



**HAL**  
open science

# Régression fonctionnelle de Poisson pour l'analyse de données de séquençage haut-débit

Cervin Guyomar

► **To cite this version:**

Cervin Guyomar. Régression fonctionnelle de Poisson pour l'analyse de données de séquençage haut-débit. Sciences du Vivant [q-bio]. 2015. dumas-01296665

**HAL Id: dumas-01296665**

**<https://dumas.ccsd.cnrs.fr/dumas-01296665>**

Submitted on 1 Apr 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AGROCAMPUS  
OUEST

- CFR Angers  
 CFR Rennes



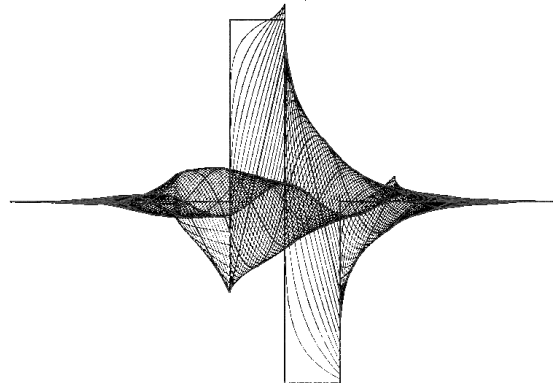
Année universitaire : 2014-2015  
Spécialité : **Statistique appliquée**

### Mémoire de Fin d'Études

- d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage  
 de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage  
 d'un autre établissement (étudiant arrivé en M2)

# Régression fonctionnelle de Poisson pour l'analyse de données de séquençage haut-débit

Par : Cervin GUYOMAR



**Soutenu à Rennes**

**le 8-09-2015**

**Devant le jury composé de :**

Président :

Autres membres du jury (Nom, Qualité)

Maître de stage : Franck PICARD

Enseignant référent : David CAUSEUR

*Les analyses et les conclusions de ce travail d'étudiant n'engagent que la responsabilité de son auteur et non celle d'AGROCAMPUS OUEST*



## Fiche de confidentialité et de diffusion du mémoire

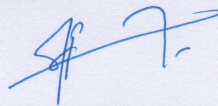
### **Confidentialité :**

Non  Oui si oui :  1 an  5 ans  10 ans

Pendant toute la durée de confidentialité, aucune diffusion du mémoire n'est possible<sup>(1)</sup>.

A la fin de la période de confidentialité, sa diffusion est soumise aux règles ci-dessous (droits d'auteur et autorisation de diffusion par l'enseignant).

Date et signature du maître de stage<sup>(2)</sup> : 31.07.15



### **Droits d'auteur :**

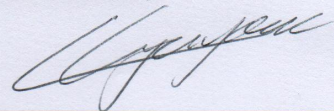
L'auteur<sup>(3)</sup> autorise la diffusion de son travail

Oui  Non

Si oui, il autorise

- la diffusion papier du mémoire uniquement(4)
- la diffusion papier du mémoire et la diffusion électronique du résumé
- la diffusion papier et électronique du mémoire (joindre dans ce cas la fiche de conformité du mémoire numérique et le contrat de diffusion)

Date et signature de l'auteur : 31/08/2015



### **Autorisation de diffusion par le responsable de spécialisation ou son représentant :**

L'enseignant juge le mémoire de qualité suffisante pour être diffusé

Oui  Non

Si non, seul le titre du mémoire apparaîtra dans les bases de données.

Si oui, il autorise

- la diffusion papier du mémoire uniquement(4)
- la diffusion papier du mémoire et la diffusion électronique du résumé
- la diffusion papier et électronique du mémoire

Date et signature de l'enseignant :

(1) L'administration, les enseignants et les différents services de documentation d'AGROCAMPUS OUEST s'engagent à respecter cette confidentialité.

(2) Signature et cachet de l'organisme

(3) Auteur = étudiant qui réalise son mémoire de fin d'études

(4) La référence bibliographique (= Nom de l'auteur, titre du mémoire, année de soutenance, diplôme, spécialité et spécialisation/Option) sera signalée dans les bases de données documentaires sans le résumé



# Remerciements

Impossible de clore ce mémoire de fin d'études sans remercier Franck PICARD, qui m'a guidé tout au long de ce stage, pour ses conseils, sa sympathie et le temps qu'il m'a accordé. Merci également à mes deux autres encadrants de stage, Philippe VEBER et Vincent RIVOIRARD, sans qui ce stage n'aurait pas été possible. Je suis ravi d'avoir pu profiter de leurs enseignements.

Merci également à l'équipe du LBBE, et en particulier à Vincent MIELE qui m'a autorisé à partager son bureau, aux autres stagiaires de M2, et à l'équipe du pôle informatique, qui ont tous rendu à un moment ou à un autre ce stage un peu plus agréable !

Je remercie enfin David CAUSEUR, mon encadrant à Agrocampus, qui m'a vivement conseillé ce stage, et plus généralement tous les étudiants et enseignants de la spécialité statistique, qui me font garder un excellent souvenir de cette troisième année.



# Sommaire

<b>1</b>	<b>Introduction : une nouvelle méthode pour l'analyse des données de ChIP-Seq</b>	<b>1</b>
1.1	Le ChIP-Seq : Une technique de séquençage haut débit pour la détection d'interactions ADN-protéine . . . . .	1
1.1.1	Principe du ChIP-Seq . . . . .	1
1.1.2	Les différents types de signaux de ChIP-Seq . . . . .	2
1.1.3	L'analyse des données de ChIP-Seq : un bref état de l'art de la détection de pics . . . . .	4
1.2	Régression fonctionnelle de Poisson par ondelettes . . . . .	4
1.2.1	Qu'est ce qu'une base d'ondelettes ? . . . . .	4
1.2.2	L'approche par dictionnaire : un problème de grande dimension . . . . .	5
1.3	Problématique du stage . . . . .	7
<b>2</b>	<b>Adaptation de la méthode proposée à des données de séquençage haut débit</b>	<b>8</b>
2.1	Les données utilisées . . . . .	8
2.2	Estimation à l'échelle d'un chromosome : sélection de régions d'intérêt . . . . .	9
2.2.1	Choix d'un package R pour la régression de Poisson . . . . .	9
2.3	Estimation à l'échelle de plusieurs milliers de bases : découpage d'un signal . . . . .	9
2.4	Recherche de zones d'intérêt dans un signal de ChIP-Seq . . . . .	9
2.5	Non-invariance par translation de la décomposition en ondelettes . . . . .	12
2.5.1	Qu'est ce que l'invariance par translation ? . . . . .	12
2.5.2	Cycle spinning complet et partiel . . . . .	12
2.6	Introduction de réplicats dans le modèle . . . . .	13
2.6.1	Construction d'une matrice de design pour plusieurs réplicats . . . . .	13
2.6.2	Calcul d'un intervalle de confiance . . . . .	14
<b>3</b>	<b>Validation de la méthode sur des données simulées et expérimentales</b>	<b>16</b>
3.1	Calibration des paramètres à partir d'un jeu de données simulé . . . . .	16
3.1.1	Protocole de simulation . . . . .	16
3.1.2	Résultats et recommandations . . . . .	18
3.2	Estimation des performances à partir de données expérimentales . . . . .	21
3.2.1	Démarche suivie . . . . .	21
3.2.2	Interprétation des résultats de la comparaison . . . . .	21
3.2.3	Limites de la comparaison . . . . .	22
<b>4</b>	<b>Perspectives et conclusion</b>	<b>23</b>
<b>A</b>	<b>Caractéristiques des jeux de données utilisés</b>	<b>26</b>
<b>B</b>	<b>Résultats de simulation</b>	<b>27</b>
<b>C</b>	<b>Quelques précisions sur l'implémentation de la sélection des zones d'intérêt du signal au LBBE</b>	<b>29</b>





# Table des figures

1.1	Principe de fonctionnement du CHIP-Seq . . . . .	2
1.2	Distribution des reads + et - autour d'un pic de CHIP-Seq, pour le facteur de transcription CTCF. . . . .	3
1.3	Trois exemples de signaux de CHIP-Seq (facteur de transcription, histone, polymerase 2) . . . . .	3
1.4	Les 7 ondelettes de Haar définies pour un signal de longueur 8 . . . . .	6
1.5	Les 7 ondelettes de Daubechies définies pour un signal de longueur 8 . . . . .	6
2.1	Illustration de la procédure de sélection de zones . . . . .	10
2.2	Distribution du nombre de reads par fenêtre de 10000 bases - origines de réplication . . . . .	11
2.3	Illustration de la non-invariance de la reconstitution par une translation d'une base . . . . .	13
2.4	Estimation pour cinq réplicats des données d'origines de réplication, et intervalles de confiance . . . . .	15
3.1	Fonction d'intensité utilisée pour les simulations . . . . .	17
3.2	Comparaison des temps de calcul pour différentes tailles de fenêtres et bases de fonctions . . . . .	18
3.3	Évolution du MSE en fonction du nombre d'itérations de cycle spinning . . . . .	19
3.4	Qualité de reconstitution en fonction du pas de cycle spinning . . . . .	19
3.5	Illustration du choix d'un pas de cycle spinning multiple d'une puissance de deux . . . . .	20
3.6	Effets de la taille de la fenêtre, et du choix de la base ou du dictionnaire sur le MSE . . . . .	21
3.7	Comparaison du signal débruité et des emplacements des origines de réplication . . . . .	22
B.1	Qualité de reconstitution en fonction du pas de cycle spinning, pour deux autres bases de fonctions . . . . .	27
B.2	Comparaison des 3 tailles de fenêtre et ensembles de fonctions . . . . .	28

# Liste des tableaux

2.1	Résultats de la sélection de régions sur deux types de données . . . . .	12
A.1	Provenance des jeux de données utilisés . . . . .	26



# Chapitre 1

## Introduction : une nouvelle méthode pour l'analyse des données de ChIP-Seq

Le modèle d'un ADN déterminant à lui seul l'ensemble des caractères d'un individu semble s'estomper et les progrès des techniques de séquençage permettent d'accéder à des composantes jusqu'ici inconnues du fonctionnement des cellules [1]. De nombreuses techniques prometteuses ne se focalisent donc pas sur la séquence d'ADN en elle-même, mais sur l'interaction de certaines molécules avec la double hélice [2]. Ainsi, s'intéresser au rôle de certaines protéines semble crucial pour mieux comprendre les phénomènes de transcription, réplication, épissage ou réparation de l'ADN [3], mais n'est pas permis par les techniques classiques d'étude de la séquence d'ADN. Le ChIP-Seq (Chromatin-Immunoprecipitation-Sequencing) permet de localiser l'emplacement sur le génome de ces protéines, en séquençant massivement les portions d'ADN sur lesquelles elles agissent. Les données générées par cette technique nouvelle posent certaines difficultés, et de nombreuses méthodes proposent déjà une manière de les analyser, et de détecter les pics dans le signal qui correspondent à une interaction ADN-protéine. Ces méthodes peinent toutefois à prendre en compte certaines des particularités du ChIP-Seq, et à donner un cadre général pour étudier ces données. Ce rapport présente dans un premier temps ces particularités ainsi qu'une méthode qui leur semble bien adaptée, traite certains des problèmes rencontrés lors de son application à des données réelles, et présente des recommandations lors de l'analyse de telles données suivant cette méthode.

### 1.1 Le ChIP-Seq : Une technique de séquençage haut débit pour la détection d'interactions ADN-protéine

#### 1.1.1 Principe du ChIP-Seq

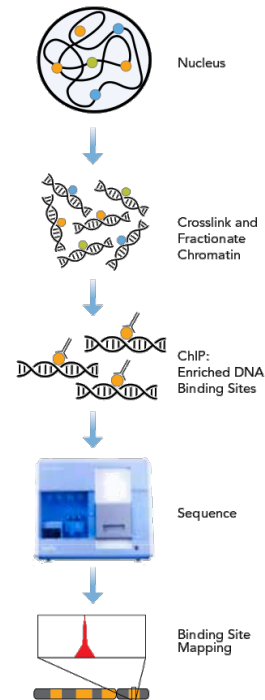
Le ChIP-Seq est une technique de séquençage haut débit qui combine l'immunoprécipitation de la chromatine, technique déjà largement employée sur des microrarrays en ChIP-chip, à une étape de séquençage massif de fragments d'ADN. La figure 1.1 détaille les grandes étapes de cette technique.

À l'issue de l'étape de séquençage, on obtient donc un ensemble de séquences d'ADN, correspondant à une extrémité d'un des fragments isolé par sonication. Ces courtes séquences, ou *reads*, sont ensuite alignées sur un génome de référence, il s'agit de l'étape de *mapping*. Dans le cas où un read correspond à plusieurs endroits du génome, il est généralement écarté de l'analyse, afin d'éviter des suraccumulations artificielles de reads, qui peuvent correspondre à des régions du génome hautement répétées. À l'issue du traitement de ces données de séquençage, les données de ChIP-Seq sont souvent exportées sous la forme d'un fichier d'alignements, reprenant les positions de début et de fin de chaque read sur le génome, au format .bam ou .bed.

Les fragments soniqués sont séquencés par leurs deux extrémités (5' et 3'). Par conséquent, en ChIP-

Le ChIP-Seq repose sur quatre étapes :

1. **Crosslink** : À l'aide de formaldéhyde, l'ADN et les protéines sont liés de manière covalentes
2. **Sonication** : L'ADN est découpé en courts segments (environ 500 paires de bases) par une étape de sonication
3. **Immunoprécipitation** : Un anticorps spécifique à la protéine étudiée est ajouté et se fixe aux complexes ADN-protéine. On peut ensuite récupérer les fragments d'intérêt
4. **Séquençage haut débit** : On isole les brins d'ADN des complexes protéine-ADN-anticorps, et on procède à leur séquençage. On obtient ainsi de courtes séquences d'ADN (de 30 à 50 bases selon le séquenceur utilisé). Chaque fragment peut être séquençé par ses deux extrémités.



