d'autre molécules, structures, ...
Dans ce cas, la représentation du système est souvent implémentée sous la forme d'un modèle individu-centré : chaque entité du système et chaque interaction entre ces entités sont calculées au cours du temps. Lors d'une simulation, on peut alors observer numériquement chacun des éléments du système. Contrairement au SSA, ce type de simulation est généralement basé sur un calcul en temps discret. La simulation individu-centrée permet de représenter des systèmes d'une complexité arbitraire mais au prix d'un coût de calculs exorbitant si le nombre de molécules présentes est élevé. Ces modèles permettent d'étudier des interactions protéines-protéines dans l'espace et de rendre compte de comportements complexes tels que la formation d'agrégats (Soula *et al.*, 2005; Bernaschi *et al.*, 2007).
Même si les modèles individu-centrés constituent une alternative intéressante à l'algorithme de Gillespie lorsque l'espace doit être pris en compte, le coût computationnel de ce type de modèle reste prohibitif. Peu de travaux crédibles portent donc sur l'étude en trois dimensions des cellules et de la stochasticité que l'interaction et le déplacement des protéines entraînent. La plupart des modèles spatialisés se concentrent donc "simplement" sur l'initiation de la transcription en modélisant la diffusion localement autour du promoteur (voir section 3.3).

3.2.3 Modèles de réseaux de gènes

Certains gènes, via les protéines qu'ils produisent, vont affecter l'expression d'autres gènes, voire affecter leur propre expression. Leurs produits peuvent en effet être des polymérases, des facteurs de transcription ou encore des molécules aidant à la maturation d'autres types de protéines. Certains micro-ARN peuvent, par exemple, capturer d'autres ARNs, les empêchant ainsi d'être traduits. Certaines protéines peuvent aussi accélérer le processus de dégradation des ARNs ou des protéines.
Il y a donc des interactions plus ou moins directes entre l'expression de différents gènes. Ces interactions sont susceptibles de moduler la stochasticité de l'expression des gènes via les interactions indirectes gènes à gènes. On représente alors ces interactions sous la forme d'un "réseau génétique".

Les réseaux génétiques ont été très étudiés au cours des quinze dernières années (voir (Lim *et al.*, 2013) pour une revue récente). Néanmoins, la plupart de ces études ne prennent en compte que la moyenne d'activité des gènes et négligent totalement la stochasticité. Cette approche déterministe permet d'étudier la dynamique et les points fixes des réseaux mais elle néglige le fait – pourtant trivial – que la stochasticité d'un gène peut modifier le comportement du réseau en le rendant, par exemple, plus robuste ou en permettant des transitions entre deux états stables séparés par une barrière énergétique. Il est donc important d'intégrer la stochasticité de l'expression des gènes dans ces modèles (Sîrbu *et al.*, 2012). Ainsi, une revue récente de Munsky et ses collaborateurs montre l'intérêt d'étudier l'impact de la stochasticité sur la régulation génique (Munsky *et al.*, 2012). Une telle étude peut par ailleurs intégrer l'environnement de la cellule. En effet, celui-ci peut aussi influer sur l'expression d'un gène et sur la stochasticité (Shahrezaei *et al.*, 2008).

3.3 Modélisation du promoteur et de l'initiation

3.3.1 Modèle à un état (modèle de poisson)

Nous avons vu que la stochasticité de l'expression génique, mesurée indirectement par les concentrations de protéines, peut varier suivant l'intensité de la transcription ou de la traduction par rapport à celle de la dégradation des ARNn et des protéines (McCullagh *et al.*, 2009). À probabilité de transcription égale, le nombre de protéines peut donc présenter une variabilité correspondant à la variabilité transcriptionnelle plus ou moins filtrée par les mécanismes post-transcriptionnels. Ce phénomène peut être modélisé par un modèle de promoteur à un état (puisque la probabilité de transcription est constante) doté d'une dynamique Poissonnienne.
Dans ce type de modèle, le promoteur est donc toujours dans l'état actif et les événements de production d'ARN se produisent avec une probabilité constante. L'expression génique suit donc un processus Poissonnien caractérisé par un taux de production d'ARN ρ. Si on considère que les ARNs R ainsi produits sont dégradés avec une probabilité elle-même constante $\tilde{\rho}$, alors on a :

$$\left\{ \frac{dR}{dt} = \rho - \tilde{\rho}R \right. \tag{I.1}$$

Ce qui est représenté dans la figure I.12 si l'on considère la chaîne post-transcriptionnelle telle que représentée dans la figure I.11. Ainsi, si l'on considère la suite de la chaîne de

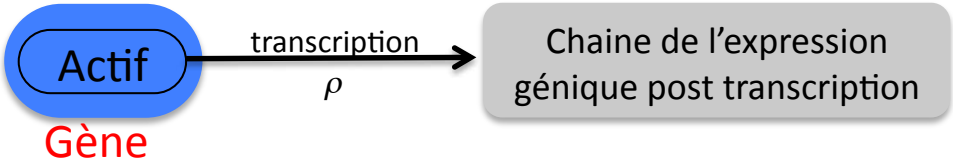

Figure I.12 – Modèle de promoteur à un état. En plus de représenter la chaîne post-transcriptionnelle telle que décrite dans la figure I.11, on rajoute le gène. Le choix de la modélisation de la chaîne post-transcriptionnelle dépend du cas étudié.

l'expression génique pour obtenir le nombre de protéines P, on aura le modèle répresenté en couplant le modèle représenté figure I.12 et celui représenté figure I.11.C.
Il y a deux manières de traiter les protéines dans ce modèle : de manière totalement déterministe (on considère alors que la concentration de protéines est directement proportionnelle au nombre d'ARN) ou, comme pour les ARN, par un processus de production-dégradation stochastique. Dans ce cas, le modèle peut s'écrire sous la forme :

$$\left\{ \begin{array}{rcl} \dfrac{dR}{dt} & = & \rho - \tilde{\rho} R \\ \dfrac{dP}{dt} & = & \gamma R - \tilde{\gamma} P \end{array} \right. \tag{I.2}$$

où γ est le taux de traduction des ARNs et $\tilde{\gamma}$ le taux de dégradation des protéines.
Le système d'équation I.2 peut être résolu analytiquement ce qui permet de calculer l'évolution du nombre d'ARNs et de protéines ($R(t)$ et $P(t)$) au cours du temps :

$$\left\{ \begin{array}{rcl} R(t) & = & (R_0 - \langle R \rangle) e^{-\tilde{\rho} t} + \langle R \rangle \\ P(t) & = & \left(P_0 - 2\langle P \rangle - \dfrac{\gamma (R_0 - \langle R \rangle)}{\tilde{\gamma} - \tilde{\rho}} \right) e^{-\tilde{\gamma} t} + 2\langle P \rangle + \dfrac{\gamma (R_0 - \langle R \rangle)}{\tilde{\gamma} - \tilde{\rho}} e^{-\tilde{\rho} t} \end{array} \right. \tag{I.3}$$

Les concentrations moyennes d'ARN ($\langle R \rangle$) et de protéines ($\langle P \rangle$) sont alors égales à :

$$\left\{ \begin{array}{rcl} \langle R \rangle & = & \dfrac{\rho}{\tilde{\rho}} \\ \langle P \rangle & = & \dfrac{\rho \gamma}{\tilde{\rho} \tilde{\gamma}} \end{array} \right. \tag{I.4}$$

En l'absence de perturbations externes, les concentrations moléculaires du système vont donc converger vers $\langle R \rangle$ et $\langle P \rangle$.
Une des caractéristiques des distributions de protéines obtenues par un tel modèle Poissonnien, à l'état stable, est que les distributions suivent une loi de Poisson et donc que leur variance et leur moyenne ($\langle P \rangle$) sont égales (Paulsson, 2005a). On a donc, pour les protéines : $\sigma^2 = \langle P \rangle$. Autrement dit, le facteur de Fano est égal à 1 ($\frac{\sigma^2}{\langle P \rangle} = 1$) et donc $\frac{\sigma^2}{\langle P \rangle^2} = \frac{1}{\langle P \rangle}$ ce qui donne $NV = \frac{1}{\langle P \rangle}$.
Dans un tel modèle, les paramètres du modèle peuvent donc modifier le nombre de protéines et sa variabilité mais ces deux observables sont liés. C'est une contrainte forte qui ne rend pas toujours compte des observations biologiques.

3.3.2 Modèle à deux états ("random telegraph")

Contrairement à ce que suppose le modèle précédent, la probabilité de transcription d'un gène peut varier dans le temps. En effet, on a vu qu'un gène et son promoteur sont codés sur l'ADN qui est lui-même plus ou moins enroulé autour des histones. La chromatine ainsi formée, suivant son niveau de compaction, peut permettre ou non la transcription de ce gène (Raser et O'Shea, 2005). Ainsi, la probabilité de transcription est susceptible de varier suivant l'état de la chromatine.
Une situation similaire peut être observée si l'on considère, par exemple, que la probabilité de transcription dépend de la présence d'un facteur de transcription sur le promoteur (Kepler et Elston, 2001; Raj *et al.*, 2006). Dans toutes ces situations, la variabilité observée au niveau des protéines est le résultat de la propagation (plus ou moins filtrée) de

la stochasticité de la transcription du gène mais celle-ci est différente suivant l'état du promoteur.

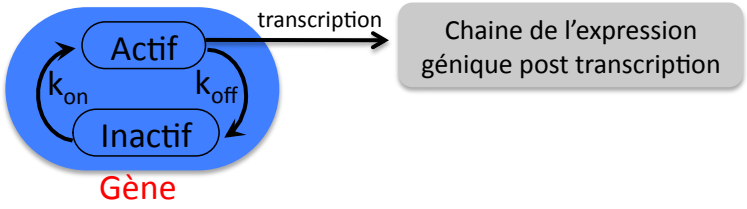

Figure I.13 – Modèle dit "random télégraph". Comparativement à la figure I.12, ici, la complexité du gène change : ce dernier n'est plus actif qu'à certains moments.

On retrouve ici le modèle de Poisson à la différence principale qu'une étape supplémentaire est ajoutée dans le modèle : on considère que le promoteur peut être en deux états (Figure I.13) :

- Le premier est un état dans lequel le gène est inaccessible à la polymérase et où la probabilité de transcription est nulle.
- Le deuxième est un état accessible où la probabilité de transcription peut alors être non nulle. Lorsque le gène est dans cet état la transcription suit une dynamique Poissonnienne.

Cette situation peut être modélisée en considérant que le promoteur du gène peut être dans un état actif ou inactif et que la probabilité de transcription est égale à ρ lorsque le promoteur est actif et 0 lorsqu'il est inactif. On parle alors d'un modèle de type "random-telegraph" (Figure I.13). Le promoteur passe de l'état actif à l'état inactif (et réciproquement) aléatoirement avec des constantes de transition k_{off} (et k_{on} réciproquement).

Si l'on considère que la chaîne post-transcriptionnelle est modélisée comme présentée figure I.12.C, ce modèle peut être étudié analytiquement. Dans ce cas, le modèle est régi par le système dynamique suivant :

$$\begin{cases} \dfrac{dR}{dt} &= \rho k_T - \tilde{\rho} R \\ \dfrac{dP}{dt} &= \gamma R - \tilde{\gamma} P \end{cases} \tag{I.5}$$

avec k_T, la proportion du temps où le gène est actif, définit ainsi :

$$\begin{cases} k_T &= \dfrac{k_{on}}{k_{on} + k_{off}} \end{cases} \tag{I.6}$$

où k_{on} est la probabilité que le gène passe de l'état inactif à l'état actif et k_{off} la probabilité de la transition inverse.
Comparativement au modèle de Poisson, les moyennes de protéines et d'ARN à l'état stable changent en fonction de k_T :

$$\begin{cases} \langle R \rangle &= \dfrac{\rho}{\tilde{\rho}} k_T \\ \langle P \rangle &= \dfrac{\rho\gamma}{\tilde{\rho}\tilde{\gamma}} k_T \end{cases} \tag{I.7}$$

Dans ce modèle, le lien entre variance et moyenne caractérisant le modèle de poisson n'est plus trivial. Néanmoins, considérant l'activation et l'inactivation du gène, la synthèse et la dégradation des ARNs et le lissage temporel des ARNs et des protéines, on peut calculer analytiquement la variance normalisée des protéines (Paulsson, 2005a) :

$$\begin{cases} NV = \frac{1}{\langle P \rangle} + \frac{1}{\langle R \rangle} \frac{\tilde{\gamma}}{\tilde{\rho}+\tilde{\gamma}} + \frac{k_{off}}{k_{on}} \frac{\tilde{\gamma}}{\tilde{\rho}+\tilde{\gamma}} \frac{1+\frac{\tilde{\rho}}{\tilde{\gamma}+k_{on}+k_{off}}}{1+\frac{k_{on}+k_{off}}{\tilde{\rho}}} \end{cases} \tag{I.8}$$

3.3.3 Modèle à n états

Les différents facteurs de transcription intervenant dans l'initiation de la transcription interagissent ensemble et avec la chromatine. Cela peut, par exemple, moduler le temps d'accroche/décroche de ces facteurs et donc le temps qu'un gène met pour passer de l'état actif à l'état inactif.
De même, une polymérase en train de transcrire un ARN (ou tout autre élément de la machinerie transcriptionnelle) ralentit probablement la compaction de la chromatine à ce niveau et module aussi très probablement le temps de désactivation du gène.
Ces temps de latence entre état ouvert et fermé peuvent être modélisés comme des états intermédiaires supplémentaires. Dans ce cas, le gène a plus de deux états, le cas le plus simple étant l'ajout d'une période réfractaire entre l'état actif et l'état inactif (Figure I.14).

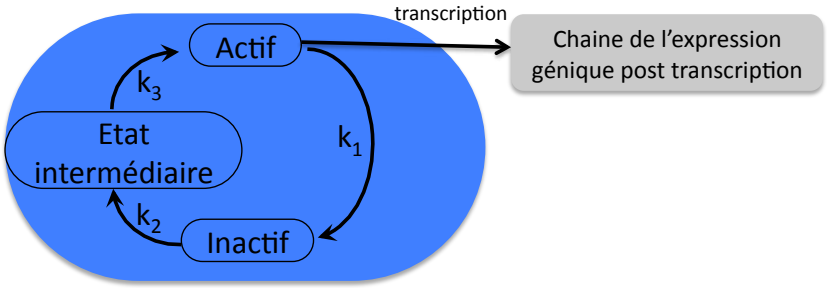

FIGURE I.14 – Modèle à période réfractaire. Comparativement à la figure I.13, la complexité du gène augmente encore : ce dernier passe par un état intermédiaire entre les états passif et actif.

Comme pour le modèle de télégraphe (Figure I.13) et avec la représentation de la chaîne post-transcriptionnelle de la figure I.12.C, ce modèle (Figure I.14) peut lui aussi être caractérisé par les équations I.5, I.4 et I.7 mais cette fois-ci, k_T respecte :

$$\begin{cases} k_t = \frac{\frac{1}{k_1}}{\sum_{i=2}^{n} \frac{1}{k_i}} \end{cases} \tag{I.9}$$

où k_1 est la probabilité que le gène quitte l'état actif et $\{k_i | i \in \{2..n\}\}$ les probabilités que le gène passe d'un état inactif à un autre état, lui-même actif ou inactif (on suppose

ici qu'un seul des états du gène permet la transcription). Dans cette équation n'est pas visible la notion de délai mais uniquement d'états intermédiaires.
Une étude récente au cours de laquelle la concentration d'une protéine a été mesurée au cours du temps par vidéo-microscopie, a mis en évidence de telles périodes réfractaires dans le processus de transcription (Suter *et al.*, 2011).
L'ensemble des états et interactions entre molécules impliquées dans la transcription (FT, polymérase, chromatine,...) peut aussi être modélisé plus finement. Il existe des modèles permettant de rajouter un très grand nombre d'états possibles du promoteur, plusieurs de ces états pouvant induire de la transcription, chacun avec leur propre probabilité. Un modèle de ce type a été développé par Antoine Coulon au cours de sa thèse. Ce modèle est illustré dans la figure I.15.

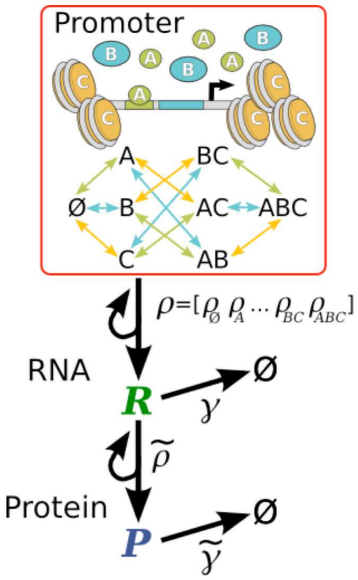

Figure I.15 – Modèle à n états : Promdyn (développé par Antoine Coulon). On y voit la chromatine et différents états possibles (i) du promoteur, chacun avec leur probabilité de transcription ρ_i. Tout passage d'un état à l'autre a aussi sa propre probabilité comme pour les modèles précédents mais les transitions sont ici représentées par un graphe, ce qui autorise un niveau de complexité arbitraire.

Il a permis de mettre en évidence que la concentration en facteurs de transcription peut réguler la stochasticité et qu'en modélisant finement la mécanique du promoteur et de l'initiation, on peut arriver à retrouver des résultats similaires aux modèles de réseaux de gènes même avec un seul gène (Coulon *et al.*, 2010).

3.3.4 Modèles spatialisés

À nouveau, les modèles de l'initiation précédemment décrits font l'hypothèse d'un nombre de molécules suffisamment grand pour que les concentrations moléculaires puissent être considérées homogènes.
Pourtant, il est possible qu'un facteur de transcription, par exemple, soit en nombre très limité dans une cellule (< 10). Dans ce cas, entre l'encombrement moléculaire restreignant

la mobilité des FT et le déplacement aléatoire des ces FT, il peut se passer un long moment avant qu'un FT ne rejoigne un site de fixation. Par contre, une fois dans son voisinage, la probabilité de recapture peut être importante.
Pour rendre compte de cela, il est nécessaire d'inclure les aspects spaciaux dans les modèle de l'initiation de la transcription. Cela permet de prendre en compte cette source de stochasticité de l'expression génique.
Voituriez et Bénichou, en collaboration, principalement, avec Coppey (Coppey *et al.*, 2004), Tejedor (Tejedor *et al.*, 2010a,b) puis Guérin (Guérin *et al.*, 2013), ont étudié la dynamique des interactions entre une protéine et sa cible dans ces conditions. Van zon (Van Zon *et al.*, 2006), quant à lui, a plus particulièrement étudié l'influence d'un faible nombre de FT sur la stochasticité de l'expression génique. Tkačik (Tkacik et Bialek, 2009) a étudié l'influence sur la SGE d'un couplage entre diffusion 3D et sliding 1D sur l'ADN d'un FT (le "sliding" correspond au fait qu'un FT, quand il rencontre un brin d'ADN, a tendance à le suivre et donc à diffuser en une dimension. Quand le FT est proche du brin d'ADN, le sliding augmente la possibilité de rencontre entre le FT et sa cible. Dans ce cas, l'expression est forte mais lorsque le FT n'est pas associé à l'ADN, il diffuse en 3D et l'expression du gène peut rester faible pendant une longue période).
Malgrès le fait qu'ils soient maintenant connus, ces effets sont encore rarement modélisés de manière explicite lors de l'étude de la stochasticité de l'expression génique en général. En effet, prendre en compte cet aspect 3D seul est déjà assez lourd en calcul et complexe à interpréter. De ce fait, lorsque l'on étudie d'autres aspects comme l'influence de l'initiation ou encore de déterminants locaux sur cette stochasticité, on néglige souvent de modéliser l'espace en trois dimension. Dans ces situations, les résultats obtenus sont alors à interpréter en prenant en compte les résultats de, par exemple, Voituriez et Bénichou.

4 Du locus à la stochasticité de l'expression du gène

Nous conclurons cette introduction en résumant la démarche expérimentale suivie au cours de notre travail. Dans cette thèse, nous allons voir à quel point et comment le locus d'un gène (localisation du gène dans le génome) influence son expression et surtout, sa stochasticité. Les méthodes et outils utilisés seront décrits plus en détails dans chaque chapitre même si, pour la plupart, ils ont déjà été présentés parmi toutes les méthodes d'observation, d'analyses ou encore de modélisations/simulations de la stochasticité de l'expression génique. Cette thèse fait en effet appel à un très grand nombre d'outils et de méthodes allant de la biologie moléculaire à la modélisation stochastique (cellules 6C2 transféctées avec un gène rapporteur, cytométrie en flux, modèle à deux états, ...).

Dans un premier temps (chapitre II), nous allons étudier expérimentalement comment la stochasticité de l'expression d'un gène change suivant son locus. De plus, nous verrons que, à locus constant, nous pouvons influencer cette stochasticité en utilisant des agents modificateurs de l'état chromatinien.

Ensuite (chapitre III), nous étudierons comment, pour un locus donné, la dynamique

chromatinienne influence la stochasticité de l'expression génique du gène.

Nous poursuivrons (chapitre IV) en regardant s'il y a, dans l'environnement génomique voisin d'un gène, des éléments caractéristiques de la séquence génomique statistiquement liés à la stochasticité de l'expression de ce gène.

Enfin, nous conclurons par une synthèse et une discussion des résultats obtenus au cours de ce travail.

Chapitre II

Influence du locus génomique sur la stochasticité de l'expression génique

Ce travail a fait l'objet d'une publication dans la revue *Progress in Biophysics and Molecular Biology* (Viñuelas *et al.*, 2012).

1 Principe de l'étude

Dans le chapitre précédent, nous avons vu qu'il existe une multitude de sources potentielles de stochasticité de l'expression génique. De plus, ces sources interagissent entre elles, avec des effets potentiellement additifs, soustractifs ou de filtrage temporel. Or, la contribution précise de chacune de ces sources est aujourd'hui très largement spéculative. Notre objectif est, par une approche de modélisation, de quantifier la contribution de certaines de ces sources – idéalement les principales. Cependant, avant de pouvoir mettre en place une approche quantitative, il est nécessaire d'identifier les sources et de vérifier que leur contribution est suffisamment significative pour permettre leur quantification. Pour cela, nous allons utiliser une approche comparative : un même système (ici une population clonale de cellules 6C2) va être placé dans des situations expérimentales différentes et la stochasticité de l'expression génique (ainsi que la moyenne de cette expression) sera mesurée pour chacune de ces situations. Si la mesure de stochasticité (typiquement la variance normalisée NV) varie d'une situation à l'autre, alors nous aurons mis en évidence une source – ou un régulateur – de stochasticité (puisque, toute chose égale par ailleurs, le changement de condition expérimentale entraine un changement de stochasticité).
Nous nous intéresserons ici à deux sources de stochasticité potentielles : la position du gène (son "locus") sur le chromosome et l'état de la chromatine autour de ce locus. Pour étudier la première source de stochasticité, nous construisons des cellules 6C2 exprimant un rapporteur fluorescent via un gène inséré dans leur génome. En construisant plusieurs populations pour lesquelles ce gène n'est pas inséré au même locus, nous pouvons observer l'influence du locus sur la stochasticité : si les caractéristiques dynamiques (moyenne, variance normalisée et, éventuellement, distribution) de l'expression de ce gène changent suivant son locus d'insertion dans le génome, alors nous pouvons faire l'hypothèse qu'une caractéristique locale du chromosome influence directement ou indirectement la stochasticité de l'expression génique.
La deuxième source de stochasticité que nous étudierons ici est liée à la première puisqu'il s'agit d'une de ces caractéristiques locales, à savoir la dynamique de la chromatine aux alentours du gène rapporteur. Comme pour le locus, l'objectif va être ici de modifier la dynamique locale de la chromatine dans des populations clonales transfectées (contrairement à la situation précédente, les populations sont ici parfaitement clonales, y compris sur le plan du gène rapporteur). Là encore, si la modification de la dynamique chromatinienne entraine une différence de stochasticité, nous pourrons faire l'hypothèse que la dynamique de la chromatine influence directement ou indirectement la stochasticité de l'expression génique.
La modification de la dynamique chromatinienne sera effectuée à l'aide de deux drogues : la triscostatine A et la 5-azacytidine.

- La trichostatin A (TSA) inhibe les HDAC (Histone DéACétylases). Or, les HDAC, en désacétylant les histones, permettent à la chromatine de se recompacter. Leur inhibition va donc accroitre (mais dans des proportions inconnues) l'ouverture de la chromatine (Yoshida *et al.*, 1995).

- La 5-azacytidine (5-AzaC) est un inhibiteur de la méthylation de l'ADN. En remplaçant la cytosine de l'ADN par un équivalent non-méthylable, elle inhibe la faculté de la chromatine à passer dans un état compact (Veselý, 1985).

Ces deux drogues permettent, par deux moyens différents, d'agir sur l'état de compaction de la chromatine. Avant d'influer sur la dynamique chromatinienne à l'aide de ces drogues, nous avons en premier lieu vérifié qu'elles ne portent pas atteinte à l'intégrité cellulaire des populations traitées. Au final, la modification de la dynamique chromatinienne par deux moyens différents permet de modifier la stochasticité de l'expression génique ce qui démontre l'importance de la première sur la seconde.

Résumé des résultats

Pour cette étude, nous avons généré six populations de 6C2 (numérotées de A à F). Ces six populations clonales ne se différencient génétiquement que par le locus de la cassette insérée dans leurs génomes[1]. Cette cassette contient un promoteur CMV (CytoMégaloVirus) et un gène codant pour la protéine fluorescente rouge *mCherry*. La distribution de fluorescence des différentes populations a ensuite été mesurée par cytométrie en flux.
La première constatation est que chacune de ces populations exprime différemment la protéine *mCherry*. En effet, les distributions de fluorescence obtenues en cytométrie en flux pour chacunes d'entre elles sont clairement différentes. Leurs moyennes et, surtout, leurs variances normalisées sont très différentes différentes une population à l'autre. Celles-ci n'ont donc pas la même stochasticité d'expression du gène *mCherry*. Puisque que la seule différence entre ces populations est le locus d'insertion du gène rapporteur, nous pouvons en déduire que celui-ci a bien une influence forte sur l'expression génique et sa stochasticité.
Nous avons ensuite utilisé deux de ces populations pour évaluer l'importance de la dynamique chromatinienne locale au site d'insertion. Chacune des ces deux populations a été traitée avec les deux drogues décrites précédemment de façon à perturber la dynamique chromatinienne. Après avoir vérifié que l'intégrité des cellules n'est pas affectée par les traitements (les populations transfectées retrouvent une activité "normale" après interruption du traitement) on mesure les variation de dynamique dans les populations perturbées par la TSA et par la 5-AzaC. Nous avons ainsi pu constater que la moyenne d'expression augmente dans les populations traitées :

- Sous 5-AzaC, elle est augmentée après 48h de traitement.
- Sous TSA, la fluorescence moyenne est aussi augmentée après 48h de traitement et ce par un facteur supérieur à celui constaté en 5-AzaC.

On montre aisément que la variance normalisée des populations évolue elle aussi au cours du traitement. En revanche, contrairement à la moyenne, la variance normalisée diminue :

- Sous 5-AzaC, elle est diminuée.
- Sous TSA, elle est encore plus diminuée.

On a donc bien montré que le locus d'un gène influence la stochasticité de l'expression de ce gène et que cette influence est, au moins partiellement, due à la dynamique chromatinienne autour de ce locus.

[1]La méthode de transfection utilisée insère la cassette à un endroit aléatoire de la cellule

2 Towards experimental manipulation of stochasticity in gene expression

Abstract

For decades, most of molecular biology was driven by the "central dogma" in which the phenotype is defined by the genotype following a fully deterministic point of view. However, during the last 10 years, a wealth of studies has demonstrated that a given genotype can generate multiple phenotypes in identical environmental conditions, mainly because of the inherently probabilistic nature of the transcription process. It has also been shown that cells can tune this variability at the molecular level. Although previously described as a useless "noise", stochastic gene expression has now been shown by many authors to be an essential part of diverse biological processes. Chromatin dynamics having a central role in higher eukaryotes, we decided to investigate its involvement in the generation and control of stochasticity in gene expression (SGE). Our experiments reveal that the chromatin environment of a gene plays an important role in regulating SGE. Indeed, we find that histone acetylation and DNA methylation significantly affect SGE, suggesting that cells are able to adjust the variability of the expression of their genes through modification of chromatin marks. Given that the alteration of chromatin marks is itself subject to the expression of chromatin modifiers, our results shed light on a complex circular causality with on the one hand, the effect of gene expression on chromatin and on the other hand, the influence of the local chromatin environment of a gene on the dynamics of its expression.

2.1 Introduction

In most living cells, DNA is transcribed into RNA molecules, which are, in turn, translated into proteins. There are two sides to this validated assertion : (i) it constitutes the basis for the "central dogma" of molecular biology, (ii) the details of its molecular realisation are still far from being completely understood. Due to recent advances in molecular and cell biology, these two sides are now colliding.

The central dogma depicts this biochemical process as a decoding process where the 1-D information that is stored in the DNA (the genotype) is decoded into a 3-D realisation (the phenotype). Although the genotype-to-phenotype information flow has been shown to be much more complex in many occasions (see e.g.(Crick, 1970) where the author "reconciles" the central dogma with the discovery of reverse transcription, or more recently the reassertion that genes, although as parts of complex networks, do control the phenotype (Davidson, 2002)), one of the central consequences of the dogma is still strongly influent in biology. Indeed, it implies full determinism of the phenotype from the genotype and – in its modern version – from the environment. In such a vision, cells that harbour the same genotype (isogenic cell lines) and are placed in the same environment (even environmental information) should display the same phenotype (Fig. II.1).

Even though such a view has been repeatedly challenged for more than 40 years (Novick et Weiner, 1957; Spudich et Koshland, 1976; Rigney et Schieve, 1977; Berg, 1978; Kupiec, 1983), the full realisation of the biological impact of cell-to-cell variability has emerged only in the last 10 years through studies of stochasticity in gene expression (Elowitz

Figure 1

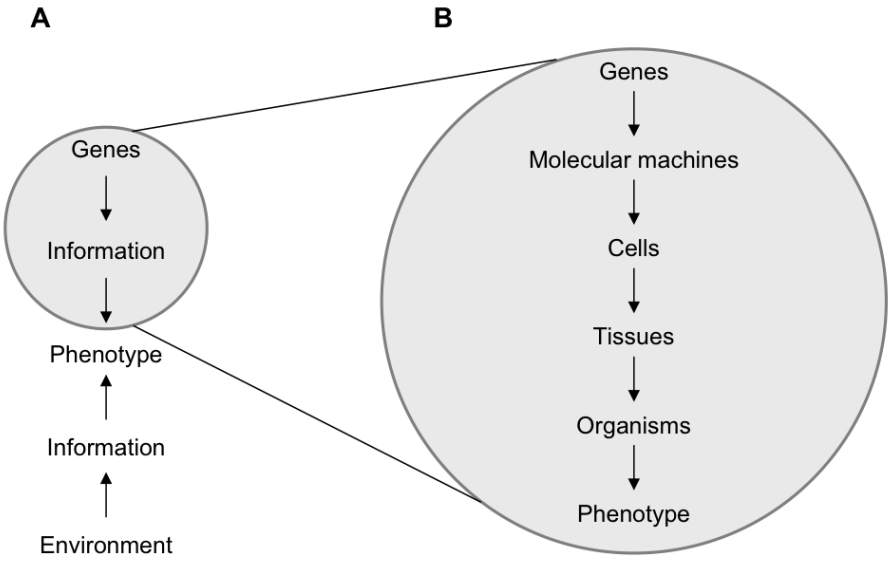

Figure II.1 – The traditional information-decoding view relating the genotype to the phenotype. Part B is a detailed view of the molecular steps involved in part A.

et al., 2002; Levsky *et al.*, 2002) ; (for recent reviews see (McCullagh *et al.*, 2009; Niepel *et al.*, 2009; Singh et Weinberger, 2009; Eldar et Elowitz, 2010; Huang, 2010)). Indeed, stochasticity in gene expression (SGE) generates dispersion of gene product concentrations within isogenic cell populations, inducing multiple phenotypes for a given genotype in a given environment.

It should be noted that stochasticity does not, by any means, imply complete randomness, or total unpredictability. Rather, constrained randomness, intermediate between rigid determinism and complete disorder is usually seen experimentally. One should therefore stress out that the stochastic/deterministic debate is confused by the improper use of those terms and that this misuse strongly biases any possible debate on those questions (Zernicka-Goetz et Huang, 2010). For example, in a recent publication (Snijder et Pelkmans, 2011), the term "deterministic" is used in the sense of the word "predictable" but has a much stronger meaning. A fully deterministic system can be entirely unpredictable (Bertin, 2012; Franceschelli, 2012; Le Bellac, 2012), whereas a probabilistic system can have a very stereotyped output, provided a sufficient number of realisations (e.g., a coin throwing experiment). It is not because such an output can be adequately described by a deterministic system, that its actual generating mechanism is deterministic. The key question in biology is precisely to understand how predictable and reproducible outputs (e.g., embryogenesis) can be generated from non-deterministic mechanisms (probabilistic molecular interactions) and, in turn, whether the stochastic nature of the underlying process contributes or not to this apparent determinism.

Since biology has always been dominated by deterministic theories, SGE raises many fundamental questions. So far, cell behaviour has been thought to be controlled by genetic

"programmes" (or their modern declinations : genetic networks) working with deterministic modulation of genes activity (often "on-off"). But, since SGE is an experimentally demonstrated fact, it is now necessary to accommodate the classical deterministic genetic programming theory with this new perspective. The central questions are therefore : How does the cellular order emerge from the molecular "chaos" ? (Levsky et Singer, 2003; Paldi, 2003; Kupiec, 2009). How and to what extent can this 'chaos" itself be modulated by the cell ? And, even more counterintuitively, how does variability participate in the robustness of the system ? (Rué *et al.*, 2012). This question can be derived at the molecular level : What are the mechanisms that do generate SGE and those that do regulate it ? Finally, the burning question concerns the biological function, if any, of SGE.
Regarding the possible biological function of SGE, one should first state that traditional molecular and cell biologists, under the influence of the 'central dogma", have often considered the 'noise" to be deleterious for cells (despite the well-known phenotypic influence of stochasticity, e.g., in the case of bet-hedging). But recent evidence has accumulated demonstrating that regulated stochasticity may play a relevant role in many biological phenomena including :

(i) a direct adaptive role, for example, by allowing bacterial strains to resist to temporarily unfavourable environmental conditions, such as the presence of antibiotic (Balaban *et al.*, 2004; Thattai et Van Oudenaarden, 2004; Stern *et al.*, 2007; Ackermann *et al.*, 2008; Beaumont *et al.*, 2009; Eldar *et al.*, 2009; Ito *et al.*, 2009; Zhang *et al.*, 2009) ;

(ii) a role in the cell cycle (Di Talia *et al.*, 2007) and circadian rhythms (Ullner *et al.*, 2009) ;

(iii) a role in decision making of the HIV virus (Weinberger *et al.*, 2005, 2008) and various prokaryotes (Maamar *et al.*, 2007; Çağatay *et al.*, 2009) that might be relevant for pathogenesis (Moxon *et al.*, 1994; Freed *et al.*, 2008; Singh et Weinberger, 2009) ;

(iv) a role during embryonic development (Arias et Hayward, 2006; Rossant et Tam, 2009) and in the process of cell differentiation (Kupiec, 1997; Hume, 2000; Paldi, 2003; Wernet *et al.*, 2006; Chang *et al.*, 2008; Halley *et al.*, 2008; Hoffmann *et al.*, 2008; Kalmar *et al.*, 2009; Wu *et al.*, 2009; Stockholm *et al.*, 2010) ;

(v) a role in the generation of cancer cells and variability in their sensitivity to treatment (Capp, 2005; Laforge *et al.*, 2005; Cohen *et al.*, 2008; Gascoigne et Taylor, 2008; Brock *et al.*, 2009; Mayburd, 2009). It has further been proposed that cancer phenotypes may result from inaccurate and aberrant patterns of gene expression generated by microenvironmental alterations (Capp, 2005) ;

(vi) a role in the ageing process (Bahar *et al.*, 2006) and during apoptosis (Spencer *et al.*, 2009).

The most clear-cut case is the stress response in Bacillus subtilis, where the transcriptional variability was shown to be positively selected for. Indeed, a recent study has demonstrated that SGE is used by this bacterium to increase its fitness in an uncertain environment (Çağatay *et al.*, 2009). Similar conclusions were also found in mycobacteria (Sureka *et al.*,

2008). These important works have demonstrated that SGE can be optimally tuned at the molecular level. It thus calls for further investigations on the molecular mechanisms at stake. Indeed, a number of molecular causes have been proposed for the generation and regulation of SGE, including :

(i) small molecule numbers (Paulsson, 2005b). Due to the existence of thousands of macromolecules present in low numbers per cell, individual chemical events, relying on collisions between randomly diffusing molecules, occur by chance and do induce SGE.

(ii) spatial aspects due to the random diffusion of molecules (Van Zon *et al.*, 2006). The dissociation/rebinding process of a repressor is a perfect example to illustrate this phenomenon. After dissociation of a repressor from the operator, it may rapidly rebind to the DNA. But, as explained by Van Zon *et al.* (2006), rebinding trajectories are so short that, on this timescale, the RNA polymerase cannot effectively compete with the repressor for binding to the promoter. Thus, a dissociated repressor molecule will rebind many times, which lowers the effective dissociation rate, and increases SGE. The crowded nature of the nuclear environment will add upon such a phenomenon (Roberts *et al.*, 2011).

(iii) the essential dynamic nature of protein-protein interactions (Coulon *et al.*, 2010). Considering the mere molecular interplay at a single promoter, we showed that a single gene can demonstrate an elaborate spontaneous stochastic activity. Based on modelling approaches, our results revealed that a periodic pattern of promoter occupancy by transcription factors and chromatin remodelling can underline a tight promoter-mediated control of SGE.

(iv) specific regulatory network architectures (Çağatay *et al.*, 2009). Analysing the negative feedback loop of B. subtilis competence circuit, based on comparative analysis of native and synthetic circuits, Cagatay et al. have found that gene circuit architecture determines the extent of the stochastic behaviour of gene expression. A vast body of literature investigates possible correlations between network wiring and the noise regulation property of that network (see e.g., (Kittisopikul et Süel, 2010) and for a recent review see (Chalancon *et al.*, 2012)).

(v) the non-specificity in protein-protein interactions (Kupiec, 2010). Due to their low level of specificity, showed in recent studies, proteins can interact with numerous molecular partners following large combinatorial interaction possibilities. Thus, molecular interactions are therefore intrinsically stochastic and will generate SGE.

(vi) properties of the RecA protein in *Escherichia coli* (Elowitz *et al.*, 2002) and of the transcriptional elongation machinery in *Saccharomyces cerevisiae* (Ansel *et al.*, 2008). By deletion/transduction experiments on the recA gene, involved in the rescue of stalled DNA replication forks, Elowitz et al. clearly demonstrated in *Escherichia coli* that lack of RecA significantly increases SGE. The progression of transcriptional elongation can also increase the level of SGE. Indeed, Ansel et al. showed that when elongating RNA polymerase II is stalled, expression of the corresponding messenger is blocked until transcriptional initiation takes place again, increasing SGE.

(vii) unequal repartition of the molecular content of the mother cell into the two daughter cells (Huh et Paulsson, 2011a). In this study, Huh and Paulsson have suggested that much of the cell-to-cell heterogeneity that has been attributed to various aspects of gene expression could as well come from random segregation events at cell division. More precisely, they showed that disorder in synthesis and degradation originates in random segregation, while disorder in segregation typically reflects spatial heterogeneity rather than randomness in synthesis and degradation. At least part of that effect may be due to unequal repartition of the mitochondria at division (Das Neves *et al.*, 2010), generating variation in intracellular concentration of ATP which in turns induces variation in RNA polymerase II processing speed (Johnston *et al.*, 2012).

(viii) chromosomal positioning of the genes in eukaryotic cells (Becskei *et al.*, 2005). According to Becskei et al., rare random events of gene activation, which is an important source of SGE, are determined primarily by the positioning of the genes along the chromosomes. Such a position effect could have been selected for during evolution (Chalancon *et al.*, 2012).

Considering the importance of the latter process in higher eukaryotic cells, we decided to investigate the role of chromatin dynamics in SGE. For this, we designed an experimental procedure to generate clones of chicken erythroid progenitors expressing a fluorescent reporter gene that has been integrated into genomic DNA in a single copy, positioned at a random locus (Fig. II.2).
In order to simplify the study, we only selected two clones showing a single reporter insertion site and different levels of SGE. By using flow cytometry measurement, we then analysed their spontaneous gene expression activity as well as their activity when perturbed by chromatin-modifying agents. A complementary version of this work harbouring an important part of chromatin dynamic modelling will appear elsewhere (Viñuelas* *et al.*, 2013). Here, we focus on the main results that we observed and draw conclusions regarding the consequences of our findings on a proposed circular causality involving SGE.

2.2 Materiel and methods

2.2.1 Cell culture

Experiments were performed on a chicken transformed erythroblast cell line (6C2) (Beug *et al.*, 1982) maintained, at a maximum density of 1×10^6 cells per ml, in alpha minimal essential medium (Gibco) supplemented with 10% (v/v) foetal bovine serum, 1% (v/v) normal chicken serum, 100 μM β-mercaptoethanol (Sigma-Aldrich), 100 units/ml penicillin and 100 μg/ml streptomycin (Gibco). Those cells are from an established cell line making the generation and selection of clones an easy task. We also have at hand a normal erythroblast cell type, names T2ECs (Gandrillon *et al.*, 1999), which will allow on the longer term to assess the validity of our conclusions in fully normal primary cells.

2.2.2 Generation of stably transfected clones

The generation of 6C2 clones expressing a fluorescent reporter stably integrated in the genome was performed by a nucleofection of the pT2.CMV-mCherry/pCAGGS-T2TP plas-

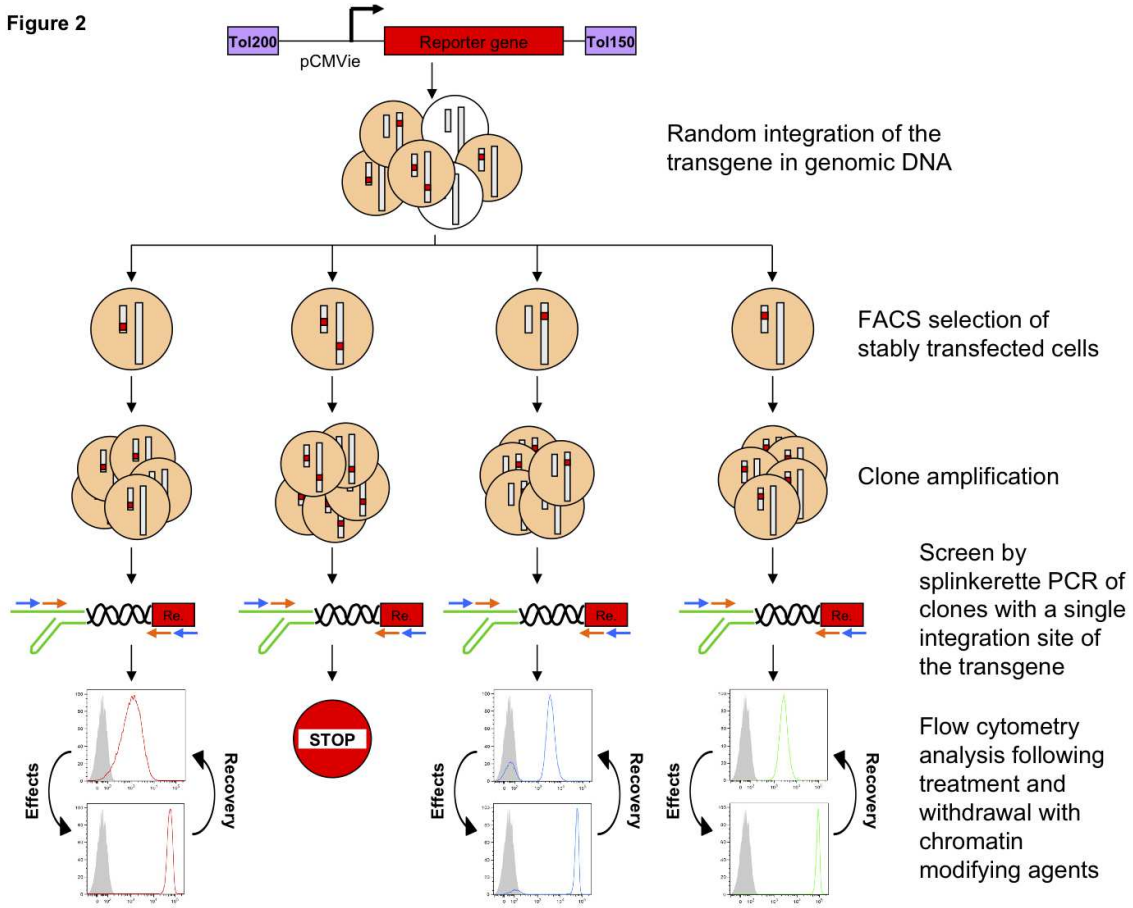

Figure II.2 – Experimental strategy used for assessing the role of the local chromatin environment on stochasticity in gene expression. After generation and sorting of clones that were stably transfected with the reporter and that showed, by splinkerette PCR, a single insertion site of the transgene, the effects of two chromatin modifying-agents were analyzed by flow cytometry. FACS : florescence activated cell sorting.

mid mix (ratio 5 :1) using the Cell line Nucleofector® Kit V (Lonza) in a Nucleofector™ II (Amaxa Nucleofector™ Technology) (T-16 program) as previously described (Mejia-Pous *et al.*, 2009). The stable genomic integration of the reporter, flanked by Tol2 motifs, is allowed by a transposase (pCAGGS-T2TP plasmid) that recognises these motifs and inserts the trangene into genomic DNA by a "cut and paste" strategy (Kawakami, 2007). After elimination of non-integrated transgenes (by dilution after 7 days of culture ; (Mejia-Pous *et al.*, 2009)), stably transfected cells, expressing red fluorescence, were sorted and individually cloned in U-shape 96-well microplates (Cellstar Greiner bio-one) using a FACSVantage SE cytometer (Becton-Dickinson). Note that transfected clones harbour the fluorescent reporter – at least in one copy – but that the insertion points are randomly distributed along the chromosome. Thus, all the clones are identical (same global genome, same promoter for the reporter gene, . . .) except on the precise locus of the reporter (i.e., the local chromatin context).

2.2.3 Molecular characterisation of clones

To only select clones with a single reporter insertion, the transgene integrations were characterised by splinkerette PCR as previously described (Mejia-Pous *et al.*, 2009). Briefly, for each stably transfected clone, genomic DNA was extracted, purified and digested overnight by a 4 bp recognition site enzyme (in our study TaiII). A ligation of splinkerette adaptors was then done for 1 h (Devon *et al.*, 1995), followed by nested PCR reactions using primers that are specific for the transgene mCherry and for the annealed splinkerette adaptor. After purification and sequencing of the PCR products, the genomic reporter insertion sites were identified by similarity searches using the sequence analysis tool iMapper (Kong *et al.*, 2008).

2.2.4 Cellular characterisation of clones

Clone fluorescence analyses were performed by flow cytometry with a FACSCanto II flow cytometer (Becton-Dickinson) on cells extemporaneously pelleted and resuspended in Dulbecco's Phosphate Buffered Saline 1 × solution (Gibco). Each sample was analysed from an acquisition of 50,000 events (gated on living cells – same gate for all experiments) and the positive fluorescence threshold was set using non-transfected cells. For each acquisition, an external calibrator (SPHERO^{TM} Rainbow Calibration Particles ; Spherotech) was systematically analysed in the cytometer, in order to compare experimental results throughout time. Flow cytometry data were extracted and analysed using FlowJo 8.8.6 software. The mean and variance of fluorescence intensity were then computed for each clone and adjusted using a non-transfected cell population to correct for cells autofluorescence. Finally, using these two parameters, the normalised variance (NV, i.e., variance/square mean), a robust estimator of SGE (Tao *et al.*, 2007) was computed.

2.2.5 Treatments with chromatin-modifying agents

To analyse the effect of chromatin on stochastic gene expression, 6C2 clones were treated with trichostatin A (TSA, a histone deacetylase inhibitor ; P5026 Sigma-Aldrich) and with 5-azacytidine (5-AzaC, an inhibitor of DNA methylation ; A2385 Sigma-Aldrich) respectively at a final concentration of 500 nM and 500 μM. In these experiments, a kinetic treatment composed of five time points (0 h, 8 h, 24 h, 32 h and 48 h) was performed for the two drugs. For each of them, 1 × 106 (for 0 h, 8 h, 24 h) or 5 × 105 (for 32 h, 48 h) cells were treated and characterised by flow cytometry.

To evaluated the potential reversibility of the response after drug treatment and therefore after modifications of chromatin marks, 8 × 106 cells for each clone were treated with each of the two drugs during 48 h prior to being rinsed and grown in drug-free medium. One day, 2 days, 5 days, 6 days and 9 days after the end of treatment, flow cytometry analyses were performed as described above. These experiments were then complemented for one clone, by another kinetics with additional time points at 3 and 4 days after the end of 5-AzaC treatment.

2.3 Results

2.3.1 Chromatin environment and stochastic gene expression

Figure 3

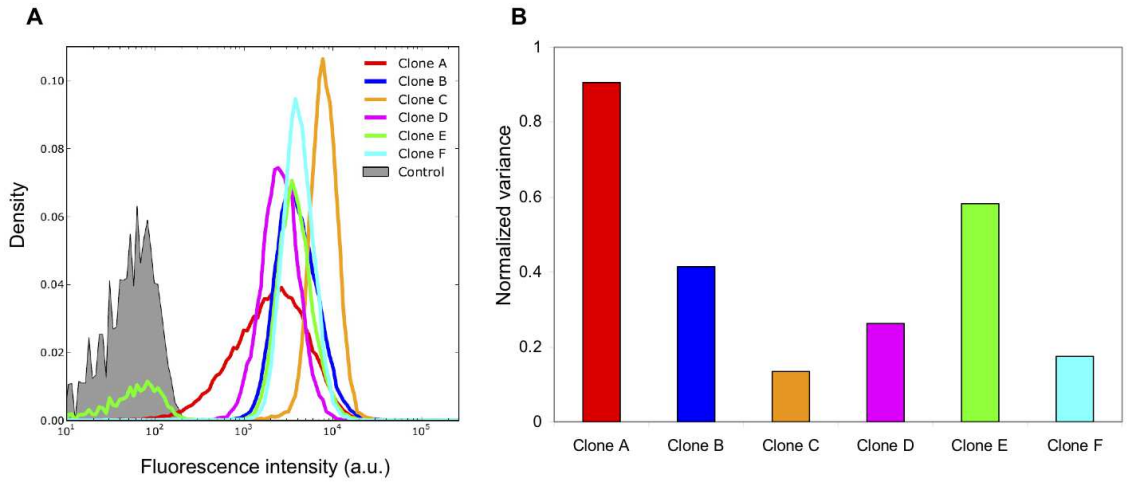

Figure II.3 – Illustration of the impact of the local chromatin environment on stochastic gene expression. Distributions of fluorescence (A) of six 6C2 clones having the reporter gene integrated at different loci, and their corresponding normalized variances (B) are shown. In panel A, cellular autofluorescence is shown in grey.