**Figure 1.1** – Principe de fonctionnement du ChIP-Seq (Illumina.com)

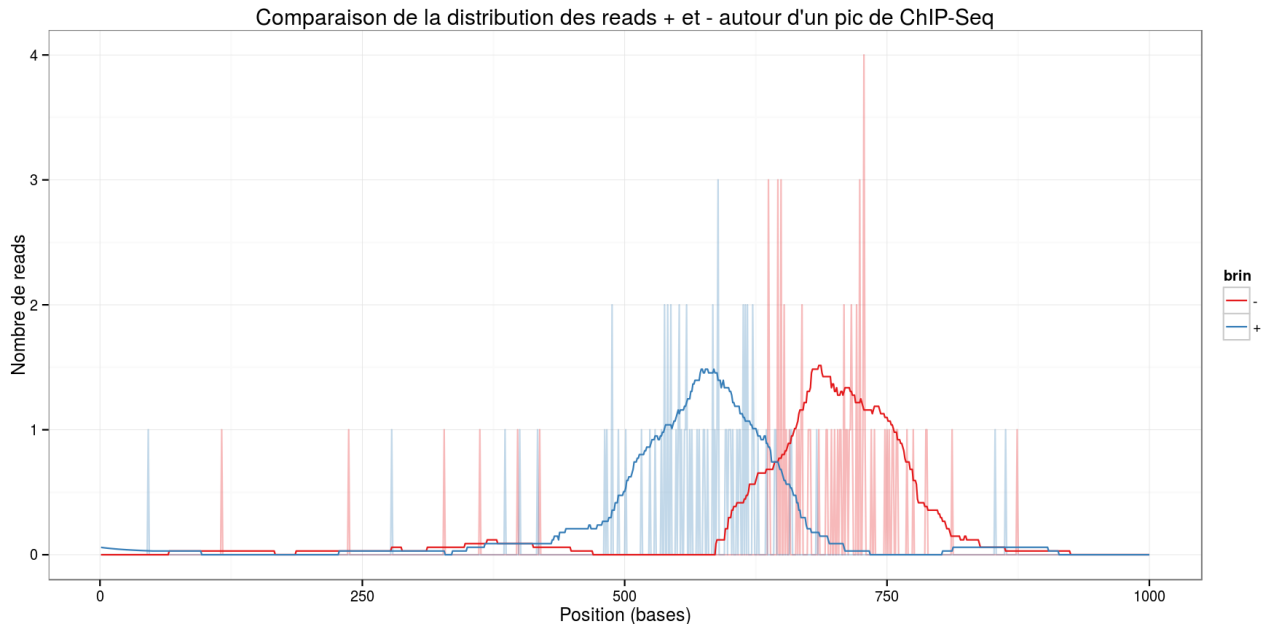
Seq, la distribution des reads autour d'un site de liaison ADN-protéine est bimodale : une partie des reads sont séquençés dans le sens positif (+), et l'autre dans le sens négatif (-), et ces deux ensembles sont distribués de part et d'autre du pic de ChIP-Seq (voir figure 1.2). Les sommets de ces deux distributions sont séparés d'une longueur qui est celle des fragments après sonication. Afin de localiser précisément les loci d'interaction, de nombreux programmes, comme MACS [4], modélisent la longueur de ces fragments, afin de superposer les distributions des reads + et -. Ces méthodes existantes n'ont pas été considérées dans ce travail : nous n'avons utilisé pour le moment qu'un seul sens de lecture des fragments, mais il sera possible ultérieurement de décaler puis d'additionner les signaux correspondant aux deux sens de lecture, afin de disposer de plus de données.

### 1.1.2 Les différents types de signaux de ChIP-Seq

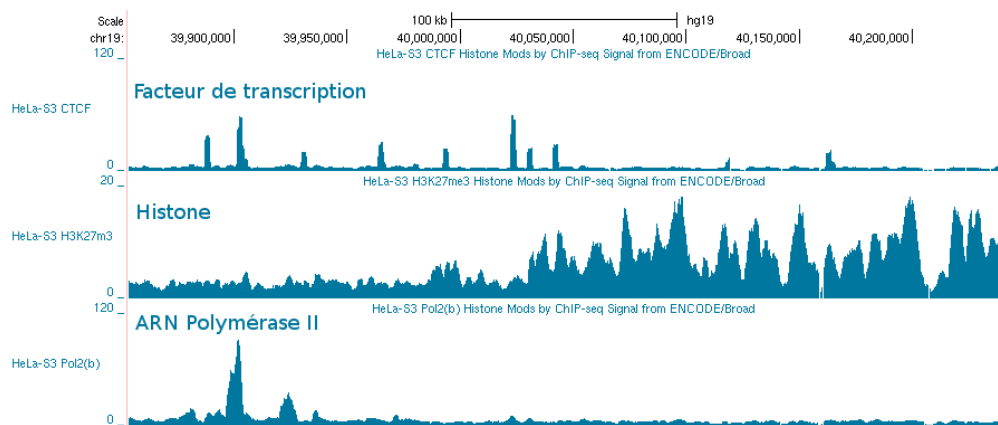
Pour réaliser une expérience de ChIP-Seq, il est nécessaire de choisir un (ou plusieurs) anticorps spécifique d'une protéine, qui est le plus souvent un facteur de transcription, une histone, ou une enzyme. Les facteurs de transcription sont impliqués dans la régulation de l'expression des gènes, tandis que les histones ont un rôle dans le maintien de la structure tridimensionnelle de l'ADN, et donc son accessibilité pour la transcription ou la réplication. Elles peuvent être modifiées de diverses manière (méthylation, acétylation...), ce qui constitue des marques épigénétiques [5]. Enfin, une enzyme couramment étudiée en ChIP-Seq est l'ARN Polymérase II, qui assure la transcription de l'ADN en ARN dans la cellule.

Un signal de ChIP-Seq se définit par les emplacements des reads le long d'un génome de référence. Les lieux d'interactions entre une protéine et l'ADN se traduisent par une suraccumulation de reads à certaines positions du génome. Mais les caractéristiques de ces interactions varient en fonction de la protéine considérée, et la forme de ces « pics » de reads peut varier. Parmi les différents types de signaux qu'il est possible d'observer, les facteurs de transcription montrent des pics très étroits (quelques dizaines à centaines de bases) et concernant une faible partie du génome, car ils se lient de manière très spécifique à des motifs sur l'ADN. À l'inverse, l'étude des modifications des histones génère des pics dont la largeur est au moins celle du nucléosome, mais qui sont généralement bien plus larges, et peuvent faire plusieurs centaines de kilobases [6]. L'étude de ces modifications a montré que leur présence sur de larges portions du génome les rendait plus facilement héréditaires [7]. Le cas des signaux d'ARN Polymérase II est particulier.

On y observe à la fois des pics étroits et des régions surenrichies plus larges. Les pics correspondent aux nombreuses enzymes concentrées sur les sites d'initiation de la transcription, tandis que les polymerases en cours de transcription sont réparties de manière plus diffuse le long de la zone transcrite et forment des zones surenrichies plus larges. La figure 1.3 illustre les différences entre des représentants de ces 3 classes de signaux.



**Figure 1.2** – Distribution des reads + et - autour d'un pic de ChIP-Seq, pour le facteur de transcription CTCF.



**Figure 1.3** – Trois exemples de signaux de ChIP-Seq (facteur de transcription, histone, polymerase 2)

### 1.1.3 L'analyse des données de ChIP-Seq : un bref état de l'art de la détection de pics

Dans les données de ChIP-Seq, les reads sont répartis spatialement le long du génome à différentes positions. Le signal biologique correspondant à des interactions entre l'ADN et une protéine se traduit par des accumulations particulièrement importantes de reads le long de ce génome. Il s'agit donc de repérer ces régions, sans accepter un trop grand nombre de faux positifs, c'est à dire de surenrichissements qui ne proviennent pas d'une réelle liaison ADN-protéine. En effet, parmi les reads alignés, un grand nombre correspondent à un bruit de fond qui peut représenter 60 à 99% des reads présents dans le jeu de données [6], et doit donc être pris en compte afin de sélectionner les régions correspondant à des pics. Si certaines méthodes considèrent ce bruit de fond comme uniforme [8, 9], il est admis que son intensité est très variable le long du génome à cause de biais, dus par exemple au séquençage ou au *mapping* des reads. Certaines méthodes prennent donc en compte un contrôle négatif pour modéliser le bruit de fond dans les données, suivant diverses méthodes [4, 10]. Enfin, la méthode ZINBA [11] prend en compte l'évolution de certaines covariables, telles que le taux de GC le long du génome, pour modéliser le bruit de fond, et rend moins indispensable le recours à un coûteux contrôle négatif.

Pour déterminer quelles régions du génome correspondent à des pics, les méthodes existantes utilisent une grande diversité de techniques dites de *peak calling* [8]. On peut toutefois distinguer trois grandes catégories : les méthodes qui définissent des fenêtres susceptibles de contenir un pic et évaluent le nombre de reads au sein de cette fenêtre (MACS [4]), celles qui recherchent des maximums locaux dans le signal (findPeaks[12]) et définissent ensuite les dimensions du pic ainsi détecté et enfin les méthodes basées sur des modèles de Markov cachés, comme Hpeak [10].

## 1.2 Régression fonctionnelle de Poisson par ondelettes

Afin de proposer une méthode plus flexible pour l'analyse des données de ChIP-Seq, ce travail propose de se pencher sur l'estimation de la fonction d'intensité associée au signal étudié. Le postulat de départ est que  $Y_t$ , le nombre de reads débutant à une position du génome est un comptage qui suit une loi de Poisson. L'objectif de cette étude est de d'approcher la fonction d'intensité  $f_0(t)$  qui, pour chaque position  $t$  du génome, donne le paramètre  $\lambda_t$  de la loi de Poisson dont est issue  $Y_t$ . Estimer cette fonction revient à débruiter le signal. La démarche proposée est d'approcher  $f_0$  par une fonction  $f$ , telle que  $\log(f)$  est une combinaison linéaire de fonctions d'une base orthonormale de fonctions. Cela revient à se placer dans le cadre du modèle linéaire généralisé, avec une fonction de lien  $\log$ . Le choix opportun de cette base de fonctions permet de s'assurer de bonnes propriétés pour la méthode. Ici,  $f_0$  est parcourue de maxima locaux qui correspondent aux pics recherchés. Ces pics se caractérisent par leur position le long du génome, leur amplitude et leur fréquence, c'est à dire leur largeur. La reconstruction par les ondelettes, dont certaines des caractéristiques sont reprises dans la section suivante, est satisfaisante lorsqu'il s'agit d'approcher une telle fonction.

### 1.2.1 Qu'est ce qu'une base d'ondelettes ?

Par famille d'ondelettes, on définit un ensemble de fonctions dérivant d'une même fonction, appelée ondelette mère. Une ondelette mère particulièrement simple, et qui constitue un bon exemple, est celle de Haar :

$$\forall x \in [0, 1], \quad \varphi(x) = \begin{cases} 1 & \text{si } x \in [0, \frac{1}{2}], \\ -1 & \text{si } x \in [\frac{1}{2}, 1], \\ 0 & \text{sinon} \end{cases}$$

À partir de cette ondelette mère, il est possible de générer, par translation, dilatation et normalisation de la fonction, toute une famille d'ondelettes :

$$\varphi_{j,k}(x) = 2^{j/2} \varphi(2^j x - k)$$
$$\varphi_{j,k}(x) = \begin{cases} 1 & x \in [\frac{k}{2^j}, \frac{k}{2^j} + \frac{1}{2^{j+1}}], \\ -1 & x \in [\frac{k}{2^j} + \frac{1}{2^{j+1}}, \frac{k}{2^j} + \frac{1}{2^j}], \\ 0 & \text{sinon} \end{cases}$$

Les ondelettes de cette famille sont caractérisées par leur échelle  $j$  (dilatation), et leur position  $k$  (translation). Une base d'ondelettes permet donc une analyse qualifiée de *temps-fréquence* [13]. En effet, une ondelette représente à la fois la fréquence (par son échelle  $j$ ) et la position ( $k$ ) d'une caractéristique du signal. La figure 1.4 illustre cette localisation temps-fréquence : chacune des fonctions est non-nulle sur seulement une partie de l'intervalle, et ces fonctions présentent différents niveaux de fréquence (3 en l'occurrence). Du fait de cette localisation temps-fréquence, estimer un signal de ChIP-Seq par une combinaison linéaire d'ondelettes permet de le représenter de manière parcimonieuse : quelques coefficients d'ondelettes permettent de représenter un pic, quelque soit sa position sur le génome et sa largeur.

Pour chaque suite de nombre de longueur  $2^J$  (comme un nombre de reads le long d'une portion de génome par exemple), il est possible, par l'intermédiaire de la transformée directe en ondelettes, de déterminer des coefficients d'ondelettes qui permettent la reconstitution de ce signal par la base. Cette base contient l'ensemble des  $\varphi_{j,k}$  avec  $j \in \{0, \dots, J\}$  et  $k \in \{0, \dots, 2^j - 1\}$ . Ces fonctions forment une base orthogonale, notée  $\Upsilon = \{\varphi_{i,j}\}$  qui permet de reconstituer un ensemble de fonctions (les fonctions de carré intégrable dans le cas des ondelettes de Haar [13]) par combinaison linéaire des ondelettes. Finalement, le nombre total de coefficients d'ondelettes nécessaires pour une telle reconstitution s'élève à  $p = 2^J - 1$ , auxquels s'ajoute un dernier coefficient d'échelle  $\beta_0$ . Le modèle s'écrit alors :

$$f = \exp(\beta_0 + \sum_{j=1}^p \beta_j \varphi_j)$$

La maximisation de la vraisemblance du modèle permet alors de trouver un vecteur  $\hat{\beta}$ , et l'estimation de la fonction d'intensité associée  $\hat{f}$  s'écrit comme le produit matriciel de la matrice de design  $A$ , avec  $A_{ij} = \varphi_j(X_i)$ , par le vecteur des coefficients  $\hat{\beta}$  maximisant la vraisemblance. L'équation 1.1 illustre ce calcul pour un signal de taille 8 avec les ondelettes de Haar.