To analyse SGE, we first measured by flow cytometry the expression of the mCherry fluorescent reporter in several clones that displayed a single insertion of the transgene (Fig. II.3 A). These clones harbour identical transgenes, controlled by identical CMV promoters, and hence only differ by the genomic location of the reporter – and therefore both by its surrounding DNA sequences as well as its chromatin context. In this study, we focussed our attention on chromatin context since it is easily accessible to experimental manipulation through the use of global chromatin modifiers (see below) and was therefore privileged as the first analysis level. A dedicated study analysing relationships between genomic DNA sequences and SGE is currently in preparation in our group (Kaneko et al., in preparation ; chapter IV of this thesis). Fig. II.3 A and B show that this difference in genomic location is sufficient to generate distributions of fluorescence that are very different in terms of mean, normalised variance, and, more globally, considering the full distribution profile (a critical information to understand the SGE in eukaryotes ; (Raj *et al.*, 2006)). As expected, each clone displays a different mean fluorescence intensity (MFI), but clones also show different levels of variability (i.e., expressed as by the normalised variance, NV) (Fig. II.3 B). Remarkably, this mere statement could be opposed to the full determinism dogma. Since cells differ only by the chromatin context of the reporter, this result demonstrates the existence of a genomic/chromatin effect on SGE in higher eukaryotic cells and supports

previous studies suggesting this hypothesis (Raj *et al.*, 2006; Voss *et al.*, 2009; Singh *et al.*, 2010a; Skupsky *et al.*, 2010; Harper *et al.*, 2011; Suter *et al.*, 2011). These differences are examples of "position effect" on SGE. It is therefore quite clear that the position of a gene on the genome has a profound influence on both its mean level of expression but also on its level of non-genetic cell-to-cell variability.

2.3.2 Chromatin dynamics and stochastic gene expression

One question that is still open at that stage is to what extent the clone-to-clone differences were actually related to the dynamics of the chromatin structure surrounding the reporter, or to some other feature that might be locally different on the genome. To analyse the effect of chromatin dynamics variations, we decided to treat two different clones showing respective high and moderate levels of SGE (clones A and B ; Fig. II.3), with two different global chromatin-modifying agents : (i) trichostatin A (TSA), which acts as a histone deacetylase inhibitor (Yoshida *et al.*, 1995) and (ii) 5-azacytidine (5-AzaC), which acts as an inhibitor of DNA methyltransferases (Veselý, 1985). Both of these substances are know to influence chromatin dynamics by producing a more open configuration. Fluorescence distributions were measured at different time points during the treatment phase – i.e., following drug addition – and the recovery phase – i.e., following drug withdrawal.

Global effect of chromatin-modifying agents on distributions of fluorescence

Treatment phase Fig. II.4 shows the evolution of the fluorescence distributions after 8 h, 24 h, 32 h and 48 h of treatment with TSA (panel A) and 5-AzaC (panel B). For the two clones, TSA treatment induces very similar gradual shifts of the distribution towards higher values of fluorescence, starting from the 8 first hours of treatment (Fig. II.4 A). Regarding the 5-AzaC treatment (Fig. II.4 B), a global shift of the distributions towards higher values of fluorescence is also observed. However, in contrast with TSA treatment this shift appears and stabilises in the very first hours of the 5-AzaC treatment. Yet, 5-AzaC treatments reveal different fluorescence distribution profiles on the two clones. While clone B displays very similar distributions over the treatment, the distribution of clone A gradually sharpens indicating a decrease of variability.
Finally, the global shift of distributions towards high levels of fluorescence observed for the two treatments could be explained by several studies which showed that inhibitors of histone deacetylase and DNA methyltransferase lead to the opening of chromatin, increasing the probability of transcription factors binding and promoter activation (Ghoshal *et al.*, 2002).

Recovery phase Then, we investigated the possible reversibility of the effects observed after TSA and 5-AzaC treatments on the stochastic expression of the transgene.
For TSA treatment, after cell rinsing, a reversibility effect appears suggesting a progressive closing of chromatin with a significant effect on SGE. For the two clones, the distributions gradually shift towards lower fluorescence values, showing a nearly total reversibility after 5 days of culture (Fig. II.5 A). These results corroborate the study of Toth et al. where they showed an almost complete reversibility of TSA treatment effect on chromatin decondensation after removal of the drug (Tóth *et al.*, 2004).

Figure 4

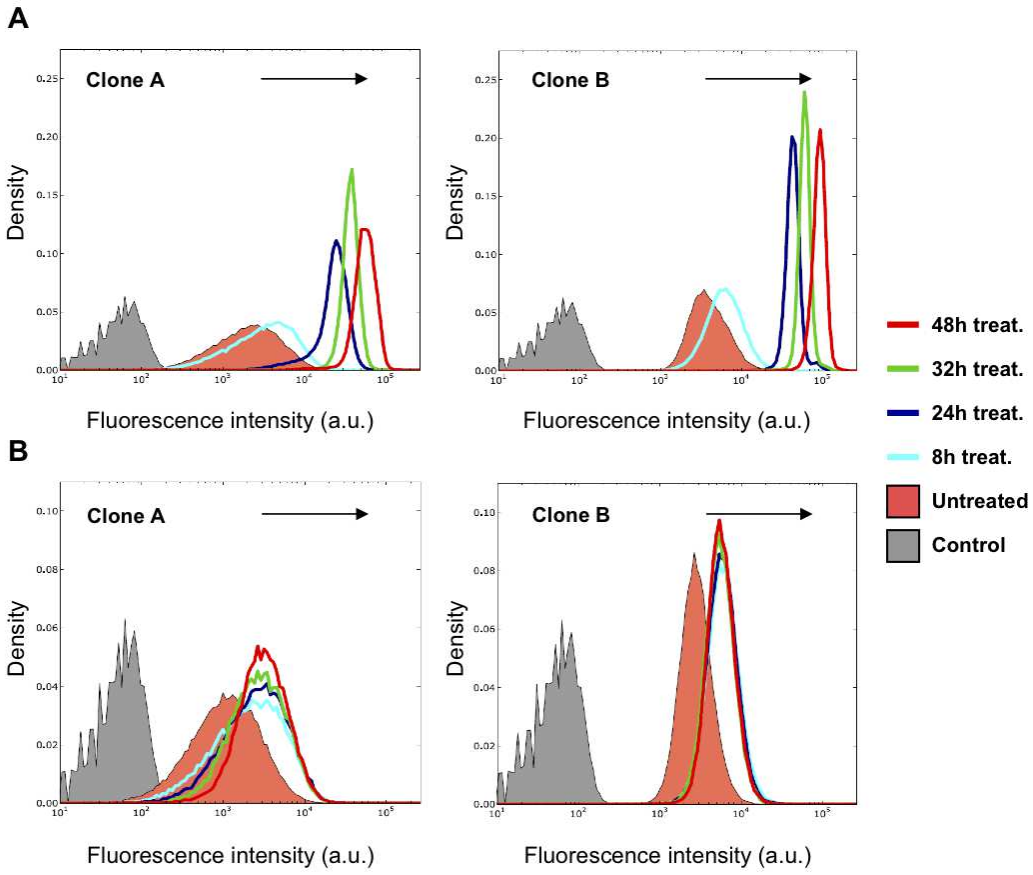

Figure II.4 – Effects of treatments with chromatin modifying-agents on the fluorescence distributions for two clones having the reporter gene integrated at different loci. Clones were treated with trichostatin A (A), and with 5-azacytidine (B), followed over time. For each of them, cell autofluorescence and untreated condition are shown respectively, in grey and in pink. Arrows indicate the global shift direction of distributions during the treatment phase.

Similarly, concerning the reversibility of the 5-AzaC treatment, a total reversibility seems to be possible 5 days after the end of treatment (Fig. II.5 B). Indeed, such reversible effects of cytidine analogue treatment on gene expression were previously observed. For example, in the study of McGarvey et al., the authors showed that a silenced DNA-hypermethylated gene can be activated by 5-aza-2'-deoxycytidine treatment of cells and can return to a gene silencing state once the drug is removed (McGarvey, 2006). But the most intriguing observation, in our study, is the distributions first progress towards higher fluorescence values 1 and 2 days after the withdrawal of the drug before decreasing back to levels observed prior to treatment (Fig. II.5 B). In order to detect the beginning of the recovery, we performed an additional kinetics of recovery, measuring the reversibility 1, 2, 3, 4 and 5 days after the end of a 5-AzaC treatment for clone A.

Figure 5

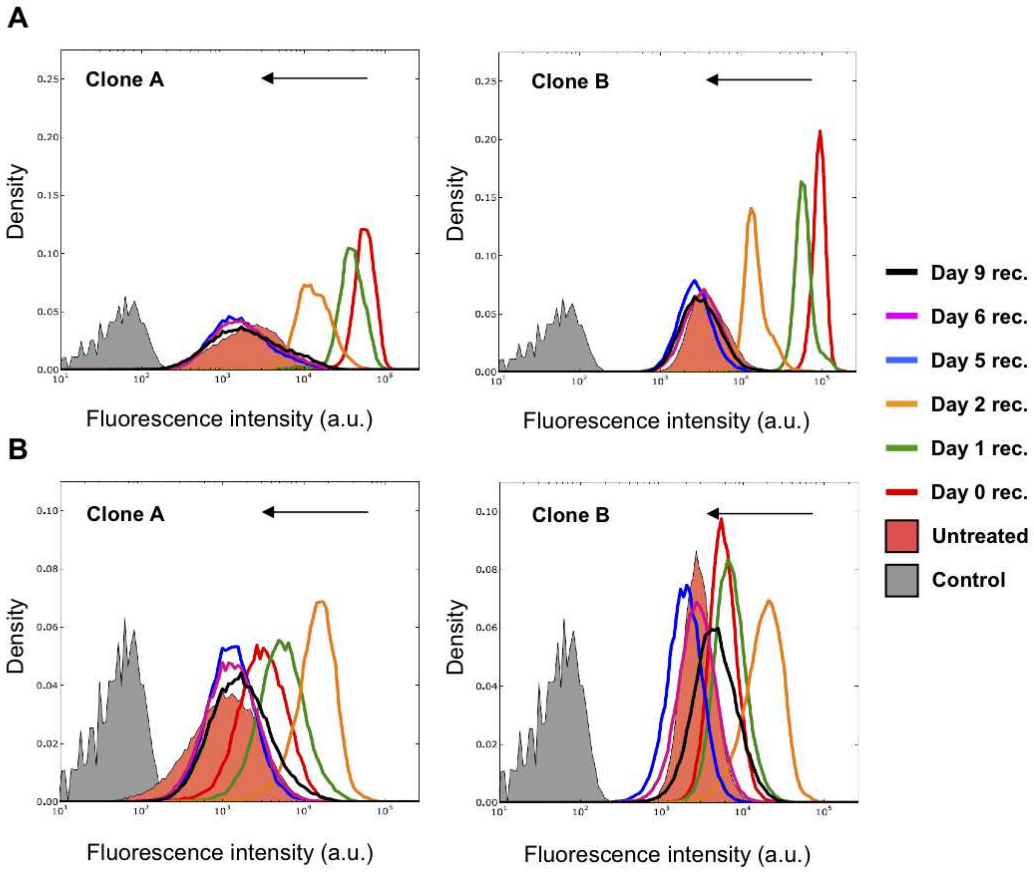

Figure II.5 – Recoveries of fluorescence distributions following withdrawal of chromatin modifying-agents for two clones having the reporter gene integrated at different loci. After 48 hours of trichostatin A treatment (A), and of 5-azacytidine treatment (B), cells were rinsed and analyzed by flow cytometry at different time points during nine days. For each clone, cell autofluorescence and untreated condition are shown respectively, in grey and in pink. Arrows indicate the global shift direction of distributions during the recovery.

This experiment demonstrated that the decrease starts from the third day after the end of treatment (Fig. II.6). Different explanations of this observation can be proposed. First, it can be due to the mechanism of action of 5-AzaC. This pyrimidine is an unmethylable analogue of cytidine and is incorporated into DNA. Consequently, DNA needs more than two duplications to eliminate all 5-AzaC, which could explain the delay between the end of treatment and the beginning of the decrease, which roughly correspond to two cell cycles. However, this result does not explain why the fluorescence levels increase for 2 days after the drug rinsing. It can also be proposed that 5-AzaC has multiple effects on the cell. Indeed, we could suggest that the opening of chromatin following a 5-AzaC treatment is delayed compared to that induced by TSA. This hypothesis is supported by the limited effect of 5-AzaC treatment during 48 h (Fig. II.4 B). A longer 5-AzaC treatment would

Figure 6

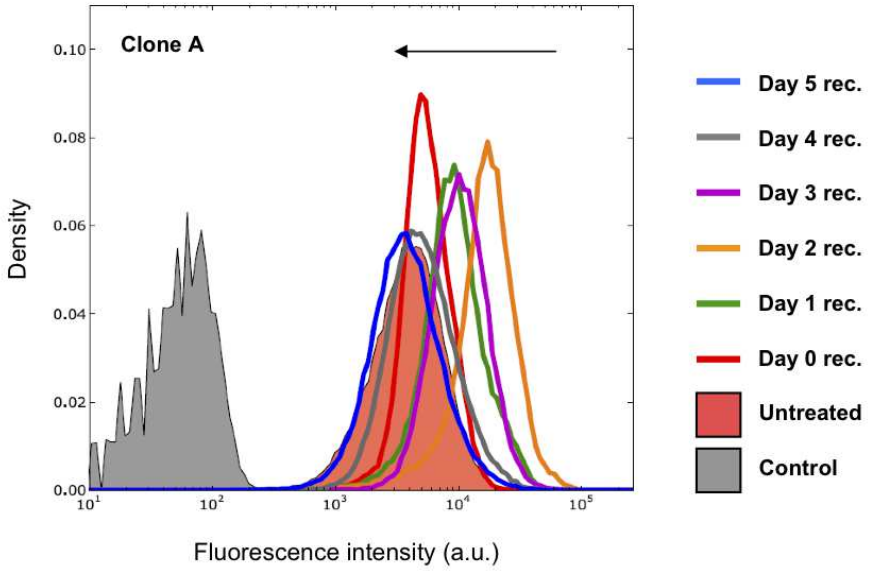

Figure II.6 – Complementary analysis of chromatin modifying-agents post-treatment recovery on fluorescence distributions. For one clone (clone A), the kinetic measures of 5-azacytidine treatment recovery were completed with two additional time points (day 3 and day 4). Cell autofluorescence and untreated condition are shown respectively, in grey and in pink. Arrows indicate the global shift direction of distributions during the recovery.

probably bring elements of response.

Effects of TSA versus 5-AzaC on mean and variability of expression levels
In terms of MFI, the patterns observed for the two clones during TSA treatment are very similar (Fig. II.7 A). A continuous increase of MFI is observed during the treatment phase, whereas a decrease immediately follows the drug withdrawal, ending with a nearly total recovery of the initial MFI value. Similarity of clone responses is also observed for 5-AzaC treatment, although the specific response to this treatment is different to that of TSA. Yet MFI is quite stable during the 5-AzaC treatment phase and significantly increases only after the rinsing (during 2 days), before returning to the initial value. Comparing the patterns observed between TSA and 5-AzaC, we could suggest that the resulting opening of chromatin, induced by these drugs, takes more time for 5-AzaC than for TSA. However, this hypothesis needs further investigations to be confirmed or refuted. Regarding the impact of these two drugs on NV, TSA and 5-AzaC treatments induce significantly different effects (Fig. II.7 B). The evolution of NV during the TSA treatment is very similar between clones and seems to be inversely related to the MFI evolution : NV gradually decreases during the treatment phase and increases progressively during the recovery period to reach the initial values and, for clone A overtakes it (note that however, the ranks of the two clones in terms of NV are the same at the beginning and at

Figure 7

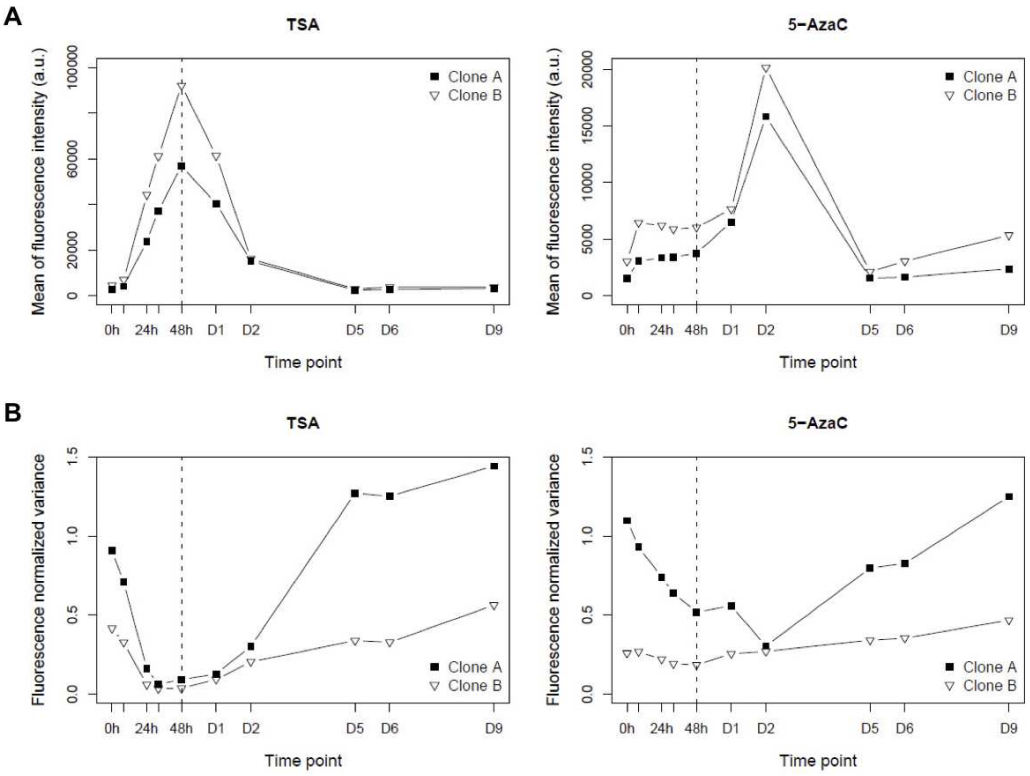

Figure II.7 – Evolution of mean and normalized variance fluorescence intensity during treatments of two clones with chromatin modifying-agents. For each clone, the mean (A) and the normalized variance (B) of fluorescence intensity after additional and withdrawal of trichostatin A and 5-azacytidine are shown over time.

the end of the experiment : clone A is more stochastic than clone B). It is known that the mean and the NV of gene expression tend to display an inverse relationship (Bar-Even *et al.*, 2006; Newman *et al.*, 2006; Neildez-Nguyen *et al.*, 2008), in line with the idea that low proteins abundance is an important contributor to the SGE (Paulsson, 2005b). This relationship is roughly observed in the progress of the TSA treatment and recovery.

The effects of 5-AzaC treatment on NV are very different than those observed for TSA. First of all, the kinetic evolution of the NV qualitatively differs from clone-to-clone, suggesting that the drug perturbs differently the two chromatin environments (Fig. II.7 B). NV for clone A decreases during all treatment phase and during 2 days after the drug withdrawal, before linearly increasing at the end of the experiment. On the opposite, NV for clone B is quite stable with a slight upward global trend. Moreover, in opposition with TSA treatments, in the case of 5-AzaC the variation of NV does not mirror MFI values. Finally, although the two treatments induce different responses in terms of NV, as for TSA, the ranks of NV values of the two clones are also conversed at the beginning and at the end of the 5-AzaC treatment : clone A is more stochastic than clone B.

2.4 Discussion

In this work, we have demonstrated that local chromatin environment could play a pivotal role in the level of SGE in higher eukaryotic cells, as proposed earlier (Kupiec, 1997; Neildez-Nguyen *et al.*, 2008). Moreover, modifications of chromatin marks by different drugs affecting chromatin dynamics induce different effects on the variability of expression although effects on MFI are quite similar. While NV seems to be inversely related to MFI for TSA treatments, the kinetic patterns of these two measures during 5-AzaC treatment appear much more independent. This shows that the regulatory mechanisms involved can control both the mean and the variability of gene expression, furthermore sometimes independently (as observed here in the 5-AzaC treatment ; Fig. II.7). However, the intrinsic properties of the chromatin environment of the reporter seem to be conserved after strong perturbations of chromatin state. Indeed, measures such as MFI, NV and the shape of fluorescence distributions tend to return to their initial values after the end of treatments with chromatin-modifying agents. This result shows a full reversibility of the cellular system after important modification of the chromatin state. Therefore, it is plausible that cells are able to temporally modify their level of SGE via histone acetylation or DNA methylation, possibly through metabolism modifications (Kupiec, 1997; Paldi, 2003; Wellen *et al.*, 2009), before returning to their initial physiological state.
Recent studies have described the biological importance of SGE in the generation of cancer cells and in their sensitivity to treatment (Capp, 2005; Laforge *et al.*, 2005; Cohen *et al.*, 2008; Gascoigne et Taylor, 2008; Brock *et al.*, 2009; Mayburd, 2009). It has, for example, been demonstrated that a drug-resistant subpopulation (arising non-genetic heterogeneity) could be ablated by using chromatin-modifying agents including TSA (Sharma *et al.*, 2010). Several studies also discussed therapeutic effects of inhibitors of DNA methyltransferases (DNMTs) and of histone deacetylases (HDACs) in cancer (Christman, 2002; Caffarelli et Filetici, 2011; Takai *et al.*, 2011), and clinical studies were performed with treatments using combination of DNMTs and HDACs inhibitors (Rudek *et al.*, 2005; Yang *et al.*, 2005). Our work links all of these aspects and suggests that epigenetic factors could be relevant targets to modify SGE, opening new perspectives for cancer therapies. One could for example use a TSA treatment for a short period of time in order to reduce the molecular variability, for a specific molecular target (say an oncogenic kinase) while treating simultaneously with a specific kinase inhibitor. Of course, this could come at the expense of a mean higher expression of the kinase, but the risk/benefit could be in favour of such a strategy, if properly handled.
Finally, what we have seen here is a complex regulation mechanism in which chromatin is involved in regulating both the mean and the variability of gene activity. Since chromatin components are themselves products of – stochastic – gene transcription, this opens the way for a study of the circular causality of gene transcription variability. Such a circular causality would be composed of two intertwined causes :

(i) a gene-to-chromatin causality where the activity of lower parts of the system (here genes, which expression modifies protein concentrations, and hence protein interactions) affects the upper levels (here the state of chromatin) ;

(ii) a chromatin-to-gene causality, whereby a very complex dynamical structure (the local chromatin environment) that is composed of a myriad of different proteins influences

the dynamical behaviour of its simpler molecular brick, the gene, and regulates both its mean expression level and its cell-to-cell non-genetic variability.

Of course, this circularity can only be induced when the genes, that are perturbed by the chromatin state, are the ones that modify (directly or indirectly) the chromatin properties. Clearly this is not the case in this study where we used a transgene and an exogenous promoter. However, the regulation of gene MFI and NV observed here leads us to propose the circular model shown in Fig. II.8.

Figure 8

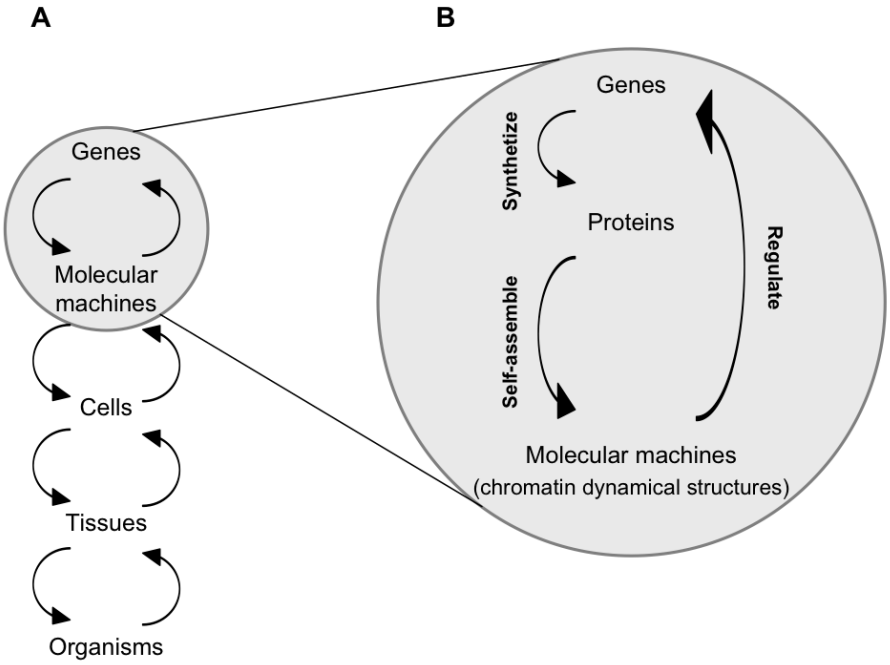

Figure II.8 – An alternative view of causality in biological systems. Although the emergence from local interactions is admittedly a reality, we also need to envision that part of the information is encoded into higher-level structures that impose specific constraints on the lower levels.

One of the consequences of this circular model is that it provides feedback loops by which, in spite of the chaotic nature of the underlying levels, the cell can control its mean transcription activities and its variability. Although our work has been focussed on trying to understand a very specific level in biological systems (genes to chromatin), we believe that such a circular causality can be demonstrated at both higher and lower levels of organisation as already proposed by other authors (see e.g., (Kupiec, 2010) ; on the higher order selection of protein-protein complexes, and (Noble, 2007) ; for the importance of the physiological downward causation on gene expression). Such a circular causality will provide the cell with many regulation loops and ultimately with a fine-tuning of its phenotype and phenotypic variability. This hypothesis could explain our results about the full reversibility of the cellular system, observed after the removal of

chromatin-modifying agents. Indeed, regulation of chromatin state, for example by the cell metabolism (Kupiec, 1997; Paldi, 2003, 2012; Wellen *et al.*, 2009), could be considered as a part of a circular causality process, but at a higher level than that observed about the influence of local chromatin environment on gene expression. Testing such a hypothesis would require to modulate the metabolism and to measure its influence on chromatin dynamics in the cells. Furthermore, since selection acts neither on genes nor on chromatin structures but on the result of their dynamical circular interactions, it may be argued that it has retained structures that do generate sufficient constraints on the stochastic lower levels to generate an output that is both predictable enough and variable enough to be compatible with life, a process that indeed was born and maintained from variation.

Acknowledgements

We thank François Chatelain, Alexandra Fuchs and Manuel Théry for helpful discussions and support during the early stages of the project. This work was supported by fundings from the Institut rhônalpin des systèmes complexes (IXXI) and from the Réseau National des Systèmes Complexes (RNSC). Part of the project was also supported by an ANR grant (ANR 2011 BSV6 014 01). JV is supported by a CNRS post-doctoral grant and GK is a PhD fellow from the Région Rhône-Alpes.

3 Conclusion

Cette étude montre que le point d'insertion du gène rapporteur influence fortement la stochasticité de l'expression de ce gène. De plus, deux traitements de la dynamique chromatienne sensés contraindre la chromatine à rester décompactée influent eux aussi sur cette stochasticité. Enfin, puisque, après arrêt du traitement, les cellules retrouvent progressivement la dynamique qu'elles avaient avant traitement, on peut supposer que la dynamique chromatinienne peut varier au cours de la vie de la cellule sans conséquence majeure sur son intégrité. Ce résultat complémentaire implique que la cellule a la capacité de modifier la stochasticité de l'expression de ces gènes en modifiant la dynamique chromatinienne aux alentours de ceux-ci. Il est alors possible que certaines protéines exprimées par la cellule soient impliquées dans la modification de l'état chromatinien de certaines zones du génome (potentiellement leur propre gène). On voit dès lors que la stochasticité n'est plus un simple bruit mais devient (ou peut devenir) un levier sur lequel la cellule peut agir si nécessaire et qu'elle peut réguler suivant ses besoins.
Un deuxième enseignement se dégage de ces résultats :
En théorie, si la chromatine est plus longtemps décompactée, les gènes présents dans la zone decompactée sont plus accessibles à la machinerie cellulaire de transcription. En conséquence, l'expression génique devrait augmenter. C'est bien ce que l'on observe ici dans le cas du traitement par la TSA ou la 5-AzaC. Néanmoins, le mode d'action exact de ces deux drogues n'est pas connu. De plus, observer la dynamique chromatinienne en temps réel dans des cellules vivantes est aujourd'hui impossible expérimentalement. On ne sait donc pas exactement à quel point la chromatine reste décompactée sous l'influence de ces drogues. On ignore aussi si la chromatine a la possibilité de se recompacter après traitement et à quelle vitesse. Les résultats présentés ici ouvrent donc au moins une nouvelle piste de recherche :

- En comparant les distributions de fluorescence des populations transfectées, traitées ou non par les drogues, nous pouvons chercher à estimer les caractéristiques locales de la dynamique chromatinienne d'une part et l'influence des drogues sur celle-ci d'autre part.

C'est ce programme scientifique que nous essayerons de suivre dans le prochain chapitre.

Chapitre III

Quantification de la dynamique chromatinienne à partir de sa contribution à la stochasticité de l'expression génique

Ce travail a fait l'objet d'une publication en 2013 dans la revue *BMC Biology* (11 : 15) :

Quantifying the contribution of chromatin dynamics to stochastic gene expression reveals long, locus-dependent periods between transcriptional bursts.

Viñuelas*, José and Kaneko*, Gaël and Coulon, Antoine and Vallin, Elodie and Morin, Valérie and Mejia-Pous, Camila and Kupiec, Jean-Jacques and Beslon, Guillaume and Gandrillon, Olivier

* : These authors contributed equally to this work.

Le texte de ce chapitre correspond intégralement au texte de cette publication, à l'exception des figures et de leurs légendes qui ont été remaniées pour des besoins de mise en page.

1 Principe de l'étude

Dans le chapitre précédent, nous avons vu qu'il était possible d'observer l'influence du locus d'un gène rapporteur sur la stochasticité de son expression. De plus, en traitant la cellule par des drogues modifiant la dynamique chromatinienne, nous avons pu modifier la stochasticité de l'expression du gène rapporteur. Ces résultats nous ont conduit à formuler l'hypothèse selon laquelle la dynamique chromatinienne locale au locus d'un gène serait un des principaux déterminants de la sa stochasticité.
Dans ce chapitre, nous allons utiliser la modélisation pour passer de cette observation à la *quantification* du lien entre dynamique chromatinienne et stochasticité. Pour cela, nous allons construire un modèle de la chaîne de transcription-traduction du gène rapporteur, coupler ce modèle à un modèle de la dynamique chromatinienne (voir chapitre I section 3) puis, par une exploration systématique des paramètres de ce modèle couplé, rechercher les ensembles de paramètres expliquant le mieux les dynamiques observées.
À cette fin nous devons construire un modèle du processus afin qu'il puisse exprimer au mieux les hypothèse formulées précédemment quant aux sources potentielles de stochasticité de l'expression génique. Ce modèle devra donc être :

- suffisamment simple pour que l'exploration paramétrique soit possible et pour permettre l'interprétation des résultats ;

- suffisamment complexe pour arriver à capturer la dynamique mesurée expérimentalement (moyenne, variance normalisée mais aussi distribution de la fluorescence). Sa complexité devra aussi permettre de rendre compte des deux caractéristiques observées précédemment, à savoir la dépendance au locus d'insertion du transgène et la dépendance à la dynamique d'ouverture-fermeture de la chromatine.

En résumé, un modèle trop simple (ou, au contraire, trop complexe) ne nous apprendrait rien sur l'objet modélisé car le lien de l'un à l'autre sera trivial ou, au contraire impossible à établir.
Cette "règle" élémentaire de la modélisation nous permet d'identifier les éléments qui devront nécessairement être présents dans le modèle. Comme la source de stochasticité étudiée sera la dynamique chromatinienne mais que la mesure de la stochasticité sera la fluorescence, il est nécessaire d'inclure dans le modèle la chaîne de transcription-traduction qui permet de passer de l'un à l'autre. En outre, cette chaîne devra être modélisée suffisamment finement pour pouvoir tenir compte de son influence sur la stochasticité de l'expression du gène. Enfin, pour modéliser la dynamique chromatinienne et l'influence des drogues sur celle-ci, le modèle devra inclure un modèle élémentaire de transition entre un état ouvert et un état fermé de la chromatine. L'effet des drogues correspondra alors à la modification des paramètres de ce modèle chromatinien.
Le deuxième enseignement est lié au mode d'implémentation du modèle. En effet, ce modèle devra permettre d'exprimer *in silico* des données de même nature que celles mesurées *in vivo*. Comme nous avons accès aux distributions de fluorescence dans différentes populations, le modèle devra permettre de produire de telles distributions. Or, cela n'est pas possible avec un modèle analytique. Nous devrons donc pouvoir simuler le modèle de façon à reproduire les distributions de fluorescence observées. Pour ce faire nous utiliserons l'algorithme SSA (voir chapitre I section 3.2.1). En répétant un grand nombre de fois la

simulation, nous pourrons générer des distributions simulées qui seront ensuite comparées aux distributions mesurées.

Résumé des résultats

En réutilisant les six populations clonales étudiées au chapitre précédent, nous constatons qu'il n'y a pas de relation évidente entre la moyenne de fluorescence $\langle P \rangle$ et la variance normalisée NV dans les données mesurées[1]. Or, un modèle de promoteur Poissonnien devrait conduire à une dynamique telle que $NV = 1/\langle P \rangle$ (Paulsson, 2005a). Compte tenu de l'absence d'une telle relation dans les données, nous avons choisi de modéliser le promoteur à l'aide d'un modèle à deux états de type "random telegraph" (voir chapitre I section 3.3.2). Ce modèle peut être interprété de la façon suivante : le promoteur CMV présent sur le chromosome en amont du gène rapporteur peut être accessible ou non par la polymérase (et, implicitement, par l'ensemble de la machinerie transcriptionnelle). Cette accessibilité conditionnelle dépend de l'état de la chromatine qui peut être ouverte (promoteur accessible, le gène peut être transcrit suivant un processus Poissonnien) ou fermée (promoteur inaccessible, le gène ne peut pas être transcrit). Il est important de mentionner qu'à ce stade, ce choix de niveau de complexité est empirique. Si nous n'étions pas, par la suite, parvenus à reproduire suffisamment fidèlement les données expérimentales avec ce modèle, nous en aurions augmenté la complexité.
Grace à une exploration paramétrique, nous avons pu identifier les paramètres permettant d'identifier les distributions expérimentales au moyen du modèle *in silico* pour cinq des six clones. L'étude du sixième clone, dont la distribution était bi-modale, nous a par la suite montré qu'une partie de la population avait perdu le transgène suite à une mutation. L'histogramme bimodal est donc issu de la superposition d'une sous-population n'exprimant que de l'autofluorescence (mode de faible intensité) et d'une population exprimant le transgène (mode de forte intensité). On notera que l'identification des distributions n'était parvenu, sur ce clone, qu'à reproduire la modalité forte.
Les résultats paramétriques obtenus pour ces cinq clones montrent qu'une différence de dynamique chromatinienne (les taux d'ouverture et de fermetude de la chromatine) est suffisante pour expliquer les différences de stochasticité de l'expressions génique constatées entre clones. De plus, les dynamiques chromatiniennes obtenues entre clones ont des caractéristiques communes :

- des temps fermés longs à très longs (plusieurs dizaines d'heures),
- des temps ouverts beaucoup plus courts.

Cette première constatation est en conformité avec la littérature du domaine : l'activité transcriptionnelle a lieu sous la forme de "bouffées d'activité" (des "bursts transcriptionnels"). L'intérêt principal de notre étude est que nous pouvons quantifier ces bursts et les dynamiques transcriptionnelles correspondantes pour chacun des clones, donc pour différents locus génomiques. Cette quantification nous permet de compléter l'information sur la dynamique : en effet, grâce à notre modèle, nous pouvons montrer que la différence

[1]Formellement, la moyenne de fluorescence est $\alpha \langle P \rangle$, ou α est un facteur linéaire quantifiant la fluorescence d'une protéine.

entre les clones peut être expliqué par de grandes variations du temps moyen fermé entre les différents locus génomiques.
Enfin, nous pouvons réutiliser notre modèle pour quantifier les effets de la TSA sur la dynamique chromatienne. Les résultats montrent qu'il est possible de simuler avec une bonne précision un traitement TSA à condition de modifier la dynamique chromatinienne du modèle. Là encore, c'est en jouant sur les temps moyens fermés que nous parvenons à identifier le plus fidèlement les distributions. Cela nous permet de conclure que la TSA agit sur la dynamique chromatinienne essentiellement en augmentant la fréquence (et non la durée) des épisodes d'ouverture.

2 Quantifying the contribution of chromatin dynamics to stochastic gene expression reveals long, locus-dependent periods between transcriptional bursts

Abstract

Background A number of studies have established that stochasticity in gene expression may play an important role in many biological phenomena. This therefore calls for further investigations to identify the molecular mechanisms at stake, in order to understand and manipulate cell-to-cell variability. In this work, we explored the role played by chromatin dynamics in the regulation of stochastic gene expression in higher eukaryotic cells.

Results For this purpose, we generated isogenic chicken-cell populations expressing a fluorescent reporter integrated in one copy per clone. Although the clones differed only in the genetic locus at which the reporter was inserted, they showed markedly different fluorescence distributions, revealing different levels of stochastic gene expression. Use of chromatin-modifying agents showed that direct manipulation of chromatin dynamics had a marked effect on the extent of stochastic gene expression. To better understand the molecular mechanism involved in these phenomena, we fitted these data to a two-state model describing the opening/closing process of the chromatin. We found that the differences between clones seemed to be due mainly to the duration of the closed state, and that the agents we used mainly seem to act on the opening probability.

Conclusions In this study, we report biological experiments combined with computational modeling, highlighting the importance of chromatin dynamics in stochastic gene expression. This work sheds a new light on the mechanisms of gene expression in higher eukaryotic cells, and argues in favor of relatively slow dynamics with long (hours to days) periods of quiet state.

Keywords

Chromatin dynamics, expression noise, gene regulation, stochastic model

2.1 Background

Although the importance of stochasticity in gene expression has been anticipated more than three decades ago (Novick et Weiner, 1957; Spudich et Koshland, 1976; Kupiec, 1983), the existence of a strong stochastic component in gene expression has only recently been experimentally demonstrated, showing that, despite constant environmental conditions, isogenic cells do show significant fluctuations in their gene-expression levels (Elowitz *et al.*, 2002; Levsky *et al.*, 2002; Black *et al.*, 2012; McCullagh *et al.*, 2009; Singh et Weinberger, 2009; Niepel *et al.*, 2009; Eldar et Elowitz, 2010). Moreover, regulated stochasticity, and its resulting phenotypic diversity, has been shown to be involved in several biological processes (Viñuelas *et al.*, 2012), including cell differentiation (Rao *et al.*, 2002; Chang *et al.*, 2008), development (Samoilov *et al.*, 2005; Boettiger et Levine, 2009) virus decision-making (Rao *et al.*, 2002; Weinberger *et al.*, 2005), and bacterial survival during environmental stress (Süel *et al.*, 2007; Mettetal et Van Oudenaarden, 2007; Veening *et al.*, 2008; Çağatay *et al.*, 2009).

Many studies have shown that the average expression level of a gene depends strongly on its genomic location (Boutanaev *et al.*, 2002; Caron *et al.*, 2001; Gierman *et al.*, 2007; Nie *et al.*, 2010; Versteeg *et al.*, 2003). In cultured cells, the silencing position effect (similar to the position effect variegation seen in Drosophila and mammals) is a well-characterized example of the influence of chromatin on gene expression ; with a stably integrated transgene, a progressive silencing of the reporter occurs, at a rate that strongly depends on the integration site (Feng *et al.*, 2001). Several studies based on treatments with 5-azacytidine (a DNA-demethylating agent (Veselý, 1985)) and with trichostatin A (a histone deacetylase inhibitor (Yoshida *et al.*, 1995)) have shown that DNA methylation and histone acetylation play a pivotal role in this process. Indeed, these treatments reverse the extinction of the transgene (Feng *et al.*, 2001; Pikaart *et al.*, 1998). Almost all of these studies, however, have focused on the mean value of gene expression, and only a few have addressed the question of the relationships between stochastic gene expression and chromatin, in either yeast (Paulsson, 2005b; Becskei *et al.*, 2005; Boeger *et al.*, 2008; Cai *et al.*, 2008; Kelemen *et al.*, 2010; Batenchuk *et al.*, 2011) or higher eukaryotes (Raj *et al.*, 2006; Voss *et al.*, 2009; Singh *et al.*, 2010a; Skupsky *et al.*, 2010).

Initially conducted in prokaryotes (Elowitz *et al.*, 2002; Ozbudak *et al.*, 2002), experiments to explore the molecular causes of stochastic gene expression were rapidly extended to yeast models (Black *et al.*, 2012; Becskei *et al.*, 2005; Blake *et al.*, 2006; Raser, 2004). These experiments suggested that, other than trivial aspects such as small molecule numbers, more sophisticated causes, such as chromatin remodeling, were important players in stochastic gene expression (Raj et Van Oudenaarden, 2008). More precisely, of the various possible sources of stochasticity, one in particular, namely locus-dependent chromatin dynamics (for example, transitions between an "open" state that allows gene transcription and a "closed" state that represses gene transcription) is a promising candidate to explain the regulation of stochastic gene expression. This role of chromatin was highlighted by the work of Becskei et al., who in 2005 showed the existence of genomic domains in the yeast genome, which produce a low transcriptional noise (that is, the part of stochastic gene expression arising from irregular transcript production) (Becskei *et al.*, 2005). The following year, by analyzing the variability of mRNA levels from tandemly and non-tandemly integrated pairs of transgenes in mammalian cells, Raj et al. identified the influence of

genomic domain on transcriptional noise, suggesting the importance of the switching rate between chromatin states via remodeling. Gene activation or inactivation would occur in cases of chromatin decondensation or condensation, respectively (Raj *et al.*, 2006). To analyze the effect of chromatin remodeling on promoter activation and therefore on stochastic gene expression, Raser and O'Shea used yeast strains lacking components of the chromatin-remodeling complexes. A major conclusion of their work was that the alteration of chromatin-remodeling enzymes resulted in changes in stochastic gene expression (Raser, 2004). However, most of these studies have tried to link chromatin dynamics to stochastic gene expression using indirect approaches (Becskei *et al.*, 2005; Raj *et al.*, 2006; Raser, 2004; XU *et al.*, 2006).

In many situations, from prokaryotes to eukaryotes, simple mathematical models describing the transcriptional dynamics as a two-state process have been shown to account effectively for the stochastic expression of a gene (Paulsson, 2005a; Larson *et al.*, 2009). Indeed, the two-state model, also known as the "random-telegraph model" (Golding *et al.*, 2005; Pedraza et Paulsson, 2008), now constitutes a standard in the field. This model assumes that the promoter switches randomly between two states, "on" and "off", with only the former allowing initiation events to occur. These transitions could correspond to several mechanisms, including assembly and disassembly of specific complexes, progression through the cell cycle, or the recruitment of the locus into transcription factories (Chubb et Liverpool, 2010). In many cases, evidence supports the hypothesis that these "on" and "off" states primarily reflect alternative chromatin configurations (Miller-Jensen *et al.*, 2011).

Recently, using a short-lived luciferase protein, Suter et al. monitored transcription at high temporal resolution in single mammalian cells, and identified bursts of transcription, a mechanism previously suggested in prokaryotes and eukaryotes (Elowitz *et al.*, 2002; Raj *et al.*, 2006). Using the random-telegraph model, they characterized the temporal patterns of transcriptional bursts for different genes, and obtained the distributions of the "on" and "off" times (Suter *et al.*, 2011). Harper et al. performed a complementary analysis of transcriptional bursting in single mammalian cells (Harper *et al.*, 2011). By quantifying the time dependence and cyclic behavior of the transcriptional pulses from the prolactin promoter, they estimated the length and variation of both transcriptionally active and inactive phases. Both studies point to the existence of a refractory "off" period, but they diverge on the role of chromatin remodeling ; in contrast to the Suter study, in which chromatin environment seemed to play a secondary role in shaping bursting patterns, Harper et al. concluded that chromatin remodeling may play an important role in the timing of transcriptional bursting. Finally, based on time-lapse fluorescence microscopy experiments, coupled with the use of the two-state model, Dar et al. gave a recent comprehensive study on noise in mammalian cells (Dar *et al.*, 2012). In their work, these authors suggested that transcriptional bursting, as opposed to constitutive expression, dominates across the human genome. Moreover, by analyzing more than 8,000 distinct genomic loci, they found that both frequency and burst size vary by chromosomal location. Therefore, the role of chromatin dynamics in the control of stochastic gene expression in higher eukaryotes remains a central matter of debate.

In a preliminary study, our group showed, using isogenic cell populations expressing a fluorescent reporter, that modification of chromatin marks, using chromatinmodifying agents such as 5-azacytidine (5-AzaC) and trichostatin A (TSA), induced significant effects on

mean fluorescence intensity (MFI) and normalized variance (NV ; that is, the variance normalized by the square of the mean) (Viñuelas *et al.*, 2012). We also showed that TSA and 5-AzaC had different effects on NV, whereas their effects on MFI were similar. Finally, investigating the possible reversibility of the effects identified by flow cytometry after the drug treatments, we found that MFI, NV, and the shape of the fluorescence distributions tended to return to their initial values after the treatment end. This result, which shows full reversibility of the cellular system after important modifications of the chromatin state, suggests that cells could be able to temporally modify their level of stochastic gene expression via modifications of chromatin marks, before returning to their initial physiological state.

To assess the possible influence of chromatin-opening/ closing dynamics on the stochasticity of gene expression, the next step was to combine biological experiments with a modeling analysis. For that purpose, we generated a series of clonal isogenic cell populations from chicken erythrocyte progenitors (6C2 cells). These populations were stably transfected with a unique copy of a reportergene coding for the red fluorescent protein *mCherry*, but the reporter was inserted at different chromosomal positions in each clone (Figure 1, left). Using flow-cytometry measurements, we found substantial clone-to-clone differences in the stochastic expression of the reporter. In particular, some of the clones had very similar MFI but different NV values. Because the only difference between these clones was the genomic location of the reporter, the observed differences in stochastic gene expression must stem from the chromosomal positioning effect, such as locus-specific dynamics of the chromatin surrounding the transgene. To evaluate whether chromatin dynamics significantly affect the stochasticity of gene expression, we treated some clones with 5-AzaC and TSA. Cell responses to these drugs clearly showed that both MFI and NV were affected, indicating that the chromatin environment of the reporter gene plays a significant role in the stochasticity of its expression. This result confirmed preliminary conclusions obtained by our team (Viñuelas *et al.*, 2012). By fitting a two-state model to the experimental data, we provided a mechanistic interpretation for the clone-to-clone diversity of expression patterns, in terms of differences in chromatin dynamics. More specifically, based on both analytical derivations (Paulsson, 2005a) and simulations (Gillespie, 1977), we explored the dynamics of the model and iteratively refined its kinetic parameters. The outcome was an accurate reproduction of the distribution of expression levels before, during, and after drug treatment.

Our current study supports the view that expression dynamics is strongly driven by short and infrequent transcriptional bursts, as previously described in other models, including mammalian models. However, the major advance of this work is that, whereas the duration and intensity of bursts did not show strong clone-to-clone differences, the time between bursts was found to depend strongly on genomic location and was broadly affected by drug treatments that affect chromatin. Hence, the position-dependent opening dynamics of chromatin emerges as a key determinant of the stochasticity in gene expression.

2.2 Results

We generated a series of clones stably transfected with the *mCherry* reporter, driven by the cytomegalovirus (CMV) promoter, then using splinkerette PCR (Devon *et al.*, 1995), we retained six clones showing a unique reporter insertion site (Table III.1).

Figure 1

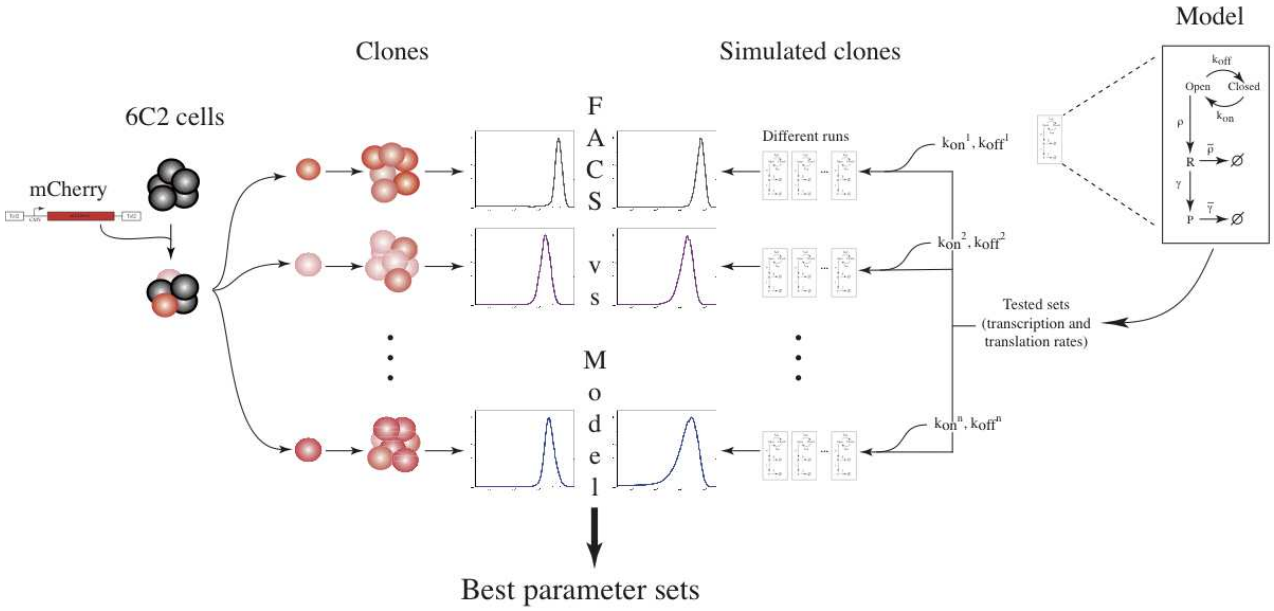

Figure III.1 – Experimental strategy used for assessing the role of chromatin environment on stochastic gene expression. After generation of cellular clones expressing the fluorescent reporter *mCherry*, stably integrated as a unique copy into the genome, the fluorescence distributions obtained by flow cytometry ("FACS") were compared with simulated distributions generated by a two-state model ("Model"). After experimental determination and exploration of transcription-translation parameters (ρ : transcription rate, γ : translation rate, $\tilde{\rho}$: mRNA degradation rate, $\tilde{\gamma}$: protein degradation rate and α : protein fluorescence coefficient), the best parameter sets were identified and then used to compute the specific chromatin dynamics (k_{on} and k_{off}, which are, respectively, the opening and closing transition rates of the chromatin at the reporter integration site) for each clone.

Clone	Chromosome	Chromosomic location	Direction
C1	Z	54910770	Reverse
C3	2	145804848	Forward
C5	15	1558243	Forward
C7	2	146026983	Forward
C11	11	12399458	Reverse
C17	2	145558788	Reverse

Table III.1 – Identification by splinkerette PCR of the *mCherry* genomic insertion sites for six 6C2 cellular clones.

These clones were then analyzed by flow cytometry, yielding for each of them the full distribution of fluorescence, and the corresponding MFI and NV (Fig. III.2 (A)). It is important to emphasize that the six clones differed only in their reporter insertion sites. Based on

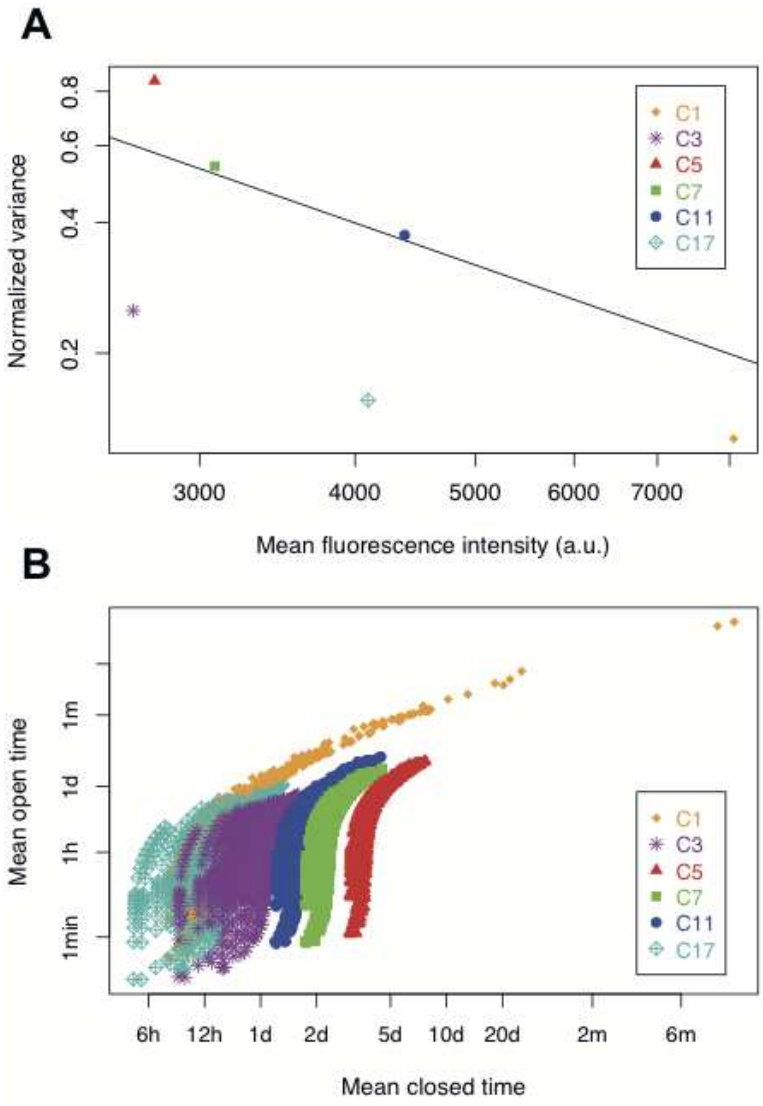

Figure III.2 – Exploration of model parameters to explain the observed stochastic gene expression for six cellular clones. (A) Relationship between normalized variance (NV) and mean fluorescence intensity (MFI) for six cellular clones (C1 to C17) stably transfected with a unique copy of the fluorescent reporter *mCherry* that was integrated at a different locus in each clone. Black line shows the relationship NV = 1/MFI (B) Distributions of the possible chromatin dynamics. For each clone, all 1,087 possible couples of $(1/k_{off}\,;1/k_{on})$ values were plotted, expressed as mean open time $(1/k_{off})$ and mean closed time $(1/k_{on})$ for all transcription-translation parameter sets explored analytically in the two-state model (see Methods). One dot therefore represents one possible analytical solution for that clone. h, hours ; d, days ; m, months.

the NV, a robust indicator of the stochasticity of gene expression (Tao *et al.*, 2007), the clones could be sorted from the most to the least stochastic, in terms of reporter-gene expression as follows : C5>C7>C11>C3>C17>C1. Moreover, analyzing the relationship between NV and MFI, we concluded that there is no direct linear relation between these

two parameters. Indeed, certain clones displayed similar MFI but very different NV values (for example, comparison of C3 with C5, or C11 with C17, Fig. III.2 (A)). This important dispersion of the points, around the inverse tendency between NV and MFI values, also suggests that mRNA abundance fluctuations were not the major source of intrinsic noise in this context.

An explanation of these observations comes from a previous preliminary study, in which we investigated whether chromatin dynamics are involved in these observed differences (Viñuelas *et al.*, 2012). Using the same cellular clones (same cell line, reporter, and environmental conditions) we performed the 5-AzaC and TSA treatments that would act directly on chromatin by two different molecular means. Our results showed that for the two drugs, modification of chromatin dynamics had clear consequences for stochastic gene expression (Viñuelas *et al.*, 2012). However, in this previous study, we did not assess how chromatin influences stochastic gene expression.

Thus, for this purpose in the current study, we fitted these data to a two-state model of gene expression, and evaluated to what extent chromatin dynamics act on stochastic gene expression. Under the assumption that all parameters but those describing the dynamics of chromatin would be identical in all the clones, we performed an iterative screening of model parameters. This allowed us to find these common parameters, and to characterize the position-specific dynamics of chromatin for each individual clone (Fig. III.1).

2.2.1 Description of the model

The choice of the model used to analyze our biological data was crucial. Two models are classically used to describe transcriptional stochasticity : 1) a Poisson model, in which the gene has, at each instant, a given chance to produce an mRNA, (McCullagh *et al.*, 2009; Golding *et al.*, 2005; Munsky *et al.*, 2012) and 2) a random-telegraph model, in which the gene additionally switches randomly between an "on" state, in which transcripts are produces in line with Poisson dynamics, and an "off" state, in which no transcripts are produced (Raj et Van Oudenaarden, 2008; Paulsson, 2005a; Munsky *et al.*, 2012). The Poisson model is known to lead to a direct linear relationship between MFI and NV on a log-log plot (that is, NV = 1/MFI) (Singh *et al.*, 2010a; Bar-Even *et al.*, 2006). Because such a relation was not sufficient to describe our data (Fig. III.2 (A)), we adopted the more general random telegraph model. It cannot be excluded that extrinsic noise may also participate to some degree in the observed fluorescence distributions. However, observing a variety of distributions for different insertion sites of the reporter (Fig. III.2 (A)) strongly suggests that the major source of noise is intrinsic. Indeed, as sources of extrinsic noise are independent of the reporter, they were expected to have somewhat similar effects in all the different clones. In addition, given the long mRNA and protein lifetimes in our system (see below), only the very slow extrinsic fluctuations are likely to affect the protein levels of the reporter.

Because flow cytometry quantifies protein fluorescence, the model must describe the expression process up to the protein level (including mRNA and protein production and degradation rates) and requires an additional parameter to convert protein quantity into fluorescence intensity (Fig. III.1, right). Thus, for each clone, the model had seven parameters : k_{on} and k_{off} , respectively describing the rates of chromatin opening and closing, ρ and γ, describing the transcription and translation rates, $\tilde{\rho}$ and $\tilde{\gamma}$, describing the trans-

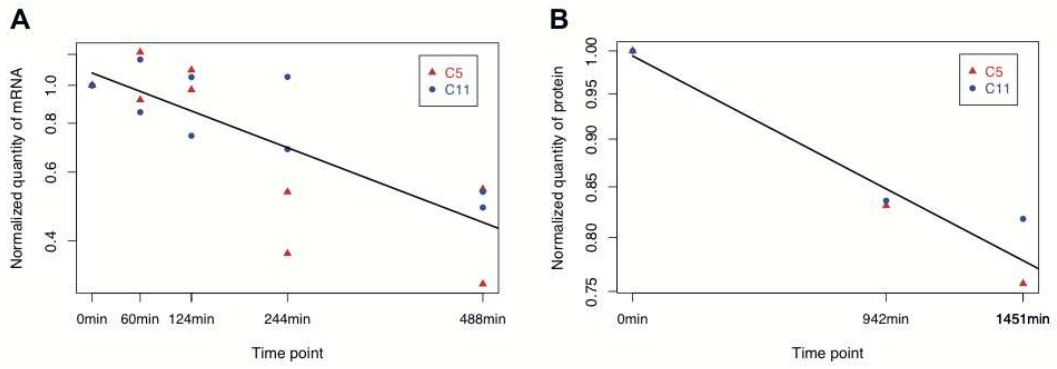

Figure III.3 – Determination of the *mCherry* reporter mRNA and protein half-lives. (A) Real-time quantitative RT-PCR measurement of *mCherry* mRNA decay after actinomycin D treatment in two different clones of 6C2 cell line. The best fitting exponential curve found minimizing least squares (between exponential curve and biological data) is shown as a dark line. Ordinates are in logarithmic scale. The deduced *mCherry* mRNA half-life is 7h04min (424min). (B) Flow cytometry measurement of *mCherry* protein fluorescence decay after cycloheximide treatment in two different clones of 6C2 cell line. The best fitting exponential curve found minimizing least squares (between exponential curve and biological data) is shown as a dark line. Ordinates are in logarithmic scale. The deduced *mCherry* protein half-life is 65h47min (3947min).

cript and protein degradation rates, and finally, a linear coefficient α, representing the fluorescence intensity of a single *mCherry* protein in the arbitrary unit measured by the flow cytometer. In order to fit the model, the optimal set of parameters must be identified, under the assumption that ρ, γ, $\tilde{\rho}$, $\tilde{\gamma}$ and α are identical in every clone, but that k_{on} and k_{off} are clonespecific. From this point, we refer to the five former parameters as the "transcription-translation parameters" and to the two latter ones as the "chromatin-dynamics parameters". Because we had six clones, we actually had to determine 17 parameters ($(6 \times 2) + 5$) in order to fully specify the model and to ultimately estimate the chromatin-dynamics parameters for each clone. For these 17 parameters, the two degradation rates ($\tilde{\rho}$ and $\tilde{\gamma}$) were determined experimentally from inhibition-based experiments (see Methods ; see Additional file 2, Fig. III.3). We found respectively that $\tilde{\rho} = 1.63\text{x}10^{-3}$ min^{-1} (mRNA half-life = 7 hours and 4 minutes) and $\tilde{\gamma} = 1.76\text{x}10^{-4}$ min^{-1} (protein half-life = 65 hours and 47 minutes). These values are consistent with average mRNAs and proteins half-lives previously measured in mammalian cells (9 and 46 hours, respectively) (Schwanhäusser *et al.*, 2011). Following this, we needed to find the optimal values of a set of 15 parameters to fit the experimentally measured fluorescence distribution of the six clones.

Several methods can be used to find such a parameter set. In particular, there are various optimization methods available, such as simulated annealing. However, because the model-experiment comparisons in our study involved stochastic simulations, the objective functions that have to be minimized (that is, some distance measure between predictions and observations) are only estimated up to a certain error level. Although small, this error level makes most optimization algorithms inadequate. Indeed, these algorithms rely on estimating the gradient or Hessian of the objective function, based on a finite diffe-

rence procedure (that is, evaluating small variations in the objective function resulting from small variations in its parameters). In a context where successive estimations of the objective function, even for the same parameters, may display random variations, these optimization algorithms are clearly doomed to failure. Overcoming this issue would require both running extremely long and computationally intensive simulations to minimize the error, and using coarse variation steps in the gradient-estimation procedure, which could result in numerical instabilities during the optimization.
For this reason, we decided to conduct a systematic parametric exploration, as this is a procedure that does not require local smoothness of the objective function. In addition, a single evaluation of the objective function represents a heavy computation load ; for example, involving thousands of realizations of a Gillespie simulation that are followed over long periods of simulated time (see Methods). In this context, a systematic parametric exploration allows massive parallelization of the computations on a grid. The sequential evaluation imposed by optimization algorithms makes this approach prohibitive. However, because the systematic exploration still requires intensive computations, we used iterative screening of the model parameters to progressively reduce the parameter space that has to be simulated.
This iterative screening was based on three steps in which we successively used analytical derivations on the model (step 1), additional experimental data (step 2), and finally, stochastic simulation (step 3). Thanks to these successive screenings, we were able to reduce by a factor of 30 the number of parameter sets to be simulated, thus making the problem computationally tractable. In the following sections, we describe the three screening steps and the results we obtained from them.

2.2.2 First screening of model parameters, based on mean and variance of fluorescence intensity

Mathematical derivations by Paulsson on the two-states model (Paulsson, 2005a) provide analytically the values of MFI and NV as a function of all parameters k_{on}, k_{off} , ρ, γ, $\tilde{\rho}$, $\tilde{\gamma}$ and α. By inverting these equations (see Methods), we are able to compute the chromatin-dynamics parameters (k_{on} and k_{off}) for each clone from : (i) its experimentally measured MFI and NV, (ii), the experimentally determined values of $\tilde{\rho}$ and $\tilde{\gamma}$ and (iii) the unknown transcription-translation parameters (ρ, γ and α). Thus, only three transcription-translation parameters remained to be determined, making their combinatorial exploration computationally tractable.
We explored wide ranges of these parameters that included all biologically relevant values (Milo *et al.*, 2009) : 20 values for ρ (from 6 to 0.00833 mRNA.min^{-1} ; that is, a transcription event occurring from every 10 seconds to every 2 hours when the chromatin is open), 15 values for γ (from 1 to 0.0003472 protein.min^{-1}.mRNA^{-1} ; that is, a translation event occurring from every 1 minute to every 2 days for each mRNA), and 12 values for α (from 0.1 to 200 fluorescence units per protein) (Rosenfeld *et al.*, 2006). See Methods for the exact tested values. For each triplet (ρ, γ and α), we computed k_{on} and k_{off} for each of the six clones from their experimental measure of MFI and NV. Of the 3,600 initial parameter sets, only 1,087 led to valid solutions, with the others leading to negative values for k_{on} or k_{off} for at least one clone. Fig. III.2 (B) shows the 1,087 possible pairs of values (k_{on} ; k_{off}) that resulted from this exploration for all the clones. It was found that, although

the chromatin-dynamics parameters could be the same order of magnitude, the mean open time ($1/k_{off}$, roughly between 1 minute and 1 day) was markedly shorter than the mean closed time ($1/k_{on}$, roughly between 6 hours and 4 days). This is characteristic of a transcriptional activity in which mRNA production events occur in brief bursts separated by longer silent periods.
The result of this first screening still produced more than 1,000 valid parameter sets, with the values of k_{on} and k_{off} spanning large intervals. This emphasizes that NV and MFI alone are not sufficient to identify, for a specific clone and therefore for a given genomic insertion site, the parameters that best explain the observed distribution of fluorescence.

2.2.3 Second screening of model parameters, based on response to treatments with chromatin-modifying agents

In order to reduce the ranges of solutions, we conducted additional experiments in which we modified the global dynamics of chromatin in both the cells and the model. We first treated three clones with the two chromatin-modifying agents TSA and 5-AzaC. As expected, TSA treatment, which leads to chromatin decondensation (Tóth *et al.*, 2004; Satoh *et al.*, 2004), induced an increase in MFI over time (Fig. III.4 (A)). 5-AzaC treatment, which inhibits chromatin condensation (Haaf et Schmid, 2000), produced the same effect as TSA treatment, but to a much lower extent. It is noteworthy that measures such as MFI, NV, and the fluorescence distributions tended to return to their initial values after removal of TSA and 5-AzaC, indicating full reversibility of the cellular system, and therefore a conservation of physiological conditions (Viñuelas *et al.*, 2012).
Based on these additional data, we could then exclude all transcription-translation parameter sets that did not account for the observed increase in expression levels even if the chromatin was considered as constantly open (see Methods). It is important to emphasize that we made the assumption that the TSA and 5-AzaC treatments affected only the chromatin-dynamics parameters. Using this strategy, we were able to reject 86% of the parameter sets, thus we kept only 114 transcription-translation parameter sets for further analyses. Fig. III.4 (B) shows the chromatin-dynamics parameter sets (that is, k_{on} and k_{off}), corresponding to the transcription-translation sets that were kept. All retained cases had in common that the mean open time $1/k_{off}$ was very short compared with any other timescale in the model (in particular both the mean closed time $1/k_{on}$ and the mean mRNA lifetime $1/\tilde{\rho}$). Hence, the actual duration of the bursts could not be estimated because two parameter sets with different k_{off} but an identical number of mRNAs produced per active period will exhibit similar distributions. For instance, if, on average, 20 mRNAs are produced during bursts that last 30 seconds or during bursts that last 10 minutes, the results will be practically identical because mRNAs decay with a half-life of more than 7 hours. Hence, as in other studies (Raj *et al.*, 2006), we could not determine the parameters k_{off} and ρ, but only their ratio ρ/k_{off}, that is, the mean number of mRNAs produced during a burst. This new effective parameter, referred to as "burst size", reduces by 1 the number of parameters in the model. At a higher level, protein synthesis/degradation noise is only important in cases where there is a low copy number (Paulsson, 2005a). Because low protein abundance would not be detected in a cytometry measurement, this source of noise is marginal compared with the noise from transcription and mRNA synthesis/degradation. For instance, even for the least variable

clone (C1, which had NV = 0.12 approximately, in Fig. III.2 (A)), a mean protein level as low as 200 copies would only contribute less than 5% to the measured NV. Hence, the parameters γ and α directly compensate for each other, and can be grouped into a single effective parameter, "$\alpha.\gamma$" (for example, producing twice as many proteins with half the fluorescence does not affect the distribution of fluorescence), reducing again by 1 the number of fitting parameters.

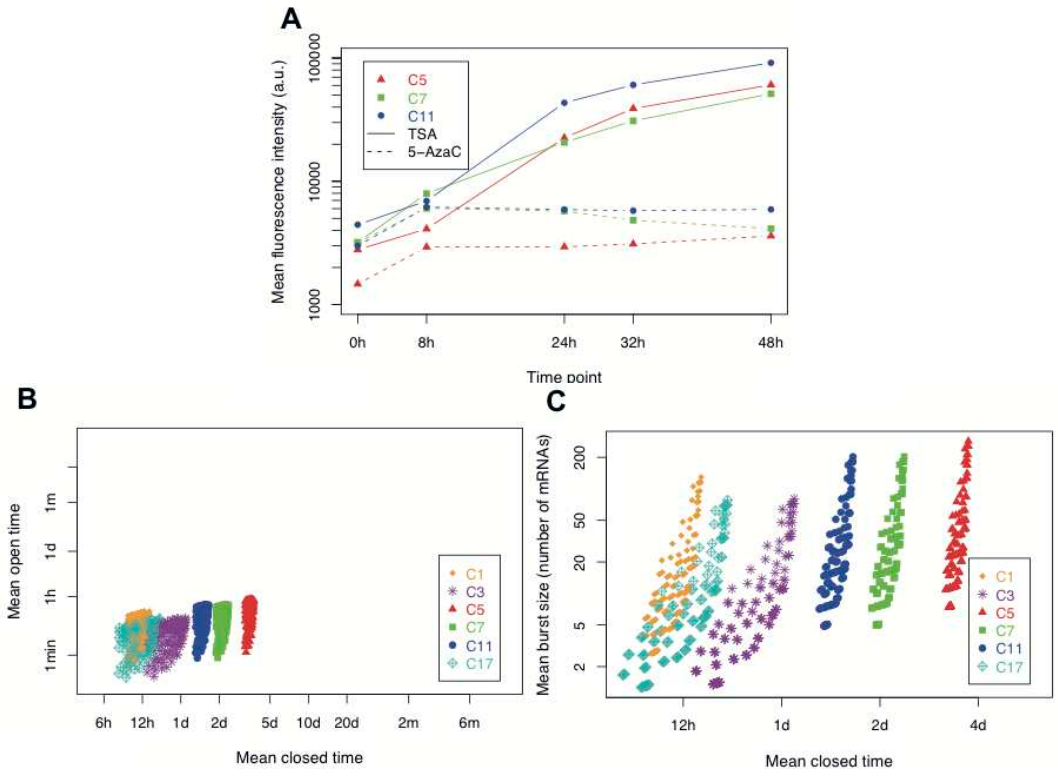

Figure III.4 – Exploration of model parameters based on treatments with chromatin-modifying agents. (A) Evolution of mean fluorescence intensity following kinetics of treatment with trichostatin A (TSA ; solid line) and 5-azacytidine (5-AzaC ; dotted line) (0 to 48 hours) for three cellular clones. (B) Distributions of the plausible chromatin dynamics. For each clone, all 114 possible couples of ($1/k_{off}$;$1/k_{on}$) values were plotted, expressed as mean open time (1/koff) and mean closed time ($1/k_{on}$), after removal of all parameter sets that were not able to account for the transcription-translation dynamics under TSA and 5-AzaC treatments. (C) This experiment was the same as for (B), except than the transcription rate (ρ) and the mean open time ($1/k_{off}$) parameters were reduced to a single effective parameter (ρ/k_{off}), representing the mean burst size. min, minutes ; h, hours ; d, days ; m, months.

Reformulating the sets of ($1/k_{off}$; $1/k_{on}$) couples retained after the second screening (Fig. III.4 (B)) in terms of (ρ/k_{off} ; $1/k_{on}$), as shown in Fig. III.4 (C), we observed relatively similar ranges of values for the mean burst size ρ/k_{off} for the six clones (although values spanned from 1 to 200 mRNAs per burst). By contrast, the mean closed time of chromatin seemed to be highly clone-dependent, ranging from 6 to 12 hours for clone C1 and C17 to more than 2 days for C5 and C7 (Fig. III.4 (C)). This suggests that the chromatin-

opening dynamics depend on the clones, and therefore on the chromatin environment of the reporter.

2.2.4 Third screening of model parameters, based on full distribution of fluorescence

To select the best parameter set from the 114 remaining sets, we simulated distributions of fluorescence corresponding to the remaining parameter sets, and compared them with the fluorescence distributions measured by flow cytometry. For each parameter set, we used a stochastic simulation algorithm (SSA) (Gillespie, 1977), to simulate 50,000 cells per clone, and then computed the resulting fluorescence distributions. Background fluorescence levels were added to the simulated distributions by convolution with the fluorescence distribution of the negative control-cell population (that is, cells that did not express any fluorescent protein). The resulting values were then compared with the six experimental distributions using a Kolmogorov-Smirnov test.

Analyzing the comparison scores (distances) from the Kolmogorov-Smirnov test of the 114 parameter sets, we were able to identify the subsets of parameters, and therefore the corresponding chromatin dynamics, that were the best fit to the distributions measured by the flow cytometer (Fig. III.5 (A)). Note that most sets correctly fit the experimental data (104 of the 114 sets corresponding to a single peak of good scores ; that is, <0.107), showing that the previous screening had already selected the correct parameter sets. The final parameter sets are shown (in black) in Fig. III.5 (B). For five of the six clones, we were able to generate distributions similar to those measured by flow cytometry (Fig. III.6). However, when analyzing the bimodal clone C7, we found that the simulated distribution fit only the high modality of the fluorescence distribution. This third screening supports our previous observation about the relatively similar mean burst size between the clones but the significantly different mean closed times (Fig. III.5 (B)). Looking at the chromatin-dynamics parameter set that best fit the flow-cytometry distributions for all clones (Fig. III.5 (B), in brown), our study revealed that for the six clones, mean burst sizes were between 30.0 and 118.9 mRNAs per burst, and mean closed times between 756.7 minutes ($\sim$12 hours) for the fastest clone to 5197.6 minutes ($\sim$3.5 days) for the slowest clone. However, it is important to note that, taking into account the full range of viable parameters (Fig. III.5 (B), in black) clone dynamics could be fit with similar values for their mean burst sizes (ranges of correct values are overlapping between the clones) whereas their mean closed time had to be different (Fig. III.5 (B)).

Fig. III.7 illustrates, for each clone, the results of simulations of the chromatin dynamics of a single cell, for the best parameter set. The best chromatin-dynamics parameters for each of the six clones are shown in Table III.2.

It is interesting to compare the differences between the different clones (that is, for the different chromatin environments) in terms of chromatin dynamics and their consequences on the transcription and translation of the *mCherry* reporter. It seems clear that the transcriptional activity of the reporter can vary from frequent bursts (C1) to rare bursts (C5), depending on the chromatin context. These important differences could very well be the dependence of the local chromatin properties at the reporter insertion site. Finally, the mRNA transcription rates and mRNA copies per cell we defined for the six clones (on average 2.1 and 21 respectively) (see Additional file 4, Table III.3) were in the same order

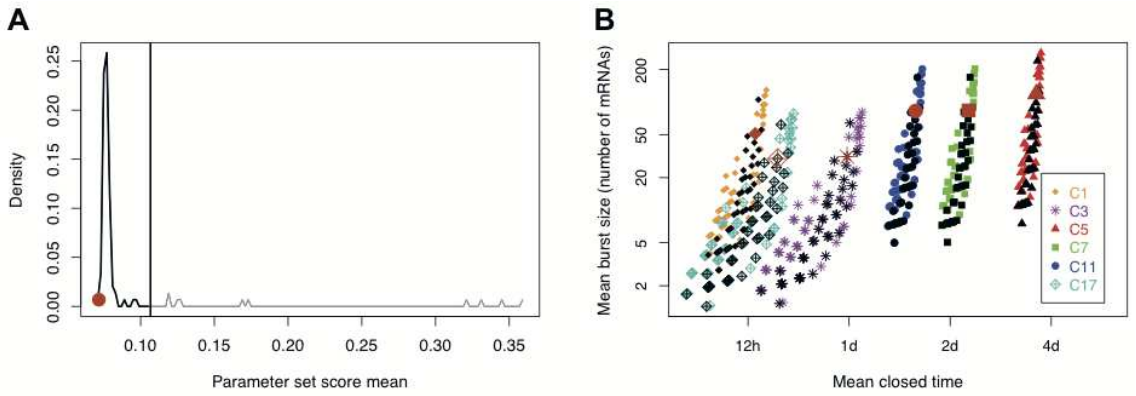

Figure III.5 – Exploration of model parameters based on a comparison of fluorescence distributions and stochastic simulation algorithm (SSA) simulations. (A) Distribution of parameter set scores. The lowest scores correspond to the better fits. These fits were obtained using values of γ and α, the parameters contained within the joint $\alpha.\gamma$ value of 0.035 a.u.min^{-1}.mRNA^{-1}. The upper limit (0.107) of the single peak showing the best scores is specified (vertical line). (B) Distribution of chromatin dynamics ("mean burst size" and "mean closed time"), obtained for the best parameter sets, after distribution comparisons for the six cellular clones. To compare with the possible chromatin dynamics presented in Fig. III.4 (B), this figure shows the chromatin dynamics obtained for the best parameter sets (black ; score means between 0.07 and 0.107 ; see panel (A)) and the optimal parameter set for each clone (brown).

Clone	$1/k_{on}$	ρ/k_{off}
C1	756.7	50.9
C3	1420.5	31.2
C5	5197.6	118.9
C7	3267.7	83.6
C11	2271.8	82.8
C17	882.7	30.0

Table III.2 – Chromatin-dynamics parameters proposed for six cellular clones. Mean closed times ($1/k_{on}$) are expressed as minutes. Mean burst sizes (ρ/k_{off}) are expressed as mRNAs.

of magnitude as those previously reported (Schwanhäusser *et al.*, 2011; Darzacq *et al.*, 2007).

2.2.5 Chromatin dynamics at genomic insertion sites and sensitivity analysis

By combining biological experiments, analytical computations and stochastic simulations, we were able to estimate all the model parameters that best fit the measured flow-cytometry distribution for the different integration sites. We now used some of these parameters (that is, $\alpha.\gamma$, $\tilde{\rho}$ and $\tilde{\gamma}$) to directly estimate the possible chromatin-dynamics parameters for any couple (MFI and NV), each corresponding to a different genomic in-

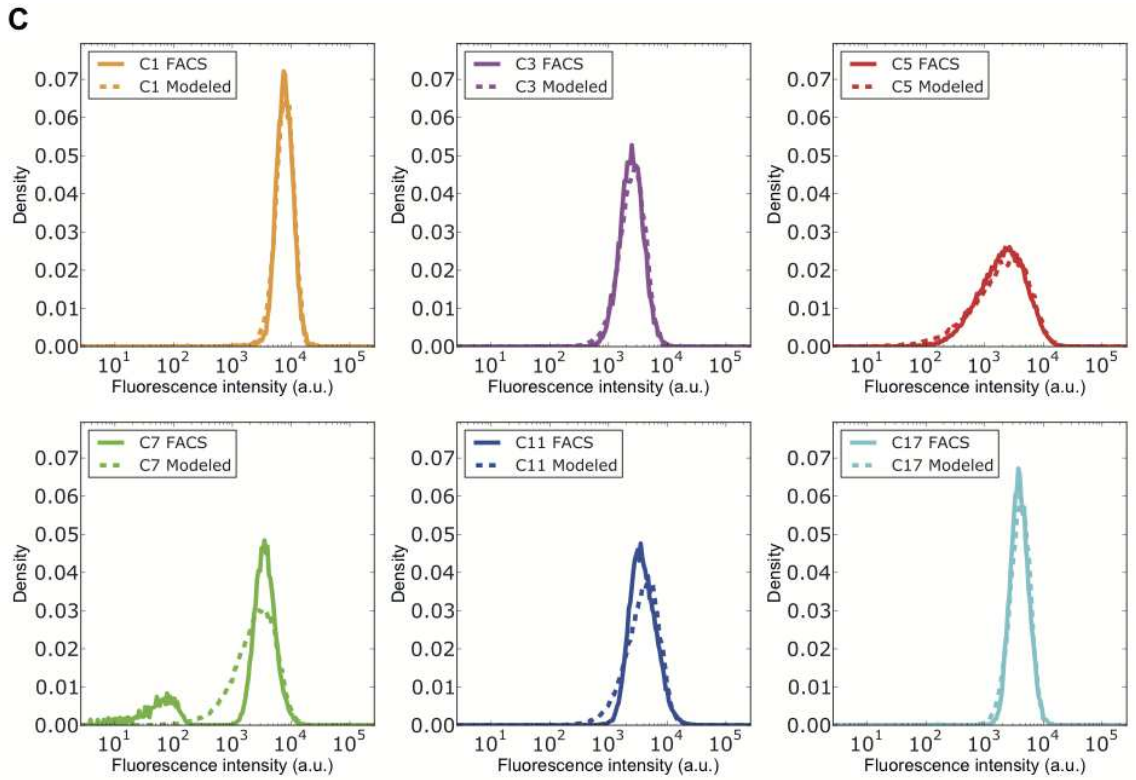

Figure III.6 – Exploration of model parameters based on a comparison of fluorescence distributions and stochastic simulation algorithm (SSA) simulations. This figure is linked with the two others with the same name : Fig. III.5 and III.7. This figure III.6 is the illustration, for the six cellular clones, of the comparison between the *mCherry* fluorescence distributions measured by flow cytometry ("FACS" ; solid line), and simulated fluorescence distributions ("Modeled" ; dotted line) obtained with the best chromatin-dynamics parameter set.

Clone	mRNA transcription rate	mRNA copies per cell
C1	3.99	40.76
C3	1.31	13.41
C5	1.37	13.95
C7	1.53	15.60
C11	2.17	22.17
C17	2.03	20.72

Table III.3 – *mCherry* transcription rates and mRNA levels for six cellular clones of the 6C2 cell line. mRNA transcription rates are expressed as $mRNA.h^{-1}$ and calculated for each clone as follow : $60\rho.k_{on}/(k_{on} + k_{off})$. mRNA copies per cell are calculated by dividing mRNA transcription rate by mRNA degradation rate.

sertion site of the reporter. We also used these parameters to estimate the sensitivity of the model (that is, the variation in the chromatin-dynamics parameters depending on the

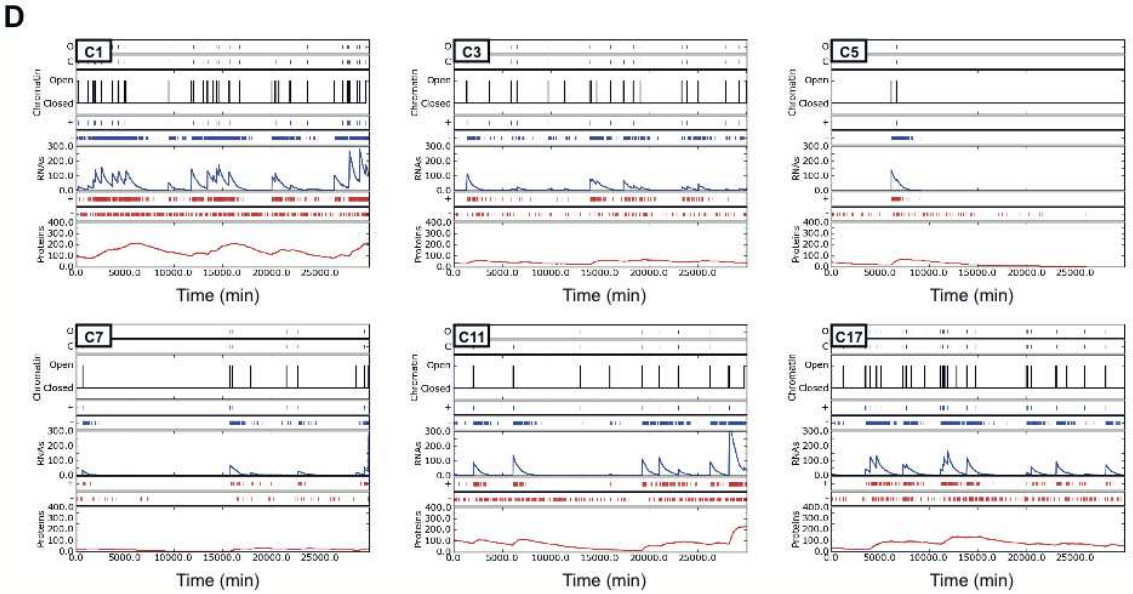

Figure III.7 – Exploration of model parameters based on a comparison of fluorescence distributions and stochastic simulation algorithm (SSA) simulations. This figure is linked with the two others with the same name : Fig. III.5 and III.6. On this figure III.7 is shown one run of Gillespie SSA per clone showing the chromatin dynamics (opening and closing chromatin events are shown in black) for one virtual cell of the isogenic population distribution (see Fig. III.6). Consequences of chromatin open/closed dynamics on mRNA transcription and protein translation are shown in blue and in red respectively. Production (+) and degradation (−) evolutions of mRNAs and proteins are also indicated. (For illustration, Fig. III.8 III.9 and III.10 show the same analysis as that presented in this figure, but for the parameter set with the highest (that is, worst) comparison score among the best ones).

two main indicators of gene expression, MFI and NV) in a biologically relevant parameter space. Indeed, we were able to use the best set of transcription-translation parameters that we obtained, along with a modified Paulsson's equation system, to determine the mean closed time of chromatin and the mean size of transcriptional bursts from the mean and NV of any similar construction (that is, the same cells but different insertion point) measured by flow cytometry (Fig. III.11). This can be represented by two three-dimensional graphs : one for the mean closed time and one for the mean burst size. It should be noted that both graphs are linked because each couple (MFI and NV) corresponded to a single couple (mean burst size and mean closed time). Two important elements could be derived from these three-dimensional graphs. First, as shown in panel A, the mean closed time was determined mainly by the NV value, whereas MFI only had a marginal contribution (at least in the activity domain of the measured clones). In other words, whatever the average transcriptional activity, the mean closed time could be derived directly from the variability in expression levels, highlighting the informational content of stochasticity in gene expression (Munsky *et al.*, 2009). By contrast, it can be seen from panel B that, to compute the mean burst size, both measures are necessary. Interestingly, the results pre-

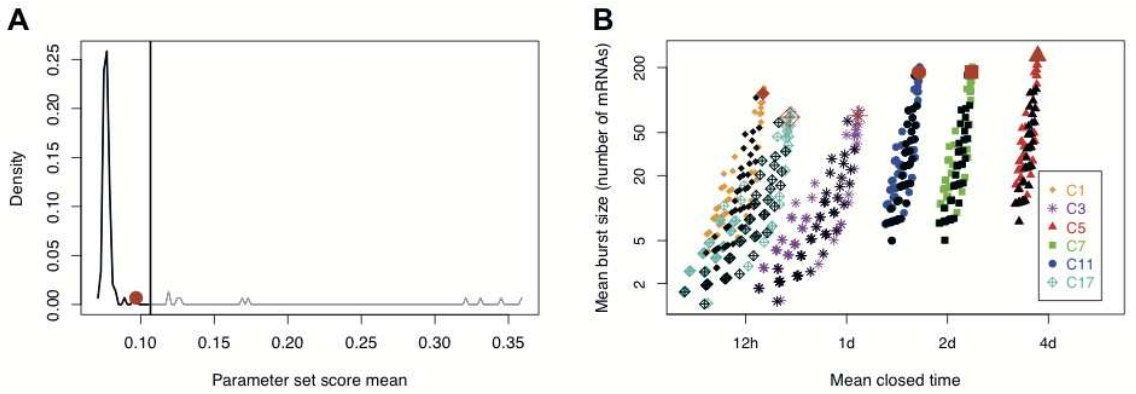

Figure III.8 – Exploration of model parameters based on a comparison of fluorescence distributions and SSA simulations. This figure is similar to the Fig. III.5 except that the selected parameter set has the upper score (worst one) among the best ones (shown as a brown circle in (A)). This figure III.8 is linked with Fig. III.9 and figure :figpcbS2d.

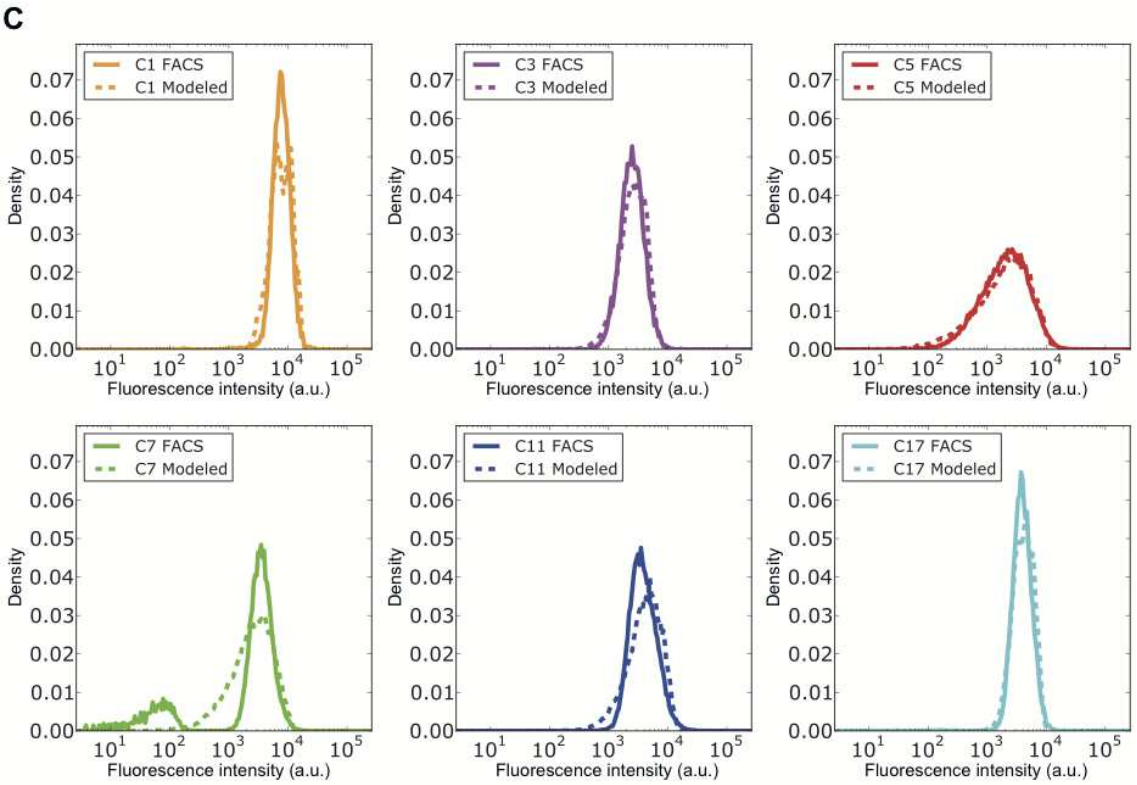

Figure III.9 – Exploration of model parameters based on a comparison of fluorescence distributions and SSA simulations. This figure is similar to the Fig. III.6 except that the selected parameter set has the upper score (worst one) among the best ones (shown as a brown circle in Fig. III.8 (A)). This figure III.9 is linked with Fig. III.8 and figure :figpcbS2d.

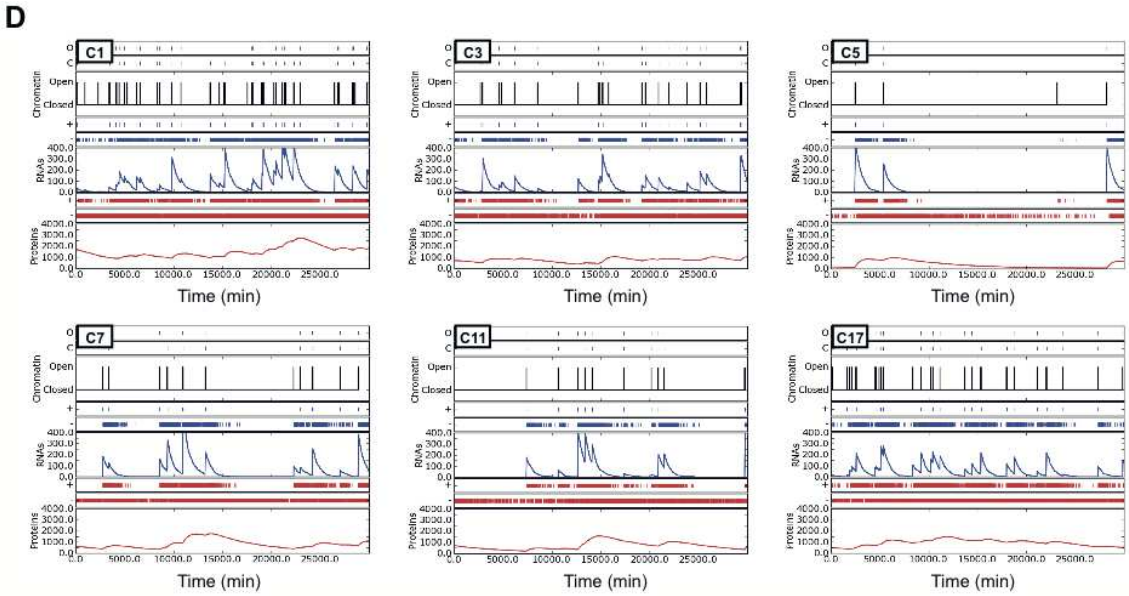

Figure III.10 – Exploration of model parameters based on a comparison of fluorescence distributions and SSA simulations. This figure is similar to the Fig. III.7 except that the selected parameter set has the upper score (worst one) among the best ones (shown as a brown circle in Fig. III.8 (A)). This figure III.10 is linked with Fig. III.8 and figure :figpcbS2c.

sented here show that, for our cell lineage, fluorescence distributions, which are relatively easy to measure by flow cytometry, coupled with a pertinent and robust analysis, allowed us to obtain valuable information about the chromatin-dynamics parameters.

Finally, we determined how the reported values (Table III.2) are affected by uncertainty in the experimentally determined mRNA and protein half-lives by conducting sensitivity analysis on equation 3 (see Methods). We found that variations of ± 5% of either mRNA or protein half-life resulted in variations in mean closed time and mean burst size that were always smaller than 5%. We therefore concluded that any experimental uncertainty in the mRNA and protein half-lives would only marginally affect the parameter values obtained through the model.

2.2.6 Testing and validation of the model following a dynamic evolution of the chromatin state

To test the contribution of chromatin dynamics to stochastic gene expression and the quality of the parameter set we obtained, we used our model to simulate a situation in which the chromatin dynamics were profoundly modified. For this, we used the flow-cytometry data from the TSA-treated clones C5 and C11 (5-AzaC was not tested because it produced less intense effects). During TSA treatment, the distributions of fluorescence, reflecting the expression of the *mCherry* reporter, gradually shifted to higher fluorescence values (Fig. III.4 (A) and III.12). According to our study, to obtain such dramatic effects, the dynamics of chromatin at the reporter insertion locus must have been modified

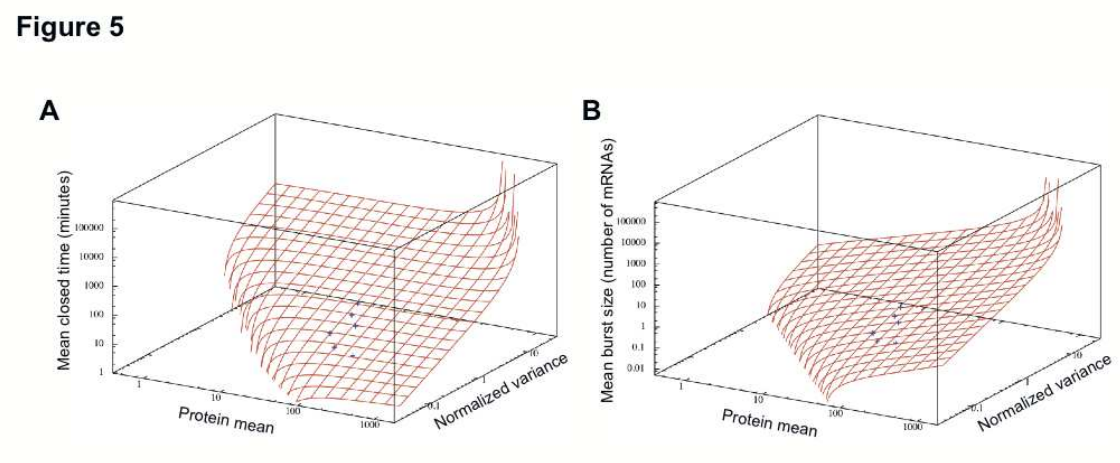

Figure III.11 – Inference of burst size and closed time from mean and normalized variance (NV) of protein levels. (A) At steady states, using the best transcription-translation parameter set (ρ, $\tilde{\rho}$, γ, $\tilde{\gamma}$ and α) and the modified Paulsson's equation system, the mean closed time could be calculated from the protein mean and protein NV (red grid). (B) Using the same data and equation system as in panel (A), the mean burst size could be calculated from the protein mean and protein normalized variance (red grid). Note that grids of both panels are linked because each value pair (protein mean and NV) corresponds to a single value pair (mean burst size and mean closed time). For both parts, clones C1, C3, C5, C7, C11, and C17 are represented as blue points on the grid, and all axes are on a logarithmic scale.

by reducing the mean closed time, increasing the mean burst size, or a combination of both. Because both parameters affect the transcriptional activity of the reporter, all the possible combinations that can account for the observed change in expression form a line in the mean closed time/mean burst size space (Fig. III.13 ; see Methods). We explored the chromatin dynamics for parameter sets lying along this line, and found the set that best fit the new flow-cytometry data. As for the previous experiments, we systematically explored the different parameter sets by sampling 11 points on the line between the two extreme situations mentioned above. It should be noted that, for this exploration, we considered the transcription-translation parameter set as constant, identical to the one computed previously (Fig. III.5 (A)). We found that the TSA treatment seems mainly to modify the chromatin mean closed time ; for the two clones used in this experiment, TSA reduced the mean closed time from more than 1 day (C11) and more than 3 days (C5) to 1 and 2.5 hours respectively (Fig. III.13). By contrast, mean burst size seemed to be increased only slightly. To support this result, we performed stochastic simulations with the retained chromatin-dynamics parameters to generate fluorescence distributions that we compared with the experimental flow-cytometry distributions (Fig. III.14). For the two clones, the simulated distributions correctly fit the flow-cytometry values at the end of the TSA treatment (48 hours). However, for the first time point (8 hours of treatment), the simulated fluorescence distribution was shifted relative to the biological experiment. The evolution of the comparison score between the measured and simulated data (Fig. III.14, insets) confirmed that during the first hours of treatment the simulation was a

poor fit to the flow-cytometry data. However, after 24 hours, a significant improvement occurred, and after 48 hours of treatment, the scores measured for the two clones were equivalent to those measured before the TSA treatment (0 hour of treatment, Fig. III.5 (A) and III.6). This clearly demonstrates that the model correctly rendered the new chromatin dynamics at steady state, although it was not able to fully reproduce the transient period. This is probably due to the kinetics of the drug effect, which was considered immediate in the model (the chromatin dynamics being changed immediately at the treatment time) whereas, in real cells, the chromatin modifications probably take place more gradually, thus delaying the activity of the drug.
To illustrate the consequences of the new chromatin dynamics on the transcription and translation induced by the TSA treatment, Fig. III.15 shows SSA simulations of the cell dynamics for the two clones before and after treatment. Owing to the low frequency of chromatin-opening events before treatment, a period of more than 10 days is shown whereas the TSA treatment was simulated for only 48 hours. The simulation clearly indicates the effects of TSA treatment on the chromatin dynamics and emphasizes the increased frequency of the chromatin-opening events, resulting in an increase in mRNA and protein concentrations.

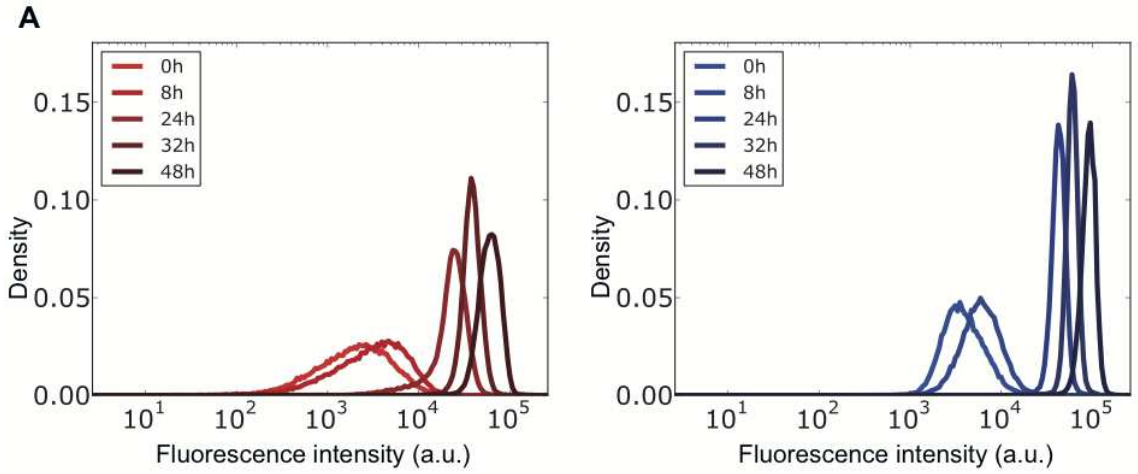

Figure III.12 – Effects of TSA-treatment kinetics on the *mCherry* fluorescence distributions for two cellular clones, C5 (red) and C11 (blue) measured by flow cytometry.

Using a two-state model, we found that the observed NVs and MFIs for each clone alone are not sufficient to identify efficiently, for a specific chromatin environment, a restricted set of parameters that best explain the observed differences between the six clones. We thus used a more complex strategy exploiting the full distribution of fluorescence as measured by flow cytometry. By mixing

2.3 Discussion

Analyzing stochastic expression of a stably integrated fluorescent reporter in six isogenic cell populations, differing only in their reporter integration site, this study provides

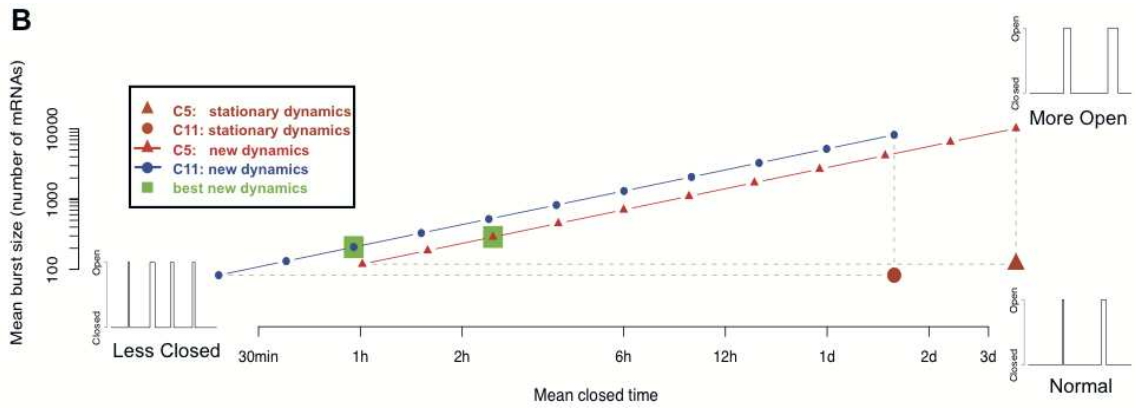

Figure III.13 – Model simulation of the perturbation of chromatin dynamics after trichostatin A (TSA) treatment. This figure III.13 is linked with Fig. III.12, III.14 and III.15. Here is represented new chromatin dynamics (mean burst size (ρ/k_{off}) and mean closed time ($1/k_{on}$)) fitting the observed fluorescence distribution evolution induced by TSA treatment. Different examples of these chromatin dynamics, inducing a higher open mean time (resulting from TSA treatment), are illustrated in the detailed view. After distribution-comparison tests, the best new chromatin dynamics (green), and those related to the steady state (brown) were ascertained. min, minutes ; h, hours ; d, days.