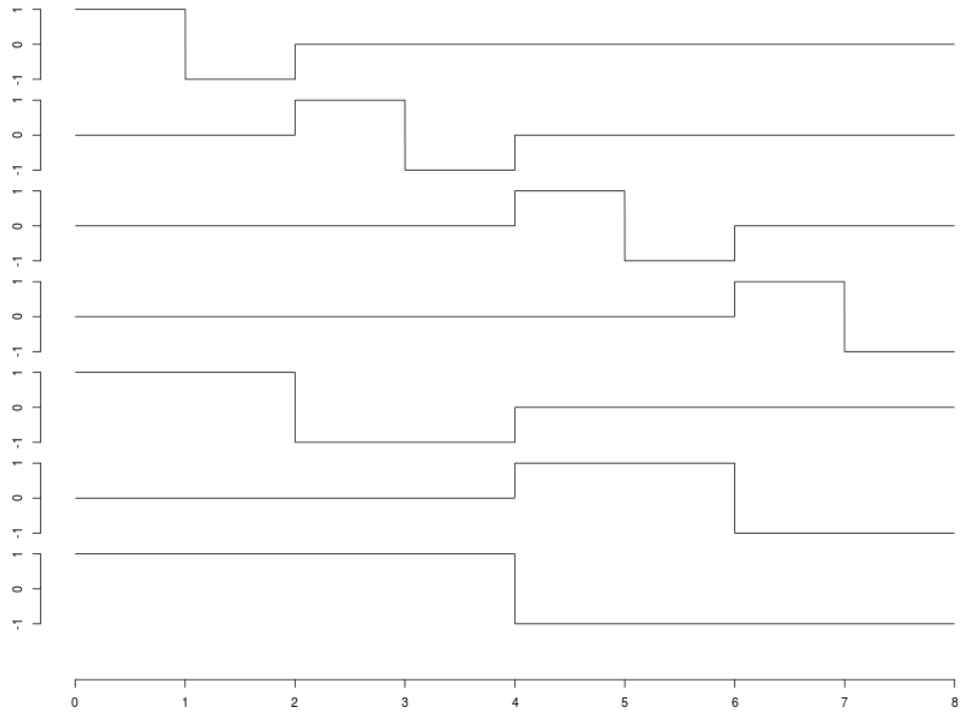
$$\log(\hat{f}_0(X)) = \begin{pmatrix} \text{intercept} & \varphi_{2,0} & \varphi_{2,1} & \varphi_{2,2} & \varphi_{2,3} & \varphi_{1,0} & \varphi_{1,1} & \varphi_{0,0} \\ 1 & 2 & 0 & 0 & 0 & 1.41 & 0 & 1 \\ 1 & -2 & 0 & 0 & 0 & 1.41 & 0 & 1 \\ 1 & 0 & 2 & 0 & 0 & -1.41 & 0 & 1 \\ 1 & 0 & -2 & 0 & 0 & -1.41 & 0 & 1 \\ 1 & 0 & 0 & 2 & 0 & 0 & 1.41 & -1 \\ 1 & 0 & 0 & -2 & 0 & 0 & 1.41 & -1 \\ 1 & 0 & 0 & 0 & 2 & 0 & -1.41 & -1 \\ 1 & 0 & 0 & 0 & -2 & 0 & -1.41 & -1 \end{pmatrix} * \begin{pmatrix} \hat{\beta}_0(X) \\ \hat{\beta}_1(X) \\ \hat{\beta}_2(X) \\ \hat{\beta}_3(X) \\ \hat{\beta}_4(X) \\ \hat{\beta}_5(X) \\ \hat{\beta}_6(X) \\ \hat{\beta}_7(X) \end{pmatrix} \quad (1.1)$$

## 1.2.2 L'approche par dictionnaire : un problème de grande dimension

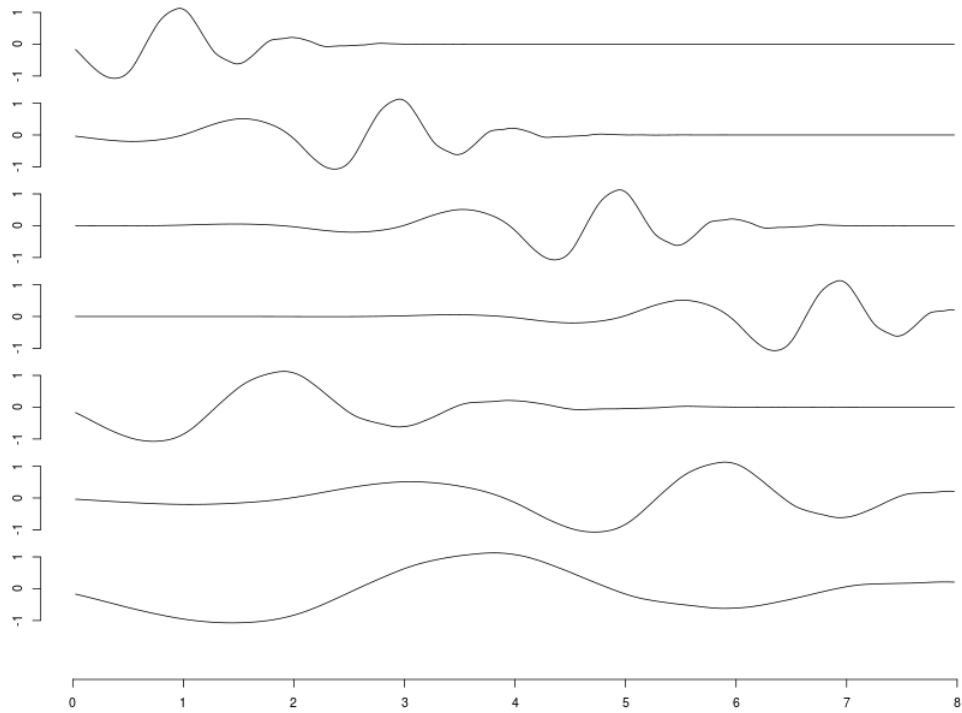
Si les ondelettes de Haar permettent de bien comprendre les avantages de cette méthode pour approcher un signal, elles ne forment peut-être pas la meilleure base de fonctions à utiliser. Ces fonctions en créneaux sont en effet assez éloignées du signal que l'on cherche à approcher. D'autres familles d'ondelettes existent, et présentent des propriétés communes. Les ondelettes de Daubechies permettent de réaliser la même reconstitution, sans être continues par morceaux comme les ondelettes de Haar. La figure 1.5 illustre les fonctions de cette base pour un court signal ( $n=8$ )

On peut imaginer que certaines parties du signal soient plus efficacement reconstruites en utilisant la base de Daubechies, car la forme de ses fonctions correspond davantage à celle attendue pour les pics. Mais rien n'empêche que les ondelettes de Haar reconstruisent plus efficacement certaines parties du signal. Prenons l'exemple d'une fonction continue par morceaux (et donc bien reconstituée avec efficacité et parcimonie par les ondelettes de Haar), mais parcourue de pics plus proches des ondelettes de Daubechies. Dans ce cas de figure, employer les deux bases simultanément permet de mieux reconstituer le signal, et de restreindre le nombre de coefficients non nuls nécessaires à cette reconstitution. Cette méthode est qualifiée d'approche par dictionnaire. Le dictionnaire contient les fonctions de plusieurs bases orthonormales, et est donc nécessairement redondant. Plus le dictionnaire est grand, plus les pics du signal peuvent être reconstitués de manière précise et parcimonieuse, car le nombre de coefficients non nuls nécessaires à une bonne reconstitution est moindre en utilisant plusieurs bases [14].





**Figure 1.4** – Les 7 ondelettes de Haar définies pour un signal de longueur 8



**Figure 1.5** – Les 7 ondelettes de Daubechies définies pour un signal de longueur 8

Néanmoins, le recours à un dictionnaire de bases d'ondelettes pour estimer un signal pose un problème de grande dimension. Avec une seule base d'ondelettes, il est nécessaire d'estimer  $n - 1$  coefficients, plus un coefficient d'échelle. Pour un dictionnaire contenant deux bases, le nombre de coefficients à estimer est de  $2 * (n - 1)$ , plus un dernier coefficient d'échelle, soit  $2n - 1$  coefficients au total. Par ailleurs, dans le cas d'une approche à une seule base, on aimerait favoriser la parcimonie des coefficients, en limitant le nombre de fonctions de la base nécessaires à l'estimation de la fonction d'intensité. Afin de répondre à ces deux problèmes, l'estimation des coefficients du modèle ne se fait pas en maximisant la log-vraisemblance du modèle, mais la log vraisemblance pénalisée d'une contrainte L1, soit une régression Lasso. Cette méthode a pour intérêt de rendre possible la régression avec un nombre de variables largement supérieur aux individus (ce qui est le cas de l'approche par dictionnaire) et de garantir le choix parcimonieux des coefficients.

Dans ce cadre, les coefficients du modèle sont estimés ainsi :

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ -l(\beta) + \sum_{j=1}^p \lambda_j |\beta_j| \right\}$$

Où  $l(\beta)$ , la log-vraisemblance dans le cas d'une régression de Poisson, se calcule ainsi :

$$l(\beta) = \sum_{i=1}^n -\mu_i + y_i * \log(\mu_i) - \sum_{i=1}^n \log(Y_i!)$$

$\mu_i$  étant le paramètre et la moyenne de la variable  $Y_i$ , on peut écrire  $\mu_i = f_i = \exp(\sum_{j \in J} \beta_j A_{ij})$ . Ce qui donne :

$$l(\beta) = - \sum_{i=1}^n e^{t A_i} + y_i^t A_i + \text{constante}$$

Par rapport à la régression Lasso dans le cas Gaussien, on remarque que la constante de pénalisation  $\lambda$  est indexée sur  $j$ , et donc propre à chaque variable du problème. Cela est dû à l'hétéroscédasticité du modèle, qui contraint à employer des contraintes différentes pour chaque variable. La méthode de calcul proposée pour calibrer ces poids est donnée par la référence [15].

### 1.3 Problématique du stage

Le ChIP-Seq s'avère être une technologie prometteuse afin de mieux comprendre de nombreux phénomènes biologiques. Cependant, les données générées contiennent beaucoup de bruit, et il est nécessaire de choisir soigneusement les outils permettant de détecter les pics. La variété des protéines cibles, et donc des signaux complique encore cette tâche, car peu de méthodes d'analyse s'adaptent à ces différents types de signaux. Le recours à une régression fonctionnelle de Poisson reposant sur une base d'ondelettes semble être un bon choix pour dépasser ces contraintes, en particulier car la localisation en temps et en échelle des ondelettes correspond bien aux pics recherchés dans le signal. Les premiers résultats sur des données simulées ou de séquençage haut débit sont prometteurs [15], mais l'application à des données complètes pose encore problème. En particulier, la méthode doit être adaptée à des signaux dont la longueur est celle d'un chromosome (plusieurs centaines de millions de bases), et il est nécessaire de pouvoir prendre en compte des réplicats, courants en ChIP-Seq, qui permettent de s'affranchir de certains biais. Ce rapport présente des solutions à apporter à ces problèmes, ainsi que des procédures de validation appliquées à cette méthode.

## Chapitre 2

# Adaptation de la méthode proposée à des données de séquençage haut débit

### 2.1 Les données utilisées

L'un des principaux objectifs de la méthode proposée est qu'elle puisse s'adapter à la diversité des signaux rencontrés dans les données de séquençage haut-débit. Par conséquent, la méthode a été évaluée à partir de données de diverses origines. Nous avons pu recourir dans un premier temps à des données simulées, mais aussi à des données expérimentales de différentes provenances. Des données ont été simulées à partir des fonctions *bumps* et *blocks* [16], très utilisées en analyse de signaux.

Plusieurs jeux de données expérimentales ont également été utilisés afin d'évaluer la méthode, et correspondent aux différents cas de figure dans lesquels cette méthode pourrait être utilisée. Tout d'abord, nous nous sommes intéressés à un facteur de transcription, CTCF, impliqué dans la structure tridimensionnelle de l'ADN [17]. Dans ces données, les pics attendus sont très étroits et restreints à une faible proportion du génome. CTCF a été choisi car il est très étudié en ChIP-Seq, et dispose d'un anticorps très spécifique, ce qui est important pour garantir la qualité des résultats de ChIP-Seq [6]. Un autre cas classique d'usage du ChIP-Seq est la détection de modifications des histones. La triméthylation de la lysine 27 de l'histone 3 est l'une des mieux étudiées de ces modifications [18]. Il a été montré que cette méthylation est associée à la répression de la transcription [19]. Cette modification peut être détectée à l'échelle d'un gène, mais aussi sur de larges régions de plusieurs centaines de kilobases [18].

Les données utilisées proviennent du projet ENCODE [20], et sont librement accessibles (voir annexe A.1). Elles ont été obtenues suivant un protocole strict, qui garantit leur qualité. En particulier, deux réplicats techniques sont systématiquement réalisés. Nous disposons ainsi de données supposées de bonne qualité, pour lesquels des réplicats existent. ENCODE propose également les coordonnées de pics dans ces données, obtenues à l'aide d'une autre méthode de détection, *scripture* [21].

Enfin, des données d'un autre type ont également été étudiées. Elles ne proviennent pas de ChIP-Seq mais de séquençage haut débit de SNS (Short Nascent Strands), qui permettent la localisation des origines de réplication [22]. Il est possible d'étudier ces données de la même manière que des données de ChIP-Seq, c'est à dire en considérant un nombre de reads commençant à une position donnée. Ces données montrent des pics de largeur moyenne, de l'ordre du kilobase. Cinq réplicats d'une même expérience ont été utilisés. Ils diffèrent grandement par couverture : le nombre de reads alignés sur le génome allant de 8 millions à plus de 60 millions. Ces données ont déjà été étudiées pour rechercher des origines de réplication suivant une autre méthode, reposant sur des fenêtres glissantes [23].

## 2.2 Estimation à l'échelle d'un chromosome : sélection de régions d'intérêt

### 2.2.1 Choix d'un package R pour la régression de Poisson

La longueur d'un chromosome peut atteindre plusieurs centaines de millions de base. La rapidité des méthodes d'analyse des données de génomique est donc cruciale. C'est pourquoi l'implémentation de la régression, étape la plus coûteuse en terme de calculs, se doit d'être le plus rapide possible. Dans cet objectif, trois packages différents permettant la régression de Poisson pénalisée ont été étudiés :

1. **GLMnet** [24] : Ce package repose sur du code Fortran, ce qui rend l'estimation très rapide. Malheureusement, l'optimisation de la perte de Poisson dans le cadre du modèle linéaire généralisé n'est pas un problème simple, et de nombreux problèmes de convergence des coefficients sont rencontrés dans certaines situations.
2. **Grplasso** [25] : Ce package permet une régression de type group-lasso et lasso, y compris dans le cas Poissonien, et présente également des problèmes de convergence, qui rendent l'estimation impossible dans certains cas de figure. De plus, cette méthode est très lente par rapport à GLMnet
3. **Penalized** [26] : Ce package permet de réaliser des régressions avec de multiples fonctions de lien et pénalités. Il ne présente pas les problèmes de convergence des autres méthodes. Il est cependant bien plus lent que glmnet

Finalement, l'optimisation des coefficients d'une régression de Poisson s'avère difficile, et parmi les packages testés, seul *penalized* semble satisfaisant, car il ne présente pas les mêmes problèmes de convergence des coefficients que les packages concurrents. Le choix de ce package rend l'estimation lente, ce qui est une contrainte importante étant donné la taille des données de ChIP-Seq.