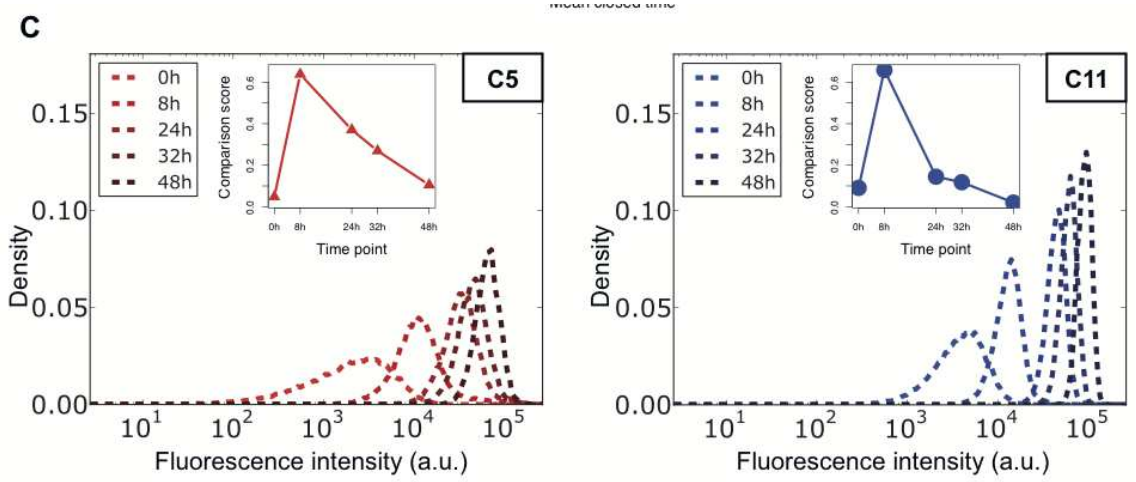

Figure III.14 – Model simulation of the perturbation of chromatin dynamics after trichostatin A (TSA) treatment. This figure III.14 is linked with Fig. III.12, III.13 and III.15. Here is represented simulated *mCherry* fluorescence distribution evolution obtained for the best new chromatin dynamics (see Fig. III.13). (Insets) Evolutions of the distribution-comparison scores (comparisons between measured distributions after TSA treatment and the simulated distributions).

new evidence suggesting that the local chromatin environment (reporter insertion site) influences stochastic gene expression. Our results are in agreement with previous studies on HIV gene expression, where it was shown that the existence of different fates for infec-

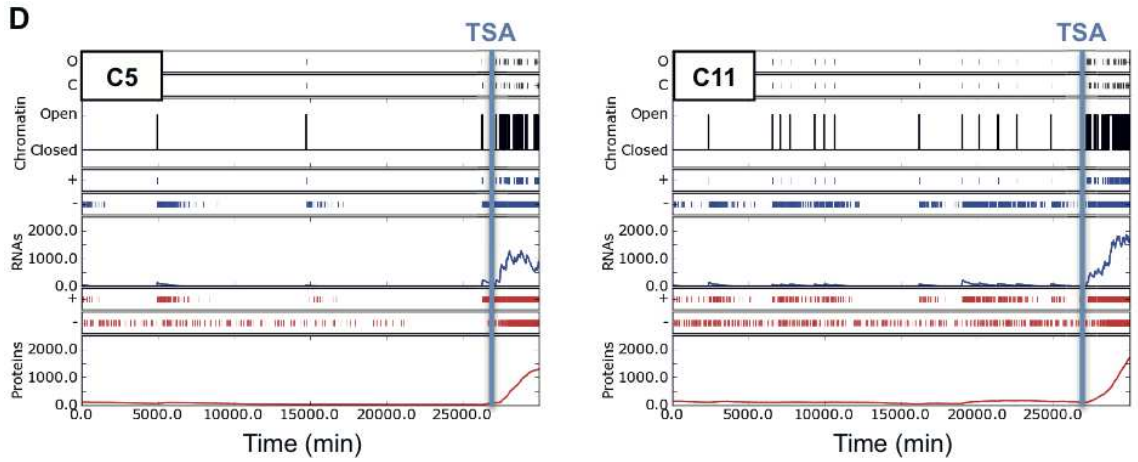

Figure III.15 – Model simulation of the perturbation of chromatin dynamics after trichostatin A (TSA) treatment. This figure III.15 is linked with Fig. III.12, III.13 and III.14. Here is represented one run of the Gillespie SSA per clone showing the dynamics of the chromatin before and during 48 hours of TSA treatment (opening and closing chromatin events are shown in black) for one virtual cell of the isogenic population distributions (see Fig. III.14). Consequences of chromatin open/closed dynamics on mRNA transcription and protein translation are shown in blue and in red respectively. Production (+) and degradation (−) evolutions of mRNAs and proteins are also shown. The beginning of TSA treatment is indicated by a vertical blue line. (For illustration, Fig. III.12, III.16, III.17 and III.18 show the same analysis as presented in this figure but for a parameter set (same as used in Fig. III.8, III.9 and III.10) showing a weaker fit).

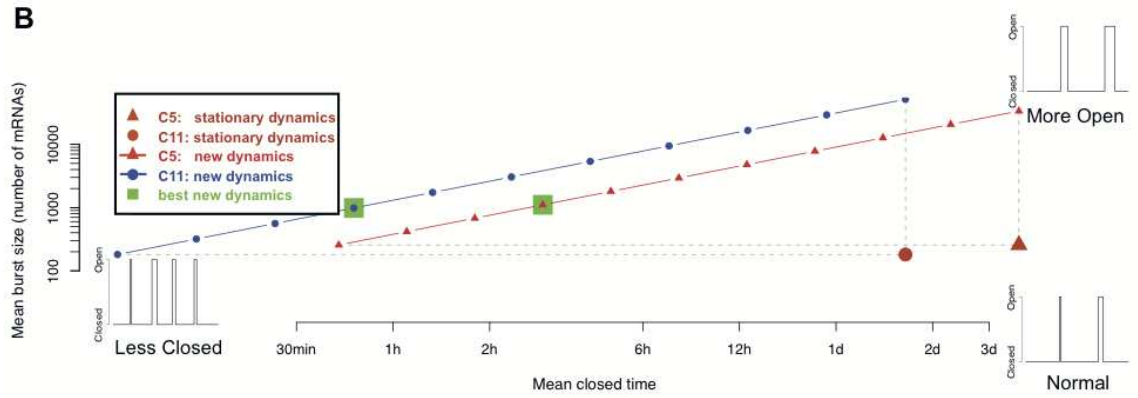

Figure III.16 – Model simulation of the perturbation of chromatin dynamics by TSA treatment. This figure is similar to the Fig. III.13 except that the best new chromatin dynamics was computed from the parameter set which had the highest (that is, worst) score (shown as a brown circle in the panel (A) of Fig. III.8) of the best scores obtained. This figure III.16 is linked with Fig. III.12, III.17 and III.18.

ted cells correlated with the virus-integration sites (Weinberger *et al.*, 2005; Miller-Jensen *et al.*, 2011), and that transcriptional burst size and burst frequency vary depending on

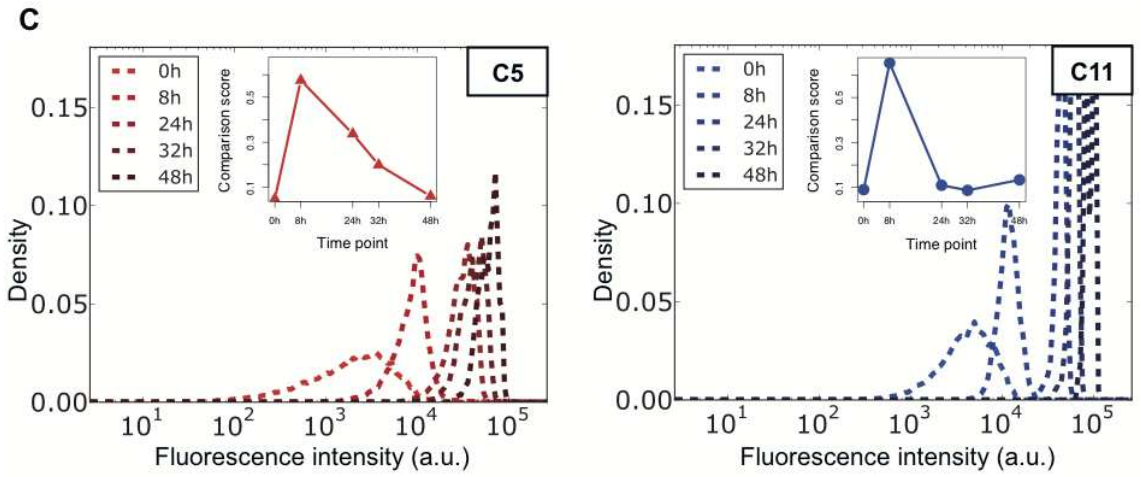

Figure III.17 – Model simulation of the perturbation of chromatin dynamics by TSA treatment. This figure is similar to the Fig. III.14 except that the best new chromatin dynamics was computed from the parameter set which had the highest (that is, worst) score (shown as a brown circle in the panel (A) of Fig. III.8) of the best scores obtained. This figure III.17 is linked with Fig. III.12, III.16 and III.18.

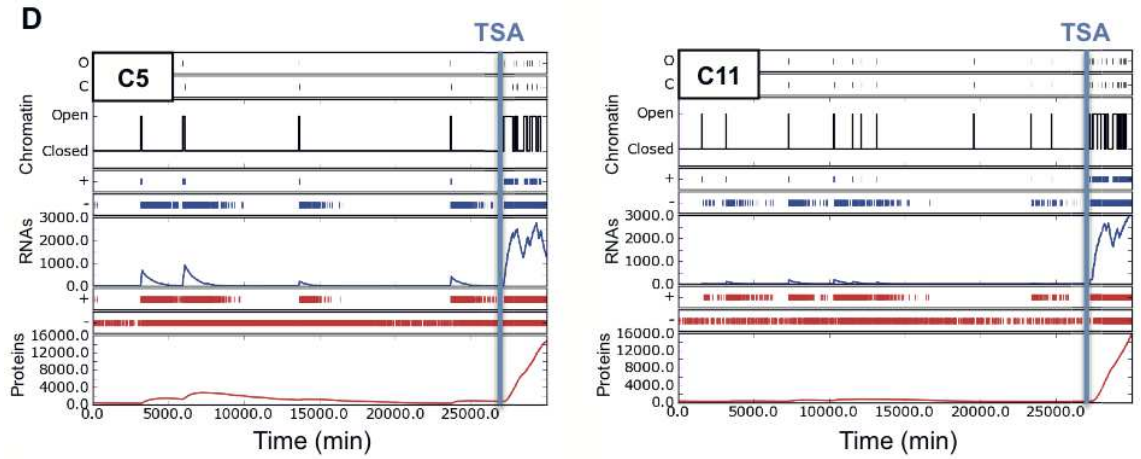

Figure III.18 – Model simulation of the perturbation of chromatin dynamics by TSA treatment. This figure is similar to the Fig. III.15 except that the best new chromatin dynamics was computed from the parameter set which had the highest (that is, worst) score (shown as a brown circle in the panel (A) of Fig. III.8) of the best scores obtained. This figure III.18 is linked with Fig. III.12, III.16 and III.17.

the virus-integration sites (Singh *et al.*, 2010a; Skupsky *et al.*, 2010). This chromosomal positioning effect on stochastic gene expression was also shown in yeast and in mammalian cells (Becskei *et al.*, 2005; Raj *et al.*, 2006), suggesting the existence of genomic local domain-level noise, probably under the control of the switching rate of chromatin between the open and closed configurations (Becskei *et al.*, 2005; Raj *et al.*, 2006; Raser, 2004). The biological function of this domain-level noise is not yet completely understood. Batada

and Hurst showed in yeast that genomic domains that enable low noise act as sinks for essential genes, for which noise is more deleterious than for nonessential ones (Batada et Hurst, 2007), suggesting an evolutionary pressure for shaping low-noise genomic domains. It is to be noted here that local chromatin dynamics is not the sole difference between the integration sites ; other genomic features, possibly correlated with chromatin states, could also be involved. We are currently investigating such a question on a genome-wide scale.

Using a two-state model, we found that the observed NVs and MFIs for each clone alone are not sufficient to identify efficiently, for a specific chromatin environment, a restricted set of parameters that best explain the observed differences between the six clones. We thus used a more complex strategy exploiting the full distribution of fluorescence as measured by flow cytometry. By mixing analytical models, complementary experiments, and stochastic simulations, we progressively identified the parameters that best fit the flow-cytometry distributions. The final set of parameters we obtained was able to reproduce accurately the experimental data for all clones except the unique bimodal one, C7, for which the simulated distribution fit only the high modality. This bimodal distribution observed for clone C7 could be due to : 1) specific chromatin dynamics related to the genomic insertion site of the reporter, or 2) a genetic mutation event affecting the reporter-gene integrity and resulting in two genetically distinct subpopulations. In the first case, if the transition rates between active and inactive states are extremely slow relative to transcript and protein degradations, each promoter state would be relatively stable, and this transcription regime could result in bimodal protein expression (Elowitz *et al.*, 2002; Levsky *et al.*, 2002; Black *et al.*, 2012). However, in the context of a two-state model, the value of the distribution between the two modes normally reflects the transient dynamics taking place after the gene switches from one state to the other, producing distribution tails from each mode towards the other. For clone C7, this part of the distribution between the two wellseparated modes was notably low, almost null. This indicates that the passage from one state to the other was extremely rare, so rare that the protein half-life (although rather long at 66 hours) is negligible in comparison. Such a slow dynamic is unlikely to be caused by chromatin dynamics. Indeed, during the submission of this work, very recent experimental evidence suggested that the bimodal distribution of clone C7 arose from a genetic mutation of the reporter (Dr Alexander Skupin, Dr Aymeric Fouquier d'Herouel and Dr Sui Huang, ISB, personal communication). This event induces extinction of the transgene in a subpopulation of C7, and the appearance of the low modality (data not shown). Including this subpopulation in the fitting process would therefore induce a bias. Consequently, we re-ran the analysis, taking into account only the five clones showing a unimodal distribution. The chromatin-dynamics parameter set that best fit the flow-cytometry distributions for all clones presented in Fig. III.5 (B) remained identical (data not shown), in accordance with the fact that the initial fitting process fit only the high modality of C7 and ignored the low one.

After selection of the best parameter sets and characterization of the chromatin dynamics for each clone, our work provided elements suggesting that the chromatin state is essentially dominated by the closed state, as previously shown (Harper *et al.*, 2011), but most importantly, that the chromatin environments of the clones clearly differed in their mean closed time. Indeed, for all clones, the mean burst sizes roughly comprised between 30 and 120 mRNAs per burst, which is consistent with previous quantifications (Raj *et al.*,

2006; Suter *et al.*, 2011; Dar *et al.*, 2012; Zenklusen *et al.*, 2008; Chubb *et al.*, 2006; Singh *et al.*, 2012), whereas the means closed times were much more markedly clone-specific (roughly distributed between 12 hours and 3.5 days). This result suggests that the duration of the chromatin closed state could explain the basal stochastic gene expression differences observed between the six clones, in contrast to the mean burst size, for which values overlapped when considering all the best parameter sets. Therefore, the mean closed time could be an essential relevant parameter involved in the regulation of stochastic gene expression. The simulation demonstrated the existence of a highly bursty transcription process. It is noteworthy that a previous study using the CMV promoter did not observe such transcriptional bursts or intervals of inactivity (Yunger *et al.*, 2010) ; however, that study used timescale analysis with a window that was significantly shorter than that used in our work. The use of the CMV promoter was essential for our study. In addition to overcome technical bias (see Methods), the fact that, using a strong promoter, we found significant differences between clones in gene-expression dynamics, and therefore genomic-integration sites, suggests that the source of the observed noise is related to the gene context (for example, chromatin state). The study strongly suggests that similar results could be obtained using a weaker endogenous promoter. Recent literature seems to corroborate this hypothesis ; in the recent work of Dar et al., the authors showed that the genomic-integration site influences burst kinetics, with a the promoter type having a marginal influence (Dar *et al.*, 2012). Understanding promoter-specific effects would require abolishing the context effect that is, performing a study using different promoters in a controlled genomic location. This is currently being addressed in our group.

The results presented here also show how, using a two-state model and fluorescence distributions measured by flow cytometry, possible chromatin-dynamics parameters can be identified. In this study, the filtering of promoter activity by mRNA and protein dynamics allows inference of temporal information from a steady-state measurement (that is, fluorescence distributions). In this regard, the mRNA and protein half-lives are the components that define the range of timescales that can be assessed from the experiment. Using destabilized reporters (Suter *et al.*, 2011; Harper *et al.*, 2011; Dar *et al.*, 2012; Singh *et al.*, 2012) would probably improve the precision of our approach towards faster timescales, provided that the fluorescence signal remains sufficiently strong to be detected by flow cytometry. In such cases, it should be possible to resolve burst duration ($1/k_{off}$) and transcription rate (ρ) separately. Note, however, that having half-lives that are too short could impair the ability to probe long timescales, such as the time between bursts. In addition, resolving experimentally the full distribution of chromatin open/closed times (that is, the distributions of k_{on} and k_{off}) is only possible with single-cell time-lapse experiments (Suter *et al.*, 2011; Harper *et al.*, 2011).

Finally, using our mathematical model, we simulated a situation in which the chromatin dynamics were directly modified by TSA. As expected, TSA treatment activated the mean reporter-gene expression (Feng *et al.*, 2001; Pikaart *et al.*, 1998; Grassi *et al.*, 2003) and seemed to increase the fraction of time spent in the "on" phase, probably as a result of a permissive chromatin state (Raser, 2004; Miller-Jensen *et al.*, 2011; Bar-Even *et al.*, 2006). The direct consequence of this treatment was a gradual shift of the distributions towards higher fluorescence values. After testing several possible chromatin dynamics leading to chromatin opening, our model was able to produce simulated distributions that efficiently fitted the flow-cytometry values during most of the TSA treatment. Moreover,

the results suggest that TSA treatment does not increase the duration of the individual "on" phase, but rather increases the frequency of these phases by reducing the duration of the "off" phase, thus globally increasing the relative proportion of "on" phases, and hence increasing the transcriptional activity. It should be notes that, owing to the instantaneous modification of the chromatin dynamics imposed in the model, the simulated distributions were a poor fit to the flow-cytometry data during the first stage of the treatment, whereas they were a perfect fit at the end of the treatment. In order to analyze the kinetics of chromatin opening, a significant improvement of our model would be to perform more precise modeling of treatment kinetics leading the new chromatin dynamics. Our study highlights the importance of chromatin-opening events in the regulation of transcription. It suggests that, to fine-tune the level of expression variability of a gene, higher eukaryotic cells might act on the chromatin mean closed time. This result provides new clues about the mechanisms involved in stochastic gene-expression regulation by chromatin remodeling.

Our work suggests that the probability of chromatin entering an open state is a key determinant of gene expression in our system. A recent study in *Escherichia coli*, using a somewhat different strategy, identified that the k_{off} parameter (probability of shifting into a transcriptionally closed state) was the main parameter used by the bacterium for gene upregulation (So *et al.*, 2011), which is therefore in sharp contrast to our own results. This might be related to the different biophysical nature of the "on" and "off" states in prokaryotes versus eukaryotes, owing to the specific nature of chromatin in eukaryotes. Finally, our results also emphasize the very slow dynamics of chromatin. Indeed, this work suggests that, depending on the genomic location of the transgene, chromatin can stay in a closed state for days, switching only occasionally to an open active state. This emphasizes the slowness of the stochastic-expression process. However, it is important to note that even if chromatin seems to be a major player in regulating gene-expression noise, we did not explore the numerous other possible sources of stochasticity such as cellular division (Huh et Paulsson, 2011a,b), elongation dynamics (Dobrzynski et Bruggeman, 2009), the combinatorial interplay of complexes at the promoter (Coulon *et al.*, 2010), presence of transcription factories (Chubb et Liverpool, 2010), and other spatial aspects (Van Zon *et al.*, 2006). Solutions for dissecting the contribution of all the components of the regulation of stochastic gene expression could be found by 1) dedicated experimental studies, as for example in the recent work by Singh et al., in which the authors proposed a method to discriminate between mRNA birth/death and promoter fluctuations as intrinsic sources of noise (Singh *et al.*, 2012), coupled with 2) a progressive increase in the model complexity based on advances in our understanding of the different mechanisms involved in the stochasticity of gene expression.

2.4 Conclusions

In this study, we highlight the importance of the dynamics of chromatin in the control of cell-to-cell variability. Our results suggest that long periods of "off" time (during which transcription does not occur) followed by brief period of "open" times (with a strong transcriptional activity) can best explain the observed difference between clones in terms of stochastic gene expression. This paves the way for future studies exploring the role of chromatin dynamics at a more local scale.

2.5 Methods

2.5.1 Cell culture

All experiments were performed on 6C2 cells, a chicken erythroblast cell line transformed by the avian erythroblastosis virus (*AEV*) (Beug *et al.*, 1979, 1982). Cells were maintained in alpha minimal essential medium (Gibco-BRL,Gaithersburg, MD, USA) supplemented with 10% (v/v) fetal bovine serum, 1% (v/v) normal chicken serum, 100 μmol/l β-mercaptoethanol (Sigma-Aldrich, St Louis, MO, USA), 100 units.ml^{-1} penicillin and 100 μg.ml^{-1} streptomycin (Gibco-BRL), at a maximum density of 1x10^6 cells per ml.

2.5.2 Generation of stably transfected clones

Stably transfected clones, expressing a fluorescent reporter, were obtained as previously described (Beug *et al.*, 1979). Briefly, 6C2 cells were nucleofected using the Cell line Nucleofector® Kit V (Lonza) in a Nucleofector™ II (Amaxa Nucleofector™ Technology) (T-16 program) by a pT2.CMV-*mCherry*/pCAGGS-T2TP plasmid mix (ratio 5/1). pT2.CMV-*mCherry* plasmid was constructed following the same strategy as described for the pT2.CMV-*hKO* plasmid (Beug *et al.*, 1979) except that the *hKO* reporter gene was replaced by *mCherry* extracted from the pRSET-B plasmid (kindly provided by Dr. Roger Tsien, University of California, San Diego). The integration into genomic DNA of the reporter is allowed by the Tol2 transposon system (Beug *et al.*, 1982) the CMV-*mCherry* sequence, flanked by Tol2 motifs, is recognized by a transposase (pCAGGS-T2TP) and randomly inserted into 6C2 genomic DNA. Seven days after transfection, stably transfected cells expressing the reporter gene were sorted and individually cloned in U-shape 96-well microplates (Cellstar Greiner bio-one) using a FACSVantage SE cytometer (Becton-Dickinson).

2.5.3 Generation of stably transfected clones

Stably transfected clones, expressing a fluorescent reporter, were obtained as previously described (Mejia-Pous *et al.*, 2009). Briefly, 6C2 cells were nucleofected in a transfection apparatus (Nucleofector™ II ; Amaxa Nucleofector™ Technology) (T-16 program) using a commercial kit (Cell Line Nucleofector® Kit V ; Lonza GmBH, Cologne, Germany) and a pT2.CMV-*mCherry*/pCAGGS-T2TP plasmid mix (ratio 5/1). The pT2.CMV-*mCherry* plasmid was constructed using the same strategy as described for the pT2.CMV-hKO plasmid (Mejia-Pous *et al.*, 2009), except that the hKO reporter gene was replaced by *mCherry*, extracted from the pRSET-B plasmid (kindly provided by Dr Roger Tsien, University of California, San Diego, CA, USA). mRNA birth/death fluctuations constitute a major source of stochasticity in gene expression because many mRNA species are present at very low molecular counts within cells (Bar-Even *et al.*, 2006; Singh *et al.*, 2012; Newman *et al.*, 2006; Taniguchi *et al.*, 2010), thus we reduced this source of intrinsic noise by using the cytomegalovirus (CMV) promoter. Obtaining a strong signal also allowed us to overcome bias caused by autofluorescence in the flow-cytometry data. The integration into genomic DNA of the reporter is allowed by the Tol2 transposon system (Kawakami et Noda, 2004) ; the CMV-*mCherry* sequence, flanked by Tol2 motifs, is recognized by a transposase (pCAGGS-T2TP), and randomly inserted into 6C2

genomic DNA. Seven days after transfection, stably transfected cells expressing the reporter gene were sorted and individually cloned in U-shaped 96-well microplates (Cellstar Greiner Bio-One GmBH, Frickenhausen, Germany) using a cytometer (FACSVantage SE ; Becton-Dickinson, Franklin Lakes, NJ, USA).

2.5.4 Molecular and cellular characterization of clones

For each clone, the genomic reporter insertion sites were identified using a splinkerette PCR method as previously described (Mejia-Pous *et al.*, 2009), in order to select only clones with a single insertion site. Briefly, genomic DNA isolated from clones expressing the gene reporter was purified by phenol extraction and ethanol precipitation, before being digested for 16 hours at 65°C with TaiI, a restriction enzyme with a 4 bp recognition site. The digested DNA was then ligated to a splinkerette adaptor for 1 hour at 22°C. After purification of the ligated product, two rounds of PCR (PCR1 and nested PCR2) were performed using primers specific for the reporter transgene *mCherry* and for the annealed splinkerette adaptor, and a commercial polymerase (AccuPrimeTM Taq DNA Polymerase High Fidelity ; Invitrogen Inc., Carlsbad, CA, USA). The PCR products were then purified and sequenced. Finally, the genomic reporter insertion sites were identified by similarity searches using the sequence analysis tool iMapper (Kong *et al.*, 2008). The identification of the insertion sites of the selected clones was confirmed using a high-throughput splinkerette-PCR method (Uren *et al.*, 2009), allowing the analyses of hundreds of clones. This work will be described in details elsewhere.

For characterization of clones and analysis of treatment effects (see below), flow-cytometry analyses were performed (FACSCanto II ; Becton-Dickinson) on cells extemporaneously pelleted and resuspended in Dulbecco's phosphate-buffered saline 1x solution (Gibco-BRL). Each sample was analyzed using an acquisition of 50,000 events (gated on living cells), and the positive fluorescence threshold was fixed using non-transfected cells. Possible variability resulting from flow-cytometer calibration was taken into account by systematically analyzing flow-calibration particles (SPHERO^{TM} Rainbow ; Spherotech Inc., Lake Forest, IL, USA), as a calibration reference.

Non-transfected cells were used to measure 6C2 native autofluorescence, and the difference between the fluorescence of transfected and non-transfected ones was used as an indicator of the transgene activity (note that autofluorescence was also systematically added to the model's output to compute the distribution distance scores).

For each clone, two indicators were systematically used : MFI (mean fluorescence intensity) and NV (the variance divided by the square mean).

For a given cell, the measured fluorescence f (from the flow cytometer) is $f = f_t + f_a$; that is, the sum of the true fluorescence f_t (coming from the reporter proteins) and the autofluorescence f_a (coming from the rest of the cell). The autofluorescence is not a constant, but has a distribution that is obtained using non-transfected cells. The two first moments of f read simply as $\langle f \rangle = \langle f_t \rangle + \langle f_a \rangle$ and $\sigma^2(f) = \sigma^2(f_t) + \sigma^2(f_a)$.

Hence, with MFI and NV being the mean and normalized variance of the true fluorescence, we get : $MFI = \langle f \rangle - \langle f_a \rangle$ and $NV = \frac{\sigma^2(f) - \sigma^2(f_a)}{(\langle f \rangle - \langle f_a \rangle)^2}$

Finally, to compare the theoretical distributions obtained from simulations (which only included the reporter fluorescence) with those obtained from experiments (which also included the autofluorescence), the model's output was first combined with the experimental

autofluorescence. This was carried out by summing each simulation result with the value of a randomly selected cell from the autofluorescence distribution. The resulting distribution was the convolution between the theoretical and the autofluorescence distributions, and was then compared with the experimental distributions using a Kolmogorov-Smirnov test.

2.5.5 Determination of *mCherry* mRNA and protein degradation rates

To determine the *mCherry* mRNA degradation rate, the mRNA concentration was estimated using quantitative reverse transcription (qRT)-PCR after transcription inactivation was achieved using actinomycin D treatment. Two clones (C5 and C11) were treated, in duplicate, for 0, 60, 124, 244 and 488 minutes with a final concentration of 10 $\mu g.ml^{-1}$ actinomycin D (A9415 ; Sigma-Aldrich), before extracting the mRNA after the instructions of RNeasy® Plus Mini Kit (Qiagen Inc., Valencia, CA, USA). To prepare the real-time PCR assay, 1 μg of total RNA from each sample was reversed transcribed using the SuperScript™ III First-Strand Synthesis System for RT-PCR (Invitrogen Inc.) in the presence of random hexamers. Quantification of mRNA levels by real-time PCR was performed in 96-well plates using a real-time PCR system (LightCycler 480 ; Roche Diagnostics, Basel, Switzerland). The measurement was performed in a final volume of 10 μl of reaction mixture (containing 2.5 μl of cDNA template diluted 1 in 5), prepared using a commercial kit (Light Cycler 480 SYBR Green I Kit ; Roche Diagnostics) in accordance with the manufacturer's instructions, and with the primer set at a final concentration of 0.5 $\mu mol.l^{-1}$ (mCher-For : CCACCTACAAGGCCAAGAA, mCher-Rev : ACTTGTACAGCTCGTCCATG). An internal standard curve was generated using serial dilutions (from 2000 to 0.02 $fg.\mu l^{-1}$) of purified PCR product. The reactions were initiated by activation of Taq DNA polymerase at 95°C for 5 minutes, followed by 45 three-step amplification cycles consisting of denaturation at 95°C for 15 seconds, annealing at 55°C for 15 seconds, and extension at 72°C for 15 seconds. The fluorescence signal was measured at the end of each extension step. After the amplification, a dissociation stage was run to generate a melting curve for verification of amplification-product specificity. The crossing point (CP) was determined by the second derivative maximum method in the LightCycler® 480 software (version 1.5.0). After normalization, taking into account cellular viability and mRNA quantity used for the retrotranscription step, the mRNA half-life was determined by fitting mRNA quantity evolution by a decreased exponential (least square) method.

To determine the *mCherry* protein degradation rate, we used flow cytometry to measure the protein half-life after translation inactivation using cycloheximide treatment. C5 and C11 clones were treated in duplicate for 0, 16, and 24 hours with a final concentration of 100 $\mu g.\mu l^{-1}$ cycloheximide (C4859 ; Sigma-Aldrich), and for each time point, the fluorescence of the treated cells was measured by flow cytometry. The autofluorescence component was removed as explained earlier. The protein half-life was determined using exponential fit of the fluorescence mean decrease curve, similarly to the procedure used for determining the mRNA half-life.

2.5.6 Treatments with chromatin-modifying agents

To analyze the effect of chromatin state on the stochasticity of gene expression, clones were treated with TSA, a histone deacetylase inhibitor (P5026 ; Sigma-Aldrich) and 5-AzaC, an inhibitor of DNA methylation (A2385 ; SigmaAldrich). For each clone, kinetic treatment experiments were performed ; clones were treated with 500 nmol/l TSA or 500 μmol/l 5-AzaC at five time points (0, 8, 24, 32, and 48 hours). For each time point, 1x10^6 cells (for 0, 8, and 24 hours) or 5x10^5 cells (for 32 and 48 hours) were treated with the relevant drug and characterized by flow cytometry.

2.5.7 Model description

The two-state model of gene expression represents the chromatin activity as an "on-off" process specified through the transition rates k_{on} and k_{off} (respectively representing the "off-on" transition and the "on-off" transition). To enable comparison with the experimental data, a simple model of mRNA and protein dynamics based on two production/degradation models completed the model. The production of mRNA was allowed only in the "on" state (open chromatin) but completely forbidden in the "off" state (closed chromatin). The model thus corresponds to the following equations :

$$\begin{cases} k_t &= \dfrac{k_{on}}{k_{on}+k_{off}} \\ \dfrac{dR}{dt} &= \rho k_T - \tilde{\rho} R \\ \dfrac{dP}{dt} &= \gamma R - \tilde{\gamma} P \\ f &= \alpha P \end{cases} \tag{III.1}$$

where, k$_{on}$ is the closed-to-open transition rate, k$_{off}$ is the open-to-closed transition rate, and k$_t$ is the resulting proportion of the "on" state ; R is the number of mRNAs, ρ is the mRNA production rate (when chromatin is open), and $\tilde{\rho}$ is the mRNA degradation rate ; P is the number of *mCherry* proteins, γ is the *mCherry* production rate (per mRNA) and $\tilde{\gamma}$ is the *mCherry* degradation rate ; f is the fluorescence intensity of the cell (after substraction of the autofluorescence) and α, is a linear proportionality coefficient to convert the number of proteins into arbitrary fluorescence measures.
This model can be simulated with the SSA (see below) to ascertain the behavior of single cells and eventually to compute the fluorescence distributions. It can also be analytically derived to compute the MFI and NV of large cell populations at steady state.

2.5.8 Analytical derivation of the model

Paulsson proposed an analytic expression of the mean quantity and NV of protein in the two-state model, as a function of chromatin-dynamics parameters and transcription-translation parameters (Paulsson, 2005a). In the case of a single gene and taking into account the parameter a, Paulsson's equation gives :

$$\begin{cases} MFI &= \alpha \dfrac{\rho}{\tilde{\rho}} \dfrac{\gamma}{\tilde{\gamma}} \dfrac{k_{on}}{k_{on}+k_{off}} \\ NV &= \left(\dfrac{\tilde{\rho}}{\rho} \dfrac{\tilde{\gamma}}{\gamma} \dfrac{k_{on}+k_{off}}{k_{on}} \right) + \left(\dfrac{\tilde{\rho}}{\rho} \dfrac{k_{on}+k_{off}}{k_{on}} \dfrac{\tilde{\gamma}}{\tilde{\rho}+\tilde{\gamma}} \right) + \left(\dfrac{k_{off}}{k_{on}} \dfrac{\tilde{\gamma}}{\tilde{\rho}+\tilde{\gamma}} \dfrac{1+\dfrac{\tilde{\rho}}{\tilde{\rho}+k_{on}+k_{off}}}{1+\dfrac{k_{on}+k_{off}}{\tilde{\rho}}} \right) \end{cases} \tag{III.2}$$

This equation can be used to express k_{on} and k_{off} as a function of MFI, NV, and the transcription-translation parameter sets. Rewriting the equation gives :

$$\begin{cases} A &= \left(1 + \dfrac{\alpha \dfrac{\rho}{\tilde{\rho}} \dfrac{\gamma}{\tilde{\gamma}} - MFI}{MFI} \right) \\ B &= \dfrac{\left(A \left(\dfrac{\tilde{\rho}}{\rho} \right) \left(\dfrac{\tilde{\gamma}}{\gamma} + \dfrac{\tilde{\gamma}}{\tilde{\rho}+\tilde{\gamma}} \right) \right) - NV}{(A-1)\dfrac{\tilde{\gamma}}{\tilde{\rho}+\tilde{\gamma}}} \\ 0 &< (A(B(\tilde{\gamma}+\tilde{\rho})+\tilde{\rho}))^2 - 4A^2B\left(\tilde{\rho}(\tilde{\rho}+\tilde{\gamma}(B+1))\right) \\ k_{on} &= \dfrac{-\left(A(B(\tilde{\gamma}+\tilde{\rho})+\tilde{\rho})\right) - \sqrt{(A(B(\tilde{\gamma}+\tilde{\rho})+\tilde{\rho}))^2 - 4A^2B\left(\tilde{\rho}(\tilde{\rho}+\tilde{\gamma}(B+1))\right)}}{2A^2B} \\ k_{on} &> 0 \\ k_{off} &= (A-1)k_{on} \end{cases} \tag{III.3}$$

2.5.9 Parametric exploration of the analytical model

Because the clonal populations differed only in their insertion points (that is, their chromatin-dynamics parameters), equation 3 enabled us to find the clone-specific parameters from MFI and NV (measured by flow cytometry) and the transcription-translation parameters ρ, $\tilde{\rho}$, γ, $\tilde{\gamma}$ and α. $\tilde{\rho}$ and $\tilde{\gamma}$ can be determined experimentally (see above) but ρ, γ and α remained unknown. We explored a wide range of these parameters (large enough to include all biologically relevant values) : ρ = 6.0, 1.0, 0.5, 0.333, 0.25, 0.200, 0.1666, 0.14286, 0.125, 0.111, 0.100, 0.0500, 0.0333, 0.0250, 0.02, 0.01666, 0.01333, 0.0111, 0.00952, and 0.00833 mRNA/min, corresponding to one mRNA produced each $1/\rho$ = 10 seconds, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, and 50 minutes, and 1, 1.25, 1.5, 1.75, and 2 hours, when chromatin is in the open state ; γ = 1.0, 0.200, 0.100, 0.0333, 0.01667, 0.006667, 0.00333, 0.00222, 0.001666, 0.001333, 0.00111, 0.0008333, 0.00069444, 0.00046296, and 0.0003472 protein/min/mRNA, corresponding to a protein produced each $1/\gamma$ = 1, 5, 10, and 30 minutes, 1, 2.5, 5, 7.5, 10, 12.5, 15, and 20 hours, and 1, 1.5 and 2 days per mRNA molecule ; and α = 0.10, 0.15, 0.50, 1.0, 1.5, 5.0, 10.0, 15.0, 50.0, 100.0, 150.0, and 200.0 arbitrary units.

Exploring all values of ρ, γ and α gave us 3600 couples (k_{on} ; k_{off}) of which only 1,047 respected the condition mentioned in equation 3 ($k_{on} > 0$).

2.5.10 Comparison between the analytical model and the trichostatin A-treated clones

Equation 1 enabled us to compute the mean mRNA number (R) and the mean protein number (P) at steady state, from the values of the chromatin-dynamics and transcription-translation parameters :

$$\begin{cases} \langle R \rangle &= \dfrac{\rho}{\tilde{\rho}} \dfrac{k_{on}}{k_{on}+k_{off}} \\ \langle P \rangle &= \dfrac{\rho}{\tilde{\rho}} \dfrac{\gamma}{\tilde{\gamma}} \dfrac{k_{on}}{k_{on}+k_{off}} \end{cases} \tag{III.4}$$

Then, assuming that, at t = 0 , the cell switches to a new chromatin dynamics (because of the TSA treatment), compute R(t), (the evolution of mRNA number), and P(t), (the evolution of protein number following the TSA treatment) can be computed. If k_{on}^{TSA} and k_{off}^{TSA} are the new chromatin-dynamics parameters induced by the treatment, the equation is :

$$\begin{aligned} R(t) = & \quad \frac{\rho}{\tilde{\rho}} & \left(\frac{k_{on}}{k_{on}+k_{off}} - \frac{k_{on}^{TSA}}{k_{on}^{TSA}+k_{off}^{TSA}} \right) e^{-\tilde{\rho}t} \\ & + \frac{\rho}{\tilde{\rho}} & \frac{k_{on}^{TSA}}{k_{on}^{TSA}+k_{off}^{TSA}} \\ P(t) = & \left(\frac{\rho}{\tilde{\rho}} \frac{\gamma}{\tilde{\gamma}} - \frac{\rho}{\tilde{\rho}} \frac{\gamma}{\tilde{\gamma}-\tilde{\rho}} \right) & \left(\frac{k_{on}}{k_{on}+k_{off}} - \frac{k_{on}^{TSA}}{k_{on}^{TSA}+k_{off}^{TSA}} \right) e^{-\tilde{\gamma}t} \\ & + \left(\frac{\rho}{\tilde{\rho}} \frac{\gamma}{\tilde{\gamma}-\tilde{\rho}} \right) & \left(\frac{k_{on}}{k_{on}+k_{off}} - \frac{k_{on}^{TSA}}{k_{on}^{TSA}+k_{off}^{TSA}} \right) e^{-\tilde{\gamma}t} \\ & + \left(\frac{\rho}{\tilde{\rho}} \frac{\gamma}{\tilde{\gamma}} \right) & \left(\frac{k_{on}^{TSA}}{k_{on}^{TSA}+k_{off}^{TSA}} \right) \end{aligned} \tag{III.5}$$

The exact values of k_{on}^{TSA} and k_{off}^{TSA} remained unknown at this stage, but we could simulate the extreme situation by assuming that, under TSA treatment, the chromatin is fully open. Analytically, this gives :

$$\frac{k_{on}^{TSA}}{k_{on}^{TSA}+k_{off}^{TSA}} = 1 \tag{III.6}$$

Note that this equation represents an extreme situation, not the exact TSA influence on chromatin.

Introducing equation 6 into the dynamics of equation 5, we were able to compute, for a given transcription-translation parameter set, the maximum rate of protein concentration increase, and thus the maximum increase of reporter fluorescence. For each parameter set, we compared the predicted fluorescence increase under the extreme condition of a fully open chromatin. We then rejected all parameter sets for which the protein number did not increase sufficiently rapidly to account for the fluorescence increase measured experimentally during TSA treatment.

2.5.11 Simulation of the model

The model can be simulated using an SSA, which is an exact continuous-time algorithm that enables simulation of chemical-reaction systems (Gillespie, 1977). Each simulation represents one of the possible realizations of the system from a specified initial state and for a given kinetic parameter set (these parameters being here considered as probabilities). Each realization depends on a pseudo-random generator, and different realizations (that is, simulations of different cells issued from the same clone) can be computed by simply initializing this random generator with different seeds. The implementation of the two-state model (equation 1) in the SSA enables simulation of the entire system dynamics and visualization of the course of chromatin state, gene transcription, number of mRNAs, mRNA translation, number of proteins and, ultimately fluorescence, in a virtual single cell. By simulating a large number of such "artificial cells", we were able to simulate "virtual flow-cytometry experiments" and to compute MFI, NV, and full distribution for a given parameter set. We simulated 50,000 virtual cells for 30,000 minutes (a sufficiently long period to ensure that all cells were at a steady state, the concentration values being initialized at the theoretical values given by the analytical model). The fluorescence of each cell was then computed, and the simulated distribution generated through convolution with the autofluorescence of the 6C2 cells measured experimentally (see above). Simulated distributions were then compared with the experimental distribution using the Kolmogorov-Smirnov test. The quality of each parameter set was then evaluated (the score of a given parameter set being the mean Kolmogorov-Smirnov score of each clone). The best parameter set was thus the one that gave the best fit for all six clonal populations.

2.5.12 Simulation of trichostatin A treatment in the model

Using the best parameter set, we simulated 50,000 cells of the two TSA-treated clones C5 and C11 for 30,000 minutes. The chromatin-dynamics parameters were then modified to account for the TSA treatment, and the two clones were simulated for a further 1,152 minutes (48 hours). For each clones, the simulated distributions were computed after 8, 24, 32 and 48 hours, and compared with the experimental distributions using a Kolmogorov-Smirnov test. The best chromatin-dynamics parameters (k_{on}^{TSA}, k_{off}^{TSA}) were those that gave the best mean score at the four time points. In total, 11 different chromatin-dynamics values were tested for each clone. Note that, knowing the MFI value of the treated clones, we could analytically compute the k_{T}^{TSA} value (using equation 5). Taking $k_T^{TSA} = \dfrac{k_{on}^{TSA}}{k_{on}^{TSA} + k_{off}^{TSA}}$ (from equation 1), we can use this analytical value to simplify the parametric exploration.

Author Contributions

JV, GK, AC, GB, and OG conceived and designed the research. JV performed all the biological experiments. GK performed all the computational analyses and model simulations. AC provided assistance for the modeling analyses. EV and VM provided technical assistance for the biological experiments. CMP and JJK advised on the design and interpretation of the experiments. JV, GK, AC, GB, and OG wrote the paper. OG and GB co-supervised the project. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank François Chatelain, Alexandra Fuchs, and Manuel Théry for helpful discussions and support during the early stages of the project. We are grateful to Denis Ressnikoff of the platform Centre Commun de Quantimétrie de Lyon (CCQ) for flow-cytometry cell-sorting assistance. We thank the Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules de Lyon (CC-IN2P3), and especially Pascal Calvat, for their computing resources. We also thank the interns who worked on this project : Mathieu Gineste, Yoann Ménière, Charles Rocabert. and Balthazar Rouberol. We thank the Andras Paldi group from the Généthon for the constructive and useful discussions on chromatin and stochasticity of gene expression.
This work was supported by funding from the Institut Rhônalpin des Systèmes Complexes (IXXI) and from the Réseau National des Systèmes Complexes (RNSC). Part of the project was supported by an ANR grant (ANR 2011 BSV6 014 01). JV is supported by a CNRS post-doctoral grant and GK is a PhD fellow from the Région Rhône Alpes and INRIA.

3 Conclusion

Cette étude nous a permis de démontrer, par un couplage étroit entre biologie expérimentale et biologie *in silico* qu'il est possible d'inférer, par l'identification paramétrique du modèle, de données de cytométrie en flux jusqu'aux paramètres dynamiques d'ouverture-fermeture de la chromatine. Elle nous a aussi permis de montrer que ces paramètres dynamiques correspondent à un régime d'intermitence caractérisé par de longues périodes durant lesquelles la chromatine interdit la transcription, entrecoupées par de brêves périodes où la transcription est autorisée. Enfin, nous avons vu que les différences inter-clones correspondent essentiellement à des variations du paramètre k_{on} (taux d'ouverture), autrement dit, à des variations du temps moyen pendant lequel la chromatine reste fermée. Ces résultats nous amènent donc à formuler l'hypothèse suivante – qu'il conviendra de vérifier : la séquence d'ADN modifie la probabilité d'ouverture de la chromatine alors que la probabilité de fermeture est relativement indépendante du locus.
Par ailleurs, une fois établi le lien entre les paramètres de la stochasticité de l'expression génique (mesurés sur les histogrammes de fluorescence) et les paramètres de la dynamique chromatinienne, nous avons pu réutiliser notre modèle pour estimer l'effet du traitement TSA sur la chromatine. Nous avons ainsi montré que la TSA agit principalement sur le paramètre k_{on}, qu'elle augmente, racourcissant ainsi les temps moyens fermés. Notre modèle n'incluant pas les détails moléculaires de l'ouverture-fermeture, il n'est pas possible de remonter aux causes moléculaires de ce racourcissement. Une hypothèse pourrait être que l'inhibition des HDAC facilite l'acétylation des histones par les HAT, ce qui permettrait de décompacter plus facilement la chromatine, mais rien ne nous permet de valider/invalider une telle hypothèse avec les outils forgés ici.
En conclusion, nos résultats suggèrent qu'une grande partie de la stochasticité de l'expression génique est explicable par des mécanismes dépendant de l'environnement du transgène via la dynamique chromatinienne. Il semble donc important d'étudier plus en détail et plus localement cet environnement pour mieux comprendre la stochasticité de l'expression génique.

Le modèle computationnel ne permet pas d'obtenir une correspondance parfaite entre les distributions simulées et celles observées expérimentalement. Cette différence peut néanmoins s'expliquer en partie par l'utilisation d'un modèle très simple. Bien entendu, il est toujours possible d'introduire plus de complexité dans le modèle de façon à mieux "coller" aux données mais il s'agit là d'une quête assez vaine. Le compromis simplicité-fidélité du modèle présenté dans ce chapitre nous paraît en effet particulièrement intéressant, ce qui – même si un tel modèle ne puisse jamais être validé à 100% – nous donne une bonne confiance dans la structure du modèle.
Une exception cependant serait le modèle du traitement à la TSA. En effet, on constate que la dynamique temporelle de l'accroissement de fluorescence dans les premières heures qui suivent le traitement n'est pas fidèlement reproduite par le modèle. Cependant, cette divergence entre les distributions *in silico* et *in vitro* semble essentiellement due à une modélisation très simplifiée de la cinétique du traitement TSA (pharmacocinétique). En effet, nous avons supposé ici que la TSA est totalement active dès le début du traitement ce qui n'est évidemment pas le cas en pratique. Cela explique très probablement l'avance des premières distributions *in silico* sur celles *in vitro*. Il est important de noter que cette

approximation concerne la cinétique du traitement et non le modèle du système biologique étudié.
Bien entendu, il serait toujours possible de complexifier le modèle de la dynamique chromatinienne. En effet, certaines études telles que celle de Suter *et al.* (2011), ont montré que l'utilisation d'un modèle à trois états (présentant une période réfractaire) permettait de mieux rendre compte des données temporelles. Le fait que cette période réfractaire n'ai pas été nécessaire ici n'implique pas qu'elle n'existe pas mais simplement que la précision de nos mesures expérimentales (réalisées en cytométrie de flux) ne permet pas de la dicerner – si elle existe. Même si on peut toujours complexifier le modèle, il convient aussi de rester raisonnable compte-tenu des données disponibles pour l'identification des paramètres. Dans notre cas, les cinq histogrammes mesurés expérimentalement (ajoutés aux traitement TSA) nous ont permis de déduire un très grand nombre d'informations mais plus de données seraient nécessaires pour aller plus loin. Complexifier le modèle permettrait certes une meilleure compréhension et plus de réalisme mais à conditions de disposer *aussi* de données expérimentales adaptées (par exemple des données time-lapse). Il en va de même pour le réalisme de la chaine de transcription-traduction, qui pourrait par exemple inclure la maturation des protéines fluorescentes (De Jong *et al.*, 2010), la croissance ou la division cellulaire (Huh et Paulsson, 2011a) mais là encore, il faudrait disposer de données supplémentaires pour pouvoir identifier les paramètres supplémentaires que cette complexification apporterait.
Plutôt que de chercher à rendre le modèle toujours plus complexe, il nous semble plus intéressant de chercher à croiser ce modèle avec d'autres données biologiques. En l'occurrence, pour chaque clone, nous disposons de trois types d'information :

Le locus du transgène Obtenu par splinkerette PCR,

La distribution de fluorescence Obtenue par Cytometrie en flux (FACS),

Les paramètres de la dynamique chromatinienne Obtenus par identification paramétrique du modèle.

À partir des données de locus, nous pouvons obtenir de nouvelles informations sur le contenu de la séquence d'ADN autour du point d'insertion. Dans le prochain chapitre, nous essayerons d'identifier les caractéristiques génomiques (plus ou moins locales) les plus à même d'expliquer la locus-dépendance de la stochasticité de l'expression génique. Nous continuerons ainsi à investiguer les diverses sous-problématiques connexes à cette question, démarche que nous estimons plus efficace que de s'en rapprocher au maximum mais en un seul point.

Chapitre IV

Effet de l'environnement génomique sur l'expression stochastique des gènes

1 Introduction

Le processus d'expression des gènes démarre par le mécanisme d'initiation de la transcription, mécanisme dont la complexité varie entre les espèces mais qui, chez les eucaryotes supérieurs, peut nécessiter un grand nombre d'étapes liées en particulier au processus de recrutement d'un très grand nombre de facteurs de transcription sur ou à proximité du promoteur (Green, 2005; Stavreva *et al.*, 2012). C'est aussi ce processus d'initiation qui est supposé influer sur les aspects quantitatifs de l'expression, y compris dans leur composante stochastique (voir section 2.2). Ceux-ci sont donc très probablement en grande partie déterminés par les caractéristiques du promoteur, de sa séquence nucléique ou des séquences qui le flanquent en amont ou en aval (chez les eucaryotes supérieurs). Même si on sait que des interactions à très longue portée peuvent fortement influer sur l'activité transcriptionnelle d'un promoteur. On peut s'attendre à trouver un lien causal entre les caractéristiques génomiques du promoteur d'un gène (élargi à la séquence proche) et l'expression de ce dernier.
De fait, il a été montré, d'abord chez la levure (Becskei *et al.*, 2005), puis chez des organismes eucaryotes supérieurs (Nie *et al.*, 2010), que le niveau moyen d'expression d'un gène dépend de sa position sur le génome. Chez l'homme, des études quantitatives par SAGE ont permis de mettre en évidence que l'expression moyenne d'un gène dépend de son locus (Versteeg *et al.*, 2003; Caron *et al.*, 2001). Chez plusieurs espèces d'eucaryotes supérieurs, il a été montré que le génome comporte des zones en moyenne plus exprimées (RIDGE[1]) tandis que d'autres sont en moyennes sous-exprimées (Anti-RIDGE). En outre, leurs caractéristiques génomiques différent. Ainsi, la densité en gènes est plus élevée et la taille moyenne des introns de ces gènes est plus courte dans les RIDGEs que dans le reste du génome (Versteeg *et al.*, 2003; Nie *et al.*, 2010).

Dans les chapitres précédents, nous avons montré que la moyenne et la stochasticité de l'expression d'un rapporteur dépend de son locus d'insertion (chapitre II) et que cette

[1]"Regions of IncreaseD Gene Expression" ; Dans un RIDGE, la mediane du niveau d'expression de tout les gènes de la région est supérieure à un seuil. Généralement, ce seuil est fixé pour que les RIDGEs représentent 10% du génome.

stochasticité pouvait dépendre de la dynamique globale d'ouverture/fermeture de la chromatine (chapitre III). Celle-ci dépend elle-même de la dynamique locale des histones au niveau du locus du gène mais aussi des histones voisins dont l'influence peut se propager le long du chromosome (partie 2.2.4). La dynamique des histones dépend en retour des contraintes physiques subies/imposées par l'ADN en raison de sa séquence et de l'activité moléculaire locale (fixation des facteur de transcription, des polymérases, élongation de la double hélice d'ADN due à la transcription, à la réplication, ...). Gierman *et al.* (2007) ont montré que, dans les RIDGEs, la chromatine est plus ouverte que dans les anti-RIDGEs ce qui permet une expression moyenne plus forte. On peut logiquement s'attendre à ce que la stochasticité de l'expression des gènes, qui dépend de l'activité chromatinienne, dépende aussi de l'environnement génomique du gène observé. Cette hypothèse a été confirmée chez la levure par (Becskei *et al.*, 2005) puis par (Wang *et al.*, 2011) mais elle reste à étudier chez les eucaryotes supérieurs.
Dans ce chapitre, nous allons profiter du matériel biologique présenté précédemment pour explorer en détails ce lien entre structure génomique et stochasticité de l'expression d'un rapporteur fluorescent dans des cellules d'eucaryote supérieur. Nous étudierons en particulier trois types de structures génomiques correspondant à des échelles d'observation différentes :

- les macro-structures (taille du chromosome correspondant au locus ; pourcentage GC du chromosome du locus,...),
- les structures locales autour du point d'insertion (taux de GC, densité en gènes ou encore densité en régions répétées),
- les structures de premier voisinage (gène(s) le(s) plus proche(s)).

Dans la section suivante, nous allons présenter ces différentes structures puis nous étudierons les corrélations entre l'activité (moyenne et stochasticité) de rapporteurs fluorescents avec les caractéristiques de la séquence au voisinage du locus d'insertion de ce rapporteur.

1.1 Structuration du génome

Chez les eucaryotes supérieurs, les gènes et les promoteurs ne représentent qu'une toute petite partie de la séquence génomique. Gènes et promoteurs sont en effet séparés par de longues séquences souvent qualifiées de "non-codantes" même si on sait désormais qu'une grande partie d'entre elles portent la signature d'une évolution adaptative et sont donc probablement fonctionnelles, par exemple en contribuant à la régulation de l'expression des gènes (Ponting et Lunter, 2006).
Outre les gènes et leurs promoteurs, les génomes eucaryotes comportent donc un très grand nombre de structures, plus ou moins fonctionnelles et à des échelles très différentes. Certaines sont maintenues par la sélection, d'autres sont liées à l'activité mutationnelle mais n'ont pas d'influence directe sur la sélection (même s'il a été proposé qu'elles puissent avoir un effet sélectif indirect en permettant la régulation de la robustesse et de l'évolvabilité de l'organisme (Knibbe *et al.*, 2007)). Dans le cadre de notre étude, nous nous concentrerons sur deux types de structures génomiques (en plus, bien entendu, des séquences codantes) : les séquences répétées et les variations du taux de GC.

1.1.1 Les séquences répétées

Ce sont de courtes séquences répétées un grand nombre de fois, éventuellement en différents endroits du génome. Ces répétitions nucléiques sont "non codantes" mais elles ne sont pas nécessairement dépourvues d'activité moléculaire : elles peuvent intervenir dans la réplication ou la réparation de l'ADN ou au contraire avoir localement une forte activité mutagène (Lai et Sun, 2003). La répartition des séquences répétées le long du génome est très variable, certaines zones étant très riches en séquences répétées tandis que d'autres en sont totalement dépourvues. Les séquences répétées étant relativement fréquentes dans les génomes mais globalement anti-corrélées avec la présence de gènes, il est intéressant d'étudier la corrélation entre leur présence et l'expression stochastique des gènes.

1.1.2 Le taux de GC

L'ADN est composé de quatre type de nucléotides : A, T, G et C. Cependant la répartition de ces quatres nucléotides le long du génome n'est pas constante. On constate en particulier que la proportion de bases GC varie fortement. On calcule le "taux de GC" le long du chromosome par :

$$\left\{ \%GC \quad = \quad \frac{C_s + G_s}{N_s} \right. \tag{IV.1}$$

où N_s, G_s et C_s représentent respectivement dans la séquence évaluée, le nombre de nucléotides totaux, le nombre de G et le nombre de C observés.
Le taux de GC n'est pas homogène sur le génome (figure IV.1). Sa variation implique entre autres que les brins complémentaires de l'ADN sont plus solidement liés dans certaines zones que dans d'autres. En effet, le nombre de liaisons hydrogènes n'est pas le même entre les bases complémentaires G et C qu'entre les bases complémentaires A et T.
Les variations du taux de GC sont corrélées avec de nombreux paramètres. Ainsi, on constate que le taux de GC est plus élevé dans les zones du génome riches en gènes (Figure IV.2).
Ce taux est en fait corrélé à beaucoup de caractéristiques génomiques. Ainsi, si la densité en régions répétées est anticorrélée avec celle des gènes et que le taux de GC est corrélé à la densité en gènes, on peut imaginer que le taux de GC est anticorrélé à la densité en régions répétées. Ces corrélations multiples en font un observable important pour notre étude. Nous nous intéresserons donc aux corrélations d'activité de transcription avec les variations locales du taux de GC.

1.2 Principe de l'étude

Dans le chapitre précédent, nous avons mis en place une série d'outils mélangeant biologie expérimentale et biologie *in silico* et permettant d'insérer un rapporteur fluorescent dans des points aléatoires du génome puis de mesurer la stochasticité de l'expression des gènes en ces différents points par cytométrie en flux (chapitres II et III). Par ailleurs, nous disposons de tous les outils pour connaître le nombre et la localisation des points d'insertion dans le génome. Nous avons donc tous les outils nécessaires pour identifier – s'ils existent – les liens entre le contexte génomique d'un gène et son expression, y compris dans sa dimension stochastique. À la seule condition de travailler sur un organisme séquencé et

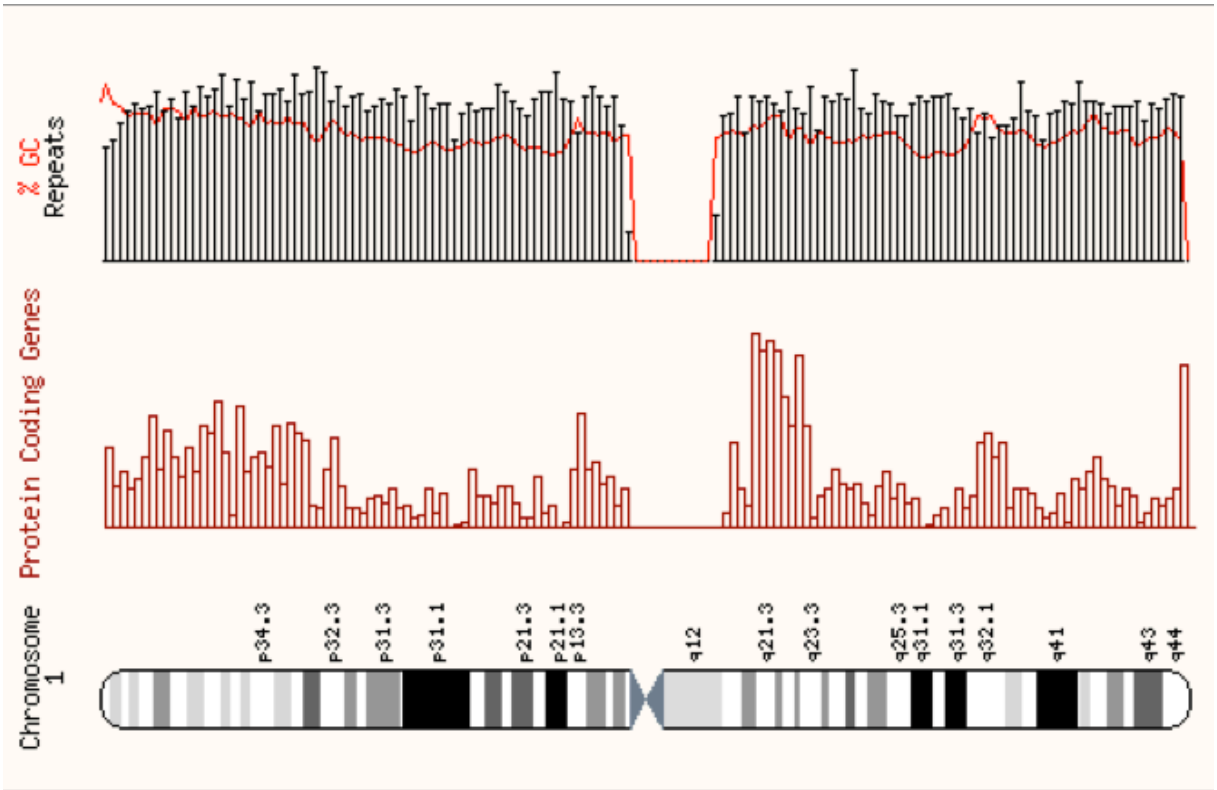

Figure IV.1 – Caractéristiques génomiques. Cette représentation du chromosome 1 de l'homme (base de données `ensembl`) montre la densité en gènes, en séquences répétées et le taux de GC le long de ce chromosome. On peut observer que les pics de densité de gènes et ceux du taux de GC sont souvent corrélés.

annoté, nous pourrons ensuite étudier les caractéristiques de la séquence aux alentours du point d'insertion et chercher à les relier à la mesure de stochasticité. Pour celà nous utiliserons la base de données `ensembl` (`http://www.ensembl.org`) et plus exactement la construction WASHUC2 (2006) du génome du poulet qui a aussi été utilisé par Nie *et al.* (2010). Cette base annotée nous permettra de connaître la position des gènes (donc leur densité locale), des séquences répétées, ainsi que le taux de GC local le long de la séquence. Toutes les informations ne seront cependant pas disponibles, entre autres parce que le génome du poulet n'est pas aussi bien annoté que, par exemple, le génome humain. Ainsi, la position des centromères des chromosomes n'est pas encore disponible pour le génome du poulet.

2 Matériels et méthodes

Afin d'étudier la stochasticité de l'expression génique et sa dépendance à son environnement génomique, nous avons utilisé, comme dans les études présentées dans les chapitres précédents, des populations clonales de progéniteurs érithrocytaires aviaires[1] transformées par le rétrovirus AEV et transfectées par un transgène composé d'un promoteur exogène

[1]Plus précisément, il s'agit de cellules issues de la lignée 6C2 (Gandrillon *et al.*, 1999)

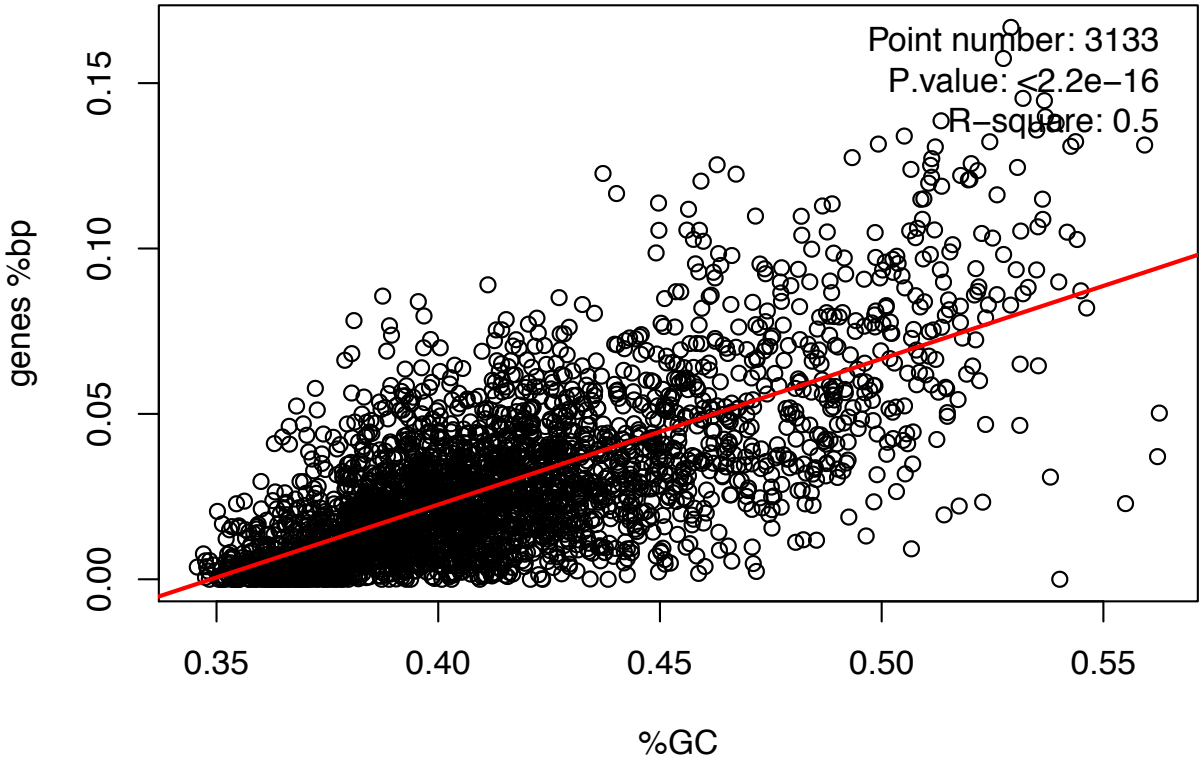

Figure IV.2 – Relation entre taux de GC et densité en gènes pour des fenêtres de 6.5×10^5 bp. Les points noirs correspondent à ces deux caractéristiques pour 3133 points positionnés à 6.5×10^5 bp les uns des autres tout au long du génome. La ligne rouge représente la corrélation entre la densité en gènes (proportion de bases codantes) et le taux de GC dans la même fenêtre (P-value= 2.2×10^{-16}, R-square= 0.5).

CMV[1] et d'un gène rapporteur fluorescent *mCherry* (On note la construction correspondante "CMV-*mCherry*").

Les cellules transfectées ont ensuite été triées et clonées pour obtenir des populations ne différant génétiquement que par le locus auquel est inséré le transgène[2]. Dans le cadre de cette étude, nous avons produit 81 populations clonales dont les loci ont ensuite été identifiés. 20 d'entre elles ont été caractérisées sans ambiguïté : "mono-insertion".

Pour chacune de ces populations, on mesure la différence d'expression génique (stochasticité inter-cellulaire) observée en fluorescence entre les cellules de populations isogéniques. Cette mesure est, comme dans les études précédente, réalisée par cytométrie de flux et traitée pour éliminer l'autofluorescence des cellules dans les données.

[1]CytoMégaloVirus – il s'agit d'un promoteur d'origine virale très actif et déconnecté du réseau de régulation de la cellule. Il est donc théoriquement insensible à l'état de la cellule et, par exemple, au cycle cellulaire.

[2]La méthodologie utilisée pour la production des populations étant globalement identique à celle utilisée pour les expérimentations précédentes, nous ne décrirons ici que son principe général et les quelques différences. Le lecteur est invité à se référer aux deux chapitres précédent pour les détails de la méthode initiale

2.1 Caractérisation des points d'insertion

Pour déterminer le locus de chaque insertion, nous avons utilisé ici une version haut débit (Uren *et al.*, 2009) de la splinkerette PCR utilisée pour les deux études précédentes. La splinkerette permet de séquencer le fragment d'ADN adjacent au point d'insertion et ce, pour chaque insertion de chaque population clonale (rappelons que les insertions ne sont pas nécessairement uniques). La différence entre la version classique et la version haut-débit est que, durant la deuxième PCR nichée permettant d'augmenter la spécificité de l'amplification génomique, nous liguons des adaptateurs 454 aux fragments amplifiés. Cela nous permet d'utiliser ensuite un séquenceur 454[1]. Tous les fragments amplifiés de toutes les populations (96 populations maximum par manipulation) sont ensuite mélangés et envoyés au séquençage mais les adaptateurs utilisés étant clones-spécifiques (ils contiennent une séquence d'identification – le "TAG" – spécifique à chaque population clonale), on pourra ultérieurement associer tous les fragments séquencés à leur clone d'origine. Nous pouvons ainsi créer un fichier fasta où toute sequence est identifiée et associée à son clone. Chacune de ces séquences est alors tronquée juste après la séquence identificatrice pour ne conserver que le fragment composé d'une partie du transgène et d'ADN génomique.
En utilisant le programme en ligne iMapper (Kong *et al.*, 2008), nous pouvons alors retrouver, pour chaque fragment, la partie du transgène qui subsiste (même s'il a subi des mutations mineures) et supprimer la séquence correspondante pour n'utiliser que l'ADN génomique. iMapper est ensuite utilisé pour aligner cette séquence courte sur l'ensemble du génome du poulet de façon à identifier le point d'insertion du transgène. Notons qu'une même séquence peut être retrouvée en plusieurs points du génome ou, inversement, ne pas permettre d'identifier de correspondance. Dans ces deux cas nous considérerons que, pour le clone correspondant, le point d'insertion est indéterminé.
A l'issue de cette procédure, on obtient, pour chaque clone, tous les sites d'insertion du transgène dans chaque clone ainsi que le sens de chaque insertion. Sur les 81 populations clonales, nous avons finalement caractérisé 136 points d'insertion différents sur le génome aviaire (Figure IV.3, tirets verts et bleus). Vingt d'entre eux correspondent à des insertions uniques (Figure IV.3, tirets verts) sans ambiguïtés.
À l'issue de l'ensemble de la procédure de transfection et d'identification des points d'insertion nous disposons donc de vingt populations clonales exploitables pour notre étude. En effet, nous ne pouvons associer la fluorescence du rapporteur à un locus particulier que dans le cas des simple-insertions. Dans le cas des multi-insertions, les signaux issus de plusieurs rapporteurs étant confondus, le lien entre l'expression génique et l'environnement génomique est impossible à établir[2].

2.2 Intégration aléatoire via le système Tol2/transposase

L'étude du lien entre caractéristiques de l'environnement génomique et stochasticité de l'expression génique nécessite que, pour chacune des caractéristiques génomiques étudiées

[1] Le séquençage dit "454" est un procédé qui permet de séquencer, en même temps, un grand nombre de fragments ADN flanqués par les adaptateurs éponymes.

[2] À noter que, même si nous n'avons pas utilisé les clones bi- ou tri- insertion, nous aurions pu les exploiter pour tester l'aspect prédictif de nos résultats. En effet, la fluorescence de ces clones pourrait théoriquement être calculée en sommant la fluorescence théorique de tous les points d'intégration qu'ils contiennent

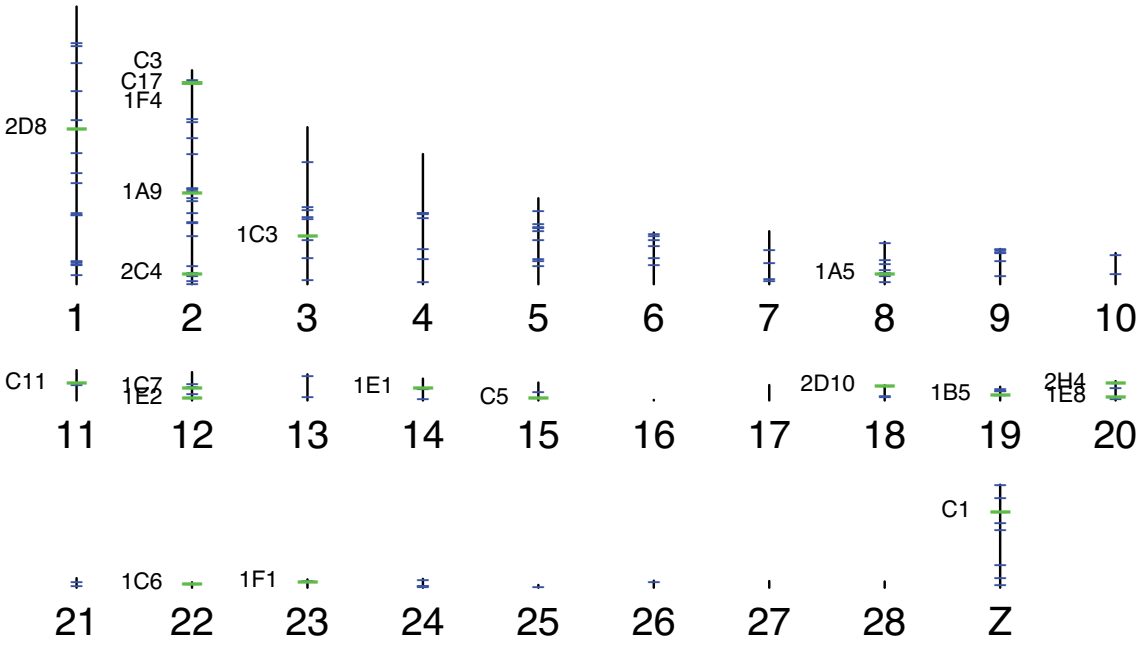

Figure IV.3 – Loci des 136 insertions du transgène mCherry. Les points d'insertion observés sur différents clones sont ici reportés sur le génome du poulet (chromosomes 1 à 28 et chromosome Z). Pour les clones où seule une insertion a été observée sans ambiguïtés, cette dernière est affichée en vert et porte le nom de son clone. Les autres insertions sont indiquées en bleu et ne sont pas nommées.

(taux de GC, densité en gènes, densité en régions répétées, ...), la mesure ne soit pas biaisée par le comportement du système Tol2/transposase utilisé pour la transfection. S'il y avait, malgrès tout, un biais (comme constaté par Huang *et al.* (2010) dans des cellules T primaires humaines), il serait important de l'identifier et de le caractériser. D'après Huang *et al.* (2010), il semble y avoir principalement une surreprésentation des insertions du système Tol2 dans les unités de transcription. Nous avons donc étudié spécifiquement la possibilité que l'insertion du transgène par le système Tol2 produise un biais en faveur d'une (ou plusieurs) caractéristique(s) génomique(s) dans nos cellules. Pour ce faire, et afin d'obtenir la plus grande précision possible, nous avons utilisé les 136 points d'insertion détectés (et non les seules mono-insertions).

2.2.1 Nombre d'intégrations par chromosome

Dans un premier temps, nous avons vérifié que le nombre d'insertions constatées par chromosome dans chaque population clonale reflète bien un comportement probabiliste (IV.4).
Nous avons ensuite vérifié que le nombre d'insertions dans chacun des 29 chromosomes de chaque population est bien proportionnel à la taille du-dit chromosome (Figure IV.4). Ceci est confirmé par un test de kolmogorov-smirnov (P.value=0.37). En revanche, la figure IV.4 montre que, compte-tenu du petit nombre de clones mono-insertion (20), leur

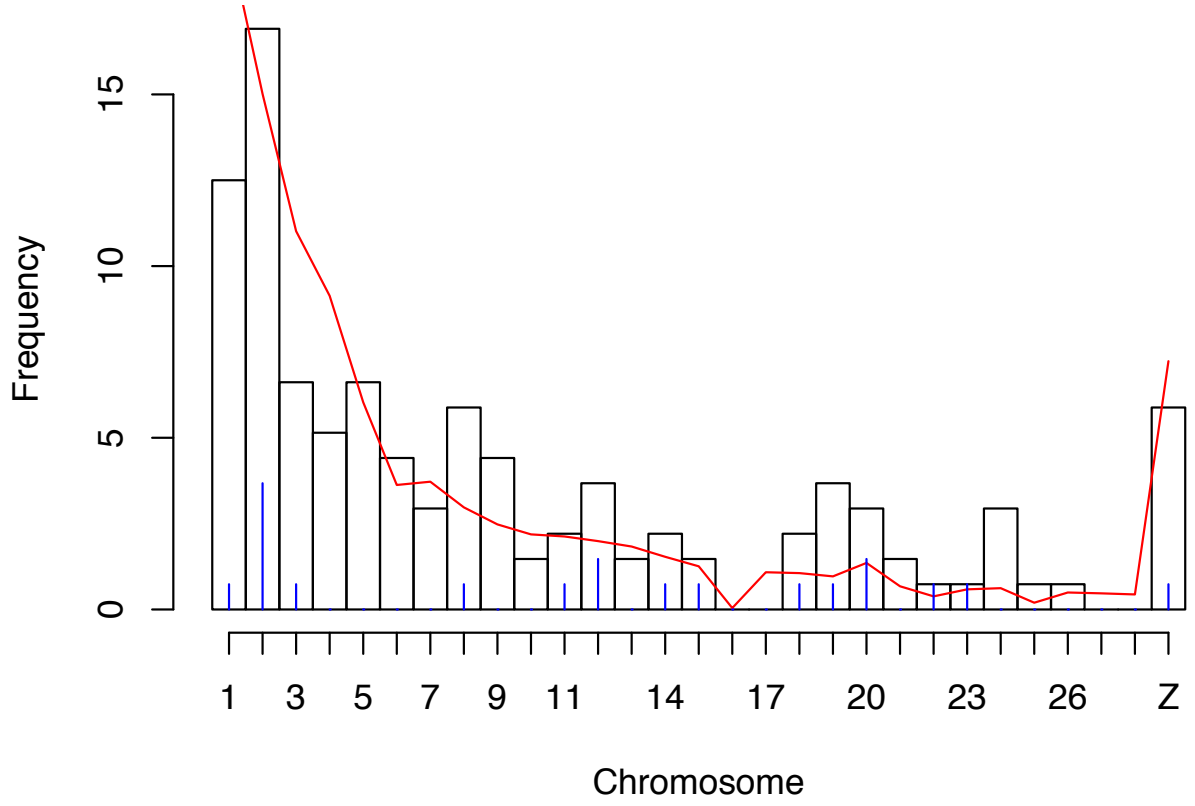

Figure IV.4 – Nature aléatoire du nombre de points d'insertion par chromosome. On représente ici le nombre de points d'insertions observés par chromosome. En noir sont regroupés l'ensemble des points observés en 454. La répartition des points des 20 clones simple insertion est représentée en bleu. La courbe théorique du nombre d'insertion par chromosome (en supposant une insertion aléatoire dépendant seulement de la longueur des chromosomes) est représentée en rouge. Il n'y a pas de différence significative entre la répartition théorique et celle des 136 points observés en 454 (test de Kolmogorov-Smirnov, P-value=0.37).

représentativité chromosomale n'est pas parfaite. Ainsi, les chromosomes 1 et 3 qui sont deux des trois plus longs ne comportent qu'un seul point d'insertion chacun alors que le chromosome 20, relativement court, en comporte 2. De même, les chromosomes 4,5,6,7 et 9, tous de taille moyenne, ne comportent aucun point d'insertion. Cependant, pour un grand nombre d'insertions, le système Tol2 ne semble pas avoir plus d'affinité pour un chromosome que pour un autre.

2.2.2 Nature aléatoire de l'insertion par rapport aux structures génomiques

Pour chacune des caractéristiques génomiques testées par la suite (séquences répétées, taux de GC et séquences codantes) nous avons testé l'éventualité que le système Tol2 introduise un biais dans la répartition des points d'insertion du transgène. Ainsi, il n'y a pas plus d'insertions dans les gènes que hors des gènes (Figure IV.5.A). Il en va de même pour les régions répétées (Figure IV.5.C). Par contre il y a une légère sur-représentation des insertions dans des zones pauvres en régions répétées (Figure IV.7), riche en GC,

denses en gènes (Figures IV.6 et IV.7) ou sur des loci proches d'un gène (Figure IV.8). Il est important de noter que ces caractéristiques ne sont pas indépendantes. Les régions riches en GC sont connues pour être aussi riches en gènes et pauvres en séquences répétées (Figure IV.2). En outre, il existe un lien très fort entre la taille d'un chromosome et son taux de GC (Figure IV.9) ainsi qu'entre la densité en séquences répétées et la densité en gènes (Figure IV.10). Toutes ces observations sont donc très probablement différentes facettes d'un même biais.

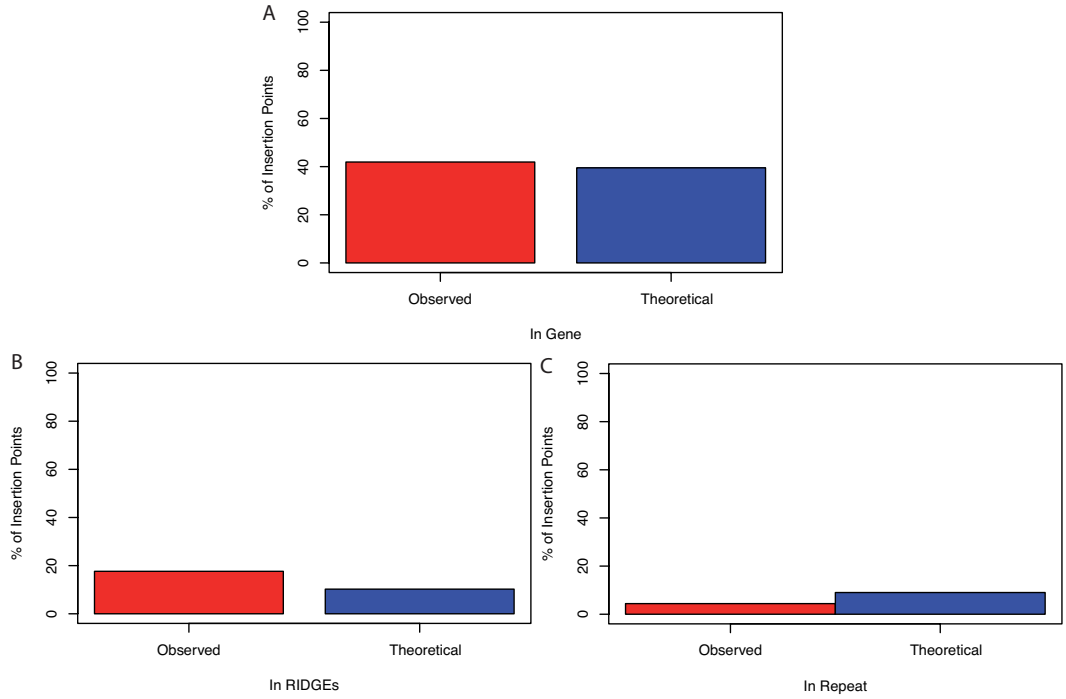

Figure IV.5 – Nature aléatoire de l'insertion par rapport aux différentes structures génomiques. (A) Nombre d'insertions dans un gène. En rouge : nombre d'insertions dans un gène parmi les 136 insertions analysées. En bleu : nombre théorique d'insertion dans un gène ($\frac{Genes(bp)}{Genome(bp)}$). Il n'y a pas de différence significative entre les nombres d'insertion dans un gène observé et théorique (P-value = 6.0×10^{-1}). (B) Nombre d'insertions dans une séquence RIDGE[1]. En rouge : nombre de points insérés dans une séquence RIDGE parmi les 136 insertions analysées. En bleu : nombre théorique d'insertions dans une séquence RIDGE ($\frac{SequenceRIDGE(bp)}{Genome(bp)}$). Le nombre d'insertions dans des séquences RIDGE est supérieur à la prédiction théorique (P-value= 9.9×10^{-3}). (C) Nombre d'insertions dans une séquence répétée. En rouge : nombre de points insérés dans une séquence répétée parmis les 136 insertions analysées. En bleu : nombre théorique d'insertions dans une séquence répétée ($\frac{Sequencerepetee(bp)}{Genome(bp)}$). Il n'y a pas de différence significative entre les valeurs observées et les valeurs théoriques (P-value=7.0×10^{-2}).

En conclusion, le système Tol2 permet une insertion relativement aléatoire avec un biais en faveur des intégrations proches des gènes.
Malheureusement, sur les vingt clones exploitables pour l'analyse (voir sections précédentes), aucun n'est inséré dans une zone "RIDGE" dense en gènes fortement exprimés

– on utilise ici le catalogue des RIDGEs proposé par (Nie *et al.*, 2010). En prenant en compte l'ensemble des 136 points d'insertion caractérisés, on constate cependant qu'il y a une différence significative entre le nombre d'insertions dans des RIDGEs et le nombre théorique d'insertions que l'on devrait y trouver (Figure IV.5.B). Ces deux informations peuvent paraître contradictoires (si les zones RIDGEs comportent plus de points d'insertion, pourquoi n'y observe-t-on pas de mono-insertion ?). Nous n'avons malheureusement pas assez de clones à notre disposition pour étudier plus précisément l'origine de ces biais. Ils posent clairement la question de l'indépendance des insertions dans un même clone dans les zones fortement exprimées.

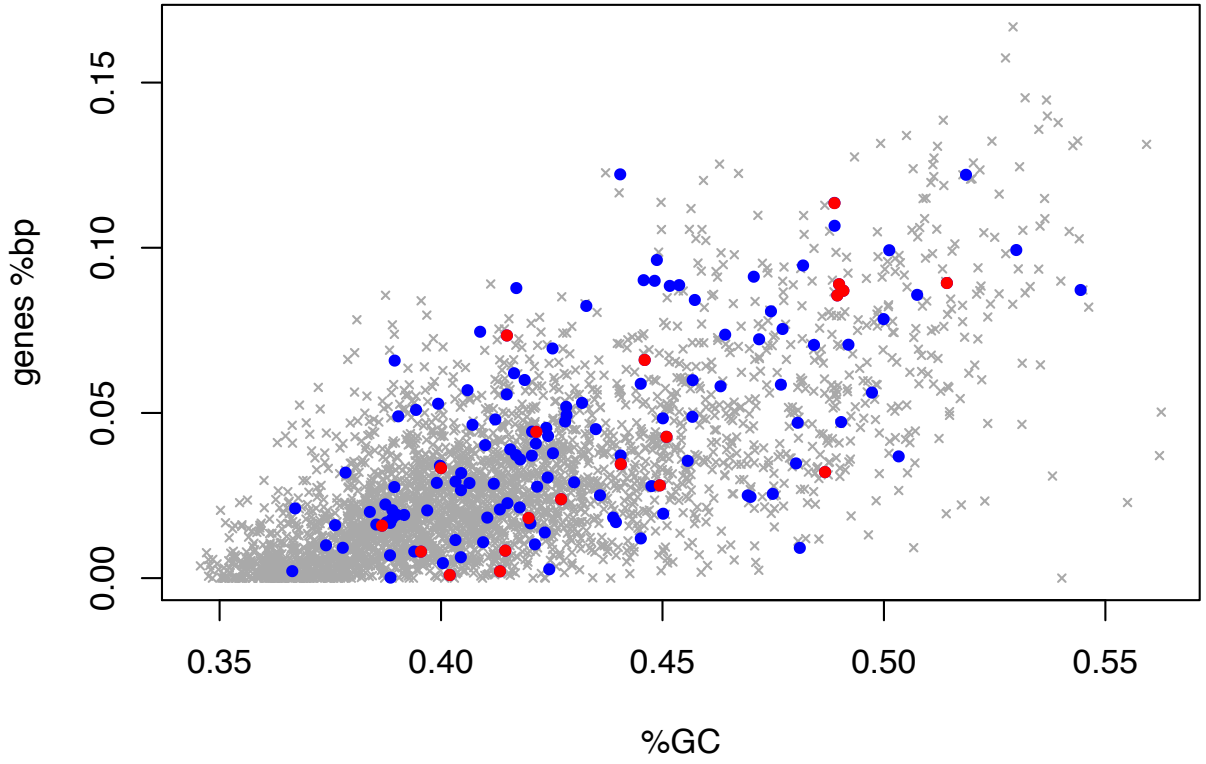

Figure IV.6 – Nature aléatoire de l'insertion par rapport au taux de GC et à la densité en gènes. La mesure correspond à des fenêtres de 6.5×10^5 bp. Les points gris représentent ces deux caractéristiques mesurées pour 3133 points positionnés à 6.5×10^5 bp les uns des autres tout au long du génome. Les points bleus représentent les caractéristiques des fenêtres prises autour de chacun des 20 points d'insertion correspondant aux clones mono-insertion caractérisés. Les points rouges représentent les caractéristiques des fenêtres prises autour de chacun des 116 points d'insertion observés dans les autres clones. Pour les 136 points et par rapport aux fenêtres prises au hasard, il y a plus d'insertions dans les zones riches en gènes (test de wilcoxon, P-value= 2.2×10^{-16}) et dans les zones GC riches (test de wilcoxon, P-value= 6.7×10^{-10}.).

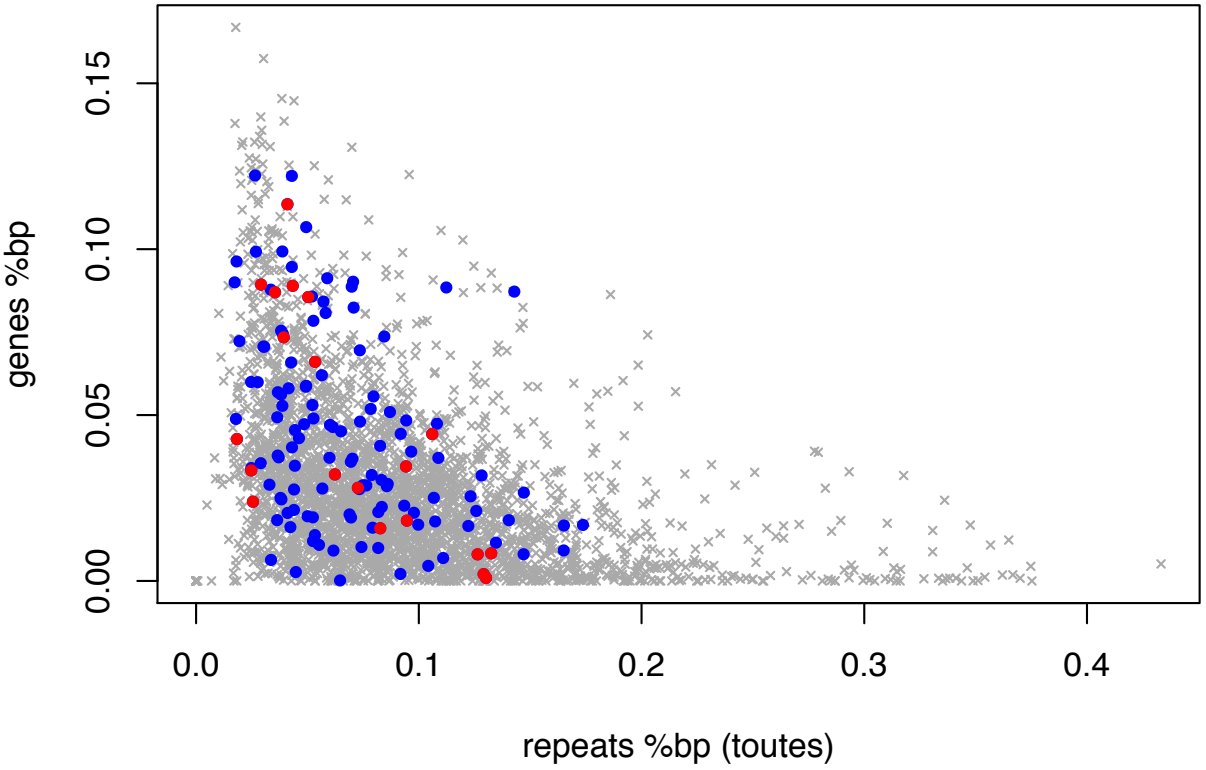

Figure IV.7 – Nature aléatoire de l'insertion par rapport à la densité en régions répétées et à la densité en gènes pour des fenêtres de 6.5×10^5 bp. Les points gris représentent ces deux caractéristiques pour 3133 points positionnés à 6.5×10^5 bp les uns des autres tout au long du génome. Les points bleus représentent les caractéristiques des fenêtres prises autour de chacun des 20 points d'insertion correspondant aux clones mono-insertion caractérisés. Les points rouges représentent les caractéristiques des fenêtres prises autour de chacun des 116 points d'insertion observés dans les autres clones. Pour les 136 points et par rapport aux fenêtres prises au hasard, il y a plus d'insertions dans des zones riches en gènes (test de wilcoxon, P-value= 2.2×10^{-16}) et dans les zones pauvres en sequences répétées (test de wilcoxon, P-value= 2.2×10^{-16}).

2.3 Mesure de la stochasticité de l'expression génique

L'expression génique des vingt clones mono-insertion fiables a été caractérisée par cytométrie en flux. La Figure IV.11 présente les vingt histogrammes correspondant aux mesures d'expression de ces populations. On constate immédiatement que l'expression génique diffère d'un clone à l'autre alors que la seule différence entre eux est le locus du point d'insertion du rapporteur. Ces résultats confirment, avec un plus grand nombre de clones, les conclusions de nos deux études précédentes (chapitres II et III). Nous pouvons caractériser ces distributions d'expression génique par les deux indicateurs utilisés précédemment : la moyenne d'expression et la variance normalisée de l'expression, ce dernier indicateur nous servant de mesure de la stochasticité de l'expression génique.

La figure IV.12 représente la variance normalisée de nos vingt clones d'intérêt en fonction de la moyenne d'expression génique. On rappelle qu'un comportement Poisonnien du

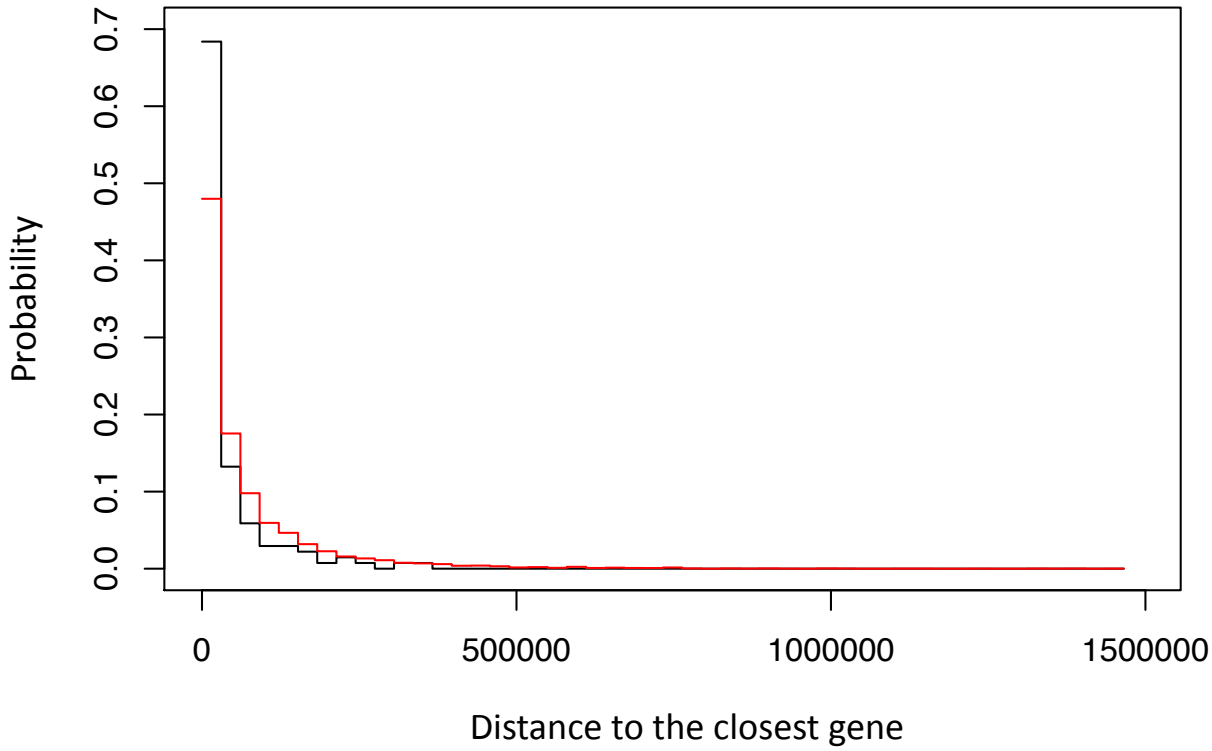

Figure IV.8 – Lien entre insertion par le système Tol2 et distance au gène le plus proche. La ligne rouge représente, pour 10000 points d'insertions virtuels pris au hazard tout au long du génome, la fonction de répartition des distances au gène le plus proche. La ligne noire correspond à la même mesure pour les 136 points d'insertion. Il y a une différence significative (test de Wilcoxon, P-value= 9.2×10^{-10}) entre les 10000 points d'insertions virtuels et les mesures expérimentales. Le système Tol2 semble favoriser les insertions à proximité immédiate des gènes (ici pour des distances inférieures à 3.0×10^4 bp).

promoteur implique l'égalité entre la moyenne et la variance normalisée de l'expression génique. Ici, pour les vingt clones, on constate une bonne corrélation entre moyenne d'expression et variance normalisée même si la moyenne n'explique que partiellement les variations de NV (corrélation log-log ; P.value = 8.3×10^{-6} ; R-square = 0.68).

2.4 Calcul des valeurs théoriques de dynamique chromatinienne

À partir des valeurs d'expression (moyenne et variance normalisée), nous pouvons réutiliser les résultats du chapitre précédent (chapitre III, Figure III.11) pour inférer les paramètres de la dynamique chromatinienne théorique au point d'insertion pour chacune des populations mono-insertion. Nous utilisons pour cela notre modèle de l'expression génique dans lequel la chromatine est supposée passer aléatoirement d'un état ouvert à un état fermé, avec des taux de transition ouvert→fermé et fermé→ouvert noté respectivement k_{off} et k_{on}.
Dans ce modèle, on considère que l'ouverture de la chromatine rend le gène accessible à toute la machinerie cellulaire nécessaire à sa transcription. Celle-ci se produit alors avec

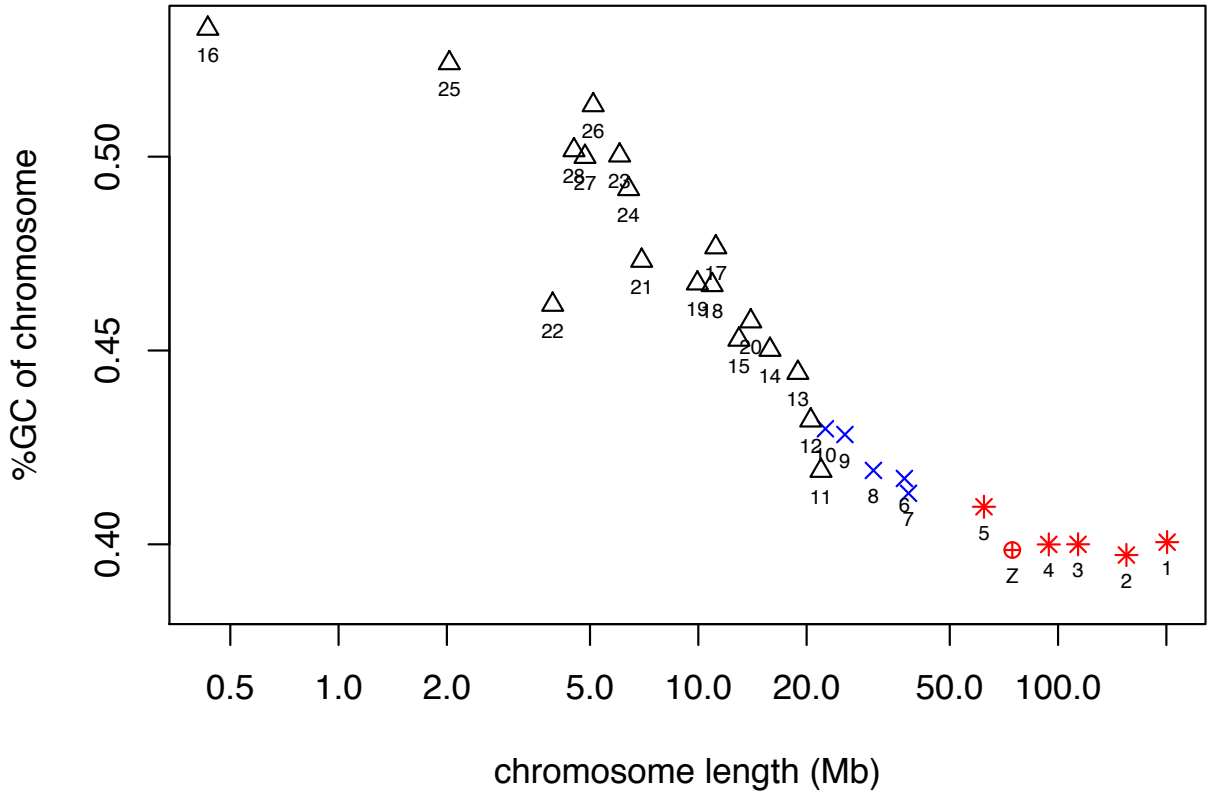

Figure IV.9 – Relation entre la taille des chromosomes et leur taux de GC pour le génome du poulet. On observe une tendance claire : plus les chromosomes sont courts, plus ils sont GC riches.

un taux ρ. Les ARN ainsi transcrits peuvent être traduits en protéines avec un taux γ ou dégradés avec un taux $\tilde{\rho}$. Ces protéines peuvent, à leur tour, être dégradées avec un taux $\tilde{\gamma}$ et elles émettent α unités arbitraires de fluorescence (unité utilisée en cytométrie en flux).
Dans le chapitre précédent, nous avons, grâce à ce modèle estimé les valeurs de $\rho, \tilde{\rho}, \gamma, \tilde{\gamma}$ et α (supposées constantes chez tous les clones[1]). En introduisant dans le modèle les valeurs d'expressions des clones (moyenne d'expression et NV), nous pouvons déterminer les paramètres de la dynamique chromatinienne aux alentours de chacun des points d'insertion, c'est à dire pour chaque clone. Conformément aux résultats du chapitre III, nous caractériserons cette dynamique par deux indicateurs : le temps moyen fermé ($1/k_{on}$) et le nombre moyen d'ARNm produits durant une période d'ouverture de la chromatine (ρ/k_{off}).
On constate (Figure IV.13) qu'il y a une corrélation entre les deux indicateurs de la dynamique chromatinienne (corrélation log-log ; P.value : 1.9×10^{-3} ; R-square : 0.42).
Les corrélations constatées entre moyenne d'expression et NV d'une part et entre nombre moyen d'ARN produits par épisode d'ouverture de la chromatine et temps moyen fermé de la chromatine au locus du gène d'autre part, sont probablement liées. Nous devrons par la

[1]Les clones utilisés ici étant totalement identiques à ceux utilisés au chapitre précédent, nous pouvons légitimement considérer que les valeurs de ces paramètres sont elle-même identiques

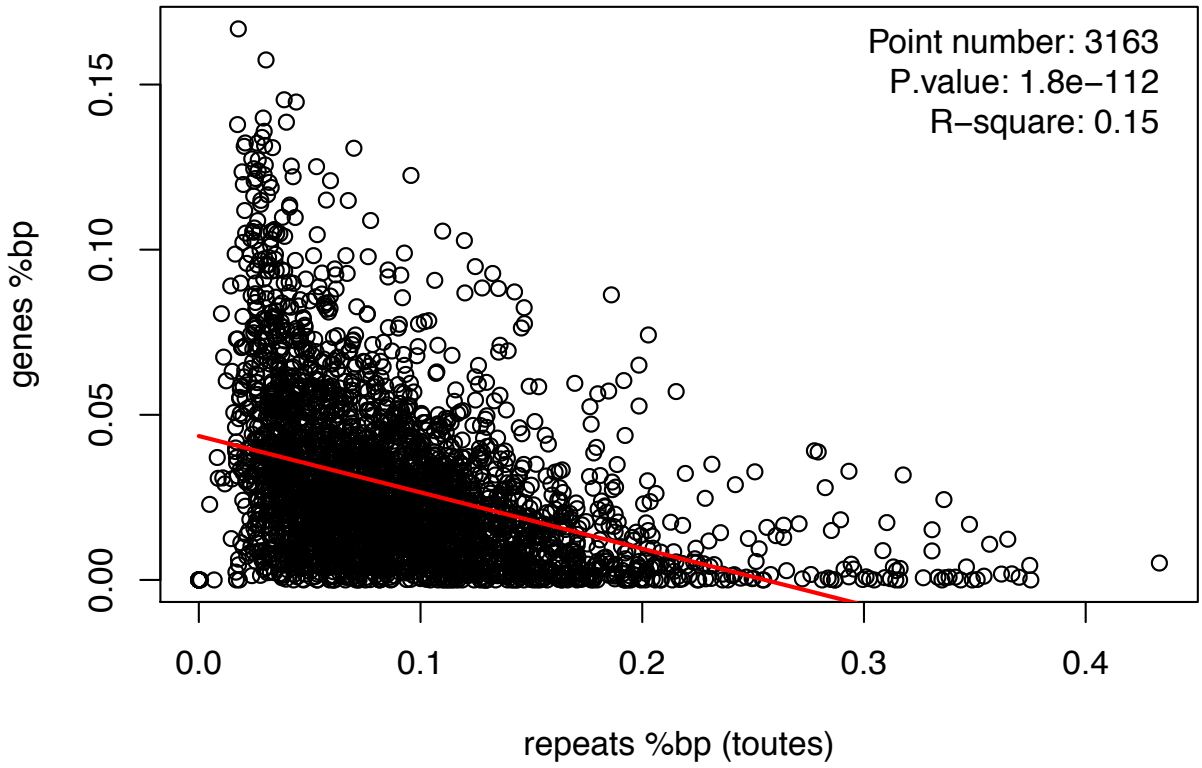

Figure IV.10 – Relation entre densité de séquences répétées et densité en gènes dans le génome du poulet. On observe ici une faible corrélation entre le pourcentage de bases occupées par des gènes et le pourcentage de bases occupées par des séquences répétées (P-value= 1.8×10^{-112}, R-square = 0.15). On remarque que les régions saturées en gènes sont pauvres en régions répétées et inversement. Cependant, les régions non saturées dans une des deux composantes ne sont pas contraintes pour l'autre composante.

suite être attentif à ces corrélations lors de l'analyse des résultats puisqu'elles entraîneront des corrélations entre plusieurs mesures lors de l'étude des caractéristiques génomiques.