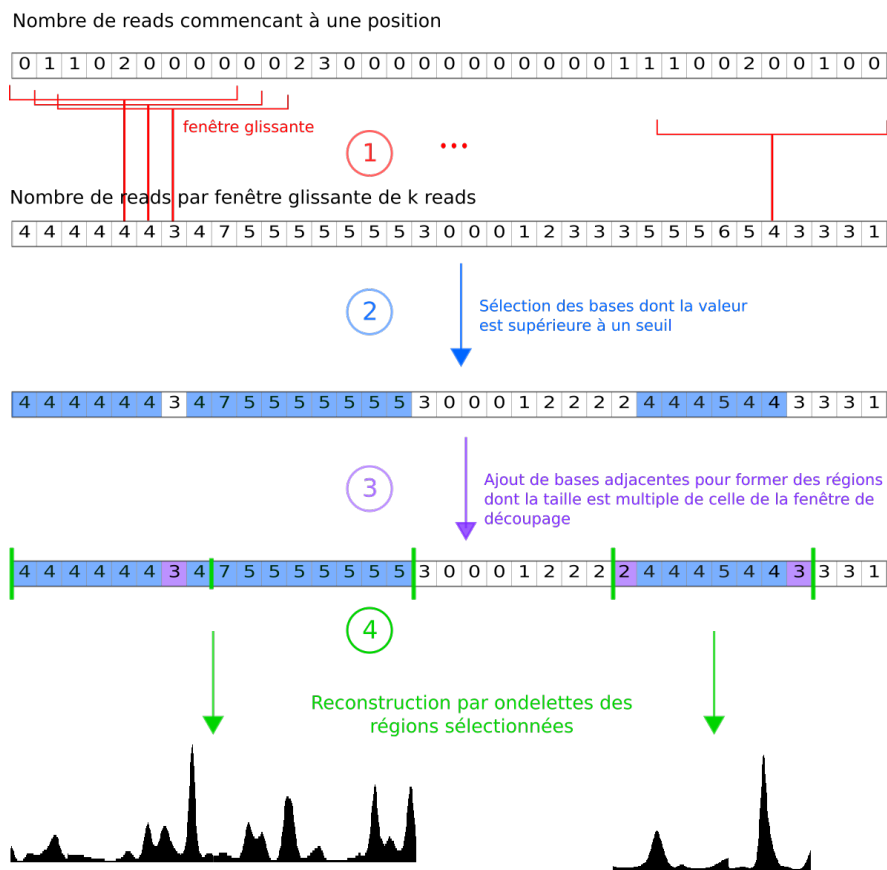
## 2.3 Estimation à l'échelle de plusieurs milliers de bases : découpage d'un signal

En pratique, l'estimation de la fonction d'intensité d'un signal ne peut se faire que pour un signal de taille restreinte. Le temps requis pour estimer la fonction d'intensité d'un signal augmente rapidement avec sa longueur, et en pratique, il n'est pas raisonnable de traiter des signaux de plus de  $2^{12}$  ou  $2^{13}$  bases. Comment alors étudier des signaux de plusieurs millions de bases ? Nous avons choisi de découper le signal suivant des fenêtres non chevauchantes dont la taille est une puissance de deux, sur lesquelles il est possible de réaliser rapidement une estimation, puis de juxtaposer ces estimations. Le signal est donc estimé en totalité, mais les coefficients correspondant à des ondelettes très grandes (plus grandes que la subdivision choisie) ne sont pas calculés. À première vue, ce découpage n'est pas sans conséquences sur la reconstitution des caractéristiques du signal : il est possible qu'un pic soit situé dans deux fenêtres différentes, ce qui risque de perturber le débruitage. La technique proposée dans la section 2.5 pour répondre au problème de la non invariance par translation résout également cette question.

## 2.4 Recherche de zones d'intérêt dans un signal de ChIP-Seq

L'estimation pour un chromosome entier en utilisant *penalized* peut prendre jusqu'à plusieurs jours, ce qui est beaucoup trop pour analyser des données de ChIP-Seq. Les résultats de ChIP-Seq ont généralement une couverture faible, et au sein d'un chromosome, seule une fraction des régions contiennent beaucoup de reads, et s'avèrent donc intéressantes [6]. On cherche donc à repérer rapidement ces zones intéressantes, avant de réaliser une estimation plus lente et précise sur ces zones. Afin de réaliser une telle sélection, il a d'abord été proposé de recourir à une régression de Poisson plus grossière, reposant par exemple sur une base de fonction moins complexe. Cette démarche s'est avérée infructueuse, car les temps de calculs n'ont pas pu être suffisamment abaissés pour la rendre profitable. Finalement, la démarche proposée dans ce paragraphe repose sur une méthode tout à fait différente. Il ne s'agit que d'une première piste pour réaliser

une telle sélection, qui n'a pas été approfondie au cours du stage. L'accent a été mis sur l'analyse fine de régions de taille moyenne, plutôt que sur la sélection de ces régions. La démarche proposée pour le moment est illustrée par la figure 2.1, et repose sur quatre étapes.



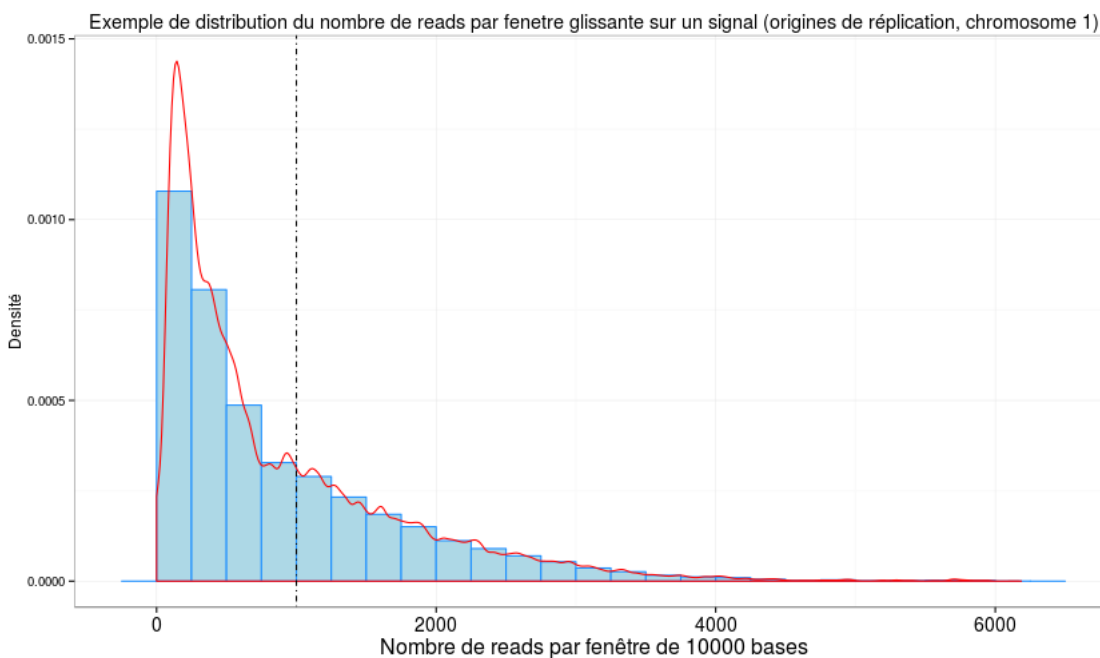
**Figure 2.1** – Illustration de la procédure de sélection de zones intéressantes dans un signal de séquençage haut débit par une fenêtre glissante

Tout d'abord, afin de sélectionner les zones du génome contenant un signal intéressant, et donc beaucoup de reads, nous appliquons une fenêtre glissante au signal, grâce à la fonction *ksmooth* de R, et comptons ainsi le nombre de reads commençant dans cette fenêtre. Pour chaque position du génome, on obtient donc le nombre de reads commençant dans les positions adjacentes. À partir de ce résultat, il reste nécessaire de définir un seuil permettant de définir les zones intéressantes du signal. Ce seuil définit le compromis entre les pics éventuellement manqués par cette sélection (pour un seuil trop haut), et un temps de calcul trop élevé (pour un seuil trop bas). Une fois des régions intéressantes repérées, elles sont étendues de manière à fusionner les zones proches, et à former des régions dont la taille est multiple de la longueur de la fenêtre de découpage vue en section 2.3. Une fois ces régions formées, il est possible de les analyser indépendamment en utilisant les ondelettes. Grâce à ce procédé, l'analyse d'une région en particulier s'avère être assez rapide, et peu coûteuse en mémoire. De plus, l'analyse de chaque zone sélectionnée se présente sous la forme d'un job qu'il est possible de soumettre indépendamment des autres au cluster de calcul du LBBE-PRABI, ce qui rend l'analyse d'un chromosome entier facilement parallélisable et très flexible. Les modalités de l'implémentation de cette technique sur le cluster de calcul du LBBE sont présentées dans l'annexe C à titre informatif.

Les valeurs des paramètres de cette procédure (largeur de fenêtre et seuil d'enrichissement) dépendent

grandement des données considérées, en particulier du type de signal et de sa couverture. Ces valeurs ont été déterminées empiriquement, et une méthode plus robuste pour les calculer serait appréciable. La taille de la fenêtre doit être choisie avec soin, afin de sélectionner les zones les plus pertinentes. En pratique, cette taille doit être comparable à la largeur des pics attendus dans le signal. Ce paramètre dépend donc fortement du type de signal étudié. Pour des signaux de séquençage haut débit correspondant à la détection d'origines de réplication, parcourue par des pics larges, une fenêtre de 10 kilobases de large a semblé pertinente. Pour un signal correspondant au facteur de transcription CTCF, 500 bases ont semblé suffisantes pour ne pas manquer de pics.

Le choix du seuil de sélection peut s'effectuer à partir de la distribution du nombre de lectures par fenêtres glissantes. La figure 2.2 représente, pour les cinq réplicats d'origines de réplifications, la distribution de ce nombre de reads par fenêtre, ainsi que le seuil retenu pour ces données, qui est de 1000 reads par fenêtre de 10000 bases. Le choix de cette valeur dépend du type de données étudiées, de leur couverture, ainsi que d'un compromis entre conservation d'un grand nombre de zones et diminution du temps de calcul. Il est toutefois préférable de placer ce seuil assez bas, afin de ne pas manquer un nombre trop important de pics. Dans le cas du facteur de transcription CTCF, en fusionnant deux réplicats, la fenêtre choisie est plus petite (500 bases), et le seuil plus faible (10 reads par fenêtre de 500 bases). Cela s'explique par la nature du signal, qui ne contient que des pics étroits, ainsi que la faible couverture de ces deux échantillons. Le tableau 2.1 reprend les caractéristiques des zones ainsi retenues. La proportion de reads retenue est comparable dans les deux cas, mais les régions retenues pour CTCF sont plus nombreuses, plus petites, et couvrent une plus petite fraction du chromosome. Ce résultat est attendu étant donné le choix d'une fenêtre glissante plus petite, et le choix d'une fenêtre de découpage plus petite (512 contre 1024 bases), qui se traduit par une extension des régions plus modérée (étape 3 figure 2.1).



**Figure 2.2** – Distribution du nombre de reads par fenetre de 10000 bases - origines de replication

Bien qu'imparfaite, car elle demande de fixer plusieurs paramètres dont les conséquences sont mal connues, la méthode proposée dans cette section permet de sélectionner les zones les plus intéressantes d'un signal très long, afin de diminuer les temps de calcul associés. Si cette sélection est indispensable pour analyser de grandes régions, l'accent au cours de ce stage a davantage été porté sur l'analyse fine de régions plus petites.

Données	Largeur de fenêtre	Seuil	Nombre de régions	Longueur sélectionnée % total	Reads sélectionnés % total
Origines de réplication <i>5 réplicats</i>	10 000	1 000	1218	18 703 360 7.5%	4 256 390 51%
CTCF (bernstein) <i>2 réplicats</i>	500	10	9247	10 986 496 4.4%	594 783 44%

**Table 2.1** – Résultats de la sélection de régions sur deux types de données

## 2.5 Non-invariance par translation de la décomposition en ondelettes

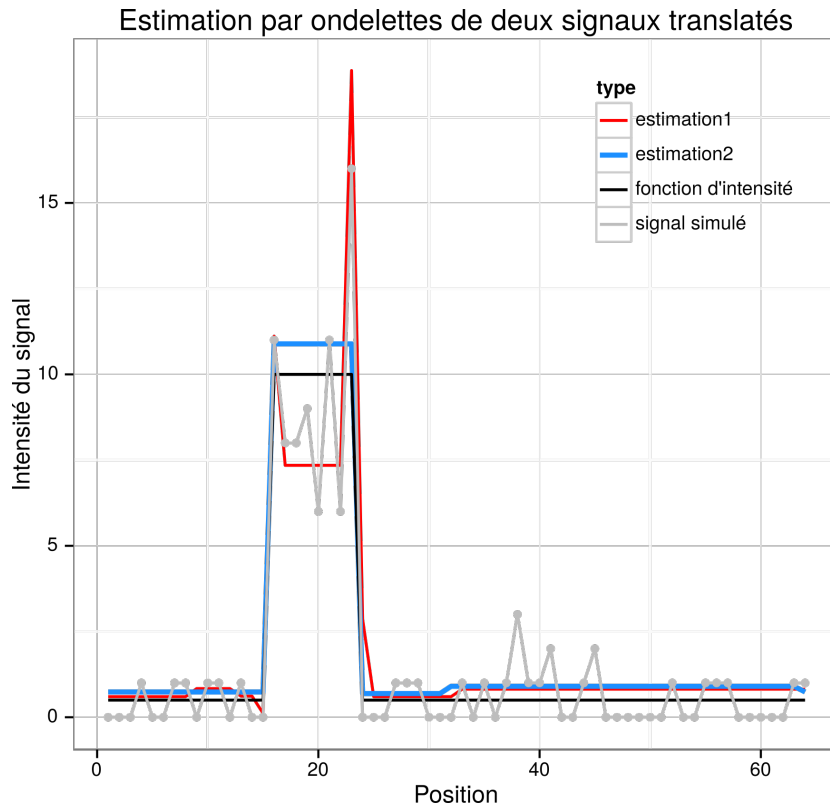
### 2.5.1 Qu'est ce que l'invariance par translation ?

Considérons un signal que nous souhaitons débruiter par des ondelettes. Parallèlement, traduisons ce signal d'une base, en faisant de la dernière base de ce signal la première du signal traduit. Les ondelettes étant localisées en temps, la couverture du signal par les différentes fonctions sera différente si l'on considère le signal d'origine ou le signal traduit [27]. Ainsi, les reconstitutions de ces deux signaux seront différentes. Ces différences peuvent être particulièrement marquées aux abords des discontinuités du signal, et une judicieuse translation peut améliorer la qualité de la reconstitution [28].

La figure 2.3 illustre ce phénomène. À partir d'une fonction d'intensité simple, comprenant un seul « pic » (en noir), on génère un signal par une loi de Poisson (en gris). La première reconstitution est réalisée sur ce signal, tandis que la seconde est réalisée sur le signal traduit d'une base, puis à nouveau traduit en sens inverse afin de pouvoir comparer les deux reconstructions. De toute évidence, la seconde estimation reproduit bien mieux la fonction d'intensité utilisée. Numériquement, la somme des carrés des erreurs pour la première estimation est de 131, contre 12 pour l'estimation du signal traduit. Cet exemple est très éloquent car l'unique pic à reconstituer sur le signal traduit est situé sur les bases 17 à 24, et est parfaitement reproduit par quelques ondelettes. Dans le premier cas, le pic est localisé sur les bases 16 à 23, et les fonctions de la base le reproduisent plus difficilement. Si dans ce cas simple, une translation particulière semble donner de meilleurs résultats, la situation est différente pour des signaux plus complexes, puisqu'un décalage du signal peut être opportun pour reconstituer un pic, mais dégrader la reconstitution d'un autre. Afin de contourner le problème de la non-invariance par translation de la transformée en ondelettes, une technique nommée *cycle spinning* a été proposée [28]. Cette méthode consiste à estimer l'intégralité des signaux qu'il est possible d'obtenir en décalant le signal, et de retenir le résultat moyen de toutes ces estimations. Par ailleurs, effectuer un cycle spinning permet de faire varier la position des fenêtres présentées dans la section 2.3, à condition de réaliser ce décalage avant le découpage, sur le signal complet, et non pas après, sur une fenêtre définie. En plus de résoudre le problème de la non-invariance par translation, le cycle spinning répond à celui des effets de bords rencontrés en découpant le signal selon des fenêtres adjacentes.

### 2.5.2 Cycle spinning complet et partiel

Bien qu'améliorant la qualité de la reconstruction par les ondelettes [28], cette technique s'avère coûteuse en termes de temps de calcul. Dans le cas d'un signal de 1024 bases, il faut faire la moyenne de 1024 estimations pour obtenir un résultat invariant par translation. Un compromis peut être de réaliser un nombre plus modeste de translations, en choisissant un pas de plus d'une base. Le nombre d'estimations diminue alors rapidement : 512 en sautant 2 bases, 256 en en sautant 4, et ainsi de suite. Dans ce cas, le résultat n'est plus invariant par translation, mais les irrégularités constatées sur les discontinuités du signal sont gommées. Le pas du cycle spinning est donc un paramètre important, permettant un compromis entre temps de calcul et meilleure estimation. Certaines modalités du choix de ce paramètre sont présentées dans la section 3.1.2.



**Figure 2.3** – Illustration de la non-invariance de la reconstitution par une translation d'une base

## 2.6 Introduction de répliqués dans le modèle

Les données de ChIP-Seq montrent souvent une grande variabilité pour différents répliqués biologiques ou techniques, et il est important de pouvoir prendre en compte ces répliqués [29].

### 2.6.1 Construction d'une matrice de design pour plusieurs répliqués

Afin de réaliser l'analyse des différents répliqués de manière conjointe, et non pas de manière séparée, il est possible d'écrire le modèle linéaire généralisé pour le répliquat  $l$  sous la forme :

$$f_l = \exp(\beta_{0,l} + \sum_{j=1}^p \beta_j \varphi_j)$$

Les répliqués peuvent différer par leur couverture, c'est à dire le nombre de reads mappés sur le génome. Le terme d'intercept  $\beta_0$  est propre au répliquat, et permet de rendre compte de ces différences de couverture entre échantillons. En revanche, les autres coefficients du modèle sont eux communs aux différents répliqués, et traduisent l'allure de la fonction recherchée. La matrice de design est ainsi construite avec une colonne d'intercept par échantillon, et les colonnes des coefficients du modèle.



$$A = \begin{pmatrix} & rep.1 & rep.2 & \varphi_{1,0} & \varphi_{1,1} & \varphi_{0,0} \\ 1 & 0 & 1.40 & 0 & 1 \\ 1 & 0 & -1.40 & 0 & 1 \\ 1 & 0 & 0 & 1.40 & -1 \\ 1 & 0 & 0 & -1.40 & -1 \\ 0 & 1 & 1.40 & 0 & 1 \\ 0 & 1 & -1.40 & 0 & 1 \\ 0 & 1 & 0 & 1.40 & -1 \\ 0 & 1 & 0 & -1.40 & -1 \end{pmatrix}$$

À partir de cette matrice de design, la démarche est identique à celle avec un seul réplicat, mais un coefficient supplémentaire est calculé pour chaque réplicat, et correspond à l'intercept de cette répétition. Ces coefficients permettent de tenir compte de l'éventuelle différence de couverture entre les échantillons.

## 2.6.2 Calcul d'un intervalle de confiance

À partir de l'estimation obtenue pour plusieurs réplicats, il semble intéressant de pouvoir mesurer la variabilité de l'estimation, en construisant des intervalles de confiance. La bibliographie sur la construction de tels intervalles pour le modèle linéaire généralisé n'abonde pas. Néanmoins, grâce à la normalité asymptotique, il est possible de construire un intervalle de confiance pour le vecteur  $\beta$  des coefficients.

$$\sqrt{N}(\hat{\beta} - \beta) \rightarrow \mathcal{N}(0, I(\beta)^{-1})$$

où  $I(\hat{\beta})$  est la matrice d'information de Fisher évaluée en  $\hat{\beta}$ . Dans le cas de la régression poissonnienne, cette matrice s'écrit :

$$I(\hat{\beta}) = A' * diag(\mu_i) * A$$

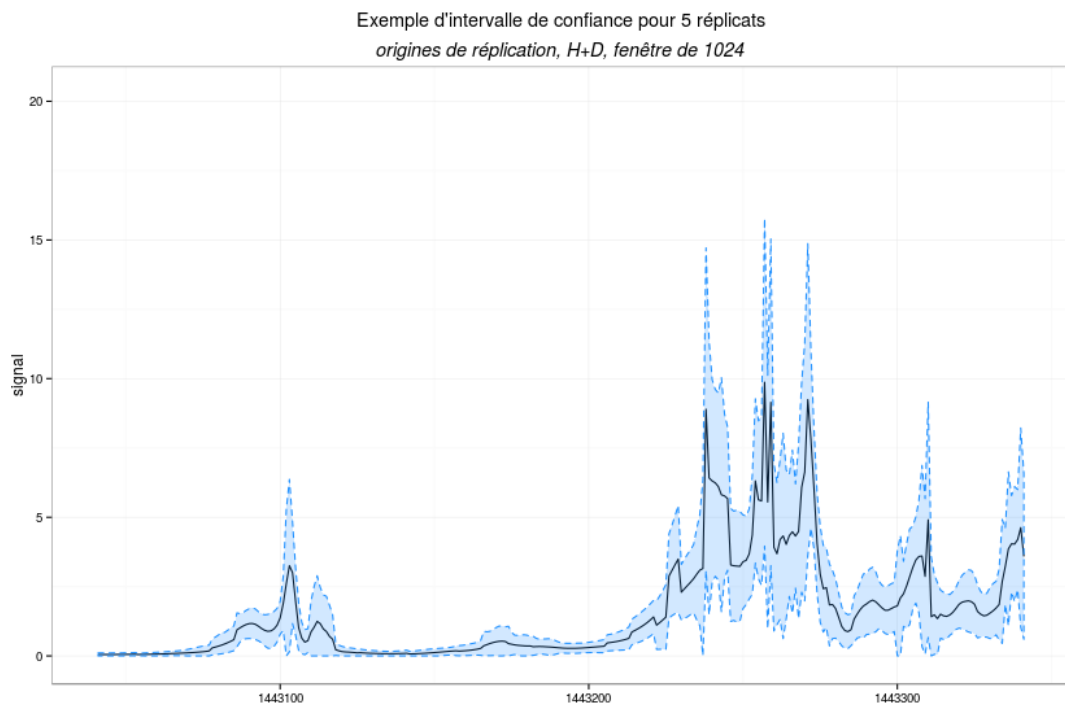
où les  $\mu_i$  sont les valeurs prédites par le modèle, et A la matrice de design.

Une fois l'intervalle de confiance sur les coefficients obtenus, deux applications successives de la delta méthode [30] permettent d'obtenir les intervalles de confiance pour le prédicteur linéaire  $\eta_i = X_i\beta$ , puis pour les valeurs prédites  $\mu_i = exp(\eta_i)$ . La delta méthode permet d'obtenir une approximation de la distribution de la transformée d'une variable aléatoire asymptotiquement normale. L'intervalle de confiance ainsi calculé pour  $\mu_i$  est :

$$g^{-1}(A'_i\hat{\beta}) \pm u_{1-\alpha/2} * h'(\hat{\eta}_i)\hat{\sigma}_{\eta_i}$$

où  $g$  est la fonction de lien du modèle linéaire généralisé, c'est à dire la fonction  $log$  dans le cas de la régression de Poisson,  $\alpha$  et le seuil de confiance choisi, et  $\hat{\sigma}_{\eta_i}$  est l'écart type du prédicteur  $\eta_i = A_i * \beta$ . Cet écart type se calcule par :  $\hat{\sigma}_{\eta_i} = \sqrt{A'_i(I(\beta))^{-1}A_i}$ .

Cet écart type est approximatif, car il repose sur des propriétés asymptotiques. De plus, d'autres écarts types peuvent être construits, à partir de la vraisemblance du modèle, ou bien comme le propose la documentation de la procédure GENMOD de SAS [31], mais c'est cette méthode qui donne les résultats les plus satisfaisants. La figure 2.4 illustre les résultats obtenus en calculant cet intervalle de confiance pour une portion des cinq échantillons du jeu de données d'origines de réplication. Les intervalles obtenus sont symétriques autour de la valeur estimée, et l'intervalle de confiance peut couvrir des valeurs d'intensité inférieures à 0, qui sont impossibles en réalité. Pour cette raison, les bornes minimales de l'intervalle qui sont négatives sont ramenées à 0.



**Figure 2.4** – Estimation pour cinq réplicats des données d'origines de réplication, et intervalles de confiance

## Chapitre 3

# Validation de la méthode sur des données simulées et expérimentales

Le chapitre précédent présente différentes techniques pour rendre possible ou améliorer l'estimation par ondelettes de long signaux, éventuellement avec plusieurs réplicats. Elles nécessitent le choix de certains paramètres, qu'il revient à l'utilisateur de choisir. Ce choix est le résultat d'un compromis entre temps d'exécution et fidélité de l'estimation. Ainsi, si idéalement, l'estimation devrait être faite avec un cycle spinning complet, de très larges fenêtres, et un dictionnaire de fonction très complet, une telle estimation prendrait un temps déraisonnable avec les moyens à notre disposition. Il est donc crucial de faire des compromis, afin de pouvoir estimer de manière satisfaisant la fonction d'intensité dans des délais raisonnables. Cette partie, à travers une étude par des simulations, s'intéresse à ce compromis, et donne des conseils pour utiliser cette méthode au mieux.

### 3.1 Calibration des paramètres à partir d'un jeu de données simulé

#### 3.1.1 Protocole de simulation

##### Signal simulé

Afin de déterminer les valeurs les plus adéquates de ces paramètres, une simulation a été conduite. Elle repose sur une fonction d'intensité choisie de manière à reproduire des caractéristiques de ChIP-Seq (voir figure 3.1). Cette fonction a une longueur de  $2^{14}$  bases (16384). Elle contient des pics étroits (200 bases) et plus larges (2000 bases), qui peuvent être seuls ou regroupés par deux ou trois. Ce signal reproduit donc les caractéristiques des facteurs de transcription par ses pics étroits et des histones par ses pics plus larges, car les largeurs de pics sont proches de celles attendues pour ces données.

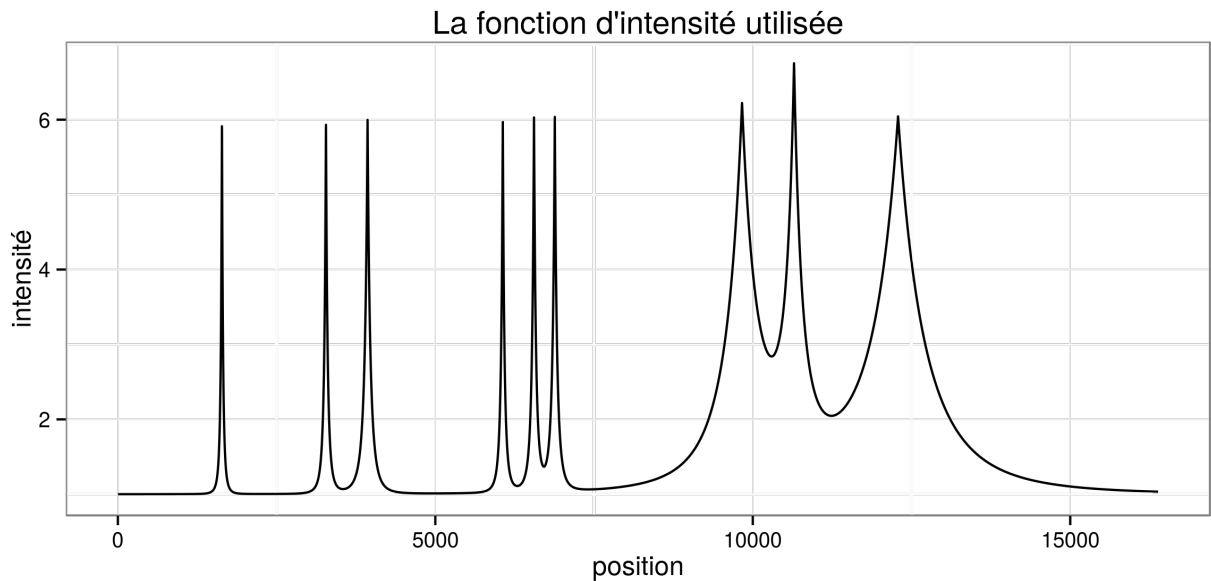
Les caractéristiques de ce signal sont donc proches de celles des facteurs de transcription comme des origines de réplifications ou histones, et proposent des largeurs de pics semblables à celles qui sont attendues dans des données expérimentales. À partir de cette fonction d'intensité, cinq signaux ont été générés en appliquant une loi de Poisson, et constitueront les cinq répétitions de cette simulation.

##### Critère de performance

Cette simulation nécessite également un critère objectif afin de déterminer les paramètres permettant la meilleure reconstitution. Dans ce but, le carré moyen des erreurs relatives (MSE) était calculé pour chaque estimation, suivant la formule :

$$MSE = \frac{1}{n} * \sum_{i=1}^n \frac{(\hat{f}(X_i) - f_0(X_i))^2}{f_0(X_i)^2} \quad \text{où } n \text{ est la longueur du signal}$$

Ce critère traduit la proximité de l'estimation avec la fonction d'intensité, et est normalisé afin de tenir compte du caractère poissonnien des données. D'autres critères moins contraignants auraient également pu être



**Figure 3.1** – Fonction d'intensité utilisée pour les simulations

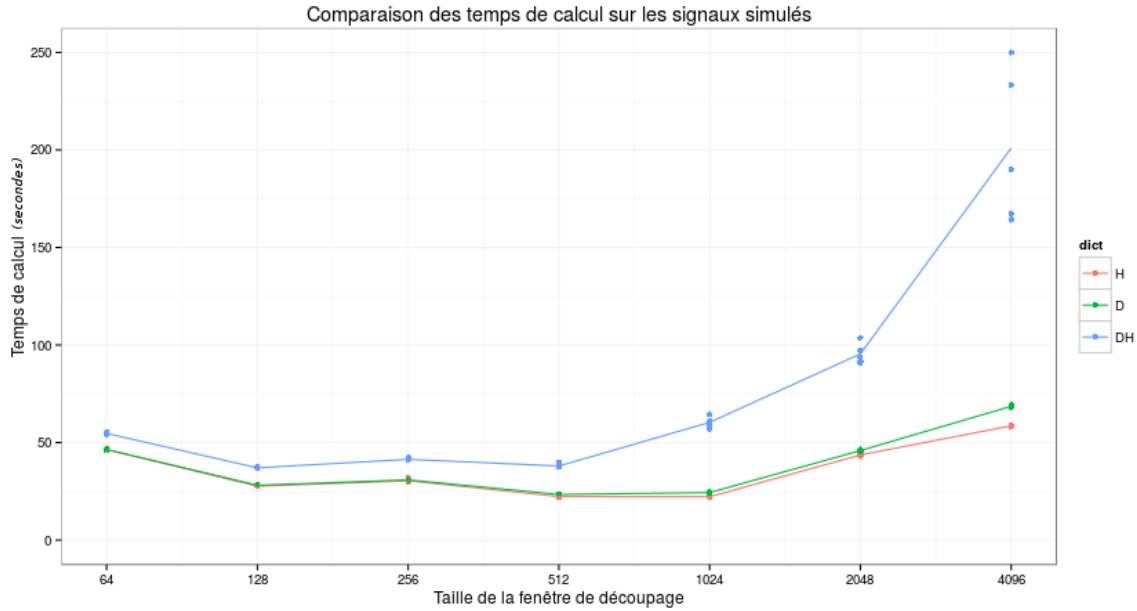
choisi, comme l'erreur dans le positionnement du pic ou la différence d'aire sous le pics, mais c'est le MSE qui traduit le mieux la proximité entre fonction d'intensité du signal et estimation.