2.5 Caractérisation des corrélations entre locus d'insertion et expression génique

Notre objectif est de mettre en évidence – s'il existe – un lien entre la stochasticité de l'expression génique et les caractéristiques locales du génome aux alentours du point d'insertion. Si nous connaissons de bons indicateurs permettant de caractériser macroscopiquement l'expression génique (la moyenne de fluorescence et la variance normalisée de la fluorescence, toutes deux mesurées sur de grandes populations de cellules), il n'en va pas de même pour les structures génomiques. Même si plusieurs auteurs font état de liens entre structures génomiques et expression génique (Wang *et al.*, 2011; Nie *et al.*, 2010), nous devons garder une certaine ouverture quant aux structures étudiées, à leurs caractéristiques et à leurs tailles.
Pour étudier les liens entre insertion et expression, nous utiliserons ici des tests statistiques. Pour comparer une caractéristique génomique avec un indicateur de l'expression

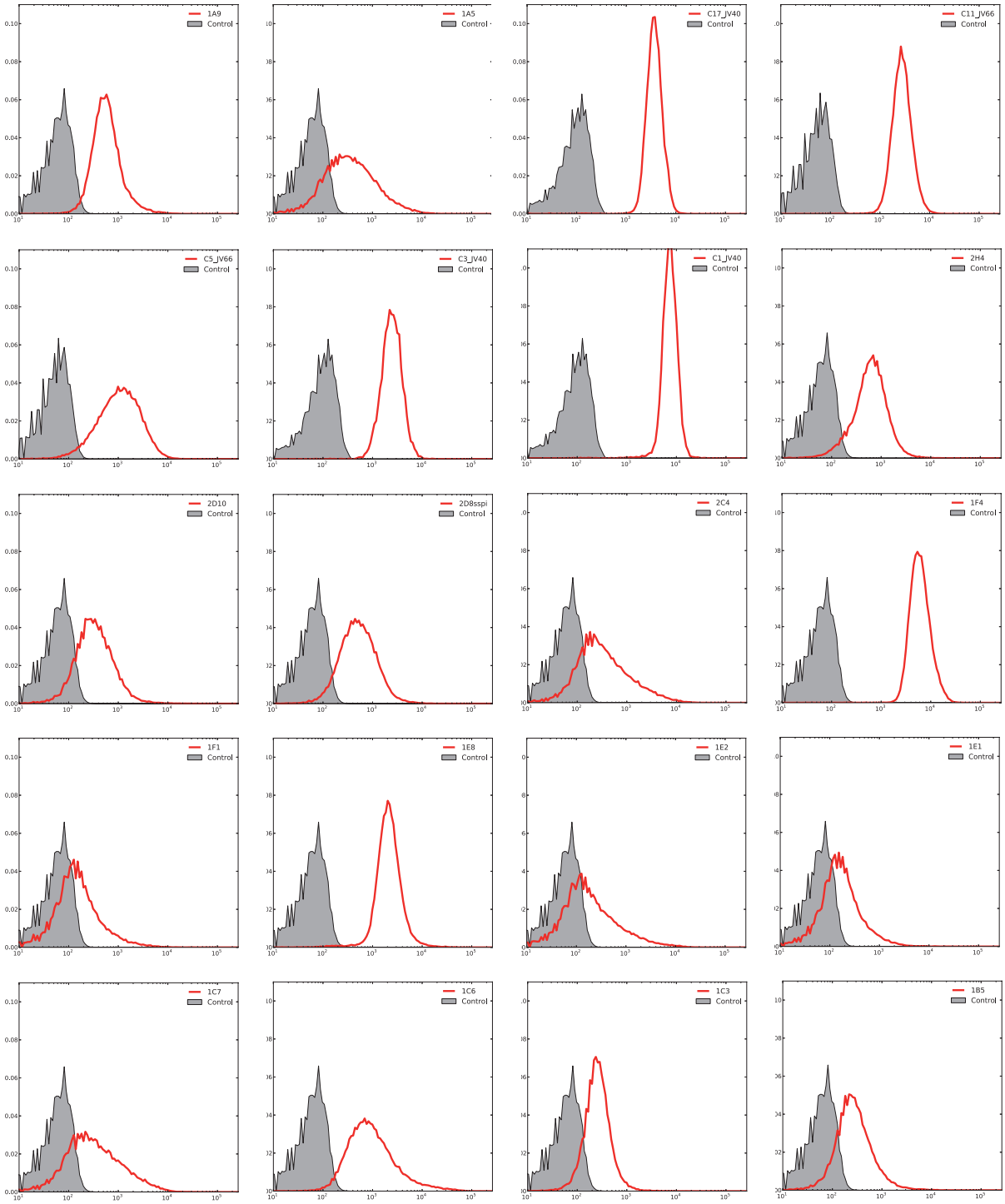

Figure IV.11 – Expression génique pour vingt différents loci du génome du poulet. On mesure ici par cytométrie en flux les distributions de fluorescence des 20 populations clonales mono-insertions (correspondant aux vingt loci indiqués en vert sur la figure IV.3). Pour chacun de ces vingt histogrammes, on représente la distribution de fluorescence du clone simple insertion correspondant (en rouge) et la distribution d'autofluorescence (en gris). Celle-ci est mesurée sur une population de même type cellulaire mais non transfectée et mesurée au même moment que le clone. Les clones ont été mesurés en trois fois.

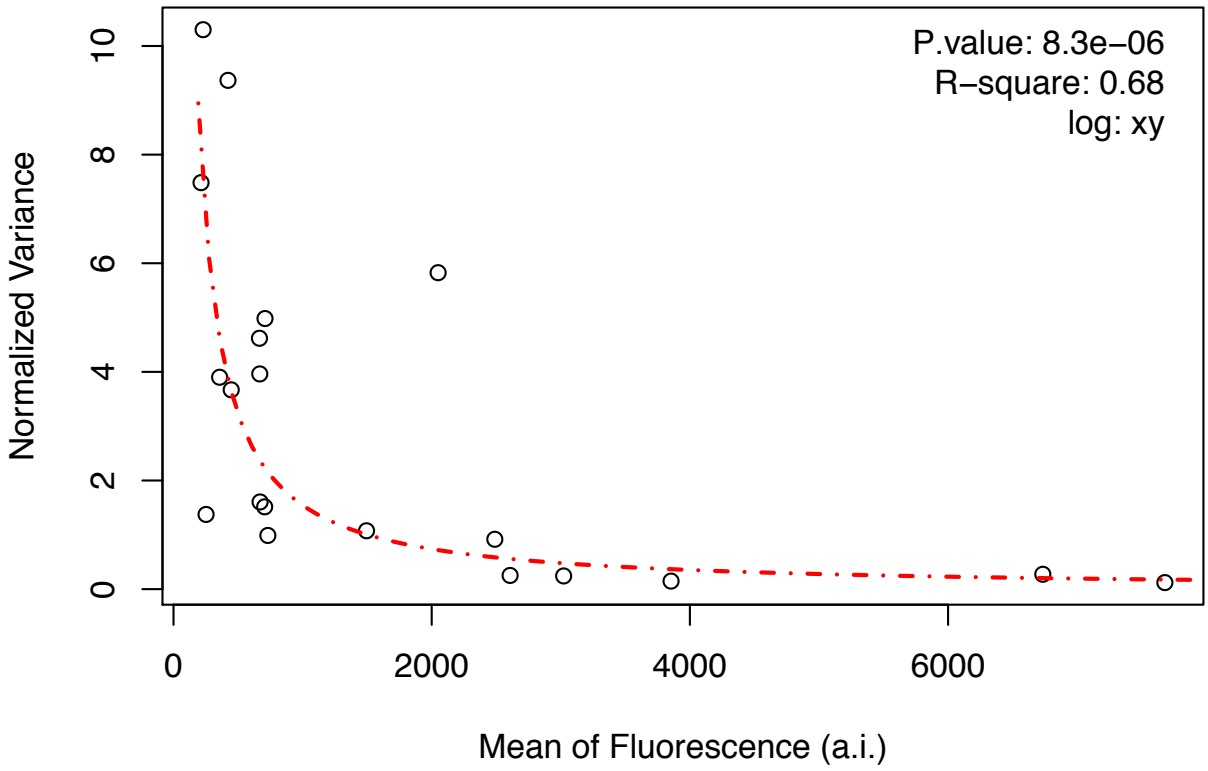

Figure IV.12 – Corrélations entre la moyenne et la variance normalisée de l'expression génique pour les distributions de fluorescence des vingt populations clonales simple-insertion. La ligne pointillée rouge représente le cas théorique d'une dynamique Poissonnienne pour laquelle la moyenne et la variance normalisée seraient corrélées. Ici cette corrélation existe mais n'explique pas entièrement les variations de NV (corrélation log-log ; P-value= 8.3×10^{-6}, R-square = 0.68).

génique, nous utiliserons un test de corrélation (le test de Pearson). Comme les valeurs de fluorescence mesurées, les distances chromosomiques étudiées s'étendent sur plusieurs ordres de grandeur. Nous effectuerons systématiquement les tests de corrélation sur des données transformées log-log (les mêmes tests ont été effectués sur les données brutes, transformées lin-log et log-lin ; les meilleures corrélations obtenues sont celles mesurées en log-log). Les tests de corrélation sont réalisés à l'aide du logiciel de statistiques R (R Development Core Team, 2010).
Dans les cas où nous recherchons des corrélations locales (donc sur des tailles de fenêtres), nous étudierons les variations de corrélation sur des fenêtres de tailles variables afin d'identifier la taille optimale. Nous avons fait ici le choix de représenter ces données sous forme graphique (voir par exemple figures IV.14 et IV.15). Si la meilleure corrélation est significative, alors elle sera représentée par une courbe rouge. Dans le cas contraire, sa courbe sera noire. Les caractéristiques de la meilleure corrélation sont affichées explicitement (nombre de clones utilisés[1], P-value, R-square).

[1]Lorsque la taille de la fenêtre de test augmente, certaines points d'insertions ne peuvent plus être utilisés en raison de leur proximité avec une des extrémité du chromosome. Le nombre de points utilisés pour calculer les corrélations diminue donc au fur et à mesure que la fenêtre de mesure s'aggrandit.

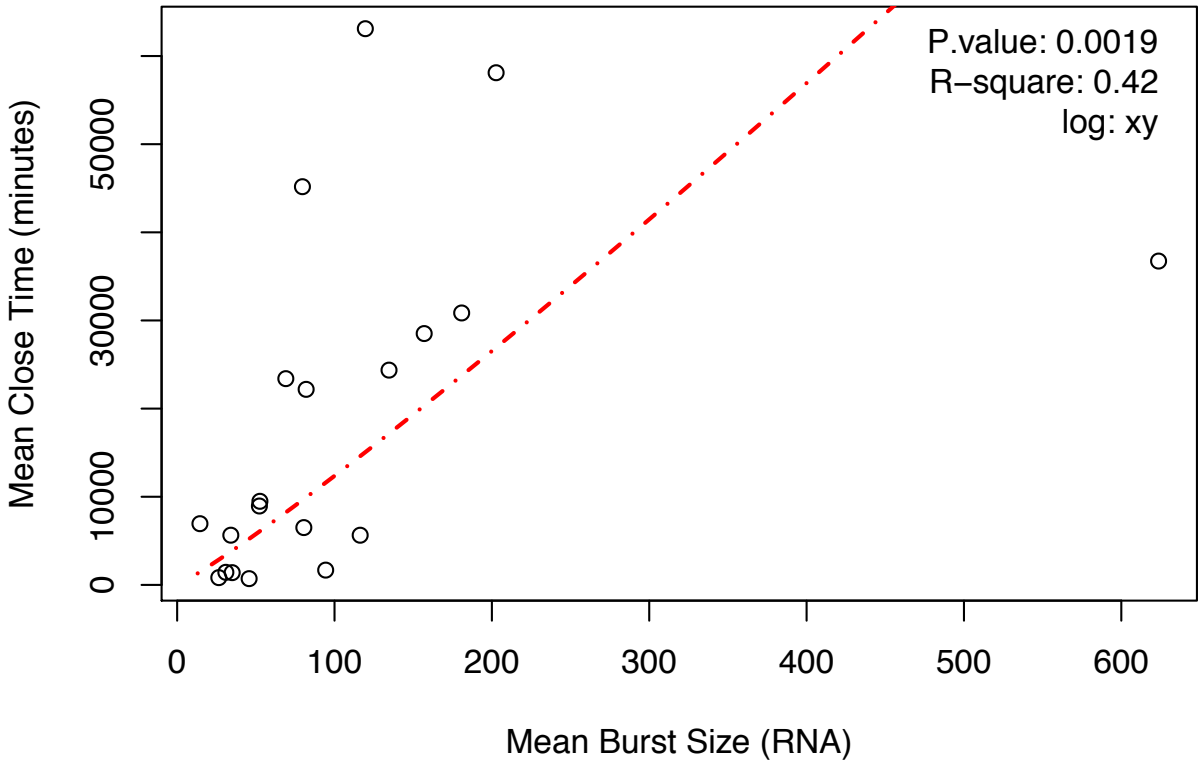

Figure IV.13 – Corrélations entre le temps moyen fermé et le nombre moyen d'ARN produits par période d'ouverture, calculés à partir des distributions de fluorescence des vingt populations clonales simple-insertion. Ces indicateurs sont estimés grace au modèle présenté au chapitre précédent. On constate une corrélation positive entre le temps moyen fermé et le nombre moyen d'ARN produits durant un épisode d'ouverture de la chromatine (corrélation log-log ; P-value= 1.9×10^{-3}, R-square = 0.42). La courbe correspondante est représentée en ligne pointillée rouge.

Une fois la taille de fenêtre identifiée, nous mesurons les valeurs correspondantes pour tous les clones et recherchons la corrélation avec les caractéristiques d'expression génique de ces clones. Notons que la phase d'identification de la fenêtre optimale ne sera pas nécessaire lorsqu'on recherchera des corrélations globales (par exemple sur la taille totale des chromosomes) ou très locale (distance au gène le plus proche).

3 Résultats

Afin de déterminer si les caractéristiques génomiques locales influencent – ou non – l'expression génique (moyenne de fluorescence et variance normalisée de la fluorescence), nous avons recherché les corrélations significatives entre l'expression génique et certaines de ces caractéristiques. Pour cela, nous avons mesuré les caractéristiques génomiques locales de trois façons :

- en considérant les caractéristiques globales du chromosome sur lequel est inséré le transgène.

- en considérant les séquences génomiques autour du point d'insertion,
- en considérant la distance entre le point d'insertion et un élément particulier (généralement le gène le plus proche)

Dans ce dernier cas, nous contrôlons si des macro-caractéristiques génomiques telles que la taille des chromosomes ou encore leur pourcentage en bases GC sont liées à la moyenne d'expression ou à la variance normalisée.

3.1 Corrélation entre expression du transgène et caractéristiques du chromosome d'insertion

Nie *et al* ont montré que plus la taille des chromosomes est grande, moins les gènes qui y sont implantés s'expriment fortement (Nie *et al.*, 2010). L'indicateur utilisé est la médiane d'expression pour tous les gènes de chaque chromosome. Dans notre cas, nous étudions 20 loci particuliers et non l'intégralité du génome. Avec ces 20 loci, la tendance observée par Nie *et al* n'est certainement pas assez forte pour compenser les effets stochastiques. Il n'est donc pas surprenant que nous n'obtenions pas de corrélation entre la moyenne d'expression et la taille du chromosome portant le transgène (Figure IV.14.B). Rappelons aussi qu'aucun des 20 loci ne se trouve au sein d'un RIDGE alors qu'il est probable que ce soient les gènes appartenant à des RIDGEs qui permettent d'observer cette tendance.

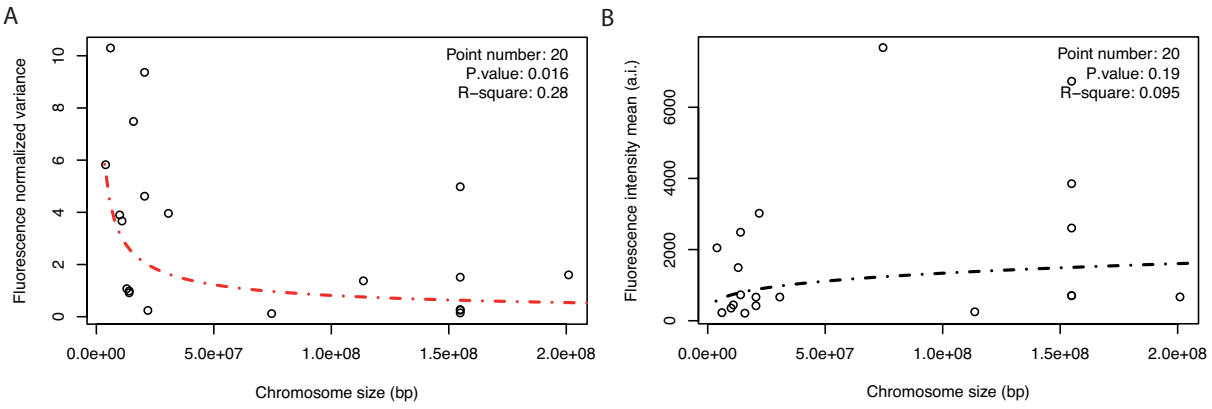

Figure IV.14 – corrélation entre taille des chromosomes et expression génique. (A) Anti-corrélation entre la taille des chromosomes et la variance normalisée des transgènes qui y sont insérés (P-value$= 1.6 \times 10^{-2}$; R-square $= 0.28$). (B) Aucune corrélation significative n'est observée entre la taille des chromosomes et la moyenne de fluorescence des transgènes qui y sont insérés.

Nous observons en revanche une corrélation faible mais significative (P-value $= 1.6 \times 10^{-2}$; R-square $= 0.28$) entre la variance normalisée (NV) de chacun des 20 clones et la taille des chromosomes portant le point d'insertion (Figure IV.14.A) : plus les chromosomes sont petits, plus la NV a tendance à être forte.
Le pourcentage GC des chromosomes étant lié à la taille des chromosomes (Figure IV.9), il n'est pas surprenant que nous observions aussi une corrélation (même faible) entre le taux de GC moyen des chromosomes et la NV (Figure IV.15.A ; P-value $= 8.6 \times 10^{-3}$;

R-square = 0.33). Les liens entre ces deux observations sont en effet probablement liées aux caractéristiques des chromosomes. Nous ne constatons pas, ici, de corrélation entre le taux de GC et la moyenne d'expression (Figure IV.15.B).

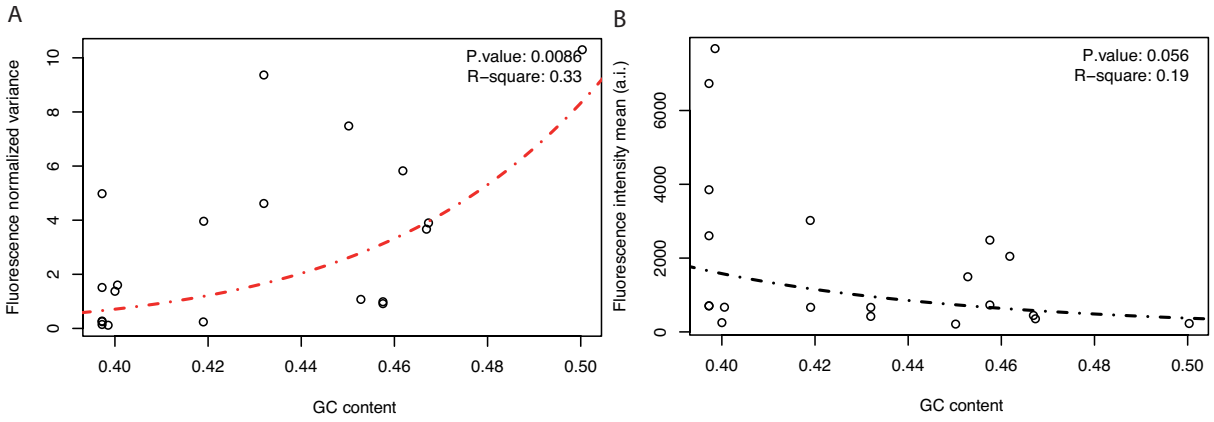

FIGURE IV.15 – lien entre pourcentage GC et expression génique. (A) Corrélation entre le taux de GC des chromosomes et la NV des transgènes qui y sont insérés (corrélation log-log ; P-value=8.6×10^{-3} ; R-square=0.33). (B) Corrélation non significative entre le taux de GC des chromosomes et la moyenne de fluorescence des transgènes qui y sont insérés.

Nous avons donc des corrélations faibles mais significatives entre des macro-caractéristiques génomiques et la variance normalisée de l'expression génique de nos vingt clones, mais pas avec la moyenne d'expression : les petits chromosomes riches en GC semblent avoir une expression génique plus bruyante.

3.2 Corrélations entre taux de GC et expression génique

Même si elles sont significatives, les corrélations entre l'expression génique et la taille ou le taux de GC des chromosomes sont très faibles. Si l'on s'intéresse à des structures plus locales, on peut rechercher les corrélations entre l'expression des gènes et le taux de GC mesuré sur des fenêtres de tailles variables (rappelons que le taux de GC est corrélé un grand nombre de caractéristiques telles que la densité en gènes).
Pour calculer un taux de GC local, il est nécessaire de définir une zone autour des points d'insertion. On peut ensuite calculer les taux de GC au voisinage de chaque point d'insertion puis calculer une éventuelle corrélation entre expression et taux de GC.
C'est pour des zones de 3.0×10^5 bp à plus de 6.0×10^5 bp (optimum à 531 248 bp – Figure IV.16.A) que nous observons les meilleures corrélations entre la NV et le taux de GC (Figure IV.16.B, P-value = 3.9×10^{-3}, R-square = 0.38). Cependant, on n'observe toujours pas de corrélation significative entre la moyenne d'expression et cette caractéristique génomique, quelle que soit la taille de la fenêtre de mesure (Figure IV.16.C). Notons que la corrélation mesurée ici entre la NV et le taux de GC local est légèrement plus forte (R-square = 0.38) qu'entre la NV et le taux de GC "global" (c'est à dire sur l'ensemble du chromosome – R-square = 0.33). En résumé, on montre ici – avec une faible corrélation – que quand le taux local de GC augmente, la stochasticité observée augmente aussi.

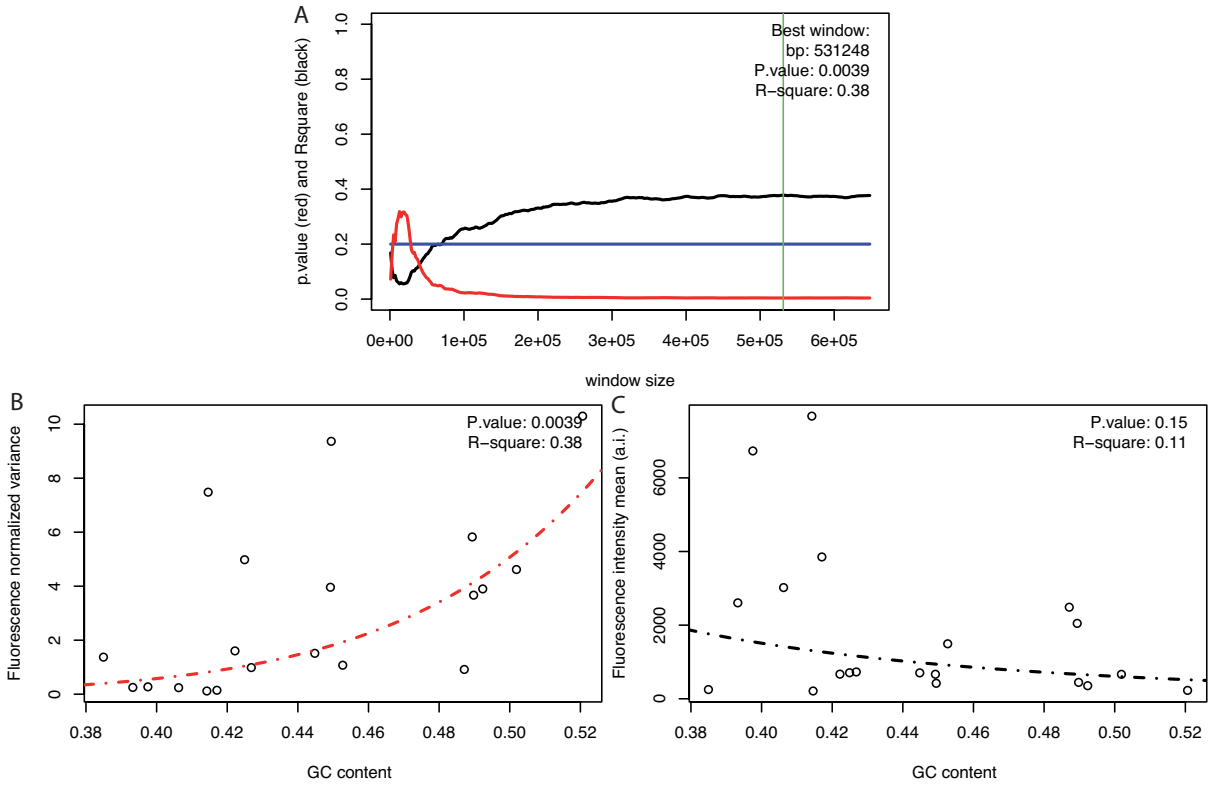

Figure IV.16 – Lien entre pourcentage GC locale et expression génique. (A) P-value et R-square du test de corrélation entre la NV d'expression d'un locus et le taux de GC pour une fenêtre de taille variable centrée sur le locus d'insertion. La meilleure corrélation est obtenue pour une fenêtre de 531 248 bp (ligne verte verticale, P-value=3.9×10^{-3} ; R-square = 0.38). Cependant, la valeur optimale est pratiquement atteinte dès 3×10^5 bp et stagne après 6×10^5 bp. (B) Corrélation entre le taux de GC autour du locus (avec la taille de fenêtre optimum) et la NV des transgènes (P-value= 3.9×10^{-3} ; R-square = 0.38). (C) Corrélation non significative entre le taux de GC autour du locus (avec la taille de fenêtre optimum) et la moyenne de fluorescence des transgènes qui y sont insérés.

3.3 Densité en séquences répétées et expression génique

En utilisant la même méthode que précédemment pour rechercher la taille de fenêtre optimale, nous pouvons rechercher les corrélations entre expression génique et densité en séquences répétées (Figure IV.17), on s'aperçoit qu'il existe un fort lien entre les deux (Figure IV.17.B, P-value= 1.2×10^{-3}, R-square = 0.45) pour une taille de fenêtre de 26 152 bp. La stochasticité décroit lorsque la densité en régions répétées augmente. Pour cette même fenêtre de 26 152 bp, on constate aussi une corrélation positive entre la moyenne d'expression du transgène et la densité en région répétées (Figure IV.17.C, P-value= 3.3×10^{-3}, R-square = 0.39).

Il semble donc que la stochasticité de l'expression génique soit fortement liée avec une ou plusieurs caractéristiques génomique locales telles que le taux de GC et la densité en régions répétées (la stochasticité augmentant quand le taux de GC augmente et que la densité en régions répétées diminue). Comme nous savons par ailleurs que les zones riches

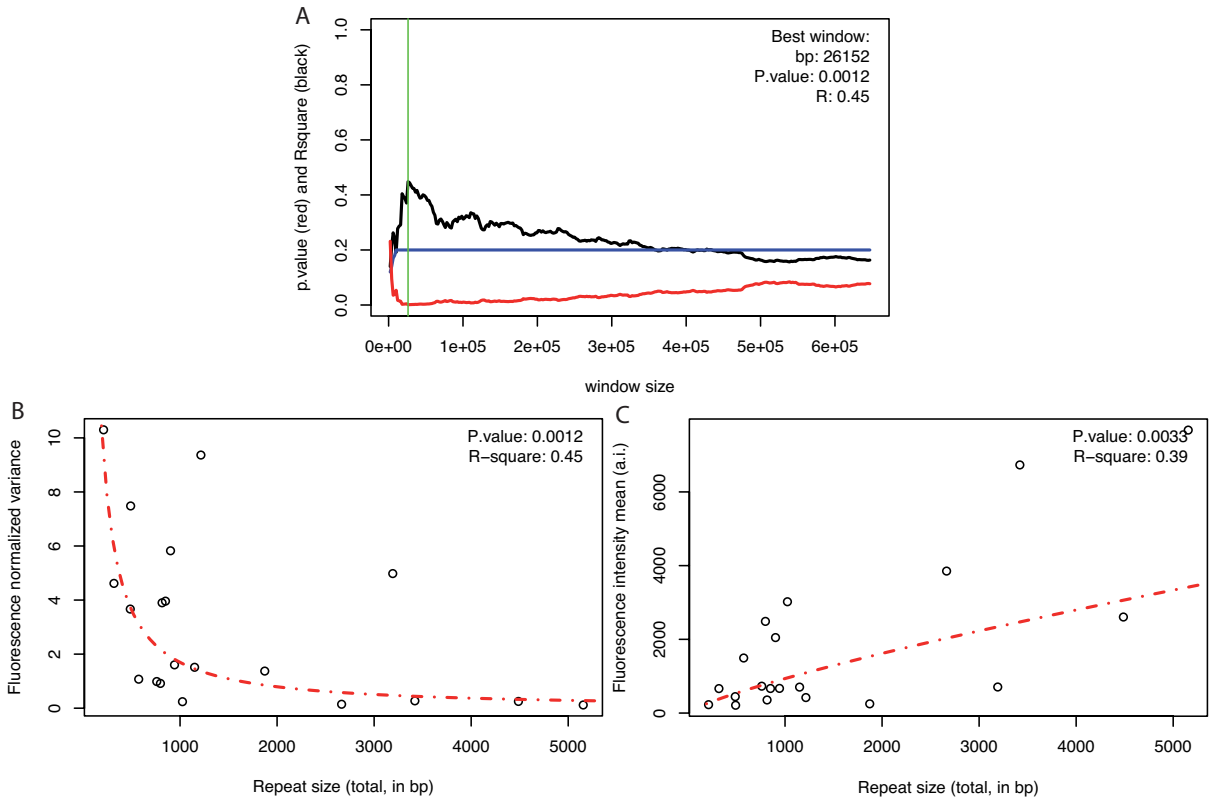

FIGURE IV.17 – Corrélation entre densité locale en séquences répétées et expression génique. (A) Variation de la corrélation entre densité en séquences répétées et stochasticité de l'expression génique en fonction de la taille de fenêtre de mesure. La fenêtres optimal est de 26 152 bp centrée sur le locus du transgène. (B) Corrélation log-log entre la variance normalisée de l'expression génique et le nombre de paires de bases occupées par des régions répétées dans des fenêtres de taille 26 152 bp centrées sur le locus du transgène (P-value= 1.3×10^{-3}, R-square = 0.45). (C) Corrélation log-log entre la moyenne d'expression et le nombre de paires de bases occupées par des régions répétées dans des fenêtres de taille 26 152 bp centrées sur le locus du transgène (P-value= 9.4×10^{-3} ; R-square = 0.39).

en gènes sont aussi riches en GC (IV.2) et pauvres en régions répétées (Figure IV.10), ce résultat nous amène naturellement a étudier le lien entre stochasticité de l'expression des gènes et densité ou proximité des séquences codantes.

3.4 Influence de la densité en gènes

Pour rechercher des corrélations entre la stochasticité de l'expression du transgène et la densité en gènes aux alentours du point d'insertion nous devons définir une taille de fenêtre sur laquelle rechercher ces corrélations. Cette taille peut être définie en "distance absolue" par rapport au transgène ou en taille de fenêtre nécessaire pour contenir N gènes. Ici, nous avons utilisé cette deuxième approche en nous inspirant des travaux de (Nie *et al.*, 2010). La taille des fenêtres utilisées n'est donc pas constante d'un clone à l'autre. En revanche, le nombre de gènes par fenêtre est constant. Cela permet d'éviter de rechercher

des corrélations liées à la densité en gènes pour des fenêtres ne contenant aucune séquence codante (ce qui est très susceptible de se produire dans le cas d'une fenêtre de taille fixe étant donnée la très grande variabilité de répartition des gènes chez les eucaryotes supérieurs). Dans un premier temps, nous avons recherché quel est le nombre N de gènes permettant d'obtenir la meilleure corrélation entre densité (ici mesurée indirectement par la taille de la fenêtre contenant les N gènes dans chaque clone) et stochasticité de l'expression génique (Figure IV.18.A). Nous avons obtenu une valeur N optimale de 37 gènes, ce qui correspond à des fenêtres de 179671 bp (pour la zone la plus dense) à 4541988 bp (pour la zone la moins dense). Il est intéressant de noter que, même si elle n'est pas optimale, nous obtenons aussi une bonne corrélation pour des fenêtres beaucoup plus petites, contenant seulement 1 à 4 gènes. Nous y reviendrons dans la section suivante.

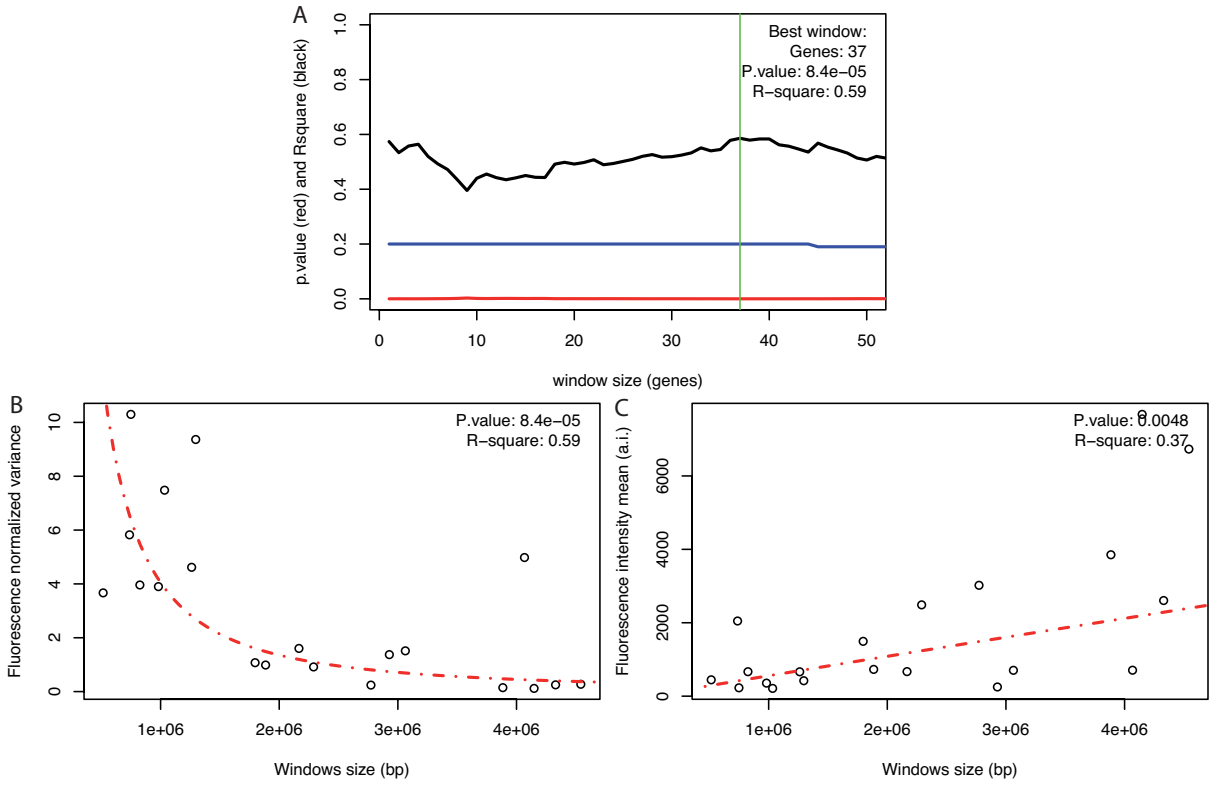

Figure IV.18 – Exploration de la densité en gènes par rapport à l'expression génique. (A) P-value et R-square du test de corrélation entre la NV d'expression de chaque clone et la taille de fenêtre (en bp) nécessaire pour contenir N gènes autour de chacun des locus d'insertion. Ces deux statistiques sont exprimées en fonction de N (donc en nombre de gènes) puisque les tailles de fenêtres ne sont pas constantes d'un clone à l'autre. La corrélation optimale est obtenue pour des tailles de fenêtre correspondant à $N = 37$ gènes (ligne verte verticale. P-value$= 8.4 \times 10^{-5}$, R-square $= 0.59$). (B) Anti-corrélation entre la NV d'expression de chaque clone et la taille de fenêtre (désormais en bp) nécessaire pour contenir $N = 37$ gènes autour de chacun des locus d'insertion (P-value $= 8.4 \times 10^{-5}$, R-square $= 0.59$). (C) Corrélation entre la moyenne d'expression de chaque locus et la taille de fenêtre (en bp) nécessaire pour contenir $N = 37$ gènes autour de chacun des locus d'insertion (P-value$= 4.8 \times 10^{-3}$, R-square $= 0.37$).

La figure IV.18.B nous montre qu'il exite une bonne corrélation entre la densité locale en gènes et la stochasticité de l'expression du transgène (P-value = 8.4×10^{-5}, R-square = 0.59) : plus la densité augmente (donc plus la taille de la fenêtre contenant 37 gènes diminue), plus la stochasticité de l'expression du transgène est forte. Dans ces mêmes conditions, on constate aussi que la moyenne d'expression diminue quand la densité en gènes augmente (Figure IV.18.C, P-value = 4.8×10^{-3}, R-square = 0.37). Sans anticiper sur la discussion, on peut d'ores et déjà noter que cette anti-corrélation entre moyenne d'expresion et densité en gènes est opposée à celle observée par Nie *et al.* (2010) (voir (Nie *et al.*, 2010), figure 7). En effet, ceux-ci montrent que l'expression moyenne des gènes augmente avec la densité en gènes. Cependant, cette tendance semble principalement due à l'activité des gènes appartenant à des RIDGEs dont la moyenne de transcription est très élevée. Ici, nous constatons donc que sans ces zones particulières que sont les RIDGEs, on observe une tendance inverse.
La méthode employée pour déterminer la taille de fenêtre optimale étant quelque peu indirecte et conduisant à de forte disparités de taille de fenêtre entre les clones, nous avons cherché à vérifier que la corrélation entre densité en gènes et stochasticité de l'expression persiste même avec une méthode plus directe. Pour cela, nous avons considéré les densités en gènes dans des fenêtres de taille fixe égale à $2 \times 6.5 \times 10^5$ bp centrées sur chacun des points d'insertion[1]. Les résultats sont présentés figure IV.19. On observe des résultats similaires à ceux obtenus pour des fenêtres de taille variable : plus le nombre de bases codantes dans la fenêtre augmente, plus la stochasticité de l'expression du transgène augmente (P-value = 1.2×10^{-4}, R-square = 0.57) tandis que la moyenne d'expression diminue (P-value = 5.6×10^{-3}, R-square = 0.35).
On notera que, quelle que soit la méthode de mesure, les corrélations entre stochasticité de l'expression et densité en éléments "fonctionnels" (les gènes) sont plus fortes que celles observées avec les taux de GC, avec la densité en régions répétées ou avec les caractéristiques globales des chromosomes. Étant données les corrélations entre la densité en gènes et ces différentes mesures, il est donc très probable que les corrélations constatées soient en fait toutes liées à une cause première qui serait l'influence de la densité locale en gènes. C'est pourquoi nous allons étudier plus finement l'influence du voisinage génique sur la stochasticité de l'expression du transgène.

3.5 Influence des gènes les plus proches

Dans la section précédente, nous avons montré que la densité en gènes autour du point d'intégration du transgène était corrélée avec la stochasticité de l'expression de ce dernier et ce pour des fenêtres de 37 gènes. Cependant, nous avons aussi vu qu'une fenêtre ne comptant que très peu de gènes pouvait aussi présenter de fortes corrélations avec l'activité du transgène (Figure IV.18.B). Ce constat nous conduit à reduire encore notre échelle d'observation et à nous focaliser non plus sur la densité mais sur la distance au gène le plus proche. Sur la figure IV.20, on peut constater que plus le locus d'insertion du transgène est proche d'un gène, plus la transcription est stochastique (i.e. la variance normalisée est

[1]La taille de fenêtre utilisée est relativement petite mais elle correspond à la plus petite distance entre les locus d'insertion et l'extrémité de leur chromosome. En utilisant une distance plus grande nous aurions soit introduit des disparités entre les clones (puisque certaines fenêtres se seraient étendues au delà du chromosome), soit dû éliminer des clones de l'étude.

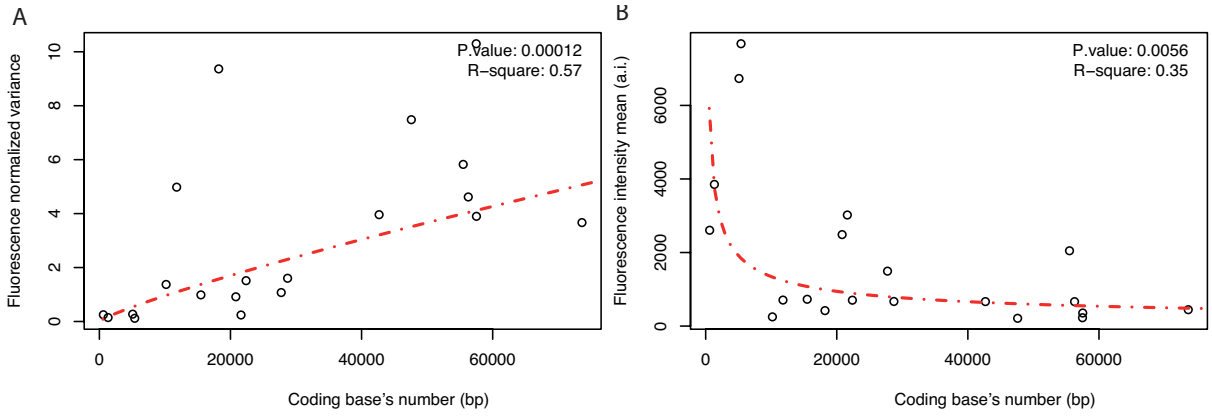

Figure IV.19 – Densité en gènes et expression génique. (A) Corrélation log-log entre le nombre de paires de bases codantes dans une fenêtre de $2 \times 6.5 \times 10^5$ bp centrée sur chaque locus et la NV d'expression de chacun de ces loci (P-value= 1.2×10^{-4}, R-square = 0.57) (B) Anti-corrélation entre le nombre de paires de bases codantes dans une fenêtre de $2 \times 6.5 \times 10^5$ bp centrée sur chaque locus et la moyenne d'expression de chacun de ces locus (P-value= 5.6×10^{-3}, R-square = 0.35).

élevée : P-value = 1.1×10^{-3}, R-square = 0.46) et plus la moyenne d'expression est basse (P-value = 8.3×10^{-3}, R-square = 0.33).

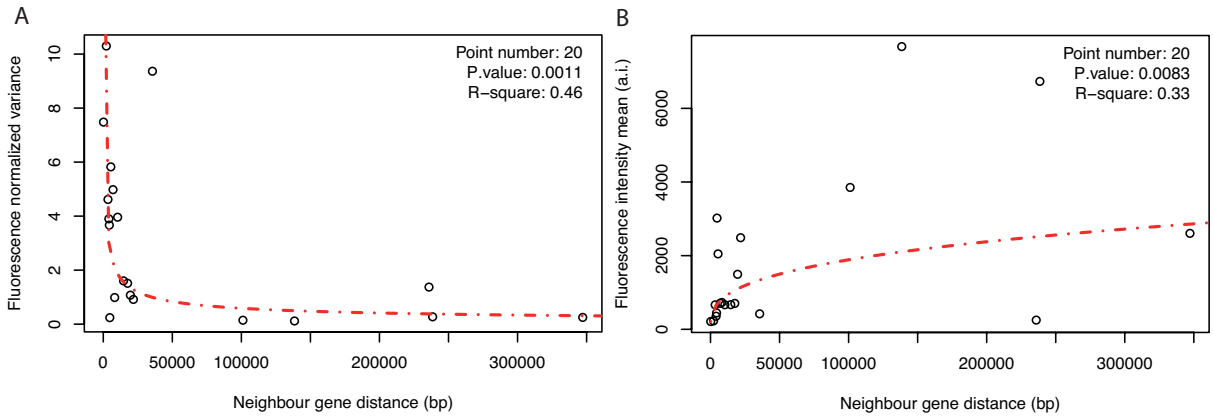

Figure IV.20 – Expression génique et la distance au gène le plus proche. (A) Anti-corrélation entre la NV d'expression du transgène et la distance au gène le plus proche (P-value= 1.1×10^{-3} ;R-square = 0.46). (B) Corrélation entre la moyenne d'expression du transgène et la distance au gène le plus proche (P-value= 8.3×10^{-3} ;R-square = 0.33).

On peut noter que la corrélation entre la stochasticité ou la moyenne d'expression et proximité d'un gène ou la densité locale en gènes sont bien de même nature : la présence de gènes proches fait croître le bruit et diminuer la moyenne d'expression.
Chez *S. cerevisiae*, Wang et al ont montré (Wang *et al.*, 2011) que le gène le plus proche influençait plus ou moins le gène de référence suivant la distance mais aussi suivant le sens relatif des deux gènes. Plusieurs cas de figures peuvent se présenter suivant le degré

de filtrage et la situation des gènes voisins du transgène (Figure IV.21).

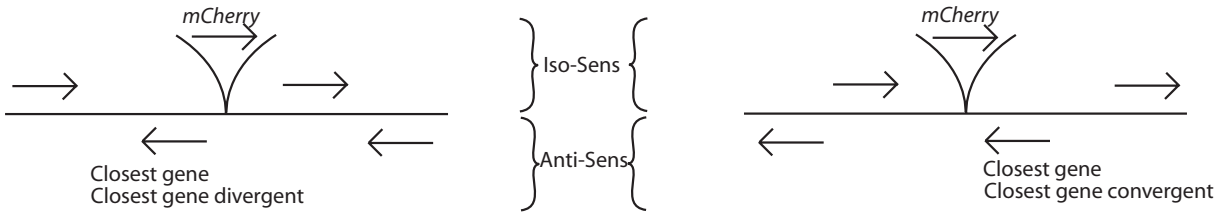

Figure IV.21 – Point d'insertion du transgène et orientations des gènes les plus proches.

On pourra ainsi s'intéresser :

- Au gène le plus proche du rapporteur (closest gene),
- Aux relations d'orientation entre les gènes (closest gene iso-sens, closest gene anti-sens),
- pour les gènes anti-sens, aux gènes convergents ou divergeants.

Comme ces différents cas de figure correspondent à des conflits moléculaires potentiels différents (par exemple entre les polymerases ou entre les propagation d'ouverture de la chromatine), nous allons les étudier séparément pour rechercher d'éventuelles corrélations. Suivant les cas de figure étudiés, les différentes configurations ne sont pas toujours réalisables. C'est pourquoi, dans toutes les figures suivantes, nous indiquerons le nombre de clones utilisés pour calculer les corrélations.
En s'intéressant aux orientations relatives des gènes, nous obtenons des tendances similaires aux précédentes, que ce soit pour les gènes les plus proches de même orientation que le transgène (Figure IV.22) ou pour les gènes d'orientation opposée (Figure IV.23). Cependant, la significativité des tests est nettement plus élevée lorsqu'on prend spécifiquement en compte l'influence de la distance au gène le plus proche anti-sens sur la stochasticité de l'expression du transgène (P-value = 1.8×10^{-4}, R-square = 0.67) par rapport à l'influence de la distance au gène le plus proche iso-sens (P-value = 7.7×10^{-3}, R-square = 0.49). Cette corrélation produit même un signal plus fort que la densité locale en gènes (R-square = 0.59, Figure IV.18). Il semble donc y avoir un effet important du sens du gène voisin le plus proche sur la stochasticité de l'expression génique.
Similairement à Wang *et al.* (2011) nous avons ensuite étudié l'effet de la position relative du gène et du transgène le long de la séquence. Pour cela, nous avons considéré les situations pour lesquelles le gène et le transgène sont dans des sens opposés, le gène le plus proche étant positionné soit avant le transgène (cas divergeant, figure IV.24) soit après le transgène (cas convergeant, figure IV.25). Là encore, nous obtenons une corrélation entre la distance au gène le plus proche et la stochasticité de l'expression du transgène nettement plus prononcée dans le cas convergeant (Figure IV.25.A : P-value = 6.1×10^{-3}, R-square = 0.63) que dans le cas divergeant (Figure IV.24.A : P-value = 2.0×10^{-2}, R-square = 0.43). On peut donc en conclure que c'est la distance avec le gène anti-sens convergeant le plus proche qui offre la meilleure corrélation avec la stochasticité de l'expression génique : plus cette distance est courte, plus la variance normalisée est élevée.

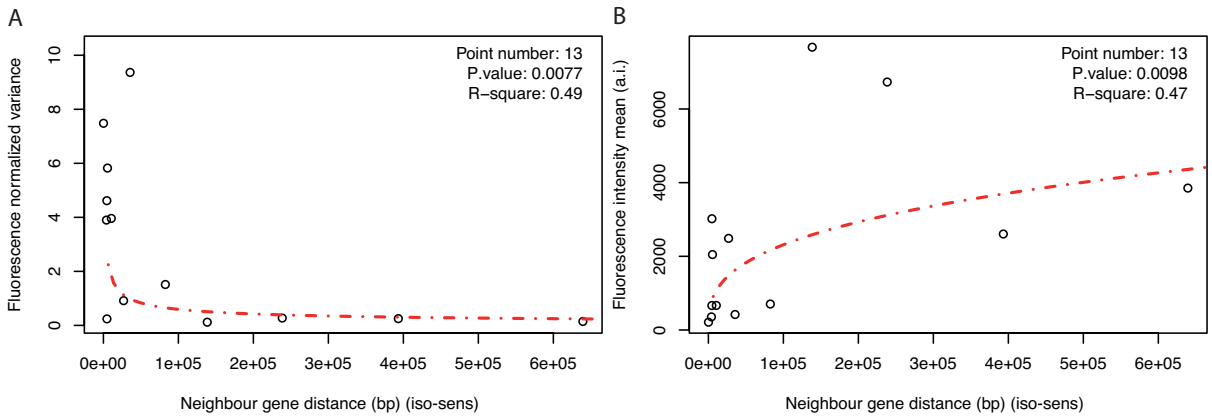

Figure IV.22 – Expression génique et la distance au gène iso-sens le plus proche (analyse portant sur 13 clones). (A) Anti-corrélation entre la NV de l'expression du transgène et la distance au gène iso-sens le plus proche (P-value= 7.7×10^{-3}, R-square = 0.49). (B) Corrélation entre la moyenne d'expression du transgène et la distance au gène iso-sens le plus proche (P-value= 9.8×10^{-3}, R-square = 0.47).

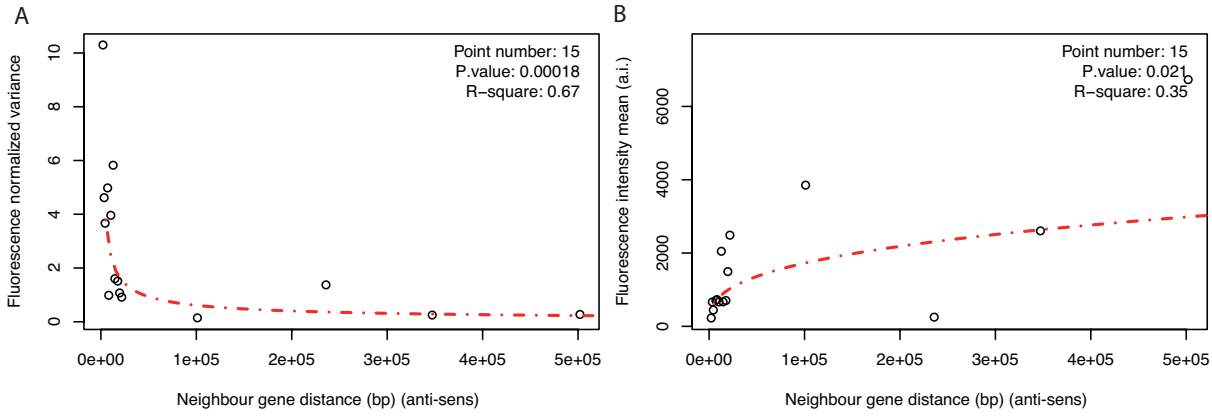

Figure IV.23 – Expression génique et la distance au gène anti-sens le plus proche (analyse portant sur 15 clones). (A) Anti-corrélation entre la NV de l'expression du transgène et la distance au gène anti-sens le plus proche (P-value= 1.8×10^{-4}, R-square = 0.67). (B) Corrélation entre la moyenne d'expression du transgène et la distance au gène anti-sens le plus proche (P-value= 2.1×10^{-2}, R-square = 0.35).

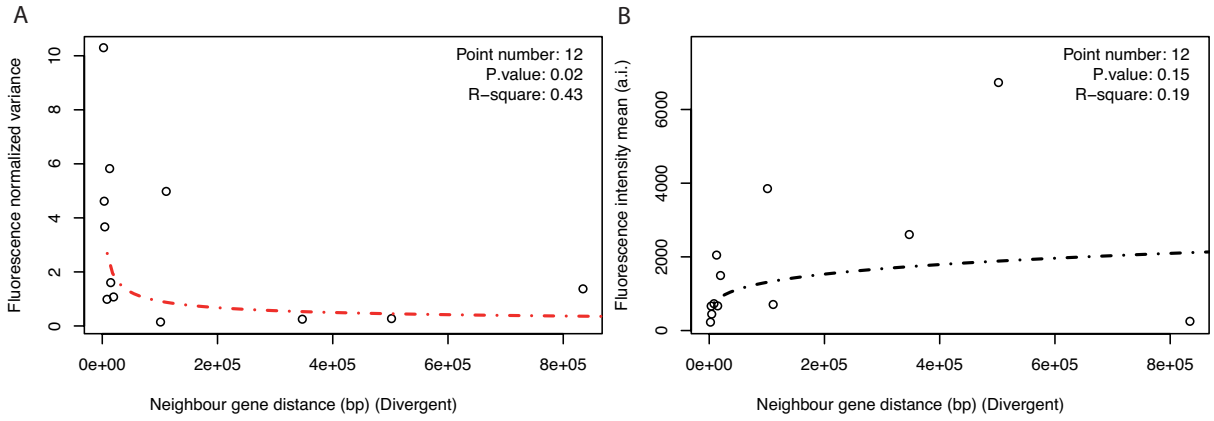

FIGURE IV.24 – Relation entre la distance au gène divergeant le plus proche et l'expression génique (analyse portant sur 12 clones). (A) Anti-corrélation entre la NV de l'expression du transgène et la distance entre le locus du transgène et le gène divergeant le plus proche (P-value= 2.0×10^{-2}, R-square = 0.43). (B) Absence de corrélation entre la moyenne d'expression du transgène et la distance entre le locus du transgène et le gène divergeant le plus proche.

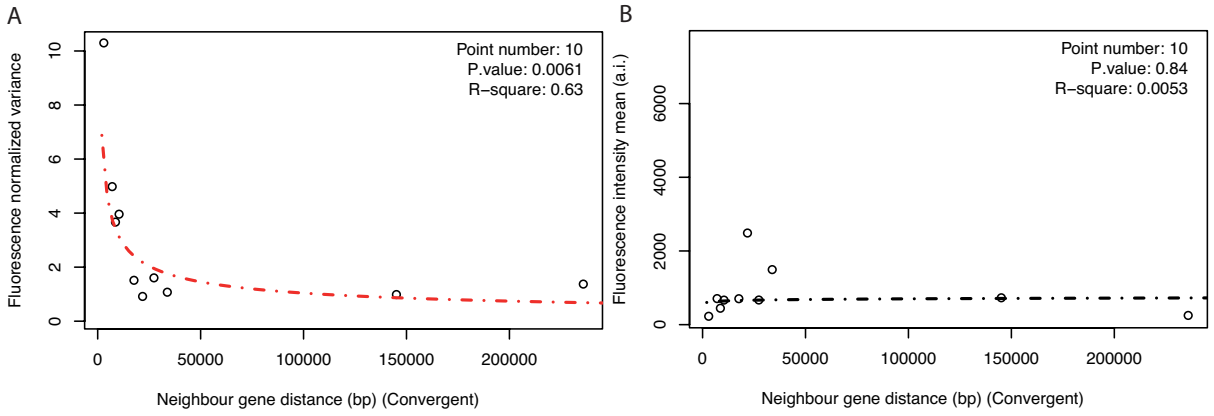

FIGURE IV.25 – Relation entre la distance au gène convergeant le plus proche et l'expression génique (analyse portant sur 10 clones). (A) Anti-corrélation entre la NV de l'expression du transgène et la distance entre le locus du transgène et le gène convergeant le plus proche (P-value= 6.1×10^{-3}, R-square = 0.63). (B) Absence de corrélation entre la moyenne d'expressiondu transgène et la distance entre le locus du transgène et le gène convergeant le plus proche.

La relation de proximité avec le gène convergeant le plus proche a donc clairement un impact fort sur la stochasticité de l'expression du transgène (alors qu'elle n'a pas d'effet statistiquement détectable sur la moyenne d'expression de ce dernier).

En utilisant les résultats du chapitre précédent, nous pouvons utiliser la moyenne et la variance normalisée d'expression du transgène pour calculer le temps moyen fermé théorique de la chromatine ainsi que la taille théorique des bursts de transcription. Les résultats de ces calculs sont présentés par rapport aux deux caractéristiques présentant la meilleure corrélation avec la stochasticité de l'expression génique : la distance au gène anti-sens le plus proche (Figure IV.23) et la distance au gène convergeant le plus proche (Figure IV.25).
En représentant les caractéristiques dynamiques de la chromatine (extrapolées à partir du modèle du chapitre III) en fonction de la distance au gène le plus proche et de son orientation, on constate une corrélation nette entre le temps moyen fermé théorique de la chromatine et la proximité du gène, dans le cas convergeant (figure IV.26.A ; P-value= 3.1×10^{-3}, R-square= 0.69) comme dans le cas anti-sens (figure IV.27.A ; P-value= 1.4×10^{-4}, R-square = 0.69).

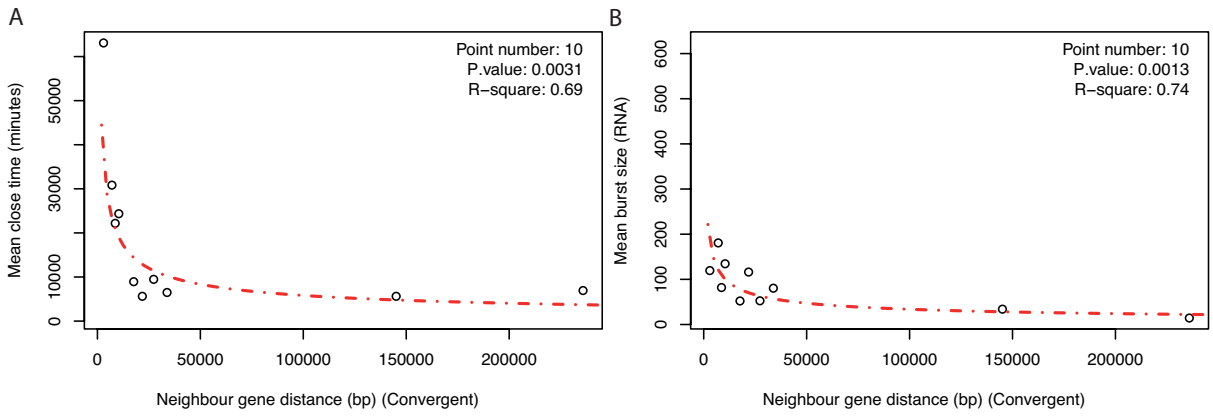

Figure IV.26 – Relation entre la distance au gène convergeant le plus proche et la dynamique chromatinienne (analyse portant sur 10 clones). (A) Anti-corrélation entre la moyenne du temps moyen fermé de la chromatine et la distance au gène convergeant le plus proche (P-value = 3.1×10^{-3}). (B) Anti-corrélation entre le nombre moyen d'ARN produits par période d'ouverture de chromatine et la distance au gène convergeant le plus proche (P-value = 1.3×10^{-3}, R-square = 0.74).

Une telle corrélation est aussi visible entre le nombre moyen théorique d'ARN produits par burst et la proximité du gène, qu'il soit convergeant (Figure IV.26.B ; P-value = 1.3×10^{-3}, R-square = 0.74) ou anti-sens (Figure IV.27.B ; P-value= 4.8×10^{-2}, R-square = 0.27).
Il est à noter que les corrélations calculées sont plus fortes dans le cas des gènes convergeants. Cependant cette analyse est réalisée avec seulement la moitié des clones, ce qui nous interdit de conclure. Reste que, pour les deux situations testées, ces corrélations montrent que la proximité d'un gène (convergeant ou anti-sens) augmente le temps moyen fermé de la chromatine et l'intensité des bursts de transcription, ce qui tend à augmenter la stochasticité de l'expression du transgène sans nécessairement augmenter la moyenne.

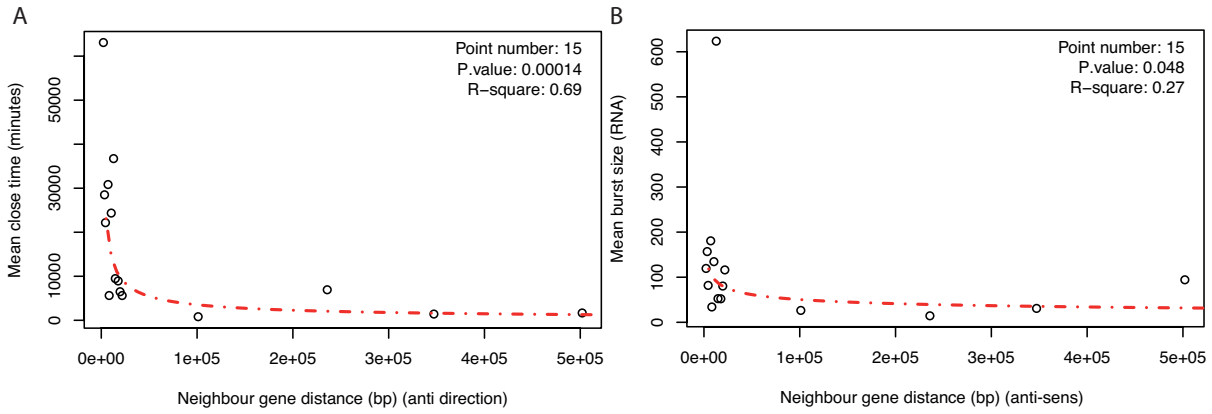

Figure IV.27 – Relation entre la distance au gène anti-sens le plus proche et la dynamique chromatinienne (analyse portant sur 15 clones). (A) Anti-corrélation entre la moyenne du temps moyen fermé de la chromatine et la distance au gène anti-sens le plus proche (P-value= 1.4×10^{-4}, R-square= 0.69). (B) Anti-corrélation entre le nombre moyen d'ARN produits par période d'ouverture de chromatine et la distance au gène anti-sens le plus proche (P-value= 4.8×10^{-2}, R-square= 0.27).

4 Discussion

4.1 Moyenne d'expression et environnement génomique

Nous constatons régulièrement, pour une même caractéristique génomique observée dans les mêmes conditions, la présence d'une corrélation significative avec la stochasticité et l'absence de corrélation significative avec la moyenne d'expression observée (Figures IV.14, IV.15 et IV.16). L'absence de corrélation entre la moyenne d'expression génique et le taux de GC local ne signifie pas forcement qu'aucune corrélation n'existe entre ces deux éléments. En effet, il est probable qu'il existe une corrélation entre la moyenne d'expression et le taux de GC local, mais pas dans les conditions d'observations présentées. En effet, cette étude est centrée sur la stochasticité de l'expression des gènes et ses liens avec leur environnement génomique. Les paramètres des analyses présentées ici ont donc été choisis afin de favoriser l'observations des liens avec la stochasticité et non avec la moyenne d'expression observée.

On sait en outre que dans les zones du génome où l'expression moyenne est plus élevée (les RIDGEs), le taux de GC est plus important (Versteeg *et al.*, 2003; Nie *et al.*, 2010). Les études de Versteeg *et al.* (2003) et Nie *et al.* (2010) portaient sur l'expression des gènes naturellement présents dans le génome (endogènes). Ces gènes différent bien sûr par leur séquence mais aussi par leur taille, leur pourcentage GC, leur promoteur, etc. Dans l'étude présentée ici, nous avons analysé l'expression d'un gène exogène, toujours de même séquence et avec le même promoteur, la seule différence entre les clones étant le contexte génomique d'insertion.

Les cellules sont le résultat d'une longue évolution. Leurs génomes ont donc été sélectionnés de sorte que le répertoire de gènes exprimés, leur niveau d'expression et de stochasticité soient adaptés à leurs besoins compte tenu de l'environnement dans lequel elles

ont évolué. Or, comme nous l'avons vu en introduction, il existe de très nombreux leviers de régulation de l'expression génique. Un gène peut par exemple être régulé par capture de ses ARNs par un micro-ARN spécifique. Il peut également être régulé par un gène codant pour un facteur de transcription spécifique d'un des sites de fixation présent sur son promoteur. Il peut être inséré à un endroit où la chromatine est globalement plus ou moins ouverte qu'ailleurs (Gierman *et al.*, 2007; Batada et Hurst, 2007). Globalement, les gènes endogènes ont donc des caractéristiques transcriptionnelles qui leur sont propres et qui permettent une régulation qui leur est spécifique. Ce n'est pas le cas du transgène utilisé ici. Celui-ci ne varie pas en séquence et, du fait de son caractère exogène, a peu de chance d'être régulé par la cellule. Les différences d'expression que l'on constate ici sont donc très probablement dues à l'environnement génomique du locus d'insertion et non à toute autre cause.

Nous avons constaté, dans les données présentées ci-dessus, que la moyenne d'expression est plus forte quand la densité en gènes est moins élevée (Figures IV.18 et IV.19) et quand la distance au gène le plus proche est plus courte (indépendamment du type de gène concerné : Figure IV.20). Nie *et al.* (2010) constatent que ce sont les zones (RIDGEs) denses en gènes qui sont en moyenne plus exprimées. Nous n'avons aucun rapporteur dans ces RIDGEs. Il y a donc probablement des variations entre les caractéristiques des gènes endogènes dans les RIDGEs et hors des RIDGEs. C'est ce que montrent Nie *et al.* (2010) : dans ces RIDGEs, les gènes sont plus riches en GC, plus courts et avec des introns plus courts.
Si l'on considère spécifiquement l'effet de la distance entre gènes, les résultats obtenus seraient en faveur d'une hypothèse de compétition pour l'accès aux ressources nécessaire à la transcription : pour un gène donné, la proximité d'un autre gène peut entraîner la consommation (ou plus simplement la mobilisation) de certains composants moléculaires, diminuant d'autant l'efficacité de la transcription et conduisant à l'augmentation de la stochasticité.

4.2 Stochasticité et environnement génomique

Les corrélations observées entre stochasticité de l'expression du transgène (ou, plus exactement, la variance normalisée de son expression) et les caractéristiques génomiques de son point d'insertion sont multiples. Cependant, au vu de l'intensité de ces corrélations et des corrélations présentes entre les différentes caractéristiques génomiques, nous ferons l'hypothèse que le paramètre explicatif principal est la proximité des gènes et que toutes les autres corrélations découlent de celle-ci. En conséquence, nous discuterons ici essentiellement de ce résultat.
Le principal enseignement de cette étude est l'augmentation de la stochasticité de l'expression du transgène lorsque celui-ci est inséré dans une zone de forte densité en gènes (Figures IV.18 et IV.19) et, en particulier, lorsqu'il est inséré à proximité d'un gène endogène (Figure IV.20). Nous avons vu que cet effet s'accompagne d'une diminution significative de l'expression et avons fait l'hypothèse que cette diminution pourrait être expliquée par un phénomène de compétition pour l'accès aux ressources moléculaires nécessaires à la transcription. Cette même hypothèse permettrait aussi d'expliquer l'augmentation de la stochasticité. En effet, ce phénomène de compétition provoquerait une diminution

apparente de la concentration locale de ces ressources moléculaires, cette diminution étant d'autant plus forte que la densité en gènes est importante. Ainsi, lorsqu'un élément de la machinerie de transcription est fixé sur un gène, il n'est pas disponible, au moins pendant un court instant, pour un autre, cet effet étant exacerbé si on assiste à des phénomènes de recapture.
Or, il est bien connu que, plus les concentrations moléculaires sont faibles, plus la stochasticité d'une réaction est importante.
La différence entre les deux observations (densité en gènes vs. distance au gène le plus proche) est la taille de la zone concernée. Cependant, lorsqu'on considère des phénomènes de diffusion, l'influence de la proximité des gènes a tendance à décroitre très rapidement avec l'augmentation de la distance entre les gènes. Il n'est donc pas étonnant qu'on observe un effet particulièrement marqué pour la distance au gène le plus proche. Cet effet est d'ailleurs plus marqué pour un voisinage inférieur à 10^5bp du transgène (Figure IV.20). On retrouve cet ordre de grandeur dans les travaux de Ebisuya *et al.* (2008). Ils montrent qu'il y a un effet de vague autour d'un gène dont on induit la transcription et ce jusqu'à environ 10^5bp du gène induit. Dans cette étude, les effets sont cependant inversés par rapport à ceux mesurés ici : Ebisuya *et al.* (2008) observent une sur-expression des gènes voisins. Les causes de ces variations n'étant pas connues, il est difficile de comparer nos études, d'autant plus que les conditions expérimentales sont très différentes, Ebisuya *et al.* (2008) mesurant l'activité de gènes endogènes quand nous mesurons l'activité d'un rapporteur exogène.

Il faut cependant noter que nous ne considérons pas ici l'espace selon ses trois dimensions (ce qui serait naturel s'agissant d'un mécanisme de diffusion de composants moléculaires) mais suivant la seule dimension de la séquence. Même si l'ADN est généralement très fortement compacté, on considère localement que la distance sur la séquence représente un bon proxy de la distance spatiale. En outre, plusieurs auteurs ont démontré la possibilité, pour les molécules impliquées dans le transcription, d'une diffusion "1D" par "sliding" sur le ruban d'ADN. L'utilisation de la distance génétique devient alors tout à fait justifée.

Un des résultats les plus intriguants est que l'augmentation de la stochasticité liée à la proximité des gènes semble fortement influencée par le positionnement relatif des gènes. En effet, la significativité des tests est plus importante quand l'on ne considère que les gènes dans le sens opposé à notre rapporteur (Figures IV.22 et IV.23). Il est même encore plus marqué si l'on considère uniquement les gènes convergeants avec notre rapporteur (Figure IV.25) par rapport aux gènes divergeants (Figure IV.24). Si cette observation n'est pas en contradiction avec un mécanisme de compétition pour les ressources, elle en précise les modalités. En effet, une "simple" compétition pour l'accès à une ressource diffusant dans le nucléoplasme n'aurait aucune raison de produire un tel effet. Celui-ci est donc plus probablement lié à une compétition pour un état chromatinien favorable à l'expression ou pour un élément lié à la séquence, qu'il s'agisse de la position des nucléosomes, de l'accessibilité du promoteur ou, là encore, d'un phénomène de sliding sur l'ADN. Il est par exemple possible que la conformation de la chromatine d'un locus exprimé influe sur la conformation de la chromatine voisine en réprimant par exemple la possibilité pour les gènes voisins d'être exprimables. Notons cependant que, si les tests liés à l'orientation des gènes sont significatifs, ils ne portent que sur un sous-ensemble des clones. Il convient donc

de considérer ces valeurs avec précaution et une vérification sur un plus grand nombre de clones semble nécessaire.

Il est à noter que l'influence de l'orientation relative des gènes a été soulignée dans d'autres études (Wang *et al.*, 2011; Woo et Li, 2011). Chez les bactéries, Wang *et al.* (2011) ont constaté une corrélation entre la distance aux gènes divergents les plus proches et la stochasticité de leur expression. En revanche, ils n'ont pas constaté de corrélation entre la distance aux gènes convergents les plus proches et la stochasticité. Woo et Li (2011) ont quant à eux constaté des corrélations aussi bien avec la distance aux gènes convergents que divergents. En outre, ils constatent que les nucléosomes sont plus fortement positionnés quand la distance est réduite, ce qui tendrait à favoriser une hypothèse de compétition pour l'ouverture de la chromatine. Une telle hypothèse serait confortée par nos prédictions théoriques. En effet, plus un nucléosome est fortement positionné, plus la dynamique chromatinienne va être ralentie (Tirosh et Barkai, 2008). Nous constatons grâce à notre modèle et aux données biologiques que, quand la distance aux gènes diminue, le temps moyen fermé augmente (Figure IV.27.A et IV.26.A), de même que le nombre d'ARN produits par burst de transcription (Figures IV.27.B et IV.26.B). Cette prédiction théorique suggère donc bien un lien entre proximité des gènes et dynamique chromatinienne. Cependant, de façon surprenante au vu des résultats, les travaux de Woo et Li (2011) ne montrent pas d'augmentation de la stochasticité avec la proximité des gènes, au contraire. Un point clé de notre étude - susceptible d'expliquer ces différences – est l'utilisation d'un rapporteur exogène (*mCherry*) transcrit par un promoteur particulièrement actif (CMV). En effet, Woo et Li (2011) constatent que la proximité des gènes s'accompagne de variations de la structure des promoteurs avec la diminution du nombre de boites TATA et du nombre de sites de fixation de facteurs de transcription, ce qui stabilise l'expression des gènes (Blake *et al.*, 2006). Choi et Kim (2009) ont par ailleurs montré que les caractéristiques des promoteurs influencent leur comportement et en particulier la stochasticité induite. De même, Batada et Hurst (2007) montrent que les gènes essentiels sont clusterisés et moins bruyants. Tous ces résultats tendent à montrer que l'évolution aurait sélectionné des structures stables dans le cas de gènes endogènes proches (en diminuant le nombre d'éléments génomiques induisant la compétition). Il serait interessant de regarder si, dans notre modèle cellulaire, les gènes naturels plus proches sont pauvres en boites TATA (à l'image de ce que voient (Woo et Li, 2011)). Lorsqu'on utilise un rapporteur exogène avec un promoteur fort, une telle optimisation du promoteur inséré n'a pas eu lieu, ce qui nous permet de mettre en évidence le rôle de la co-localisation des gènes indépendamment de l'optimisation de leurs promoteurs. Notre étude permettrait donc de révéler des effets de proximité qui seraient dans d'autres cas masqués par l'évolution des séquences promotrices.

L'hypothèse de "compétition" est confortée par d'autres études qui semblent révéler un mécanisme similaire. Ainsi, au cours de sa thèse, Guillaume Corre (2012) a constaté que deux rapporteurs fluorescents (exogènes donc) placés dans une cellule sont d'autant plus anticorrélés qu'ils sont proches. Globalement les résultats présentés ici, couplés aux

résultats publiés par ailleurs, conduisent à dresser un tableau complexe des interactions entre gènes proches. En effet, ces interactions pourraient être dues à des phénomènes de

compétition mais ceux-ci pourraient être réduits, voir supprimés, par la modification de la structure des promoteurs. Pour tester une telle hypothèse, plusieurs expérimentations sont possibles. Ainsi, il conviendrait de :

- s'assurer que la stochasticité de deux gènes en compétition est bien augmentée lorsque la distance diminue. Une telle expérience permettrait de confirmer les effets de compétition locale. Il faudrait pour cela utiliser différentes constructions génomiques comportant deux couples promoteur/rapporteur fluorescent et dont la distance entre les couples varie suivant la construction. En les insérant au même endroit, l'une après l'autre, on pourrait alors constater les différences induites sur leur stochasticité.

- s'assurer que, suivant les caractéristiques du promoteur (nombre de boites TATA et de sites de fixation de TF), le gène sera plus ou moins bruyant. Pour ce faire, il faudrait utiliser différents promoteurs à un même locus, par exemple, par recombinaison homologue.

Nous sommes en train de mettre au point ces techniques (particulièrement de recombinaison homologue) dans notre modèle cellulaire. Cependant, pour compléter cette étude et avoir plus d'informations la liant aux travaux d'autres chercheurs comme Woo et Li (2011), on pourrait utiliser une nouvelle technologie (commercialisée par la société fluidigm) qui permet de caractériser l'expression de 96 gènes dans chacune des cellules d'une population de 96 cellules. En effet, cela permettrait d'obtenir la distribution d'expression de gènes précis dans une petite population cellulaire. Cela nous permettrait en particulier de cibler notre étude sur différents groupes de gènes endogènes :

- Les gènes endogènes situés dans un RIDGE,

- Les gènes endogènes situés dans un anti-RIDGE,

- Les gènes endogènes situés dans des zones d'expression "moyenne".

En effet, au cours de cette étude, nous n'avons étudié que des gènes exogènes situés hors des RIDGEs alors que, par définition, la transcription des gènes endogènes situés dans les RIDGEs suit une dynamique très spécifique.

Nous avons donc montré qu'il existe un lien fort entre stochasticité et caractéristiques génomiques et plus précisément, que la stochasticité dépend de la distance entre les gènes, peut-être du fait d'un mécanisme de compétition. En outre, nous suggérons que ce mécanisme de compétition soit multifactoriel et qu'il puisse dépendre de la distance entre les gènes, de la structure des promoteurs en compétition, mais aussi du positionnement local des nucléosomes. Il semble donc que la stochasticité de l'expression puisse être régulée de très nombreuses manières, que ce soit en modifiant la séquence localement, en modifiant la dynamique chromatinienne locale ou, plus globalement, en changeant la position relative des gènes et donc l'activité transcriptionnelle locale (Kaplan *et al.*, 2009; Tillo et Hughes, 2009; Choi et Kim, 2009; Cairns, 2009).

Chapitre V

Conclusion

Au cours de ce travail, nous avons utilisé une approche couplée, mixant biologie expérimentale, biologie *in silico* et bioinformatique afin d'étudier les origines moléculaires de la stochasticité de l'expression génique. Pour cela, nous avons utilisé un matériel biologique inédit, à savoir un ensemble de clones ne différants que par les points d'insertion d'un rapporteur fluorescent exogène. Ce matériel biologique nous permet donc d'avoir accès, par des approches comparatives, à l'influence du locus chromosomique sur la stochasticité de l'expression des gènes. Nous revenons ici sur les deux principaux enseignements que nous pouvons tirer de ce travail avant de discuter plus largement des perspectives ouvertes et des analyses complémentaires qui pourraient être entreprises.

1 Locus génomique et stochasticité de l'expression génique

Le premier enseignement que nous pouvons tirer de notre travail est qu'il y a, chez les eucaryotes supérieurs, un effet du locus sur la stochasticité de l'expression génique (chapitre II). Un tel effet était déjà documenté pour la moyenne d'expression (Nie *et al.*, 2010; Versteeg *et al.*, 2003; Caron *et al.*, 2001) mais nous avons montré (comme précédemment documenté (Neildez-Nguyen *et al.*, 2008)) que, non seulement la stochasticité de l'expression d'un même gène change en fonction de son locus d'insertion, mais aussi que ces variations sont indépendantes de variations de la moyenne d'expression entre les différents loci. Par ailleurs, nous avons pu montrer que la stochasticité d'expression d'un locus donné était sensible à l'action de perturbateurs chromatiniens. Ce résultat nous a mis sur la voie d'une cause moléculaire probable de ces variations, à savoir la dynamique chromatinienne. Nous avons donc pu formuler l'hypothèse que cette dynamique n'est pas uniforme tout au long du génome, ce qui expliquerait d'une part les différences de moyenne d'expression (la proportion de temps ouvert diffèrerait selon les zones du génome), les différences de stochasticité de l'expression des gènes (la dynamique d'ouverture-fermeture dépendrait des zones du génome) mais aussi le découplage entre ces deux indicateurs (le temps moyen ouvert et le temps moyen fermé pouvant varier différemment suivant les zones du génome).
Dans un second temps nous sommes partis de cette hypothèse et avons utilisé l'information portée par la dynamique de l'expression du transgène (moyenne, variance normalisée et

histogramme) pour remonter à la dynamique chromatinienne au point d'insertion. Nous avons ainsi pu confirmer (chapitre III) que les données mesurées dans les différents clones pouvaient être expliquées par des variations locales des temps moyens fermés et de la taille des bursts de transcription produits lors des périodes d'ouverture de la chromatine. Un enseignement très surprenant de ce travail est que, au moins dans les cellules étudiées ici, les temps moyens fermés sont très longs (de l'ordre de plusieurs jours) alors que les temps moyens ouverts sont courts[1]. Nous avons aussi pu montrer que les différences d'expression liées au locus d'insertion du transgène s'expliquent en grande partie par les différences de taux d'ouverture de la chromatine (donc de temps moyen fermé).
Pour conclure, nous avons pu modéliser l'expérience première de ce travail (la perturbation de la stochasticité par des modificateurs chromatiniens), ce qui nous a permis de montrer que la Trichostatin A (TSA) agissait sur la chromatine essentiellement en augmentant la probabilité des bursts de transcription, ce qui explique les variations constatées de la moyenne et de la variance normalisée de l'expression du transgène lors des traitements TSA. Ce résultat nous montre en outre que le couplage entre les approches de biologie expérimentale et de biologie *in silico* permet d'exploiter beaucoup plus en profondeur les données expérimentales : ici, nous avons pu remonter à la dynamique chromatinienne à partir des "seules" données de cytométrie en flux, données faciles à acquérir en grande quantité. Néanmoins, un tel couplage ne permet que de formuler des hypothèses qui doivent toujours être mises à l'épreuve.

2 Voisinage génique et stochasticité de l'expression génique

Dans le dernier chapitre de ce manuscrit, nous avons utilisé le même matériel biologique. Cependant, l'évolution des techniques expérimentales accessibles durant la thèse nous a permis d'obtenir un plus grand nombre de clones et donc de réaliser des études statistiques plus classiques. En utilisant simultanément les données d'expression mesurées en cytométrie de flux et le locus d'insertion caractérisé par séquençage, nous avons pu montrer que le voisinage génique du transgène (distance et orientation relative du plus proche gène endogène) avait une influence importante sur la stochasticité, même à promoteur et à gène constant (la construction CMV-mCherry utilisée étant strictement la même pour tous les clones). En outre, en couplant ces mesures aux résultats précédents, nous avons pu montrer que ce voisinage intervient essentiellement en perturbant la dynamique des transitions actif-inactif du gène (ou, plus précisément, la taille des bursts de transcription et le temps moyen fermé de la chromatine). Au vu de ces résultats, nous pouvons émettre l'hypothèse que la proximité d'un gène endogène perturbe d'activité du transgène du fait de conflits pour l'accès à une ou plusieurs ressources nécessaires à la transcription (qu'il s'agisse de réactants critiques ou de l'accessibilité de ces derniers aux gènes). Plus des gènes sont proches du site d'intégration du rapporteur, plus ces conflits rendent l'expression du rapporteur stochastique. Cet effet n'est pas visible dans d'autres études (Woo et Li, 2011; Choi et Kim, 2009), probablement parce que celles-ci n'utilisent pas une

[1] Ce qui nous interdit de séparer les paramètres "temps moyens ouverts" et "taux de transcription" pour ne retenir que le paramètre aggrégé correspondant à la taille des bursts.

contruction identique mais étudient la proximité de gènes différents (endogènes). Ceux-ci ont donc probablement des caractéristiques dynamiques très différentes les uns des autres (puisque les promoteurs et les gènes sont différents), ce qui masque probablement l'effet du voisinage génique.
Cette dernière hypothèse demanderait cependant à être testée directement. Pour cela, nous pourrions étudier l'influence du voisinage génique sur différentes constructions. Nous pourrions par exemple tester des constructions comportant un promoteur différent (composition en boîtes TATA et en sites de fixation des facteurs de transcription par exemple). Cette étude permettrait de vérifier si l'impact du voisinage génique peut être masqué par les différences de dynamiques entre promoteurs.

3 Perspectives

Notre travail ouvre plusieurs perspectives que nous diviserons ici en deux groupes. D'une part il est toujours possible de rechercher des biais expérimentaux et, donc, de conduire des expériences et analyses complémentaires destinées à les écarter. D'autre part, les pistes élaborées ici quant à l'origine moléculaire de la stochasticité de la régulation génique peuvent être creusées aux moyens d'expériences et ou de modèles nouveaux.

3.1 Biais expérimentaux et expériences complémentaires

L'approche expérimentale utilisée ici est basée sur la mesure de fluorescence des cellules par cytométrie de flux. Cette mesure est susceptible d'être entachée de différents biais, depuis la génération des clones jusqu'à la mesure de fluorescence. Bien que nous ayons essayé de nous affranchir de la plupart de ces biais en générant des populations de clones ne différant les uns des autres par une seule caractéristique, il subsiste toujours un risque d'interaction entre l'un de ces biais et nos résultats.

3.1.1 Biais du système Tol2-transposase

La génération de nos clones est basée sur une méthode de biologie moléculaire éprouvée permettant une insertion à priori aléatoire. Cependant, plusieurs études ont montré une sur-représentation des insertions dans les unités de transcription (Huang *et al.*, 2010; Kondrychyn *et al.*, 2009). Dans nos données, nous avons effectivement constaté une faible sur-représentation des insertions dans les zones denses en gènes, proches d'un gène, riches en base GC, pauvres en régions répétées ou encore dans des zones en moyenne plus fortement exprimées (voir chapitre IV, section 2.2). Même si le nombre de clones traités (en particulier dans le chapitre IV) peut nous permettre d'avoir, malgré ce biais, un bon échantillonnage des points d'insertion, il serait intéressant de compléter nos données, soit en multipliant les clones (ce qui permettrait de mieux échantillonner les points d'insertion), soit en utilisant une autre technique d'insertion telle que le système PiggyBac (Huang *et al.*, 2010) à supposer qu'il permette une insertion complètement aléatoire dans notre modèle cellulaire.

3.1.2 Biais de sélection des clones

Dans toutes nos études, l'insertion aléatoire du système Tol2 ne nous permet pas de garantir la présence d'un point d'insertion dans une cellule ni, lorsqu'un point d'insertion est avéré, son unicité. Il serait évidemment possible de sélectionner des clones aléatoirement puis de les trier par séquençage, cependant une telle approche représenterait un investissement très élévé tant au niveau financier que humain. Nous avons utilisé une approche plus rapide : les clones sont sélectionnés lorsqu'ils présentent une fluorescence minimale lors d'une l'étape de tri cellulaire (réalisée au FACS). Lors de cette étape, nous éliminons les cellules dont la fluorescence n'est pas différenciable de l'autofluorescence naturelle des 6C2 et considérons ces cellules comme non stablement transfectées par le gène codant pour *mCherry*. Cette sélection par la fluorescence élimine potentiellement les cellules où le transgène se serait inséré dans une zone "silencieuse" du génome. Il est donc possible que, par cette méthode, on élimine certaines zones du génome de nos études. Comme nous venons de le dire, une approche par séquençage serait plus précise mais le nombre de clones à produire serait rédhibitoire. Pour rechercher expérimentalement des clones transfectés dans des zones silencieuses, quatre approches pourraient être utilisées :

- Lors de l'étape de tri par fluorescence, pré-sélectionner des cellules non fluorescentes puis les laisser chacune s'expendre en clone. Caracteriser ensuite chacun de ces clones par RT-qPCR puis par splinkerette PCR pour identifier ceux comportant une et une seule insertion. La préselection permettrait de ne tester qu'une partie des clones. Néanmoins l'investissement resterait probablement ici très important.

- Sélectionner des clones silencieux soumis à un traitement TSA. En effet, si leur faible (ou "non"-) expression est due à la chromatine qui est localement plus fermée, la TSA peut révéler leur expression. Cela permettrait alors de confirmer que certaines zones sont très fermées et de comparer, sans l'effet chromatine, leur stochasticité à celle de zones statistiquement plus ouvertes.

- Ouvrir localement la chromatine en utilisant pour la transfection un transgène combinant gène et promoteur et flanqués d'insulateurs[1]. Une fois les clones sélectionnés, le transgène pourrait être remplacé par la construction initiale par recombinaison homologue. Cette approche permettrait en outre de confirmer le mode d'action des insulateurs et de vérifier que les zones silencieuses sont bien dues à une dynamique chromatinienne particulière. Ce travail est actuellement en cours dans l'équipe.

- Analyser l'expression des clones par des méthodes plus sensibles que la cytométrie de flux. On pourrait ainsi utiliser une PCR pour doser l'expression du rapporteur dans les cellules et ce, clone par clone. Ainsi, on pourrait vérifier si les clones "silencieux" le sont parce que le transgène est totalement absent du génome ou parce que son expression est trop faible pour être distinguée de l'autofluorescence.

[1]Les insulateurs sont des séquences génétiques supposées bloquer la propagation de la dynamique chromatinienne le long du génome (Gaszner et Felsenfeld, 2006; Pikaart *et al.*, 1998). Ils sont supposés être particulièrement actifs sur les zones silencieuses (Gaszner et Felsenfeld, 2006).

3.1.3 Rapporteur fluorescent et nombre de protéines

Nous avons étudié la stochasticité de l'expression génique par le biais d'un rapporteur fluorescent en faisant l'hypothèse que la quantité de lumière observée dans une cellule est linéairement corrélée à la quantité de protéines fluorescentes présentes. Il est cependant possible que la concentration de ces protéines et/ou la forme de la cellule étudiée puissent induire des variations de fluorescence, même pour une quantité constante de protéines. Pour s'assurer du nombre réel de protéines dans les cellules étudiées, nous pourrions utiliser d'autres techniques comme celle utilisée et développée au CBS à Montpellier qui permet d'estimer le nombre de protéines de chaque cellule en observant les variations locales de fluorescence de très nombreux points de la cellule (Ferguson *et al.*, 2011). Des mesures sur les 6C2 transfectées (avec les protéines *mCherry* déstabilisées ou non) ont été effectuées par Ophélie Arnaud à l'aide de cette technique mais les résultats sont encore trop préliminaires pour être intégrés à nos modèles et analyses

3.1.4 Complexification du modèle

Il est bien évidemment toujours possible de supposer que notre modèle n'est pas assez "fidèle" à la réalité et qu'intégrer une ou plusieurs étapes supplémentaires dans la modélisation de la chaine de transcription-traduction permettrait de mieux rendre compte de la réalité biologique ou de nouveaux types de données plus riches en informations comme les données de vidéo-microscopie. Cependant, la complexité d'un modèle n'est en rien un gage de qualité et il est souvent préférable d'utiliser un modèle le plus parcimonieux possible pour expliquer au mieux les données. Néanmoins, il est utile d'identifier le ou les éléments les plus susceptibles de perturber nos conclusions. Nous pouvons ici citer deux éléments importants : d'une part, le modèle proposé au chapitre III considère que tous les événements moléculaires se produisent en un temps négligeable au regard de la durée de vie des molécules incriminées, d'autre part, il suppose que, hormis la stochasticité de l'expression génique, les cellules sont globalement stables dans le temps. Chacun de ces deux éléments est susceptible d'influencer nos résultats, même si nos études préliminaires et bibliographiques nous ont conduit à ne pas les prendre en compte.

Dynamique de la traduction et stochasticité de l'expression Dans le modèle proposé dans notre chapitre III, les étapes de production des protéines sont supposées instantanées. Or, dans la réalité, cela n'est évidemment pas le cas, d'une part du fait que, une fois transcrit, l'ARN messager doit être exporté vers le cytoplasme et d'autre part, parce que les protéines, une fois traduites, peuvent mettre plusieurs heures à atteindre leur configuration spatiale. Pour ce qui est de l'export des ARNm, Singh et Bokes (2012) montrent que le temps de transport des ARNs entre le noyau et le cytoplasme n'influe pas sur la stochasticité protéique, raison pour laquelle nous n'avons pas inclus de mécanisme de retard lié à ce phénomène.
Si l'on ne prend pas en compte la maturation des protéines, une partie du bruit observé, et qui a été ici interprété comme un bruit de transcription, pourraît être dû à des variations de temps de maturation (De Jong *et al.*, 2010). Les inférences faites alors sur la dynamique chromatine, seraient alors biaisées. Estimer ce biais et le compenser permettrait d'avoir une idée plus précise quant aux résultats des inférences (Komorowski *et al.*, 2010). Dans notre étude sur la dynamique chromatinienne, les inférences ne prennent pas en compte ce

type de biais. En effet, nous supposons que le temps de maturation (ou, plus exactement, la distribution des temps de maturation) est identique pour tous les clones. Si cette hypothèse est valide (ce qui est probable compte-tenu de la méthode de production des clones), la contribution de la dynamique de maturation à la stochasticité aura été intégrée à la stochasticité liée à la durée de vie des protéines. En conséquence, l'inférence de la dynamique chromatinienne réalisée au chapitre III ne devrait pas avoir été altérée par ce biais[1].

Cycle cellulaire et stochasticité de l'expression Dans toutes les études réalisées ici nous considérons que l'expression génique n'est pas influencée par le cycle cellulaire ou par la division cellulaire. Ce choix peut paraître surprenant au regard de la littérature puisque Huh et Paulsson (2011a) montrent que la division cellulaire joue un rôle dans la variabilité intercellulaire (en outre, il est connu que le temps de division cellulaire n'est pas identique pour toutes les cellules, ce qui peut aussi contribuer à la stochasticité observée). La prise en compte du cycle cellulaire et/ou de la division cellulaire est difficile, voire impossible, à partir de données de cytométrie de flux et nécessite un suivi des cellules par des approches de cytométries par microscopie ou, plus efficacement, par des approches de vidéo-microscopie. Dans notre cas, la décision d'ignorer la division et le cycle cellulaire a été prise suite à des études préliminaires durant lesquelles des progéniteurs érythropoiaires aviaires transfectés ont été observés pendant plusieurs heures en vidéo-microscopie. Lors de ces études, nous n'avons détecté aucun changement brutal du signal lors de la division cellulaire. De plus, nous n'avons observé aucune corrélation entre la taille de nos cellules et leur fluorescence en les observant par cytométrie en flux. Cela nous a permis de réaliser un modèle simplifié. Cependant, il paraît difficile d'éliminer définitivement un effet de partitionnement du rapporteur lors de la division cellulaire et ce, quel que soit le type de cellule.
Pour confirmer ce choix et mesurer la contribution éventuelle du cycle cellulaire, il serait donc nécessaire d'utiliser à grande échelle des approches de vidéo-microscopie, ce qui pose des problèmes expérimentaux importants[2]. De telles approches permettraient de synchroniser "virtuellement" les cellules et d'analyser la variation de stochasticité en fonction de la proximité de la division cellulaire (en recalant toutes les données d'expression temporelle de manière à ce que les cycles soient synchrones).

3.2 Vers la vidéomicroscopie

Dans toutes les études présentées ici, notre outil de mesure de l'expression génique était basé sur la cytométrie en flux, que ce soit pour des cellules en état "stabilisé" ou sous traitement TSA. Comme énoncé ci-dessus, la cytométrie de flux est fondamentalement limitée par l'impossibilité de mesurer un signal temporel. La vidéomicroscopie "timelapse"

[1]Par ailleurs, comme nous avons utilisé des protéines *mCherry* non déstabilisées – donc avec une très longue demi-vie – il est probable que les biais liés aux délais de maturation des protéines soient négligeables au regard de la durée de vie du rapporteur. Cet argument doit cependant être utilisé avec précaution car de plus en plus d'analyses utilisent des protéines, voire des ARNm, déstabilisés.

[2]Une part du travail réalisé au cours de cette thèse a contribué à la mise en place d'un workflow d'analyse de telles données mais les problèmes techniques rencontrés au niveau expérimental sont tels que nous n'avons pas encore pu l'exploiter efficacement.

permettrait de réintroduire une dimension temporelle fine dans nos analyses et donc de mieux comprendre les effets dynamiques en jeu. En effet, la cytométrie en flux permet d'observer une population à un instant donné alors que le timelapse permet d'étudier le comportement individuel d'une ou plusieurs cellules sur un laps de temps plus ou moins long. Il serait ainsi possible de suivre une cellule pendant des étapes majeures de son cycle de vie : la division, la différenciation, le basculement dans un processus tumoral, ... On pourrait alors étudier spécifiquement la stochasticité de l'expression des gènes au cours de ces différents événements.
Enfin, l'utilisation du timelapse pourrait permettre de répondre à la question de l'ergodicité et de répondre à une question aujourd'hui ouverte : est-ce que tous les niveaux d'expression d'un gène ont la même probabilité d'être observé à la fois à un instant donné dans une population cellulaire et au cours du temps dans une des cellules de cette population ? Moyennant des outils d'analyse précis , on pourrait aussi étudier plus finement les variations de stochasticité dans une telle lignée et étudier l'héritabilité (ou non) de la stochasticité de l'expression des gènes.
Clairement, même si nous pensons avoir ici tiré le meilleur parti possible de données de cytométrie en flux, le passage au timelapse constitue la perspective la plus intéressante ouverte par ce travail.

3.3 Vers l'élucidation des mécanismes de régulation de la stochasticité de l'expression

Aux perspectives "méthologiques" présentées ci-dessus viennent s'ajouter des perspectives "scientifiques". En effet, les résultats obtenus dans ce travail ouvrent au moins deux questions majeures, l'une orientées vers la biologie dite humide, l'autre plus vers la biologie *in silico* :

Biologie humide : à la recherche des "réactants critiques" Nos résultats amènent naturellement à l'hypothèse de la présence d'un réactant critique pour lesquels les gènes proches seraient en compétition. Dans un premier temps, ce mécanisme de compétition de gène à gène demande à être confirmé par des mesures alternatives. En outre, la très forte influence observée des gènes voisins n'a pas pu être expliquée ici. La confirmation de cette hypothèse et, surtout, l'identification de ces réactants critiques (au sens large) permettrait de mieux comprendre cet effet de compétition. Suite à nos études, la seule conclusion que nous puissions tirer est que cette possible compétition est très probablement due à un effet de compétition lié à la diffusion 1D de certaines structures (molécules, assemblages, chromatine, ...) le long de l'ADN. En effet, des études préliminaires en modélisation nous montrent que les effets de compétitions liés à la diffusion 3D sont très rapidement négligeables lorsque les gènes s'éloignent les uns des autres. En outre, une diffusion 3D ne permet pas d'expliquer simplement pourquoi l'orientation relative des gènes influe sur la compétition entre gènes endogènes et gène rapporteur. Cet effet appelle clairement une explication "basée séquence" (au sens de la linéarité du chromosome, pas nécessairement au sens de la séquence nucléotidique elle-même). Nous avons, dans notre équipe, intégré à un même locus (mais dans différents clones), des transgènes en présence ou absence d'insulateur(s). En "bloquant" ainsi la chromatine (ce qui coupe les influences locales

de la chromatine), nous pourrons étudier une partie des causes probables de cet effet. Ces travaux sont actuellement au stade de recueil des données et n'ont donc pas été présentés ici.

Biologie *in silico* : vers une modélisation intégrée Nous avons vu plusieurs causes de stochasticité dans cette thèse et en avons modélisé une petite partie. L'analyse de données biologiques et de simulations nous a permis de mieux comprendre le fonctionnement sous-jacent à la stochasticité observée. Il serait intéressant de continuer cette étude en modélisant à la fois la dynamique chromatinienne, la compétition entre gènes et, plus finement, la dynamique du promoteur. Un tel modèle "intégré" permettrait d'étudier des situations dans lesquels les clones ne differeraient pas seulement en termes de locus d'insertion d'une construction identique pour chaque clone mais aussi (ou alternativement) en termes de promoteur, de perturbateur chromatinien ou de facteurs de transcription. Pour ce faire, l'outil Promdyn développé par Antoine Coulon (Coulon, 2010; Coulon *et al.*, 2010, 2013), combiné à des expériences "humides", pourrait nous permettre d'affiner nos hypothèses. Dans un tel modèle, l'état actif du gène ne correspondrait pas "simplement" au fait que la chromatine soit ouverte mais aussi que l'ensemble de la machinerie transcriptionnelle soit bien présente sur le promoteur. Dans un second temps, la prise en compte d'effets spatiaux 3D et/ou 1D dans le modèle pourrait nous permettre de mieux comprendre les effets de compétition liés à la distance intergénique. Le développement de modèles intégrés permettrait de palier les limites actuelles des techniques expérimentales. En effet, il est difficile de suivre en même temps différentes molécules dans le noyau cellulaire dans une zone proche d'un locus donné, tout comme il est difficile de suivre l'état de la chromatine au cours du temps à un locus donné.

4 Evolution génomique et stochasticité

Les cellules telles que nous les connaissons aujourd'hui sont l'aboutissement d'un processus d'évolution constitué d'un mécanisme de variation et d'un mécanisme de sélection. Cette dernière porte sur le phénotype et ce sont donc les variations héritables de ce dernier qui sont susceptibles d'être sélectionnées. Or, le phénotype n'est pas seulement lié à la présence d'un ensemble de gènes dans le génome mais aussi à la façon dont ceux-ci sont exprimés, y compris dans la composante stochastique de cette expression (Wang et Zhang, 2011).

Ainsi, contrairement à ce qui est souvent implicitement admis, un réseau de gènes donné, censé remplir une fonction métabolique donnée, ne peut pas être uniquement décrit par la moyenne d'expression des gènes du réseau : les variations autour de cette moyenne doivent être prises en compte pour comprendre totalement le fonctionnement du réseau (Çağatay *et al.*, 2009). En effet, la stochasticité a aussi son rôle à jouer dans le phénotype et/ou dans les transitions d'état du réseau. Il est donc plus que probable que le niveau de stochasticité de l'expression des gènes ait été finement sélectionné au cours de la phylogénèse des organismes et ce en lien avec la fonction métabolique du réseau elle-même, donc avec son environnement.

Ainsi, au moins dans certaines situations simples où l'environnement varie aléatoirement entre deux états, il est connu qu'un niveau optimum de stochasticité favorise la sur-

vie d'une espèce aux dépends de la survie de tout les individus de cette espèce (voir la section 2.3 de l'introduction pour le phénomène de bet-hedging). Cette sélection de la stochasticité a été montrée théoriquement (Kussell et Leibler, 2005; Ito *et al.*, 2009) et expérimentalement (Beaumont *et al.*, 2009; Veening *et al.*, 2008). Beaumont *et al.* (2009) en particulier montrent par une approche d'évolution expérimentale *in vitro* que la stochasticité peut effectivement être sélectionnée dans ce contexte. Il a par ailleurs été montré que la stochasticité pouvait être sélectionnée à l'échelle des gènes. Ainsi, chez la levure, de nombreux gènes essentiels à la croissance cellulaire sont regroupés dans les zones les moins "bruyantes" du génome (Batada et Hurst, 2007).

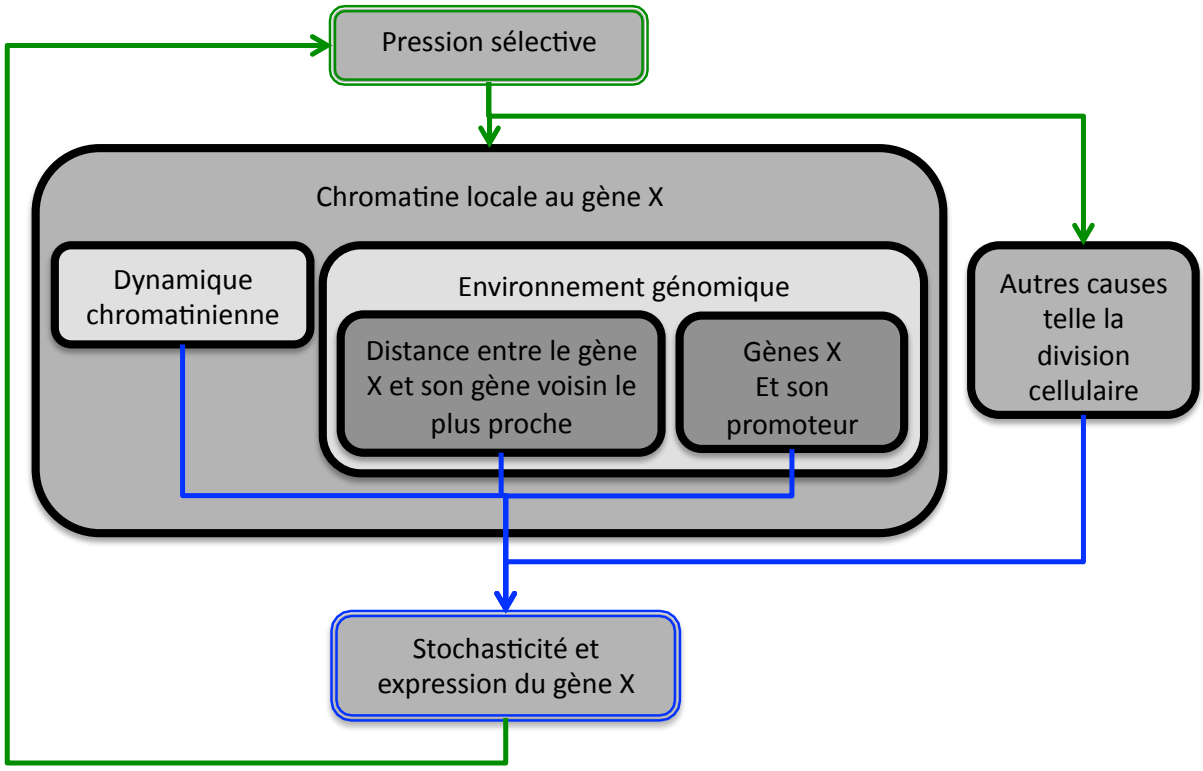

Figure V.1 – Chromatine locale, stochasticité et évolution. Cette figure repose sur les notions introduites par la section 4.2 du chapitre IV. La sélection naturelle des phénotypes les plus adaptés impacte le génome et la dynamique chromatinienne qui impactent en retour sur l'expression et la stochasticité de chaque gène. Ces dernières caractérisent les phénotypes qui vont alors à nouveau être ou non sélectionnés.

La possibilité d'une selection gène à gène du niveau de stochasticité résonne particulièrement avec nos résultats. En effet, nous avons montré que la stochasticité de l'expression d'un gène pouvait être régulée de plusieurs façons différentes, toutes interragissant les unes avec les autres (dynamique chromatinienne locale et/ou proximité d'autres gènes, ... voir figure V.1). Mais nous avons aussi montré qu'elle pouvait être modulée indépendamment de la moyenne d'expression de ce même gène. Même si l'on peut raisonnablement admettre que la moyenne d'expression est, au moins dans le cas général, le déterminant premier de la sélection, la possibilité de découpler ces deux facteurs permet bien d'envisager la

sélection spécifique du niveau de stochasticité de l'expression.
Cette possibilité ouvre de nouvelles perspectives de recherche : si la stochasticité de l'expression des gènes peut évoluer et si on constate, chez les organismes contemporains, un niveau non négligeable de stochasticité, au moins pour certains gènes, alors le rôle de cette stochasticité doit pouvoir être étudié en termes évolutifs (qu'il s'agisse d'un rôle sélectif, de dérive ou de sélection indirecte). Une telle approche amène de nombreuses questions : quelle est la dynamique évolutive de la stochasticité de l'expression des gènes ? (a-t-elle augmenté ou diminué au cours de l'évolution des espèces ?), comment la stochasticité a-t-elle contribué (ou non) aux transitions évolutives telles que le passage de la vie mono-cellulaire à la vie multi-cellulaire ? ... Réciproquement, l'influence de la stochasticité de l'expression des gènes *sur* l'évolution (via la mise en place de mécanismes de variation plus ou moins héritables) apporte un cortège de questions que nous ne pouvons probablement qu'ébaucher aujourd'hui.

Bibliographie

Ackermann, M., Stecher, B., Freed, N. E., Songhet, P., Hardt, W.-D. et Doebeli, M. (2008). Self-destructive cooperation mediated by phenotypic noise. *Nature*, 454(7207):987–990.

Aitken, C. E., Petrov, A. et Puglisi, J. D. (2010). Single ribosome dynamics and the mechanism of translation. *Annual Review of Biophysics*, 39:491–513.

Alon, U. (2007). Network motifs : theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461.

Ansel, J., Bottin, H., Rodriguez-Beltran, C., Damon, C., Nagarajan, M., Fehrmann, S., François, J. et Yvert, G. (2008). Cell-to-Cell Stochastic Variation in Gene Expression Is a Complex Genetic Trait. *PLoS Genetics*, 4(4):e1000049.

Arias, A. M. et Hayward, P. (2006). Filtering transcriptional noise during development : concepts and mechanisms. *Nature Reviews Genetics*, 7(1):34–44.

Bahar, R., Hartmann, C. H., Rodriguez, K. A., Denny, A. D., Busuttil, R. A., Dollé, M. E. T., Calder, R. B., Chisholm, G. B., Pollock, B. H., Klein, C. A. et Vijg, J. (2006). Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature*, 441(7096):1011–1014.

Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L. et Leibler, S. (2004). Bacterial persistence as a phenotypic switch. *Science*, 305(5690):1622–1625.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'shea, E., Pilpel, Y. et Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. *Nature Genetics*, 38(6):636–643.

Batada, N. N. et Hurst, L. D. (2007). Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nature Genetics*, 39(8):945–949.

Batenchuk, C., St-Pierre, S., Tepliakova, L., Adiga, S., Szuto, A., Kabbani, N., Bell, J. C., Baetz, K. et Kærn, M. (2011). Chromosomal Position Effects Are Linked to Sir2-Mediated Variation in Transcriptional Burst Size. *Biophysj*, 100(10):L56–L58.

Beaumont, H. J. E., Gallie, J., Kost, C., Ferguson, G. C. et Rainey, P. B. (2009). Experimental evolution of bet hedging. *Nature*, 461(7269):90–93.

Becskei, A., Kaufmann, B. B. et Van Oudenaarden, A. (2005). Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nature Genetics*, 37(9):937–944.

Berg, O. G. (1978). A model for the statistical fluctuations of protein numbers in a microbial population. *Journal of Theoretical Biology*, 71(4):587–603.

Bernaschi, M., Castiglione, F., Ferranti, A., Gavrila, C., Tinti, M. et Cesareni, G. (2007). ProtNet : a tool for stochastic simulations of protein interaction networks dynamics. *BMC Bioinformatics*, 8(Suppl 1):S4.

Bertin, E. (2012). How far can stochastic and deterministic views be reconciled ? *Progress in Biophysics and Molecular Biology*, 110(1):11–16.

Beug, H., Doederlein, G., Freudenstein, C. et Graf, T. (1982). Erythroblast cell lines transformed by a temperature-sensitive mutant of avian erythroblastosis virus : a model system to study erythroid differentiation in vitro. *Journal of cellular physiology. Supplement*, 1:195–207.

Beug, H., von Kirchbach, A., Döderlein, G., Conscience, J. F. et Graf, T. (1979). Chicken hematopoietic cells transformed by seven strains of defective avian leukemia viruses display three distinct phenotypes of differentiation. *Cell*, 18(2):375–390.

Bigger, J. (1944). Treatment of staphylococcal infections with penicillin by intermittent sterilisation. *The Lancet*, 244(6320):497–500.

Black, J. C., Van Rechem, C. et Whetstine, J. R. (2012). Histone lysine methylation dynamics : establishment, regulation, and biological impact. *Molecular Cell*, 48(4):491–507.

Blake, W. J., Balázsi, G., Kohanski, M. A., Isaacs, F. J., Murphy, K. F., Kuang, Y., Cantor, C. R., Walt, D. R. et Collins, J. J. (2006). Phenotypic Consequences of Promoter-Mediated Transcriptional Noise. *Molecular Cell*, 24(6):853–865.

Boeger, H., Griesenbeck, J. et Kornberg, R. D. (2008). Nucleosome Retention and the Stochastic Nature of Promoter Chromatin Remodeling for Transcription. *Cell*, 133(4):716–726.

Boettiger, A. N. et Levine, M. (2009). Synchronous and stochastic patterns of gene activation in the Drosophila embryo. *Science*, 325(5939):471–473.

Boutanaev, A. M., Kalmykova, A. I., Shevelyov, Y. Y. et Nurminsky, D. I. (2002). Large clusters of co-expressed genes in the Drosophila genome. *Nature*, 420(6916):666–669.

Brock, A., Chang, H. et Huang, S. (2009). Non-genetic heterogeneity–a mutation-independent driving force for the somatic evolution of tumours. *Nature Reviews Genetics*, 10(5):336–342.

Caffarelli, E. et Filetici, P. (2011). Epigenetic regulation in cancer development. *Frontiers in bioscience : a journal and virtual library*, 16:2682–2694.

Çağatay, T., Turcotte, M., Elowitz, M. B., Garcia-Ojalvo, J. et Süel, G. M. (2009). Architecture-dependent noise discriminates functionally analogous differentiation circuits. *Cell*, 139(3):512–522.

Cai, L., Dalal, C. K. et Elowitz, M. B. (2008). Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature*, 455(7212):485–490.

Cairns, B. R. (2009). The logic of chromatin architecture and remodelling at promoters. *Nature*, 461(7261):193–198.

Capp, J.-P. (2005). Stochastic gene expression, disruption of tissue averaging effects and cancer as a disease of development. *BioEssays*, 27(12):1277–1285.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M. C., van Asperen, R., Boon, K., Voûte, P. A., Heisterkamp, S., van Kampen, A. et Versteeg, R. (2001). The human transcriptome map : clustering of highly expressed genes in chromosomal domains. *Science*, 291(5507):1289–1292.

Chalancon, G., Ravarani, C. N. J., Balaji, S., Martinez-Arias, A., Aravind, L., Jothi, R. et Babu, M. M. (2012). Interplay between gene expression noise and regulatory network architecture. *Trends in genetics*, 28(5):221–232.

Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. et Huang, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194):544–547.

Choi, J. K. et Kim, Y.-J. (2009). Intrinsic variability of gene expression encoded in nucleosome positioning sequences. *Nature Genetics*, 41(4):498–503.

Christman, J. K. (2002). 5-Azacytidine and 5-aza-2'-deoxycytidine as inhibitors of DNA methylation : mechanistic studies and their implications for cancer therapy. *Oncogene*, 21(35):5483–5495.

Chubb, J. R. et Liverpool, T. B. (2010). Bursts and pulses : insights from single cell studies into transcriptional mechanisms. *Current Opinion in Genetics & Development*, 20(5):478–484.

Chubb, J. R., Trcek, T., Shenoy, S. M. et Singer, R. H. (2006). Transcriptional Pulsing of a Developmental Gene. *Current Biology*, 16(10):1018–1025.

Claverie, J. M. (2001). Transcriptome analysis in cancerology : bioinformatics aspects. *Bulletin du cancer*, 88(3):269–276.

Cohen, A. A., Geva-Zatorsky, N., Eden, E., Frenkel-Morgenstern, M., Issaeva, I., Sigal, A., Milo, R., Cohen-Saidon, C., Liron, Y., Kam, Z., Cohen, L., Danon, T., Perzov, N. et Alon, U. (2008). Dynamic proteomics of individual cancer cells in response to a drug. *Science*, 322(5907):1511–1516.

Coppey, M., Bénichou, O., Voituriez, R. et Moreau, M. (2004). Kinetics of target site localization of a protein on DNA : a stochastic approach. *Biophysj*, 87(3):1640–1649.

Corre, G. (2012). *Hétérogénéité phénotypique dans les populations d'origine clonale : Origine et conséquences*. Thèse de doctorat, Généthon.

Coulon, A. (2010). *Stochasticité de l'expression génique et régulation transcriptionnelle : Modélisation de la dynamique spatiale et temporelle des structures multiprotéiques*. Thèse de doctorat, INSA de Lyon.

Coulon, A., Chow, C. C., Singer, H. R. et Larson, R. D. (2013). Eukaryotic transcriptional dynamics : from single molecules to cell populations. *Nature Reviews Genetics*. Epub ahead of print.

Coulon, A., Gandrillon, O. et Beslon, G. (2010). On the spontaneous stochastic dynamics of a single gene : complexity of the molecular interplay at the promoter. *BMC Systems Biology*, 4:2.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Crick, F. et Watson, J. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356): 737–738.

Dar, R. D., Razooky, B. S., Singh, A., Trimeloni, T. V., McCollum, J. M., Cox, C. D., Simpson, M. L. et Weinberger, L. S. (2012). Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences*, 109(43):17454–17459.

Darzacq, X., Shav-Tal, Y., De Turris, V., Brody, Y., Shenoy, S. M., Phair, R. D. et Singer, R. H. (2007). In vivo dynamics of RNA polymerase II transcription. *Nature Structural & Molecular Biology*, 14(9):796–806.

Das Neves, R. P., Jones, N. S., Andreu, L., Gupta, R., Enver, T. et Iborra, F. J. (2010). Connecting variability in global transcription rate to mitochondrial variability. *PLoS Biology*, 8(12):e1000560.

Davidson, E. H. (2002). A Genomic Regulatory Network for Development. *Science*, 295(5560):1669–1678.

De Jong, H., Ranquet, C., Ropers, D., Pinel, C. et Geiselmann, J. (2010). Experimental and computational validation of models of fluorescent and luminescent reporter genes in bacteria. *BMC Systems Biology*, 4:55.

De Krom, M., van de Corput, M., von Lindern, M., Grosveld, F. et Strouboulis, J. (2002). Stochastic patterns in globin gene expression are established prior to transcriptional activation and are clonally inherited. *Molecular Cell*, 9(6):1319–1326.

Devon, R. S., Porteous, D. J. et Brookes, A. J. (1995). Splinkerettes–improved vectorettes for greater efficiency in PCR walking. *Nucleic Acids Research*, 23(9):1644–1645.

Di Talia, S., Skotheim, J. M., Bean, J. M., Siggia, E. D. et Cross, F. R. (2007). The effects of molecular noise and size control on variability in the budding yeast cell cycle. *Nature*, 448(7156):947–951.

Discher, D. E., Mooney, D. J. et Zandstra, P. W. (2009). Growth factors, matrices, and forces combine and control stem cells. *Science*, 324(5935):1673–1677.

Dobrzynski, M. et Bruggeman, F. J. (2009). Elongation dynamics shape bursty transcription and translation. *Proceedings of the National Academy of Sciences*, 106(8):2583–2588.

Ebisuya, M., Yamamoto, T., Nakajima, M. et Nishida, E. (2008). Ripples from neighbouring transcription. *Nature Cell Biology*, 10(9):1106–1113.

Eldar, A., Chary, V. K., Xenopoulos, P., Fontes, M. E., Loson, O. C., Dworkin, J., Piggot, P. J. et Elowitz, M. B. (2009). Partial penetrance facilitates developmental evolution in bacteria. *Nature*, 460(7254):510–514.

Eldar, A. et Elowitz, M. B. (2010). Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–173.

Elowitz, M. B., Levine, A. J., Siggia, E. D. et Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186.

Feng, Y. Q., Lorincz, M. C., Fiering, S., Greally, J. M. et Bouhassira, E. E. (2001). Position Effects Are Influenced by the Orientation of a Transgene with Respect to Flanking Chromatin. *Molecular and Cellular Biology*, 21(1):298–309.

Ferguson, M. L., Le Coq, D., Jules, M., Aymerich, S., Declerck, N. et Royer, C. A. (2011). Absolute quantification of gene expression in individual bacterial cells using two-photon fluctuation microscopy. *Analytical Biochemistry*, 419(2):250–259.

Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I. K., Sharon, E., Lubling, Y., Widom, J. et Segal, E. (2008). Distinct Modes of Regulation by Chromatin Encoded through Nucleosome Positioning Signals. *PLoS Computational Biology*, 4(11):e1000216.

Figueirêdo, P. H., Moret, M. A., Coutinho, S. et Nogueira, E. (2010). The role of stochasticity on compactness of the native state of protein peptide backbone. *Journal of Chemical Physics*, 133(8):085102.

Franceschelli, S. (2012). Some remarks on the compatibility between determinism and unpredictability. *Progress in Biophysics and Molecular Biology*, 110(1):61–68.

Freed, N. E., Silander, O. K., Stecher, B., Böhm, A., Hardt, W.-D. et Ackermann, M. (2008). A Simple Screen to Identify Promoters Conferring High Levels of Phenotypic Noise. *PLoS Genetics*, 4(12):e1000307.

Friedman, N., Linial, M. et Nachman, I. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620.

Gandrillon, O., Schmidt, U., Beug, H. et Samarut, J. (1999). TGF-beta cooperates with TGF-alpha to induce the self-renewal of normal erythrocytic progenitors : evidence for an autocrine mechanism. *The EMBO Journal*, 18(10):2764–2781.

Garai, A., Chowdhury, D. et Ramakrishnan, T. V. (2009). Fluctuations in protein synthesis from a single RNA template : stochastic kinetics of ribosomes. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 79(1 Pt 1):011916.

Gascoigne, K. E. et Taylor, S. S. (2008). Cancer Cells Display Profound Intra- and Interline Variation following Prolonged Exposure to Antimitotic Drugs. *Cancer Cell*, 14(2):111–122.

Gaszner, M. et Felsenfeld, G. (2006). Insulators : exploiting transcriptional and epigenetic mechanisms. *Nature Reviews Genetics*, 7(9):703–713.

Ghoshal, K., Datta, J., Majumder, S., Bai, S., Dong, X., Parthun, M. et Jacob, S. T. (2002). Inhibitors of Histone Deacetylase and DNA Methyltransferase Synergistically Activate the Methylated Metallothionein I Promoter by Activating the Transcription Factor MTF-1 and Forming an Open Chromatin Structure. *Molecular and Cellular Biology*, 22(23):8302–8319.

Gierman, H. J., Indemans, M. H. G., Koster, J., Goetze, S., Seppen, J., Geerts, D., Van Driel, R. et Versteeg, R. (2007). Domain-wide regulation of gene expression in the human genome. *Genome Research*, 17(9):1286–1295.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361.

Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115(4):1716–1732.

Golding, I., Paulsson, J., Zawilski, S. M. et Cox, E. C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036.

Grassi, G., Maccaroni, P., Meyer, R., Kaiser, H., D'Ambrosio, E., Pascale, E., Grassi, M., Kuhn, A., Di Nardo, P., Kandolf, R. et Küpper, J.-H. (2003). Inhibitors of DNA methylation and histone deacetylation activate cytomegalovirus promoter-controlled reporter gene expression in human glioblastoma cell line U87. *Carcinogenesis*, 24(10):1625–1635.

Green, M. R. (2005). Eukaryotic Transcription Activation : Right on Target. *Molecular Cell*, 18(4):399–402.

Guérin, T., Bénichou, O. et Voituriez, R. (2013). Reactive conformations and non-Markovian cyclization kinetics of a Rouse polymer. *Journal of Chemical Physics*, 138(9): 094908.

Haaf, T. et Schmid, M. (2000). Experimental condensation inhibition in constitutive and facultative heterochromatin of mammalian chromosomes. *Cytogenetics and cell genetics*, 91(1-4):113–123.

Halley, J. D., Winkler, D. A. et Burden, F. R. (2008). Toward a Rosetta stone for the stem cell genome : Stochastic gene expression, network architecture, and external influences. *Stem Cell Research*, 1(3):157–168.

Harper, C. V., Finkenstädt, B., Woodcock, D. J., Friedrichsen, S., Semprini, S., Ashall, L., Spiller, D. G., Mullins, J. J., Rand, D. A. et Davis, J. R. (2011). Dynamic analysis of stochastic transcription cycles. *PLoS Biology*, 9(4):e1000607.

Hilfinger, A. et Paulsson, J. (2011). Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 108(29):12167–12172.

Hoek, K. S., Eichhoff, O. M., Schlegel, N. C., Döbbeling, U., Kobert, N., Schaerer, L., Hemmi, S. et Dummer, R. (2008). In vivo switching of human melanoma cells between proliferative and invasive states. *Cancer research*, 68(3):650–656.

Hoffmann, M., Chang, H. H., Huang, S., Ingber, D. E., Loeffler, M. et Galle, J. (2008). Noise-Driven Stem Cell and Progenitor Population Dynamics. *PLoS ONE*, 3(8):e2922.

Huang, S. (2010). Cell lineage determination in state space : a systems view brings flexibility to dogmatic canonical rules. *PLoS Biology*, 8(5):e1000380.

Huang, X., Guo, H., Tammana, S., Jung, Y.-C., Mellgren, E., Bassi, P., Cao, Q., Tu, Z. J., Kim, Y. C., Ekker, S. C., Wu, X., Wang, S. M. et Zhou, X. (2010). Gene transfer efficiency and genome-wide integration profiling of Sleeping Beauty, Tol2, and piggyBac transposons in human primary T cells. *Molecular Therapy*, 18(10):1803–1813.

Huh, D. et Paulsson, J. (2011a). Non-genetic heterogeneity from stochastic partitioning at cell division. *Nature Genetics*, 43(2):95–100.

Huh, D. et Paulsson, J. (2011b). Random partitioning of molecules at cell division. *Proceedings of the National Academy of Sciences*, 108(36):15004–15009.

Hume, D. A. (2000). Probability in transcriptional regulation and its implications for leukocyte differentiation and inducible gene expression. *Blood*, 96(7):2323–2328.

Ito, Y., Toyota, H., Kaneko, K. et Yomo, T. (2009). How selection affects phenotypic fluctuation. *Molecular Systems Biology*, 5(264):1–7.

Johnston, I. G., Gaal, B., Neves, R. P. d., Enver, T., Iborra, F. J. et Jones, N. S. (2012). Mitochondrial variability as a source of extrinsic cellular noise. *PLoS Computational Biology*, 8(3):e1002416.

Kærn, M., Elston, T. C., Blake, W. J. et Collins, J. J. (2005). Stochasticity in gene expression : from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464.

Kalmar, T., Lim, C., Hayward, P., Muñoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J. et Martinez-Arias, A. (2009). Regulated Fluctuations in Nanog Expression Mediate Cell Fate Decisions in Embryonic Stem Cells. *PLoS Biology*, 7(7):e1000149.

Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., LeProust, E. M., Hughes, T. R., Lieb, J. D., Widom, J. et Segal, E. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236):362–366.

Kaufmann, B. B. et Van Oudenaarden, A. (2007). Stochastic gene expression : from single molecules to the proteome. *Current Opinion in Genetics & Development*, 17(2):107–112.

Kawakami, K. (2007). Tol2 : a versatile gene transfer vector in vertebrates. *Genome biology*, 8(Suppl 1):S7.

Kawakami, K. et Noda, T. (2004). Transposition of the Tol2 element, an Ac-like element from the Japanese medaka fish Oryzias latipes, in mouse embryonic stem cells. *Genetics*, 166(2):895–899.

Kelemen, J. Z., Ratna, P., Scherrer, S. et Becskei, A. (2010). Spatial epigenetic control of mono-and bistable gene expression. *PLoS Biology*, 8(3):e1000332.

Kepler, T. B. et Elston, T. C. (2001). Stochasticity in transcriptional regulation : origins, consequences, and mathematical representations. *Biophysj*, 81(6):3116–3136.

Kittisopikul, M. et Süel, G. M. (2010). Biological role of noise encoded in a genetic network motif. *Proceedings of the National Academy of Sciences of the United States of America*, 107(30):13300–13305.

Knibbe, C., Coulon, A., Mazet, O., Fayard, J. M. et Beslon, G. (2007). A Long-Term Evolutionary Pressure on the Amount of Noncoding DNA. *Molecular biology and evolution*, 24(10):2344–2353.

Komorowski, M., Finkenstädt, B. et Rand, D. (2010). Using a single fluorescent reporter gene to infer half-life of extrinsic noise and other parameters of gene expression. *Biophysical Journal*, 98(12):2759–2769.

Kondrychyn, I., Garcia-Lecea, M., Emelyanov, A., Parinov, S. et Korzh, V. (2009). Genome-wide analysis of Tol2 transposon reintegration in zebrafish. *BMC Genomics*, 10:418.

Kong, J., Zhu, F., Stalker, J. et Adams, D. J. (2008). iMapper : a web application for the automated analysis and mapping of insertional mutagenesis sequence data against Ensembl genomes. *Bioinformatics*, 24(24):2923–2925.

Kumar, R., Kuniyasu, H., Bucana, C. D., Wilson, M. R. et Fidler, I. J. (1998). Spatial and temporal expression of angiogenic molecules during tumor growth and progression. *Oncology research*, 10(6):301–311.

Kupiec, J. J. (1983). A probabilist theory for cell differentiation, embryonic mortality and DNA C-value paradox. *Specul. Sci. Technol*, 6:471–478.

Kupiec, J. J. (1997). A Darwinian theory for the origin of cellular differentiation. *Molecular & general genetics : MGG*, 255(2):201–208.

Kupiec, J.-J. (2009). *The origin of individuals*. ISBN-13 : 978-981-270-499-3.

Kupiec, J.-J. (2010). On the lack of specificity of proteins and its consequences for a theory of biological organization. *Progress in Biophysics and Molecular Biology*, 102(1):45–52.

Kussell, E. et Leibler, S. (2005). Phenotypic Diversity, Population Growth, and Information in Fluctuating Environments. *Science*, 309(5743):2075–2078.

Laforge, B., Guez, D., Martinez, M. et Kupiec, J.-J. (2005). Modeling embryogenesis and cancer : an approach based on an equilibrium between the autostabilization of stochastic gene expression and the interdependence of cells for proliferation. *Progress in Biophysics and Molecular Biology*, 89(1):93–120.

Lai, Y. et Sun, F. (2003). The relationship between microsatellite slippage mutation rate and the number of repeat units. *Molecular biology and evolution*, 20(12):2123–2131.

Larson, D. R., Singer, R. H. et Zenklusen, D. (2009). A single molecule view of gene expression. *Trends in cell biology*, 19(11):630–637.

Lavelle, C. (2009). Forces and torques in the nucleus : chromatin under mechanical constraints. *Biochemistry and cell biology*, 87(1):307–322.

Le Bellac, M. (2012). The role of probabilities in physics. *Progress in Biophysics and Molecular Biology*, 110(1):97–105.

Levsky, J. M., Shenoy, S. M., Pezo, R. C. et Singer, R. H. (2002). Single-cell gene expression profiling. *Science*, 297(5582):836–840.

Levsky, J. M. et Singer, R. H. (2003). Gene expression and the myth of the average cell. *Trends in cell biology*, 13(1):4–6.

Li, B., Carey, M. et Workman, J. L. (2007). The role of chromatin during transcription. *Cell*, 128(4):707–719.

Lim, W. A., Lee, C. M. et Tang, C. (2013). Design principles of regulatory networks : searching for the molecular algorithms of the cell. *Molecular Cell*, 49(2):202–212.

Locke, J. C. W., Young, J. W., Fontes, M., Hernández Jiménez, M. J. et Elowitz, M. B. (2011). Stochastic Pulse Regulation in Bacterial Stress Response. *Science (New York, NY)*, 334(6054):366–369.

Maamar, H., Raj, A. et Dubnau, D. (2007). Noise in gene expression determines cell fate in Bacillus subtilis. *Science*, 317(5837):526–529.

Mayburd, A. L. (2009). Expression variation : its relevance to emergence of chronic disease and to therapy. *PLoS ONE*, 4(6):e5921.

McCullagh, E., Farlow, J., Fuller, C. et Girard, J. (2009). Not all quiet on the noise front. *Nature chemical Biology*, 5:699–704.

McGarvey, K. M. (2006). Silenced Tumor Suppressor Genes Reactivated by DNA Demethylation Do Not Return to a Fully Euchromatic Chromatin State. *Cancer research*, 66(7):3541–3549.

Mejia-Pous, C., Viñuelas, J., Faure, C., Koszela, J., Kawakami, K., Takahashi, Y. et Gandrillon, O. (2009). A combination of transposable elements and magnetic cell sorting provides a very efficient transgenesis system for chicken primary erythroid progenitors. *BMC Biotechnology*, 9(1):81.

Mellor, J. (2006). Dynamic nucleosomes and gene transcription. *Trends in genetics*, 22(6):320–329.

Mettetal, J. T. et Van Oudenaarden, A. (2007). Microbiology. Necessary noise. *Science*, 317(5837):463–464.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31–46.

Miller-Jensen, K., Dey, S. S., Schaffer, D. V. et Arkin, A. P. (2011). Varying virulence : epigenetic control of expression noise and disease processes. *Trends in Biotechnology*, 29(10):517–525.

Milo, R., Jorgensen, P., Moran, U., Weber, G. et Springer, M. (2009). BioNumbers–the database of key numbers in molecular and cell biology. *Nucleic Acids Research*, 38(Database):D750–D753.

Moxon, E. R., Rainey, P. B., Nowak, M. A. et Lenski, R. E. (1994). Adaptive evolution of highly mutable loci in pathogenic bacteria. *Current biology*, 4(1):24–33.

Moyed, H. S. et Broderick, S. H. (1986). Molecular cloning and expression of hipA, a gene of Escherichia coli K-12 that affects frequency of persistence after inhibition of murein synthesis. *Journal of bacteriology*, 166(2):399–403.

Munsky, B., Hernday, A., Low, D. et Khammash, M. (2005). Stochastic modeling of the pap-pili epigenetic switch. *Proc FOSBE*, pages 145–148.

Munsky, B., Neuert, G. et Van Oudenaarden, A. (2012). Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187.

Munsky, B., Trinh, B. et Khammash, M. (2009). Listening to the noise : random fluctuations reveal gene network parameters. *Molecular Systems Biology*, 5:318.

Nachman, I., Regev, A. et Ramanathan, S. (2007). Dissecting timing variability in yeast meiosis. *Cell*, 131(3):544–556.

Neildez-Nguyen, T. M. A., Parisot, A., Vignal, C., Rameau, P., Stockholm, D., Picot, J., Allo, V., Le Bec, C., Laplace, C. et Paldi, A. (2008). Epigenetic gene expression noise and phenotypic diversification of clonal cell populations. *Differentiation*, 76(1):33–40.

Newman, J. R. S., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L. et Weissman, J. S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature Cell Biology*, 441(7095):840–846.

Nie, H., Crooijmans, R. P. M. A., Bastiaansen, J. W. M., Megens, H.-J. et Groenen, M. A. M. (2010). Regional regulation of transcription in the chicken genome. *BMC Genomics*, 11:28.

Niepel, M., Spencer, S. L. et Sorger, P. K. (2009). Non-genetic cell-to-cell variability and the consequences for pharmacology. *Current Opinion in Chemical Biology*, 13(5-6):556–561.

Noble, D. (2007). Claude Bernard, the first systems biologist, and the future of physiology. *Experimental Physiology*, 93(1):16–26.

Novick, A. et Weiner, M. (1957). Enzyme induction as an all-or-none phenomenon. *Proceedings of the National Academy of Sciences of the United States of America*, 43(7): 553–566.

Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. et Van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. *Nature Genetics*, 31(1):69–73.

Paldi, A. (2003). Stochastic gene expression during cell differentiation : order from disorder ? *Cellular and Molecular Life Sciences*, 60(9):1775–1778.

Paldi, A. (2012). What makes the cell differentiate ? *Progress in Biophysics and Molecular Biology*, 110(1):41–43.

Paulsson, J. (2004). Summing up the noise in gene networks. *Nature*, 427:415–418.

Paulsson, J. (2005a). Models of stochastic gene expression. *Physics of life reviews*, 2(2):157–175.

Paulsson, J. (2005b). Prime movers of noisy gene expression. *Nature Genetics*, 37(9): 925–926.

Pedraza, J. M. et Paulsson, J. (2008). Effects of Molecular Memory and Bursting on Fluctuations in Gene Expression. *Science*, 319(5861):339–343.

Pikaart, M. J., Recillas-Targa, F. et Felsenfeld, G. (1998). Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators. *Genes & Development*, 12(18):2852–2862.

Ponting, C. P. et Lunter, G. (2006). Signatures of adaptive evolution within human non-coding sequence. *Human molecular genetics*, 15(Spec No 2):R170–R175.

Querido, E. et Chartrand, P. (2008). Using fluorescent proteins to study mRNA trafficking in living cells. *Methods in cell biology*, 85:273–292.

R Development Core Team (2010). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. et Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):e309.

Raj, A. et Van Oudenaarden, A. (2008). Nature, nurture, or chance : stochastic gene expression and its consequences. *Cell*, 135(2):216–226.

Rao, C. V., Wolf, D. M. et Arkin, A. P. (2002). Control, exploitation and tolerance of intracellular noise. *Nature*, 420(6912):231–237.

Raser, J. M. (2004). Control of Stochasticity in Eukaryotic Gene Expression. *Science*, 304(5678):1811–1814.

Raser, J. M. et O'Shea, E. K. (2005). Noise in gene expression : origins, consequences, and control. *Science*, 309(5743):2010–2013.

Rigney, D. R. et Schieve, W. C. (1977). Stochastic model of linear, continuous protein synthesis in bacterial populations. *Journal of Theoretical Biology*, 69(4):761–766.

Roberts, E., Magis, A., Ortiz, J. O., Baumeister, W. et Luthey-Schulten, Z. (2011). Noise contributions in an inducible genetic switch : a whole-cell simulation study. *PLoS Computational Biology*, 7(3):e1002010.

Roesch, A., Fukunaga-Kalabis, M., Schmidt, E. C., Zabierowski, S. E., Brafford, P. A., Vultur, A., Basu, D., Gimotty, P., Vogt, T. et Herlyn, M. (2010). A temporarily distinct subpopulation of slow-cycling melanoma cells is required for continuous tumor growth. *Cell*, 141(4):583–594.

Rosenfeld, N., Perkins, T. J., Alon, U., Elowitz, M. B. et Swain, P. S. (2006). A fluctuation method to quantify in vivo fluorescence data. *Biophysj*, 91(2):759–766.

Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. et Elowitz, M. B. (2005). Gene regulation at the single-cell level. *Science*, 307(5717):1962–1965.

Rossant, J. et Tam, P. P. L. (2009). Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse. *Development*, 136(5):701–713.

Rudek, M. A., Zhao, M., He, P., Hartke, C., Gilbert, J., Gore, S. D., Carducci, M. A. et Baker, S. D. (2005). Pharmacokinetics of 5-azacitidine administered with phenylbutyrate in patients with refractory solid tumors or hematologic malignancies. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 23(17):3906–3911.

Rué, P., Domedel-Puig, N., Garcia-Ojalvo, J. et Pons, A. J. (2012). Integration of cellular signals in chattering environments. *Progress in Biophysics and Molecular Biology*, 110(1):106–112.

Samoilov, M., Plyasunov, S. et Arkin, A. P. (2005). Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2310–2315.

Satoh, T., Yamamoto, K., Miura, K. F. et Sofuni, T. (2004). Region-specific chromatin decondensation and micronucleus formation induced by 5-azacytidine in human TIG-7 cells. *Cytogenetic and genome research*, 104(1-4):289–294.

Schwabe, A., Rybakova, K. N. et Bruggeman, F. J. (2012). Transcription stochasticity of complex gene regulation models. *Biophysical Journal*, 103(6):1152–1161.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. et Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342.

Shahrezaei, V., Ollivier, J. F. et Swain, P. S. (2008). Colored extrinsic fluctuations and stochastic gene expression. *Molecular Systems Biology*, 4:196.

Sharma, S. V., Lee, D. Y., Li, B., Quinlan, M. P., Takahashi, F., Maheswaran, S., Mcdermott, U., Azizian, N., Zou, L., Fischbach, M. A., Wong, K.-K., Brandstetter, K., Wittner, B., Ramaswamy, S., Classon, M. et Settleman, J. (2010). A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell*, 141(1):69–80.

Singh, A. et Bokes, P. (2012). Consequences of mRNA transport on stochastic variability in protein levels. *Biophysical Journal*, 103(5):1087–1096.

Singh, A., Razooky, B., Cox, C. D., Simpson, M. L. et Weinberger, L. S. (2010a). Transcriptional bursting from the HIV-1 promoter is a significant source of stochastic noise in HIV-1 gene expression. *Biophysical Journal*, 98(8):L32–L34.

Singh, A., Razooky, B. S., Dar, R. D. et Weinberger, L. S. (2012). Dynamics of protein noise can distinguish between alternate sources of gene-expression variability. *Molecular Systems Biology*, 8:607.

Singh, A. et Weinberger, L. S. (2009). Stochastic gene expression as a molecular switch for viral latency. *Current Opinion in Microbiology*, 12(4):460–466.

Singh, D. K., Ku, C.-J., Wichaidit, C., Steininger, R. J., Wu, L. F. et Altschuler, S. J. (2010b). Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities. *Molecular Systems Biology*, 6:369.

Sîrbu, A., Ruskin, H. J. et Crane, M. (2012). Integrating heterogeneous gene expression data for gene regulatory network modelling. *Theory in biosciences = Theorie in den Biowissenschaften*, 131(2):95–102.

Skupsky, R., Burnett, J. C., Foley, J. E., Schaffer, D. V. et Arkin, A. P. (2010). HIV promoter integration site primarily modulates transcriptional burst size rather than frequency. *PLoS Computational Biology*, 6(9):e1000952.

Smith, B. C. et Denu, J. M. (2009). Chemical mechanisms of histone lysine and arginine modifications. *Biochimica et biophysica acta*, 1789(1):45–57.

Snijder, B. et Pelkmans, L. (2011). Origins of regulated cell-to-cell variability. *Nature Reviews Molecular Cell Biology*, 12(2):119–125.

So, L.-H., Ghosh, A., Zong, C., Sepúlveda, L. A., Segev, R. et Golding, I. (2011). General properties of transcriptional time series in Escherichia coli. *Nature Genetics*, 43(6):554–560.

Soula, H., Robardet, C., Perrin, F., Gripon, S., Beslon, G. et Gandrillon, O. (2005). Modeling the emergence of multi-protein dynamic structures by principles of self-organization through the use of 3DSpi, a multi-agent-based software. *BMC Bioinformatics*, 6:228.

Spencer, S. L., Gaudet, S., Albeck, J. G., Burke, J. M. et Sorger, P. K. (2009). Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature*, 459(7245):428–432.

Spudich, J. L. et Koshland, D. E. (1976). Non-genetic individuality : chance in the single cell. *Nature*, 262(5568):467–471.

Stavreva, D. A., Varticovski, L. et Hager, G. L. (2012). Complex dynamics of transcription regulation. *Biochimica et biophysica acta*, 1819(7):657–666.

Stern, S., Dror, T., Stolovicki, E., Brenner, N. et Braun, E. (2007). Genome-wide transcriptional plasticity underlies cellular adaptation to novel challenge. *Molecular Systems Biology*, 3:106.

Stockholm, D., Edom-Vovard, F., Coutant, S., Sanatine, P., Yamagata, Y., Corre, G., Le Guillou, L., Neildez-Nguyen, T. M. A. et Paldi, A. (2010). Bistable cell fate specification as a result of stochastic fluctuations and collective spatial cell behaviour. *PLoS ONE*, 5(12):e14441.

Süel, G. M., Kulkarni, R. P., Dworkin, J., Garcia-Ojalvo, J. et Elowitz, M. B. (2007). Tunability and noise dependence in differentiation dynamics. *Science*, 315(5819):1716–1719.

Sureka, K., Ghosh, B., Dasgupta, A., Basu, J., Kundu, M. et Bose, I. (2008). Positive feedback and noise activate the stringent response regulator rel in mycobacteria. *PLoS ONE*, 3(3):e1771.

Suter, D. M., Molina, N., Gatfield, D., Schneider, K., Schibler, U. et Naef, F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474.

Takai, N., Kira, N., Ishii, T., Nishida, M., Nasu, K. et Narahara, H. (2011). Novel chemotherapy using histone deacetylase inhibitors in cervical cancer. *Asian Pacific journal of cancer prevention*, 12(3):575–580.

Taniguchi, Y., Choi, P. J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A. et Xie, X. S. (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538.

Tao, Y., Zheng, X. et Sun, Y. (2007). Effect of feedback regulation on stochastic gene expression. *Journal of Theoretical Biology*, 247(4):827–836.

Tejedor, V., Bénichou, O., Voituriez, R., Jungmann, R., Simmel, F., Selhuber-Unkel, C., Oddershede, L. B. et Metzler, R. (2010a). Quantitative analysis of single particle trajectories : mean maximal excursion method. *Biophysical Journal*, 98(7):1364–1372.

Tejedor, V., Bénichou, O., Voituriez, R. et Moreau, M. (2010b). Response to targeted perturbations for random walks on networks. *Physical Review E*, 82(5 Pt 2):056106.

Thattai, M. et Van Oudenaarden, A. (2004). Stochastic gene expression in fluctuating environments. *Genetics*, 167(1):523–530.

Tillo, D. et Hughes, T. R. (2009). G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics*, 10:442.

Tirosh, I. et Barkai, N. (2008). Two strategies for gene regulation by promoter nucleosomes. *Genome Research*, 18(7):1084–1091.

Tkacik, G. et Bialek, W. (2009). Diffusion, dimensionality, and noise in transcriptional regulation. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 79(5 Pt 1):051901.

Tóth, K. F., Knoch, T. A., Wachsmuth, M., Frank-Stöhr, M., Stöhr, M., Bacher, C. P., Müller, G. et Rippe, K. (2004). Trichostatin A-induced histone acetylation causes decondensation of interphase chromatin. *Journal of Cell Science*, 117(Pt 18):4277–4287.

Ullner, E., Buceta, J., Díez-Noguera, A. et Garcia-Ojalvo, J. (2009). Noise-induced coherence in multicellular circadian clocks. *Biophysical Journal*, 96(9):3573–3581.

Ungrin, M. D., Joshi, C., Nica, A., Bauwens, C. et Zandstra, P. W. (2008). Reproducible, ultra high-throughput formation of multicellular organization from single cell suspension-derived human embryonic stem cell aggregates. *PLoS ONE*, 3(2):e1565.

Uren, A. G., Mikkers, H., Kool, J., van der Weyden, L., Lund, A. H., Wilson, C. H., Rance, R., Jonkers, J., van Lohuizen, M., Berns, A. et Adams, D. J. (2009). A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nature Protocols*, 4(5):789–798.

Van Zon, J. S., Morelli, M. J., Tănase-Nicola, S. et Ten Wolde, P. R. (2006). Diffusion of transcription factors can drastically enhance the noise in gene expression. *Biophysj*, 91(12):4350–4367.

Veening, J.-W., Stewart, E. J., Berngruber, T. W., Taddei, F., Kuipers, O. P. et Hamoen, L. W. (2008). Bet-hedging and epigenetic inheritance in bacterial cell development. *Proceedings of the National Academy of Sciences*, 105(11):4393–4398.

Verdone, L. et Caserta, M. (2005). Role of histone acetylation in the control of gene expression. *Biochemistry and Cell Biology*, 83(3):344–353.

Versteeg, R., van Schaik, B. D. C., van Batenburg, M. F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H. J. et van Kampen, A. H. C. (2003). The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Research*, 13(9):1998–2004.

Veselý, J. (1985). Mode of action and effects of 5-azacytidine and of its derivatives in eukaryotic cells. *Pharmacology & therapeutics*, 28(2):227–235.

Viñuelas, J., Kaneko, G., Coulon, A., Beslon, G. et Gandrillon, O. (2012). Towards experimental manipulation of stochasticity in gene expression. *Progress in Biophysics and Molecular Biology*, 110(1):44–53.

Viñuelas*, J., Kaneko*, G., Coulon, A., Vallin, E., Morin, V., Mejia-Pous, C., Kupiec, J.-J., Beslon, G. et Gandrillon, O. (2013). Quantifying the contribution of chromatin dynamics to stochastic gene expression reveals long, locus-dependent periods between transcriptional bursts. *BMC biology*, 11:15.

Voliotis, M. et Bowsher, C. G. (2012). The magnitude and colour of noise in genetic negative feedback systems. *Nucleic Acids Research*, 40(15):7084–7095.

Voss, T. C., Schiltz, R. L., Sung, M. H., Johnson, T. A., John, S. et Hager, G. L. (2009). Combinatorial probabilistic chromatin interactions produce transcriptional heterogeneity. *Journal of Cell Science*, 122(3):345–356.

Waghray, A., Schober, M., Feroze, F., Yao, F., Virgin, J. et Chen, Y. Q. (2001). Identification of differentially expressed genes by serial analysis of gene expression in human prostate cancer. *Cancer research*, 61(10):4283–4286.

Waidmann, M. S., Bleichrodt, F. S., Laslo, T. et Riedel, C. U. (2011). Bacterial luciferase reporters : the Swiss army knife of molecular biology. *Bioengineered bugs*, 2(1):8–16.

Wang, G.-Z., Lercher, M. J. et Hurst, L. D. (2011). Transcriptional coupling of neighboring genes and gene expression noise : evidence that gene orientation and noncoding transcripts are modulators of noise. *Genome Biology and Evolution*, 3:320–331.

Wang, L. (2013). Identification of cancer gene fusions based on advanced analysis of the human genome or transcriptome. *Frontiers of Medicine*. Epub ahead of print.

Wang, Z. et Zhang, J. (2011). Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proceedings of the National Academy of Sciences*, 108(16):E67–E76.

Weinberger, L., Voichek, Y., Tirosh, I., Hornung, G., Amit, I. et Barkai, N. (2012). Expression noise and acetylation profiles distinguish HDAC functions. *Molecular Cell*, 47(2):193–202.

Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P. et Schaffer, D. V. (2005). Stochastic gene expression in a lentiviral positive-feedback loop : HIV-1 Tat fluctuations drive phenotypic diversity. *Cell*, 122(2):169–182.

WEINBERGER, L. S., DAR, R. D. et SIMPSON, M. L. (2008). Transient-mediated fate determination in a transcriptional circuit of HIV. *Nature Genetics*, 40(4):466–470.

WELLEN, K. E., HATZIVASSILIOU, G., SACHDEVA, U. M., BUI, T. V., CROSS, J. R. et THOMPSON, C. B. (2009). ATP-citrate lyase links cellular metabolism to histone acetylation. *Science*, 324(5930):1076–1080.

WERNET, M. F., MAZZONI, E. O., ÇELIK, A., DUNCAN, D. M., DUNCAN, I. et DESPLAN, C. (2006). Stochastic spineless expression creates the retinal mosaic for colour vision. *Nature Cell Biology*, 440(7081):174–180.

WOO, Y. H. et LI, W.-H. (2011). Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes. *Proceedings of the National Academy of Sciences*, 108(8):3306–3311.

WU, J. et TZANAKAKIS, E. S. (2012). Contribution of Stochastic Partitioning at Human Embryonic Stem Cell Division to NANOG Heterogeneity. *PLoS ONE*, 7(11):e50715.

WU, Z., LUBY-PHELPS, K., BUGDE, A., MOLYNEUX, L. A., DENARD, B., LI, W.-H., SÜEL, G. M. et GARBERS, D. L. (2009). Capacity for stochastic self-renewal and differentiation in mammalian spermatogonial stem cells. *The Journal of Cell Biology*, 187(4):513–524.

XU, E., ZAWADZKI, K. et BROACH, J. (2006). Single-Cell Observations Reveal Intermediate Transcriptional Silencing States. *Molecular Cell*, 23(2):219–229.

YANG, H., HOSHINO, K., SANCHEZ-GONZALEZ, B., KANTARJIAN, H. et GARCIA-MANERO, G. (2005). Antileukemia activity of the combination of 5-aza-2'-deoxycytidine withth valproic acid. *Leukemia Research*, 29(7):739–748.

YOSHIDA, M., HORINOUCHI, S. et BEPPU, T. (1995). Trichostatin A and trapoxin : novel chemical probes for the role of histone acetylation in chromatin structure and function. *BioEssays*, 17(5):423–430.

YUNGER, S., ROSENFELD, L., GARINI, Y. et SHAV-TAL, Y. (2010). Single-allele analysis of transcription kinetics in living mammalian cells. *Nature Methods*, 7(8):631–633.

ZENKLUSEN, D., LARSON, D. R. et SINGER, R. H. (2008). Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural & Molecular Biology*, 15(12):1263–1271.

ZERNICKA-GOETZ, M. et HUANG, S. (2010). Stochasticity versus determinism in development : a false dichotomy ? *Nature Reviews Genetics*, 11(11):743–744.

ZHANG, Z., QIAN, W. et ZHANG, J. (2009). Positive selection for elevated gene expression noise in yeast. *Molecular Systems Biology*, 5:299.