### Plan d'expérience

Plusieurs paramètres ont été étudiés au cours de cette simulation, selon plusieurs modalités. Chacun de ces paramètres peut offrir des avantages en termes de qualité de reconstruction, mais à également un coût en termes de temps de calcul.

- **Choix de la base ou du dictionnaire** : L'estimation peut se faire en utilisant les bases de Haar ou de Daubechies, ou bien un dictionnaire combinant les deux. Plus les fonctions utilisées pour l'estimation sont nombreuses, meilleure est la fonction d'intensité estimée. Cependant, cette augmentation du nombre de fonctions utilisées est couteuse en temps de calcul. La figure 3.2 représente les temps d'estimation pour différentes bases et taille de fenêtre. En moyenne, recourir à l'approche par dictionnaire double les temps d'estimation.
- **Choix de la taille de la fenêtre fixe** : En augmentant la taille de la fenêtre, des coefficients supplémentaires sont calculés. Ils correspondent aux ondelettes les plus larges, et on peut s'attendre à ce que ces coefficients améliorent le débruitage. Du point de vue du temps de calcul, la figure 3.2 montre qu'il augmente rapidement avec la taille de la fenêtre, sauf pour les fenêtres les plus petites, qui s'avèrent être assez peu intéressantes. L'estimation étant difficile au delà de 2048 bases, et pas plus rapide en dessous de 512 bases, des tailles de fenêtre de 512, 1024 et 2048 bases ont été choisies pour les simulations.
- **Choix du pas de cycle spinning** : Le cycle spinning améliore la qualité de reconstitution, mais le temps d'estimation augmente rapidement quand le pas choisi diminue, et il est nécessaire de trouver un compromis entre ces deux contraintes. À partir du cycle spinning complet, qui nécessite d'estimer la fonction d'intensité pour l'ensemble des translations possibles, il est possible de reconstituer le résultat pour n'importe quel pas, en calculant la moyenne de certaines translations seulement. Tous les pas ont donc été testés, de 1 (cycle spinning complet) à la longueur de la fenêtre (aucun cycle spinning).

Des simulations ont été effectuées sur les cinq signaux de test en combinant les modalités de ces différents facteurs.



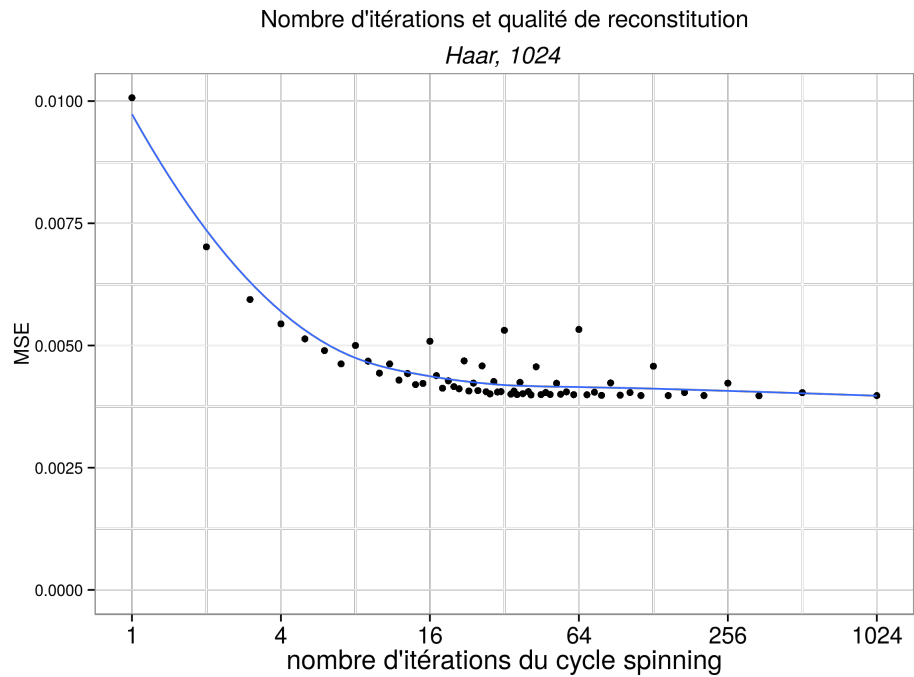
**Figure 3.2** – Comparaison des temps de calcul pour différentes tailles de fenêtres et bases de fonctions, sur les 5 signaux simulés.

### 3.1.2 Résultats et recommandations

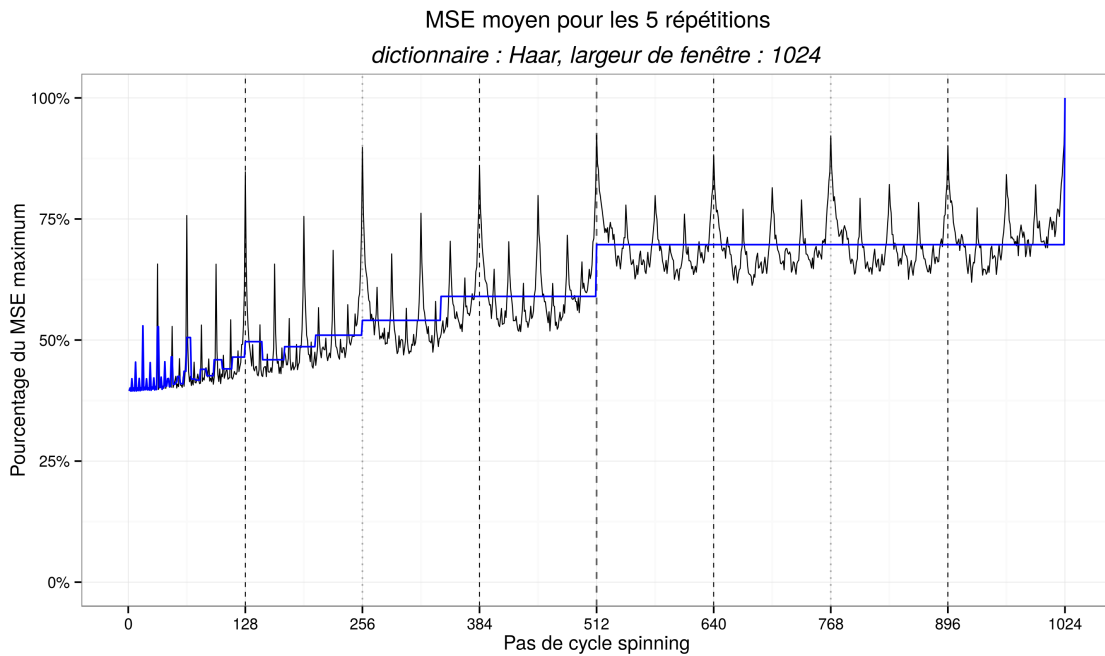
#### Choix du pas de cycle spinning

La première question soulevée par le cycle spinning est celle du compromis entre qualité de reconstruction et temps de calcul. Pour un pas de cycle spinning donné, le signal est analysé plusieurs fois. Ce nombre d'itérations de l'estimation augmente quand on se rapproche du cycle spinning complet, pour lequel il y a autant d'itérations que de bases dans le signal, et vaut un quand il n'y a pas de cycle spinning. Deux questions peuvent donc se poser au moment de choisir le pas du cycle spinning : est-il bénéfique de multiplier les itérations ? et tous les pas conduisant au même nombre d'itérations conduisent-ils aux mêmes performances de reconstruction ? Les figures 3.3 et 3.4 apportent une réponse à ces questions. La première montre que la diminution du MSE est forte pour un faible nombre d'itérations, mais le MSE tend rapidement vers une limite inférieure, alors que les temps de calcul augmentent exponentiellement. Une première recommandation à donner serait donc que le cycle spinning complet n'est pas intéressant, et que le choix d'un nombre d'itérations faible suffit à assurer une bonne reconstitution.

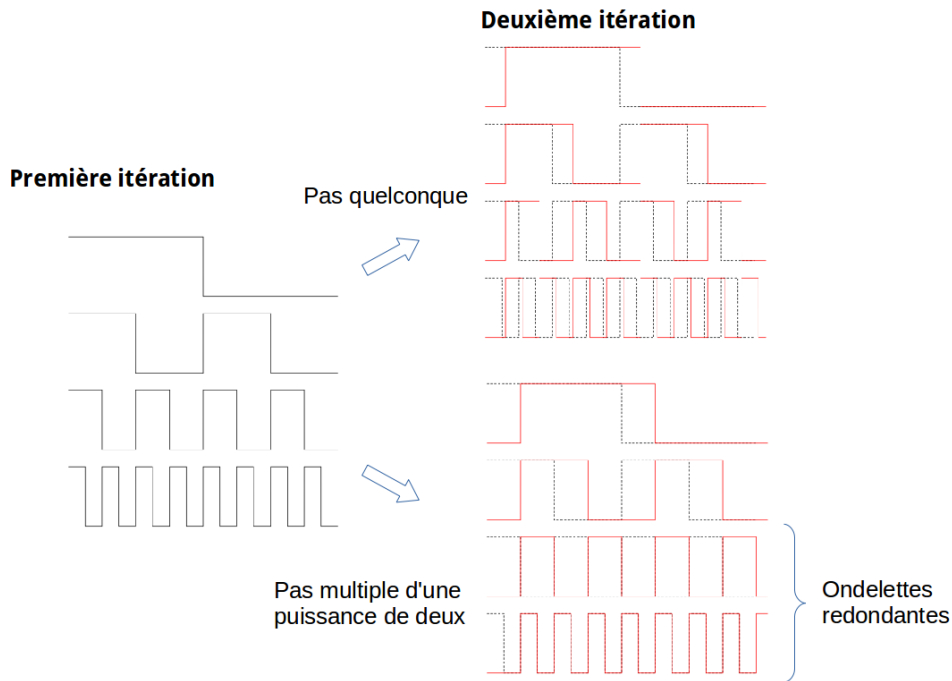
Cependant, pour un nombre d'itérations donné, les performances sont également très variables. La figure 3.4 représente la qualité de reconstitution en fonction du pas de cycle spinning pour les fonctions de Haar. Les résultats pour Daubechies et Daubechies+Haar sont disponibles en annexe B. La ligne brisée bleue donne le MSE moyen des pas correspondant à un même nombre d'itérations. On observe que pour un même segment de cette ligne, les valeurs de MSE oscillent fortement autour de la valeur moyenne. Tous les pas correspondant à un même temps de calcul ne sont donc pas équivalents du point de vue de la qualité de l'estimation. En particulier, la courbe est parcourue de pics réguliers, dont les emplacements correspondent à des pas multiples d'une puissance de deux. Ce phénomène s'explique par le fait qu'une translation d'un pas multiple d'une puissance de deux entraîne une redondance de certaines ondelettes utilisées, celles dont la taille est inférieure au pas choisi. Cette redondance est visible sur la figure 3.5. Ainsi, dans le cas où le pas est multiple d'une puissance de deux, un certain nombre des ondelettes ont déjà été utilisées pour étudier le signal lors d'une autre itération, et le bénéfice du cycle spinning est moindre. En moyenne, en utilisant la base de Haar est une fenêtre de 1024 base, le MSE est 19% plus important pour les pas multiples d'une puissance de deux que pour les autres. Il convient donc d'éviter de tels paramètres.



**Figure 3.3** – Évolution du MSE en fonction du nombre d'itérations de cycle spinning



**Figure 3.4** – Qualité de reconstitution en fonction du pas de cycle spinning, les pas multiples de 128 sont indiqués en trait pointillé.

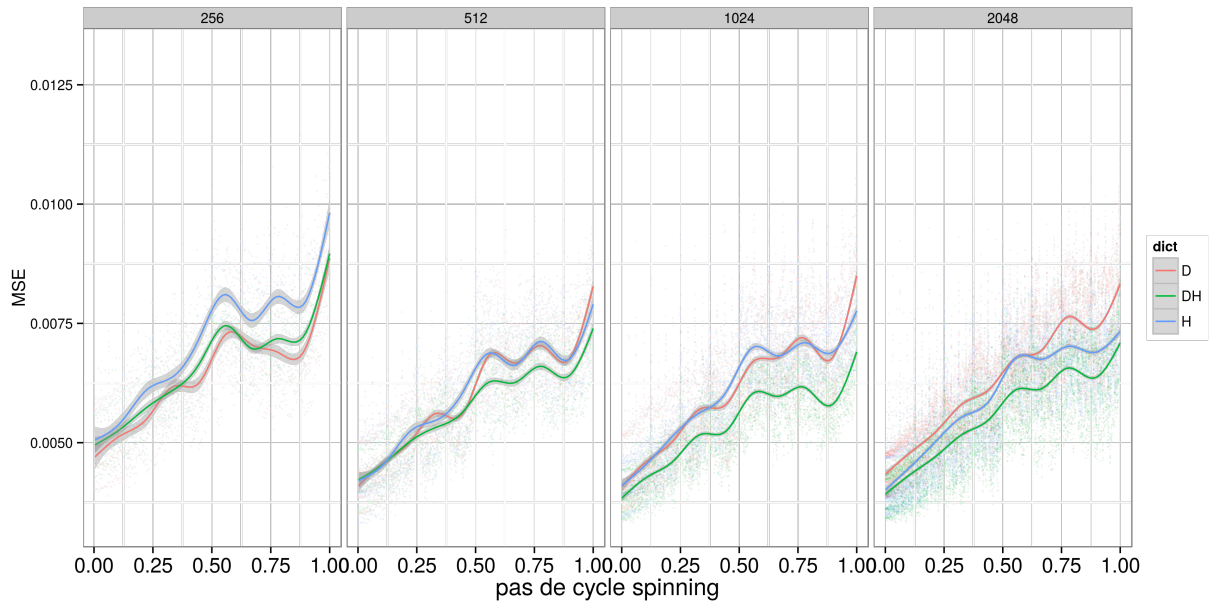


**Figure 3.5** – Illustration du choix d'un pas de cycle spinning multiple d'une puissance de deux sur les bases d'ondelettes.

### Choix du dictionnaire et de la taille de fenêtre

La figure 3.6 présente les MSE obtenus pour les différents pas de cycle spinning pour les 3 tailles de fenêtre, et en comparant les bases de Haar et Daubechies, ainsi que l'approche par dictionnaire (figure en plein format en annexe B.2). Premièrement, l'importance du choix de la taille de fenêtre semble à modérer. À l'exception de la fenêtre de 256 bases, tous les choix proposent des performances similaires. Étant donné le surcoût en temps de calcul dû à l'utilisation d'une fenêtre plus grande, il faut donc privilégier des fenêtres de taille modérée, de 512 à 1024 bases.

Comme attendu, l'approche par dictionnaire Daubechies + Haar est presque systématiquement plus performante que l'usage d'une seule base, l'erreur dans l'estimation de la fonction d'intensité est moindre lorsque l'on recourt à l'approche par dictionnaire. Cette remarque est valable pour toutes les tailles de fenêtre supérieures à 256, et toutes les configurations de cycle spinning. Dans le but d'estimer le plus finement possible cette fonction d'intensité, il semble donc préférable de recourir à ce dictionnaire de fonctions. À l'inverse, si l'objectif est de réaliser une estimation rapide, il est possible de se restreindre à une seule base. Les deux bases de Haar et Daubechies offrent alors des performances comparables, mais l'allure du signal est différente : la reconstitution par Haar de la fonction d'intensité est une fonction constante par morceaux, tandis que celle par Daubechies est plus lisse.



**Figure 3.6** – Effets de la taille de la fenêtre, et du choix de la base ou du dictionnaire sur le MSE

## 3.2 Estimation des performances à partir de données expérimentales

Afin d'évaluer l'efficacité de notre méthode pour débruiter des signaux de séquençage haut débit, nous l'avons appliquée à des données expérimentales, et avons comparé les résultats obtenus à ceux d'une autre méthode.

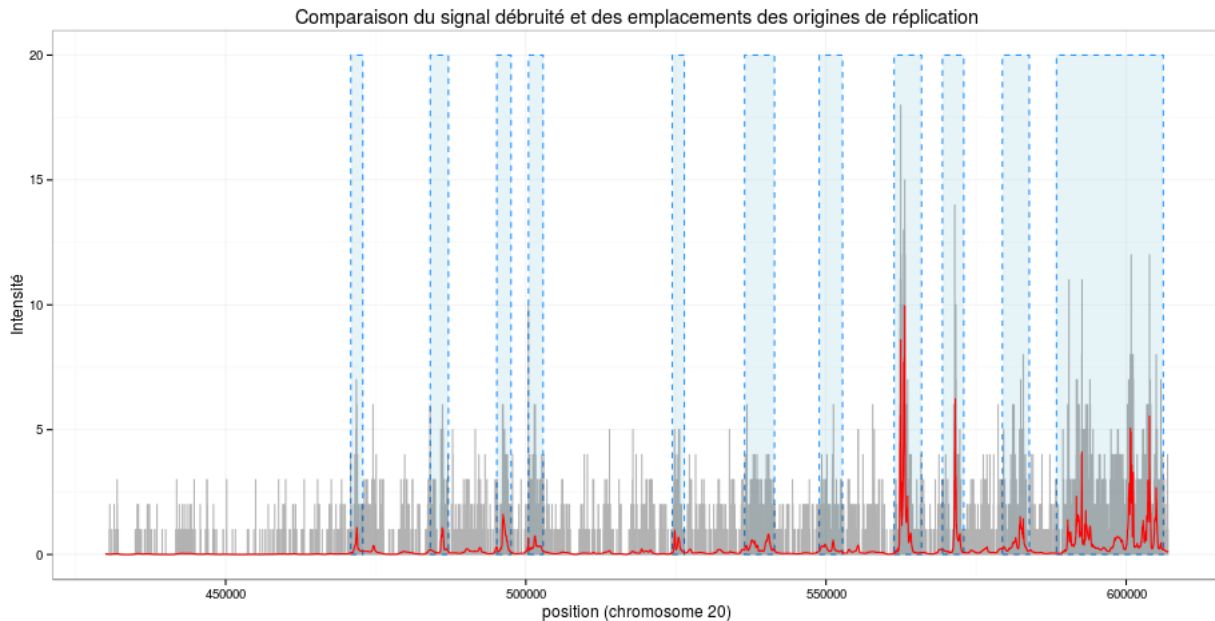
### 3.2.1 Démarche suivie

Cette comparaison a été effectuée sur des données de séquençage de *Short Nascent Strands* qui permettent de localiser les origines de réplifications (Voir Section 2.1). À partir de l'alignement des reads, nous avons estimé la fonction d'intensité du signal, en utilisant l'approche par dictionnaire, des fenêtres de 1024 bases et un pas de cycle spinning de 100 bases. Ce résultat a été comparé avec une autre analyse de ces données, par une technique de fenêtres glissantes [23]. La figure 3.7 présente ces résultats sur une portion de génome d'environ 150 kilobases. Le signal de départ est représenté en gris, la fonction débruitée en rouge, et les origines de réplification précédemment détectées en bleu.

### 3.2.2 Interprétation des résultats de la comparaison

Tout d'abord, le débruitage du signal semble efficace : bien que des reads y soient présents, les régions pour lesquelles aucune origine n'a été détectée ont une intensité estimée très faible, ce qui est l'objectif recherché. À l'inverse, là où des origines de réplification ont été détectées, la fonction d'intensité estimée prend des valeurs non nulles. À partir de cette fonction d'intensité, il est possible de déduire certaines informations supplémentaires sur ces origines de réplification. On peut les localiser de manière plus précise, car les pics dans le signal sont généralement très bien localisés. Par exemple, l'origine à l'extrême droite du graphique est parcourue par deux pics distincts dans le signal débruité. Il est également possible de quantifier la force d'une origine, par exemple à partir de l'aire sous la courbe dans la région considérée. Cette force traduit la proportion de la population de cellules séquencée pour laquelle cette origine est active, ou l'affinité de la protéine pour l'ADN en ChIP-Seq. Cependant, comparer ainsi ces pics est délicat, car des biais influent sur l'abondance des reads, et ne sont pas pris en compte par le modèle. C'est le cas du taux





**Figure 3.7** – Comparaison du signal débruité et des emplacements des origines de réplication

de GC le long du génome, qui peut influencer sur le nombre de reads associés à une position [32], et qui n'est pas pris en compte dans notre modèle actuellement.

### 3.2.3 Limites de la comparaison

Ainsi, cette première comparaison est encourageante, car la régression fonctionnelle semble apte à débruirer des signaux de séquençage haut débit, et à détecter des zones de surenrichissement notable. Elle mériterait toutefois d'être complétée par une analyse plus exhaustive. Il serait opportun d'utiliser une plus grande variété de données. Par exemple, l'étude des signaux d'ADN Polymérase serait très intéressante, car ils comportent des pics de différentes largeurs, que les ondelettes pourraient facilement reproduire. Si l'importance de l'usage de réplicats dans les analyses de CHIP-Seq est prouvée [33], étudier leur effet sur cette méthode sur des signaux complets serait également intéressant. Enfin, afin de valider l'intérêt de cette méthode, il faudrait la comparer avec plusieurs techniques de détection de pics de la littérature. Des études portant sur la comparaison de méthodes de détection de pics en CHIP-Seq existent déjà [34], et il pourrait être intéressant de reproduire cette démarche avec notre méthode.

## Chapitre 4

# Perspectives et conclusion

Si ces résultats préliminaires sont prometteurs, beaucoup de chemin reste à parcourir avant de rendre cette méthode disponible pour analyser des données de séquençage haut-débit. En premier lieu, les temps d'estimation sont longs, malgré les procédés mis au point pour les réduire. Plusieurs axes d'amélioration semblent envisageable. Tout d'abord, l'implémentation du modèle linéaire généralisé qui a été retenu (*penalized*) est particulièrement lente. Le développement d'un package assurant à la fois la convergence des coefficients et la rapidité d'exécution serait très appréciable. Des collaborations sont d'ailleurs déjà prévues dans cet objectif. En parallèle, la démarche suivie pour sélectionner des zones d'intérêt est très perfectible. Une méthode plus efficace pourrait permettre de sélectionner un nombre plus restreint de positions du génome, tout en détectant une grande majorité des pics.

Outre la réduction des temps de calcul, on pourrait également espérer une amélioration de l'estimation de la fonction d'intensité du signal. Une étape importante serait l'intégration au modèle du taux de GC le long du génome, qui est connu pour avoir un effet sur l'abondance des reads [11], ou d'autres covariables, telles que la présence de certains motifs dans la séquence qui peuvent correspondre à des emplacements de fixation des protéines sur l'ADN. Enfin, si ce stage s'est concentré sur le débruitage du signal étudié, la méthode proposée ne permet pas à l'heure actuelle de détecter des pics ou de juger leur significativité, et il nécessaire de réfléchir à un moyen de détecter ces pics à partir de la fonction d'intensité débruitée. Il serait également intéressant de tenter de bénéficier du caractère multi-échelles des ondelettes, par exemple en privilégiant les coefficients des ondelettes dont la taille est proche de celle des pics attendus.

Ce travail est une étape dans l'élaboration d'une nouvelle méthode d'analyse de données de séquençage haut débit. Il a consisté à se familiariser avec le cadre théorique de la régression de Poisson par ondelettes, à proposer certaines adaptations pour permettre l'analyse de données de séquençage à haut débit, et à juger la pertinence de cette méthode et de ces adaptations. À l'issue de ce stage, la régression par ondelettes apparaît très indiquée pour analyser de telles données, car la nature des ondelettes les rend aptes à rendre compte de la diversité des signaux rencontrés. Des applications intéressantes se profilent, car ces propriétés semblent bien adaptées aux signaux comprenant différents types de pics, tels que ceux associés à l'ADN polymérase par exemple. Il est possible d'enrichir cette méthode, en permettant l'intégration de répliquats, ou en prenant en compte la non-invariance par translation de la transformée discrète en ondelettes. Le choix des paramètres associés à ces améliorations est toutefois important, et une étude par simulations a permis d'établir quelques bonnes pratiques, en particulier dans le choix des valeurs de pas de cycle spinning. Beaucoup de travail reste à fournir avant d'utiliser plus largement cette méthode avec des données de séquençage haut débit, mais les résultats obtenus au cours de ce stage sont encourageants, et confirment l'intérêt porté aux ondelettes pour analyser des données de séquençage haut débit.

# Bibliographie

- [1] Albert Jeltsch. Beyond watson and crick : Dna methylation and molecular enzymology of dna methyl-transferases. *Chembiochem*, 3(4) :274–293, 2002.
- [2] Alla Katsnelson. Genomics goes beyond dna sequence. *Nature News*, 465(7295) :145–145, 2010.
- [3] Terrence S Furey. Chip–seq and beyond : new and improved methodologies to detect and characterize protein–dna interactions. *Nature Reviews Genetics*, 13(12) :840–852, 2012.
- [4] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). *Genome biology*, 9(9) :R137, 2008.
- [5] Robin Holliday. Epigenetics : a historical overview. *Epigenetics*, 1(2) :76–80, 2006.
- [6] Shirley Pepke, Barbara Wold, and Ali Mortazavi. Computation for chip-seq and rna-seq studies. *Nature methods*, 6 :S22–S32, 2009.
- [7] Bradley E Bernstein, Alexander Meissner, and Eric S Lander. The mammalian epigenome. *Cell*, 128(4) :669–681, 2007.
- [8] Teemu D Laajala, Sunil Raghav, Soile Tuomela, Riitta Lahesmaa, Tero Aittokallio, and Laura L Elo. A practical comparison of methods for detecting transcription factor binding sites in chip-seq experiments. *BMC genomics*, 10(1) :618, 2009.
- [9] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830) :1497–1502, 2007.
- [10] Zhaohui S Qin, Jianjun Yu, Jincheng Shen, Christopher A Maher, Ming Hu, Shanker Kalyana-Sundaram, Jindan Yu, and Arul M Chinnaiyan. Hpeak : an hmm-based algorithm for defining read-enriched regions in chip-seq data. *BMC bioinformatics*, 11(1) :369, 2010.
- [11] Naim U Rashid, Paul G Giresi, Joseph G Ibrahim, Wei Sun, and Jason D Lieb. Zinba integrates local covariates with dna-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol*, 12(7) :R67, 2011.
- [12] Anthony P Fejes, Gordon Robertson, Mikhail Bilenky, Richard Varhol, Matthew Bainbridge, and Steven JM Jones. Findpeaks 3.1 : a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15) :1729–1730, 2008.
- [13] Guy Nason. *Wavelet methods in statistics with R*. Springer Science & Business Media, 2010.
- [14] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1) :33–61, 1998.
- [15] I Ivanoff, F Picard, and Vincent Rivoirard. Adaptive lasso and group-lasso for functional poisson regression. *arXiv preprint arXiv :1412.6966*, 2014.
- [16] David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455, 1994.
- [17] Jennifer E Phillips and Victor G Corces. Ctf : master weaver of the genome. *Cell*, 137(7) :1194–1211, 2009.
- [18] Matthew D Young, Tracy A Willson, Matthew J Wakefield, Evelyn Trounson, Douglas J Hilton, Marnie E Blewitt, Alicia Oshlack, and Ian J Majewski. Chip-seq analysis reveals distinct h3k27me3 profiles that correlate with transcriptional activity. *Nucleic acids research*, 39(17) :7415–7427, 2011.

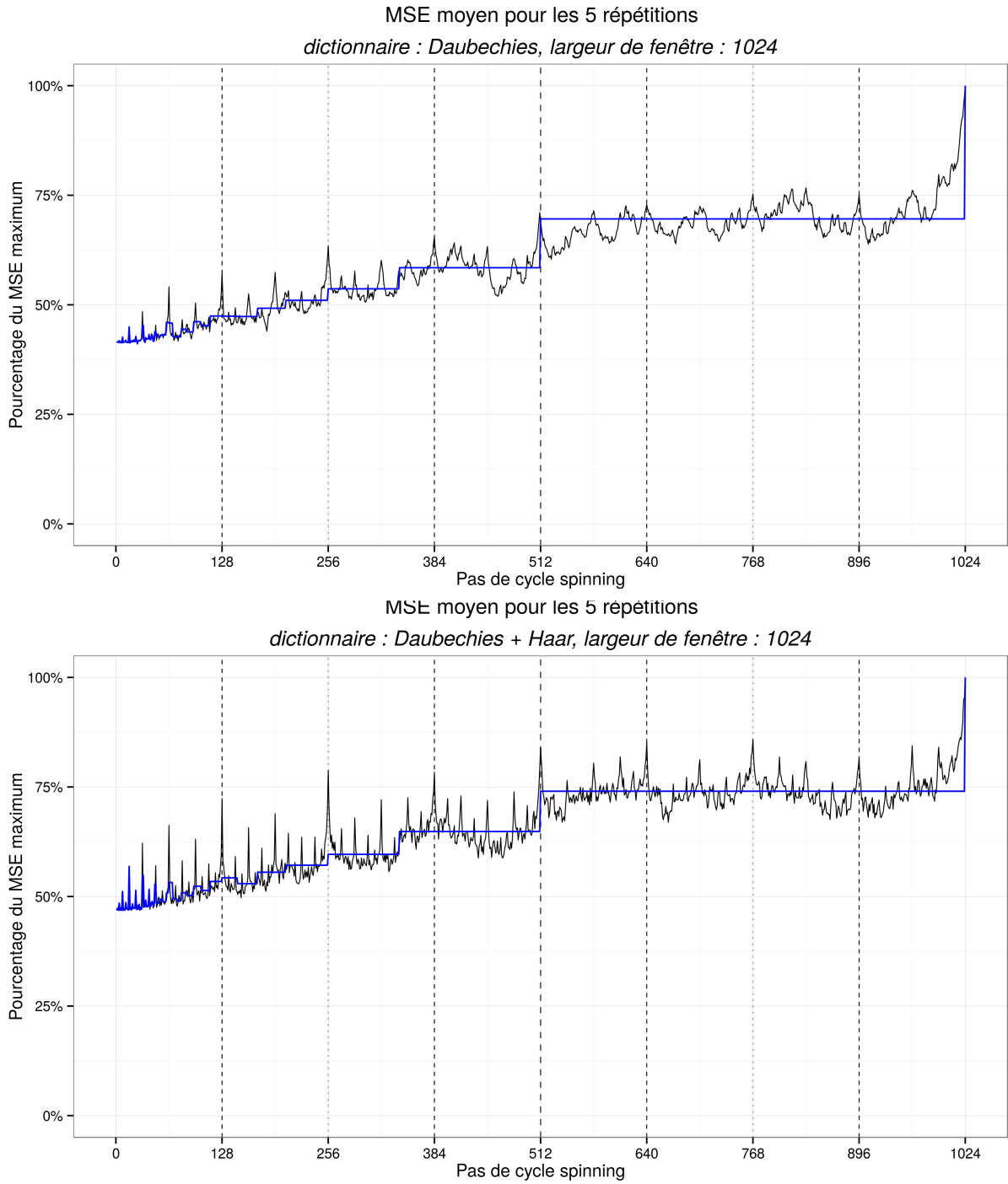
- [19] Tarjei S Mikkelsen, Manching Ku, David B Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae-Kyung Kim, Richard P Koche, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153) :553–560, 2007.
- [20] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414) :57–74, 2012.
- [21] Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature biotechnology*, 28(5) :503–510, 2010.
- [22] Jean-Charles Cadoret, Françoise Meisch, Vahideh Hassan-Zadeh, Isabelle Luyten, Claire Guillet, Laurent Duret, Hadi Quesneville, and Marie-Noëlle Prioleau. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proceedings of the National Academy of Sciences*, 105(41) :15837–15842, 2008.
- [23] Franck Picard, Jean-Charles Cadoret, Benjamin Audit, Alain Arneodo, Adriana Alberti, Christophe Battail, Laurent Duret, and Marie-Noelle Prioleau. The spatiotemporal program of dna replication is associated with specific combinations of chromatin marks in human cells. *PLoS Genet*, 10(5) :e1004282, 2014.
- [24] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. glmnet : Lasso and elastic-net regularized generalized linear models. *R package version*, 1, 2009.
- [25] L Meier. grplasso : Fitting user specified models with group lasso penalty. *R package version 0.4-2*, 2009.
- [26] Jelle Goeman, Rosa Meijer, and Nimisha Chaturvedi. penalized : L1 (lasso and fused lasso) and l2 (ridge) penalized estimation in glms and in the cox model. URL <http://cran.r-project.org/web/packages/penalized/index.html>, 2012.
- [27] Stéphane Mallat. *A wavelet tour of signal processing*. Academic press, 1999.
- [28] Ronald R Coifman and David L Donoho. *Translation-invariant de-noising*. Springer, 1995.
- [29] Yanxiao Zhang, Yu-Hsuan Lin, Timothy D Johnson, Laura S Rozek, and Maureen A Sartor. Pepr : a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated chip-seq data. *Bioinformatics*, page btu372, 2014.
- [30] Vincent Rivoirard and Gilles Stoltz. *Statistique mathématique en action*. 2012.
- [31] SAS Institute. *Sas/stat 9.2 user's guide.*, 2008.
- [32] Yuval Benjamini and Terence P Speed. Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic acids research*, page gks001, 2012.
- [33] Timothy Bailey, Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, and Jie Zhang. *Practical guidelines for the comprehensive analysis of chip-seq data*. 2013.
- [34] Morten Beck Rye, Pål Sætrom, and Finn Drabløs. A manually curated chip-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic acids research*, page gkq1187, 2010.
- [35] Aaron R Quinlan and Ira M Hall. *Bedtools : a flexible suite of utilities for comparing genomic features*. *Bioinformatics*, 26(6) :841–842, 2010.

## Annexe A. Caractéristiques des jeux de données utilisés

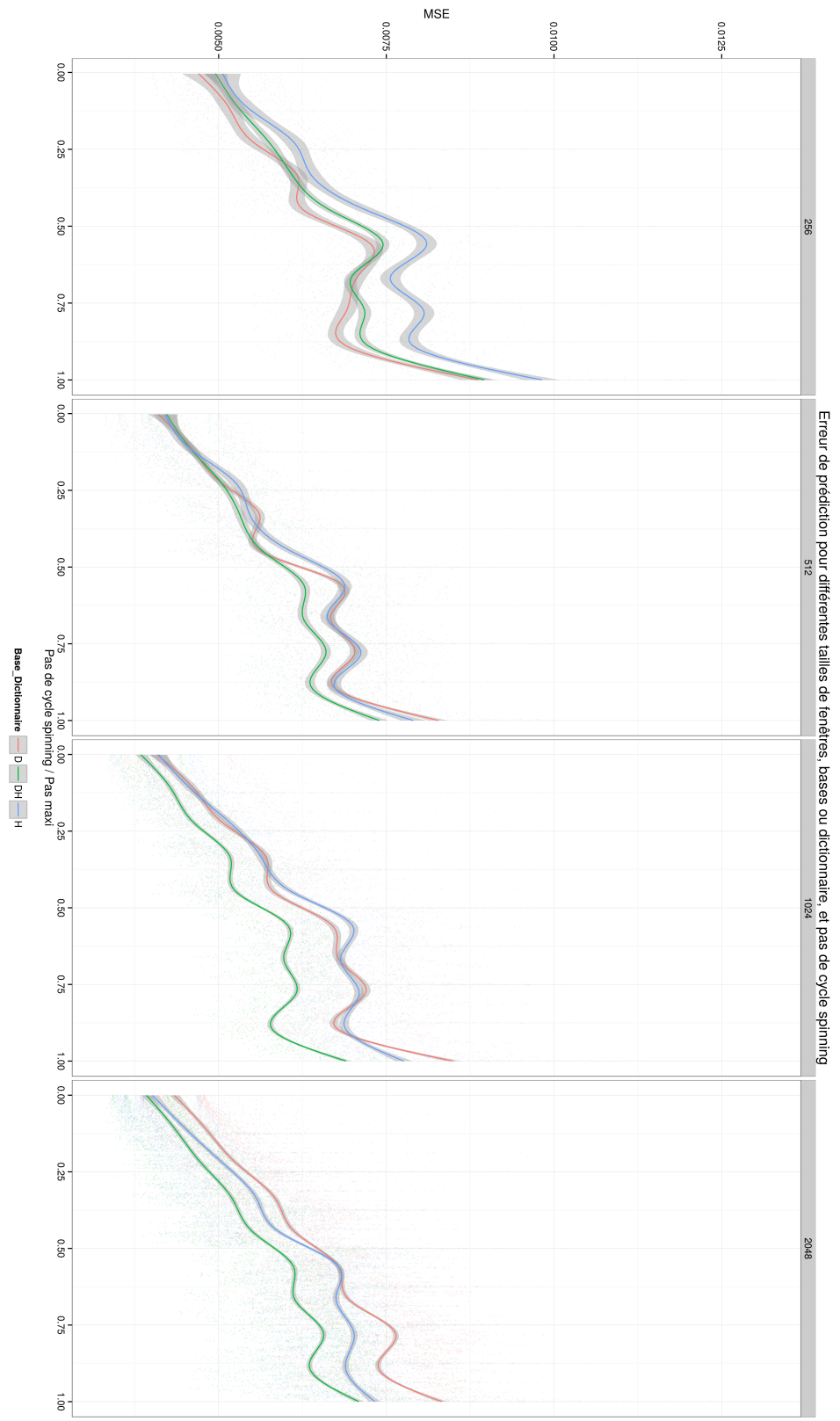
Type de cellule	Laboratoire	Réplicats biologiques	Accession des fichiers .bam
CTCF Cellule souche (H1-hESC)	Bernstein, Broad Institute	2	ENCFF000AVI ENCFF000AVK
CTCF Cellule souche (H1-hESC)	Myers, HAIB	2	ENCFF001HOZ ENCFF001HPC
CTCF Cellule épithéliale	John Stamatoyannopoulos, UW	2	ENCFF000ONK ENCFF000ONN
Histone Cellules HeLa S3	Bernstein, Broad Institute	2	ENCFF000BBS ENCFF000BBV
Origines de réplication IMR90	Montpellier GenomiX	5	GSM927235 (ncbi)

**Table A.1** – Provenance des jeux de données utilisés

## Annexe B. Résultats de simulation



**Figure B.1** – Qualité de reconstitution en fonction du pas de cycle spinning, pour deux autres bases de fonctions



**Figure B.2** – Comparaison des 3 tailles de fenêtre et ensembles de fonctions

## Annexe C. Quelques précisions sur l'implémentation de la sélection des zones d'intérêt du signal au LBBE

Ce paragraphe résume les étapes suivies pour obtenir une estimation de la fonction d'intensité à partir de fichiers .bed d'alignement de reads provenant d'une expérience de séquençage haut débit. La première étape consiste à passer d'un alignement de reads à un vecteur de couverture, qui donne le nombre de reads commençant à chaque position du génome. ce vecteur pourra être utilisé pour détecter les zones d'intérêt du signal. Les fichiers sont d'abord filtrés de manière à ne conserver que les reads correspondant au chromosome considéré, puis concaténés afin de calculer la couverture pour tous les réplicats. Enfin, la couverture le long du génome est calculée grâce à l'outil *genomecov* de la suite *bedtools* [35], selon le code suivant.

```
#!/bin/bash
grep -w chr1 reads_ori_5rep.bed > reads_ori_5rep.chr1.bed
bedtools genomecov -i reads_ori_5rep.chr1.bed -dz -5 -g chr1.genome > reads_ori_5rep.chr1.bedcov
```

Le fichier .genome utilisé en entrée est un fichier standard contenant la longueur du chromosome 1 (ou de tout autre chromosome), et s'écrit ainsi :

```
chr1 249250621
```

Le fichier .bedcov en sortie donne la position et le nombre de reads de chaque base du chromosome où commence au moins un read. Ce dernier fichier peut être lu dans R, où un vecteur contenant le nombre de reads commençant à chaque position du chromosome peut être construit. Le filtrage par fenêtre glissante décrit dans la section 2.4 peut alors être appliqué, en particulier à l'aide de la fonction *ksmooth*. Une fois des zones d'intérêt sélectionnées, pour chaque zone ainsi sélectionnée, l'appel de code bash depuis R permet :

- de filtrer les fichiers .bed de chaque réplicat, afin de ne conserve que les reads représentés dans la zone sélectionnée. Charger en mémoire l'ensemble des reads du chromosome pour en analyser une petite zone est inutile et consomme beaucoup de mémoire. Cette opération est permise par *bedtools intersect*
- d'écrire un fichier .pbs, qui permet de soumettre au cluster de calcul du LBBE une tâche, qui consiste à analyser la zone en question. Les informations fournies dans le fichier .pbs sont les coordonnées de la zone, les paramètres choisis (largeur de fenêtre, cycle spinning, nombre de coeurs affectés à la tâche), le fichier de reads crée dans le point précédent, et le script R qui permet la régression par ondelettes.

Grâce à la sélection de zones d'intérêt, l'analyse d'un chromosome ne se fait donc pas sous la forme d'une seule tâche, très couteuse en temps et en mémoire, mais sous forme d'une multitude de *jobs* assez peu gourmands et assez courts, qu'il est possible de soumettre au cluster du LBBE-PRABI. Cette approche est ainsi hautement parallélisable, et bien plus facile à gérer qu'une unique tâche.



	Diplôme : d'Ingénieur de l'Institut Supérieur des Sciences agronomiques agroalimentaires, horticoles et du paysage. Spécialité : <b>Statistique Appliquée</b> Enseignant référent : David CAUSEUR
Auteur(s) : GUYOMAR Cervin  Date de naissance* : 13/01/1994	Organisme d'accueil : UMR CNRS 5558 – LBBE (« Biométrie et Biologie évolutive ») Adresse : UCB Lyon 1 - Bât. Grégor Mendel 43 bd du 11 novembre 1918 69622 VILLEURBANNE cedex
Nb pages : 25                      Annexe(s) : 4	Maître de stage : Franck PICARD
Année de soutenance : 2015	
Titre français : <b>Régression fonctionnelle de Poisson pour l'analyse de données de séquençage à haut-débit</b>  Titre anglais : <b>Poisson functional regression for the analysis of next generation sequencing data</b>	
Résumé (1600 caractères maximum) : Les techniques de séquençage à haut-débit se sont rapidement développées pour l'étude à l'échelle du génome de phénomènes moléculaires, mais les données qu'elles génèrent posent des problèmes méthodologiques. Elles sont organisées spatialement le long du génome, et sont constituées de comptes de reads qui peuvent être modélisés par une loi de Poisson. Ce stage a porté sur une méthode statistique pour débruiter de tels signaux, et ainsi faciliter la détection de surenrichissements biologiquement significatifs, en utilisant des bases d'ondelettes. Il a consisté en l'appropriation du cadre statistique associé, qui inclut une composante de régression en grande dimension en cadre hétéroscédastique et l'application de cette méthode à des données expérimentales (en particulier de ChIP-Seq). Cette seconde étape a nécessité de mettre en place certaines techniques comme le cycle spinning, ou d'intégrer au modèle des répliquats. Ce stage illustre l'intérêt des ondelettes dans l'analyse de données de NGS, et ouvre la voie au développement d'une nouvelle méthode de détection de pics dans les données de ChIP-Seq.	
Abstract (1600 caractères maximum) : Next Generation Sequencing (NGS) is now widespread for the analysis of genome-wide molecular phenomenon, but also raises methodologicals issues, since reads counts are following a Poisson distribution, and are spatially arranged along the genome. This work introduces a new framework to denoise such data, based on the discrete wavelet transform. It consists in an investigation of the associated stistical issues, especially heteroscedastical high dimension regression, and the application of this method to experimental data, through the use of cycle spinning and the integration of replicates. This work shows the value of wavelets for analysing NGS data, and may lead to the developpementof a new method for peak calling in Chip-Seq experiments.	
Mots-clés : CHIP-Seq, NGS, ondelettes, régression poissonnienne, régression lasso Key Words: CHIP-Seq, NGS, wavelets, Poisson regression, Lasso regression	

\* Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires