HAL open science

Modèles à variables latentes pour des données issues de tiling arrays : Applications aux expériences de ChIP-chip et de transcriptome

Caroline Berard

▶ To cite this version:

Caroline Berard. Modèles à variables latentes pour des données issues de tiling arrays : Applications aux expériences de ChIP-chip et de transcriptome. Biologie végétale. AgroParisTech, 2011. Français. NNT : 2011AGPT0067 . tel-03689397

HAL Id: tel-03689397

https://pastel.hal.science/tel-03689397

Submitted on 7 Jun 2022

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.

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Ecole Doctorale ABIES
Agriculture Agriculture Alimentation Food Biologie Biology Environnement Environment Santé Health

Doctorat ParisTech

THÈSE

pour obtenir le grade de docteur délivré par

L'Institut des Sciences et Industries du Vivant et de l'Environnement

(AgroParisTech)

présentée et soutenue publiquement le 30 novembre 2011 par

Caroline BERARD

Modèles à variables latentes pour des données issues de tiling arrays

Applications aux expériences de ChIP-chip et de transcriptome

Directeur de thèse : **Stéphane ROBIN**
Co-encadrement de thèse : **Marie-Laure MARTIN-MAGNIETTE**

Jury

M. Philippe BESSE,	Professeur, INSA	Rapporteur
M. Gilles CELEUX,	DR, INRIA	Rapporteur
M. Christophe AMBROISE,	Professeur, Université d'Evry	Président du jury
Mme Anne-Laure BOULESTEIX,	Assistant professor, LMU (Allemagne)	Examinatrice
M. Laurent JOURNOT,	DR, CNRS	Examinateur
M. Stéphane ROBIN,	DR, INRA	Directeur de thèse
Mme Marie-Laure MARTIN-MAGNIETTE,	CR, INRA	Co-directrice de thèse
M. Sébastien AUBOURG	CR, INRA	Invité

AgroParisTech
UMR MIA 518 AgroParisTech / INRA
16, rue Claude Bernard 75231 Paris Cedex 05

Remerciements

Je tiens tout particulièrement à remercier mes directeurs de thèse Stéphane Robin et Marie-Laure Martin-Magniette. Travailler avec vous a été très enrichissant, merci de m'avoir encouragé tout au long de ma thèse. Marie-Laure, merci de m'avoir prise sous ton aile et de m'avoir proposé cette thèse, merci pour ton aide précieuse sur le plan scientifique et administratif, merci pour ta patience et ta disponibilité, même pendant ton congé maternité ou pendant tes vacances ! Stéphane, merci pour toute l'aide que tu m'as apportée, pour ton enthousiasme permanent, ton incroyable disponibilité et surtout tes qualités humaines exceptionnelles qui font de toi le directeur de thèse dont tout le monde rêve !

Je tiens à adresser mes plus sincères remerciements à Gilles Celeux et Philippe Besse pour avoir accepté d'être rapporteurs de mon manuscrit et de participer au jury de ma soutenance, ainsi qu'à Anne-Laure Boulesteix et Laurent Journot d'assumer les rôles d'examinateurs, et à Christophe Ambroise de présider ce jury.

Je remercie également Marie-Anne Poursat, Stéphane Le Crom et Gérard Goavert d'avoir participé à mon comité de thèse. C'est avec grand plaisir que je travaille cette année avec Marie-Anne en tant qu'ATER, qui plus est dans la filière BIBS !

Un grand merci à toute l'équipe de l'URGV, en particulier Sébastien, Sandra, Véronique, Sandrine et Alain pour toutes les informations biologiques utiles. Notre collaboration a été très enrichissante pour moi. Merci Sébastien pour ton accueil, tes réponses rapides et tes explications toujours claires. Merci Sandra pour les innombrables aller-retour de mails et pour ton efficacité à chaque fois que je te demandais un nouveau fichier !

Merci aussi à Vincent et François de l'ENS de m'avoir fourni des données intéressantes qui ont donné à cette thèse le côté appliqué si important pour moi. Merci de l'intérêt que vous avez porté à mes travaux, et surtout merci d'avoir accepté d'être les premiers testeurs de mes packages R !

Je voudrais dire un grand merci à tous les membres des équipes *Statistique et Génome* et *MORSE* pour leur soutien, leur accueil chaleureux et leur sympathie. Je garde d'excellents souvenirs de toutes ces années passées avec vous !
Je remercie mes co-bureaux Baba, Guillem, Jean-Baptiste et Stevenn qui m'ont supportée et qui ont écouté toutes mes histoires... J'en profite pour m'excuser si je les ai parfois un peu saoulés ! Un merci particulier à Stevenn avec qui j'ai eu la chance de collaborer, c'est un réel plaisir de travailler avec toi et merci pour ton calme et ta sérénité qui ont souvent aidé à me déstresser ! Je tiens également à remercier Tristan, Emilie et Marie sans qui le laboratoire ne serait pas le même. Ils sont toujours là pour compatir avec les (petites) difficultés des thésards et nous redonner le moral ! Merci Tristan de m'avoir fait profiter de ton expérience de l'enseignement, merci pour le temps passé à répondre à mes nombreuses questions pour mes premiers TDs, tu expliques trop bien ! Merci Emilie pour ton entrain, tes anecdotes toujours marrantes et pour les parties de squash ! Merci aussi à

vous deux pour la relecture de cette thèse. Merci Marie pour ta bonne humeur inaltérable et ton dynamisme, et aussi pour ton haut niveau en pâtisserie ! Mes remerciements vont également à Jean-Baptiste pour son initiation au rugby (je crois qu'il en avait marre du foot), à Antoine, mon partenaire de ping-pong et d'équa diff, à Alain pour ses conseils avisés et rassurants, à Liliane qui partage mes soucis de RER, au psychorigide Michel qui en plus n'aime pas le foot mais qui est très sympa quand-même, ainsi qu'à Jean-Jacques, Gabriel, Julie, Nathalie, Pierre, Alice, Artémio et tous les autres.

Un merci particulier à Odile, Carole, Sophie et maintenant Francine qui s'occupent de toute la partie administrative avec gentillesse et efficacité, et aussi à Hamid qui m'a gentiment autorisé un espace mémoire illimité pour mon répertoire N !

Enfin, pour leur soutien non-scientifique mais non moins significatif, je tiens à remercier mes parents et mes soeurs qui ont aussi joué le rôle de relecteur du document final.

Et toi Julien, merci pour ton soutien inconditionnel durant toutes ces années.

Table des matières

Table des matières i

Table des figures v

Liste des tableaux ix

Résumé 1

Abstract 3

Présentation générale 5

1 Contexte biologique 7
- 1 Notions de biologie moléculaire . . . 8
  - 1.1 L'ADN . . . 8
  - 1.2 Séquençage des génomes, Annotation . . . 10
  - 1.3 L'expression et le contrôle de l'expression des gènes . . . 12
- 2 Évolution des technologies haut-débit . . . 12
  - 2.1 Principe des puces à ADN . . . 13
  - 2.2 Les *tiling arrays* . . . 15
- 3 Utilisation des *tiling arrays* . . . 15
  - 3.1 Expériences de Transcriptome . . . 16
  - 3.2 Expériences de ChIP-chip . . . 17
- 4 Problématique . . . 19
- Bibliographie . . . 21

2 Modèles à variables latentes 25
- 1 Introduction . . . 26
- 2 Modèle . . . 27
  - 2.1 Loi de la variable latente . . . 28
  - 2.2 Loi d'émission . . . 28
- 3 Inférence . . . 29
  - 3.1 Présentation de l'algorithme EM . . . 30
  - 3.2 Étape E . . . 31
  - 3.3 Étape M . . . 34
  - 3.4 Initialisation et arrêt de l'algorithme EM . . . 36
  - 3.5 Variantes et extensions de l'algorithme EM . . . 37
- 4 Sélection de modèles . . . 38
- 5 Annexes . . . 41
  - 5.1 Définitions et propriétés des chaînes de Markov . . . 41
  - 5.2 Démonstrations des formules de l'algorithme Forward/Backward . . . 42

Bibliographie 44

3 Modèles de Markov cachés pour les données *tiling arrays* 47
1 Modélisation de la loi de la variable latente 48
2 Modélisation de la loi d'émission 51
2.1 Régressions linéaires 51
a) Modèle 52
b) Inférence 53
2.2 Loi gaussienne bidimensionnelle 53
a) Modèle 54
b) Inférence 57
2.3 Mélange de gaussiennes 58
a) Modèle 59
b) Inférence 62
c) Sélection de modèles 64
d) Initialisation de l'algorithme EM 64
e) Étude de simulation. 67
3 Annexes 72
3.1 Estimateurs du modèle gaussien bidimensionnel sous contraintes . . 72
3.2 Estimateurs du modèle gaussien unidimensionnel sous contraintes de colinéarité 74
Bibliographie 76

4 Classification 79
1 Règle du MAP 81
1.1 Seuil de classification 81
1.2 Association de groupes 82
2 Contrôle des faux-positifs 83
3 Classification de régions 86
3.1 Probabilités *a posteriori* pour une région 86
3.2 Règle de classification 88
Bibliographie 89

5 Applications 91
1 Contexte et outils 92
1.1 Plante modèle *Arabidopsis thaliana* 92
1.2 Caractéristiques de la puce 93
1.3 Normalisation des données 95
1.4 Visualisation dans FLAGdb++ 96
2 Analyse de données de ChIP-chip 97
2.1 Étude de H3K9me3 98
a) Analyse avec ChIPmix 98
b) Comparaison de méthodes 100
2.2 Étude de l'épigénome d'*Arabidopsis thaliana* avec MultiChIPmix . . 101
2.3 Package ChIPmix 102
3 Analyse de données de ChIP-chip IP/IP - Étude de H3K9me2 103
3.1 Analyse avec le modèle gaussien bidimensionnel $\mathcal{M}_2$ 104
3.2 Comparaison avec la méthode de Johannes *et al.* (2010) 106
3.3 Analyse avec le modèle de mélanges de mélange 108
a) Méthode avec contraintes de colinéarité 108
b) Méthode sans contraintes de colinéarité 111

4 Analyse de données transcriptome - Étude des données graine vs feuille . . 113
4.1 Comparaison de modèles 113
4.2 Analyse avec le modèle gaussien bidimensionnel $\mathcal{M}_4$ 116
4.3 Classification par gène 117
4.4 Détection de nouveaux transcrits 118
4.5 Package TAHMMAnnot 120
5 Étude de simulations - Comparaison de méthodes 121
Bibliographie 124

Conclusion et perspectives **127**

A Article publié dans *Bioinformatics* **131**

B Article publié dans *La Revue de Modulad* **139**

C Article publié dans *The Plant Journal* **155**

D Article publié dans *EMBO* **167**

E Article publié dans *SAGMB* **179**

Bibliographie **205**
Bibliographie 205

Table des figures

1.1 Complémentarité des bases dans la double hélice d'ADN. 9
1.2 Repliement (condensation) de la molécule d'ADN par spiralisation autour des histones jusqu'à obtention de chromosomes. 9
1.3 Structure d'un gène eucaryote. 11
1.4 Fabrication des puces à ADN. 14
1.5 Acquisition des données transcriptome. 16
1.6 Description de la technique de ChIP-chip. 17
1.7 Description de la technique de ChIP-chip IP/IP. 18
1.8 Représentation schématique des deux groupes à définir dans une expérience de ChIP-chip. 19
1.9 Représentation schématique des quatre groupes à définir dans une expérience de transcriptome. 19
1.10 Visualisation de l'intensité du signal. 20
1.11 Exemple d'annotation structurale sur une région d'un genome. 20

2.1 Illustration du problème de classification. 28
2.2 Graphe des dépendances conditionnelles dans un modèle de mélange. . . . 29
2.3 Graphe des dépendances conditionnelles d'une chaîne de Markov cachée d'ordre 1. 29

3.1 Définition des sous-modèles. 48
3.2 Graphe des dépendances conditionnelles markoviennes d'ordre 1 avec prise en compte de l'annotation. 50
3.3 Données de ChIP-chip : IP contre INPUT. 52
3.4 Bimodalité de la distribution des log-ratios. 53
3.5 Représentation schématique des quatre groupes dans le cas d'une expérience de transcriptome et de ChIP-chip IP/IP. 54
3.6 Isodensité et classement des sondes en quatre groupes pour le modèle $\lambda_k D_k A_k D_k^T$. 55
3.7 Isodensité et classement des sondes en quatre groupes pour le modèle $\lambda D_k A_k D_k^T$. 55
3.8 Explication schématique de la modélisation. 56
3.9 Distribution des données projetées sur les grands axes du groupe identique et des groupes différentiellement hybridés. 59
3.10 Représentation schématique des mélanges de gaussiennes dans chaque groupe, le long des trois axes. 61
3.11 Représentation schématique des composants gaussiens sphériques sans contraintes. 61
3.12 Exemple de données simulées. 67
3.13 Valeurs du MSE pour chaque condition de simulation. 69
3.14 Écart-type du MSE pour chaque condition de simulation. 69

4.1 Classification à deux groupes où les sondes sont colorées en fonction des probabilités *a posteriori* . . . 80
4.2 Classement des sondes en quatre groupes avec un seuil de classification égal à 0.7. . . . 81
4.3 Classement des sondes en 4 groupes, 3 groupes et 2 groupes. . . . 82
4.4 Classement des sondes en 4 groupes, 3 groupes et 2 groupes avec un seuil de classification égal à 0.7. . . . 83
4.5 Représentation schématique d'un gène avec Q exons . . . 87

5.1 *Arabidopsis thaliana.* . . . 93
5.2 Histogramme des Tm. . . . 94
5.3 Description des différentes possibilités d'annotation pour une sonde. . . . 95
5.4 Capture d'écran de FLAGdb++. . . . 97
5.5 Visualisation des résultats statistiques sous forme de couleur sur chaque sonde. . . . 97
5.6 Résultats de la méthode de régression pour le chr4 (Rep1), avec les droites de régression estimées. . . . 99
5.7 Résultats de la méthode de régression pour le chromosome chloroplastique. 99
5.8 Comparaison des résultats avec $SignalMap^{TM}$ sur une région du chromosome 4. . . . 100
5.9 Diagramme de Venn résumant les résultats des trois méthodes. . . . 101
5.10 Graphe H3K9me2 obtenu avec le modèle HMM $\mathcal{M}_2$ après classification par la règle du MAP. . . . 105
5.11 Représentation de l'élément transposable META1 à l'aide des résultats du modèle $\mathcal{M}_2$. . . . 106
5.12 Comparaison de la classification entre les deux modèles de Johannes *et al.* (2010) et le modèle de mélange $\mathcal{M}_1$ et le HMM $\mathcal{M}_2$ sur le jeu de données H3K9me2. . . . 107
5.13 Distribution des données projetées sur les petits axes du groupe identique et des groupes différentiellement hybridés. . . . 108
5.14 Distribution des données projetées sur les grands axes des groupes bruit, identique et différentiellement hybridés avec les ajustements des densités estimées. . . . 109
5.15 Graphe H3K9me2 obtenu avec le modèle sous contraintes de colinéarité avec une variance résiduelle différente et avec six composants par groupe, après classification par la règle du MAP. . . . 110
5.16 Différence de classement entre les modèles $\mathcal{M}_2$ et 1666-$\sigma_1^2\sigma_2^2$. . . . 111
5.17 Représentation de l'élément transposable META1 à l'aide des résultats du modèle 1666-$\sigma_1^2\sigma_2^2$. . . . 111
5.18 Représentation du HMM initial avec 40 composants. . . . 112
5.19 Représentation de l'élément transposable META1 . . . 112
5.20 Représentation de l'ajustement des densités pour chaque groupe. . . . 113
5.21 Exemple d'un gène déclaré sur-exprimé, où les sondes introniques ont tendance à être lissées avec le modèle HMM. . . . 115
5.22 Comparaison des quatre modèles sur deux régions où les gènes sont exprimés. 115
5.23 Visualisation des résultats statistiques avec le logiciel FLAGdb++. . . . 117
5.24 Exemple de trois introns déclarés sous-exprimés avec une probabilité *a posteriori* supérieure à 0.75. . . . 118
5.25 Exemple d'une région correspondant à une extension de gène. . . . 120

5.26 Exemple d'une région correspondant à une région identiquement exprimée dans l'intergénique. 120

Liste des tableaux

3.1 Taux de bonne classification et leur écart-type. 70
3.2 Pourcentage d'estimations correctes du nombre de groupes pour chaque condition de simulation. 71

5.1 Tableau des estimateurs pour le chromosome 4, pour les deux réplicats biologiques. 98
5.2 Estimation des paramètres de la loi d'émission gaussienne du modèle HMM $\mathcal{M}_2$. 105
5.3 Comparaison des modèles en fonction du nombre de composants par groupe. 108
5.4 Comparaison des modèles avec une variance résiduelle différente, en fonction du nombre de composants par groupe. 109
5.5 Table de contingence entre le modèle HMM $\mathcal{M}_2$ et le modèle sous contraintes de colinéarité avec une variance résiduelle différente et avec six composants par groupe. 110
5.6 Ajustement des quatre modèles. 114
5.7 Proportions de sondes dans les quatre groupes pour chaque type d'annotation. 116
5.8 Matrice de transition de la catégorie intergénique. 116
5.9 Matrice de transition de la catégorie intronique. 116
5.10 Matrice de transition de la catégorie exonique. 117
5.11 Exemple de classification de gènes. 118
5.12 Résultats obtenus sur le jeu de données H3K9me2 avec une forte proportion de sondes différentiellement exprimées. 124
5.13 Résultats obtenus sur le jeu de données issu de Penterman *et al.* (2007) avec une faible proportion de sondes différentiellement exprimées. 124

Résumé

Les puces *tiling arrays* sont des puces à haute densité permettant l'exploration des génomes à grande échelle. Elles sont impliquées dans l'étude de l'expression des gènes et de la détection de nouveaux transcrits grâce aux expériences de transcriptome, ainsi que dans l'étude des mécanismes de régulation de l'expression des gènes grâce aux expériences de ChIP-chip. Dans l'objectif d'analyser des données de ChIP-chip et de transcriptome, nous proposons une modélisation fondée sur les modèles à variables latentes, en particulier les modèles de Markov cachés, qui sont des méthodes usuelles de classification non-supervisée. Les caractéristiques biologiques du signal issu des puces *tiling arrays* telles que la dépendance spatiale des observations le long du génome et l'annotation structurale sont intégrées dans la modélisation. D'autre part, les modèles sont adaptés en fonction de la question biologique et une modélisation est proposée pour chaque type d'expériences. Nous proposons un mélange de régressions pour la comparaison de deux échantillons dont l'un peut être considéré comme un échantillon de référence (ChIP-chip), ainsi qu'un modèle gaussien bidimensionnel avec des contraintes sur la matrice de variance lorsque les deux échantillons jouent des rôles symétriques (transcriptome). Enfin, une modélisation semi-paramétrique autorisant des distributions plus flexibles pour la loi d'émission est envisagée. Dans un objectif de classification, nous proposons un contrôle de faux-positifs dans le cas d'une classification à deux groupes et pour des observations indépendantes. Puis, nous nous intéressons à la classification d'un ensemble d'observations constituant une région d'intérêt, telle que les gènes. Les différents modèles sont illustrés sur des jeux de données réelles de ChIP-chip et de transcriptome issus d'une puce NimbleGen couvrant le génome entier d'*Arabidopsis thaliana*.

Abstract

Tiling arrays make possible a large scale exploration of the genome with high resolution. Biological questions usually addressed are either the gene expression or the detection of transcribed regions which can be investigated *via* transcriptomic experiments, and also the regulation of gene expression thanks to ChIP-chip experiments. In order to analyse ChIP-chip and transcriptomic data, we propose latent variable models, especially Hidden Markov Models, which are part of unsupervised classification methods. The biological features of the *tiling arrays* signal, such as the spatial dependence between observations along the genome and structural annotation are integrated in the model. Moreover, the models are adapted to the biological question at hand and a model is proposed for each type of experiment. We propose a mixture of regressions for the comparison of two samples, when one sample can be considered as a reference sample (ChIP-chip), and a two-dimensional Gaussian model with constraints on the variance parameter when the two samples play symmetrical roles (transcriptome). Finally, a semi-parametric modeling is considered, allowing more flexible emission distributions. With the objective of classification, we propose a false-positive control in the case of a two-cluster classification and for independent observations. Then, we focus on the classification of a set of observations forming a region of interest such as a gene. The different models are illustrated on real ChIP-chip and transcriptomic datasets coming from a NimbleGen tiling array covering the entire genome of *Arabidopsis thaliana*.

Présentation générale

Depuis une quinzaine d'années, l'étude exhaustive de l'activité des génomes est envisageable grâce à l'émergence des technologies à haut débit. Ces technologies sont sans cesse perfectionnées et l'apparition des puces *tiling arrays* permet une exploration du génome entier à grande échelle. Les *tiling arrays* sont des puces à haute densité dont les sondes sont régulièrement espacées, couvrant l'intégralité du génome d'un organisme et ne nécessitant aucune connaissance biologique *a priori*. Les puces *tiling arrays* sont utilisées dans de nombreuses expériences, pour étudier les aberrations chromosomiques (CGH), les conditions d'expression des gènes (transcriptome) et les mécanismes de contrôle de l'expression des gènes (ChIP-chip). Elles sont des outils d'investigation de choix, et participent aussi à la mise en évidence de nouvelles unités transcriptionnelles. La finalité des expériences de ChIP-chip et de transcriptome est de comprendre les phénomènes de régulation et d'expression de gènes et d'améliorer l'annotation structurale de l'organisme étudié.

Compte tenu de la quantité d'informations générées et de l'abondance des mesures (environ 1 million de sondes par puce), une analyse manuelle devient très rapidement fastidieuse et source d'erreurs. Il est nécessaire de tenir compte des nombreux biais techniques existants, ainsi que de la variabilité biologique inhérente. L'exploitation des données ne peut se faire sans l'aide de procédures automatiques efficaces. Le recours aux moyens informatiques pour gérer et exploiter cette multitude de données est devenu indispensable et les outils bioinformatiques de visualisation des résultats sont essentiels pour faire face à la dimension du problème. Les méthodes mathématiques et statistiques sont devenues incontournables pour l'analyse et l'interprétation des données.

L'objectif de ma thèse est de proposer une modélisation précise des intensités d'hybridation générées par une puce *tiling array* pour caractériser les différences existant entre deux échantillons biologiques. En considérant une approche de classification non supervisée, j'ai développé des modèles à variables latentes dédiés à l'analyse de données de ChIP-chip et de transcriptome. Ils permettent la prise en compte de l'information biologique disponible : l'organisation séquentielle des sondes (dépendance spatiale) et l'annotation structurale du génome.

Les chapitres 1 et 2 sont des chapitres introductifs.

Le chapitre 1 expose les notions de biologie moléculaire utilisées dans ce travail ainsi que l'enjeu de l'étude de l'expression et du contrôle de l'expression des gènes. Nous expliquons le principe des *tiling arrays* et les différents types d'application, puis nous nous intéressons plus particulièrement à la reformulation de la question biologique en une question statistique.

Le chapitre 2 donne les fondements statistiques sur lesquels sont établis les différents

modèles développés dans cette thèse. Nous présentons les modèles à variables latentes, en particulier les modèles de Markov cachés, ainsi que l'algorithme EM utilisé pour l'estimation des paramètres.

Les chapitres 3 et 4 présentent les modèles à variables latentes développés dans cette thèse, en précisant la modélisation, l'inférence puis la classification.

Le chapitre 3 est consacré à la modélisation des deux signaux générés par une puce *tiling array* avec une approche de classification non supervisée. Nous modélisons la loi de la variable latente en prenant en compte toute l'information disponible concernant les sondes (dépendance spatiale et annotation), puis nous nous intéressons à la loi d'émission, qui sera choisie en fonction de la question biologique. Pour les expériences de ChIP-chip, la méthode ChIPmix (Martin-Magniette *et al.*, 2008) modélise par un mélange de deux régressions linéaires le signal de l'échantillon immuno-précipité conditionnellement au signal de référence de l'ADN génomique. Cette méthode a ensuite été généralisée pour prendre en compte l'information de plusieurs réplicats biologiques simultanément (MultiChIPmix). Pour l'analyse des expériences de transcriptome (différence entre deux conditions d'expression) ou pour la comparaison de deux échantillons immuno-précipités d'une expérience de ChIP-chip, nous proposons une modélisation jointe des deux signaux avec un mélange gaussien bidimensionnel (Bérard *et al.*, 2011). La connaissance biologique des données est prise en compte sous forme de contraintes sur les paramètres du modèle. Cette méthode répond simultanément aux deux questions soulevées lors d'une expérience transcriptome : la détection de régions transcrites et l'étude de la différence d'expression entre deux conditions. Enfin, nous proposons de nous affranchir de l'hypothèse de distribution gaussienne pour la loi d'émission en considérant une modélisation plus flexible, où la loi d'émission est un mélange de distributions gaussiennes. L'objectif est ainsi d'obtenir une meilleure estimation et donc de mieux définir la frontière de classification entre les groupes.

Le chapitre 4 s'intéresse à la classification des observations. Nous présentons d'abord les différentes possibilités de classification des sondes, puis nous proposons un contrôle de faux-positifs dans le cas d'observations indépendantes et pour une classification qui ne comporte que deux groupes. Sous hypothèse de dépendance markovienne, nous nous intéressons à la classification d'un ensemble d'observations constituant une région d'intérêt en généralisant la formule des probabilités *a posteriori* pour une observation. Cette procédure est utilisée pour classer des régions telles que les gènes, qui couvrent plusieurs sondes à la fois.

Le chapitre 5 concerne les applications des différentes méthodes proposées sur des jeux de données réelles issus de la puce permettant d'analyser le génome de la plante *Arabidopsis thaliana*. Nous avons analysé principalement trois types d'expériences (étude d'une marque chromatinienne pour le ChIP-chip et le ChIP-chip IP/IP, et étude de deux conditions d'expression pour le transcriptome), et nous présentons les résultats obtenus. Les apports de la méthode sont mis en évidence dans une étude de simulation, en comparaison avec d'autres méthodes proposées pour analyser les données *tiling arrays*. Nous présentons également les deux packages R associés aux méthodes. L'utilisation de ces méthodes par l'équipe épigénétique et épigénomique végétales de l'ENS dans un projet biologique d'analyse de différentes marques chromatiniennes chez *Arabidopsis thaliana* a donné lieu à deux publications (Moghaddam *et al.*, 2011 ; Roudier *et al.*, 2011).

Chapitre 1

Contexte biologique

Sommaire

1 Notions de biologie moléculaire 8

1.1 L'ADN 8

1.2 Séquençage des génomes, Annotation 10

1.3 L'expression et le contrôle de l'expression des gènes 12

2 Évolution des technologies haut-débit 12

2.1 Principe des puces à ADN 13

2.2 Les *tiling arrays* 15

3 Utilisation des *tiling arrays* 15

3.1 Expériences de Transcriptome 16

3.2 Expériences de ChIP-chip 17

4 Problématique 19

Bibliographie 21

Ce chapitre introductif présente les notions de biologie qui seront utiles dans cette thèse. La première partie est consacrée aux notions de biologique moléculaire telles que l'ADN, les gènes et les génomes (Section 1). La Section 2 s'intéresse au développement des technologies haut-débit et en particulier des puces à ADN et *tiling arrays* (puces de couverture ou pavages en français) qui permettent d'étudier l'activité d'un génome. Les différentes applications des puces *tiling arrays* ainsi qu'une revue des méthodes d'analyse font l'objet de la Section 3. La question biologique considérée dans cette thèse et le problème statistique associé sont présentés dans la Section 4.

1 Notions de biologie moléculaire

La biologie moléculaire est apparue au XXe siècle, à la suite de l'élaboration des lois de la génétique, la découverte des chromosomes et l'identification de l'acide désoxyribonucléique (ADN) comme support chimique de l'information génétique. Après la découverte de la structure en double hélice de l'ADN en 1953 par James Watson (1928-) et Francis Crick (1916-2004), la biologie moléculaire a connu d'importants développements pour devenir un outil incontournable de la biologie moderne à partir des années 1970. Elle a pour objectif la compréhension des mécanismes de fonctionnement de la cellule au niveau moléculaire. Les processus étudiés sont la réplication, la transcription et la traduction du matériel génétique.

1.1 L'ADN

L'acide désoxyribonucléique (ADN) est une molécule allongée, pouvant mesurer plusieurs centimètres de long. L'ADN est constitué de deux brins antiparallèles se faisant face, qui s'associent en formant une double hélice. Ces deux brins sont des séquences ordonnées de nucléotides complémentaires. Un nucléotide est composé d'un ou plusieurs groupements phosphates, d'un sucre (désoxyribose) et d'une base azotée. Il existe quatre bases azotées différentes : l'adénine (notée A) et la guanine (notée G) de la famille des Purines, et la thymine (notée T) et la cytosine (notée C) de la famille des Pyrimidines. Les bases azotées sont complémentaires deux à deux, une purique s'associant toujours à une pyrimidique. Ainsi, l'adénine est complémentaire à la thymine et la guanine est complémentaire à la cytosine. Les bases azotées complémentaires sont reliées entre elles par des liaisons hydrogènes. Il y a deux liaisons hydrogène entre A et T et trois entre C et G. Ces quatre bases azotées assurent donc la complémentarité des deux brins de la molécule d'ADN (cf. Figure 1.1). Grâce à l'alternance non aléatoire des quatre bases azotées A, C, T, G, toutes les séquences nucléotidiques constituent un message codé portant les informations génétiques.

Chez les procaryotes (comme les bactéries par exemple), l'ADN est généralement présent sous la forme d'un seul chromosome circulaire libre dans le cytoplasme de la celulle. Chez les eucaryotes, l'ADN est présent dans le noyau cellulaire, sous forme linéaire et scindé en plusieurs chromosomes. Pour permettre sa compaction, l'ADN est conditionné en chromatine. La chromatine correspond à l'association de l'ADN et de protéines qui sont principalement des histones. Dans un premier niveau d'organisation, quelques paires de bases d'ADN s'enroulent autour d'un octamère d'histones pour former un nucléosome. Dans un second niveau d'organisation, les nucléosomes se compactent et forment une hélice. Cette hélice est finalement condensée en euchromatine (condensation légère) ou en hétérochromatine (condensation prononcée) constituant le troisième

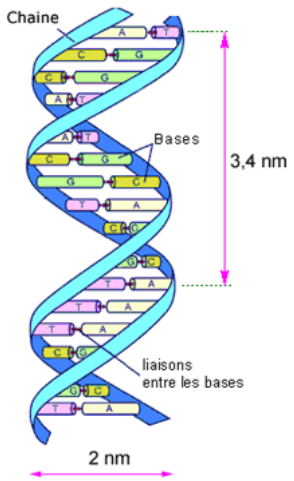

Fig. 1.1: Complémentarité des bases dans la double hélice d'ADN.

et dernier niveau d'organisation (cf. Figure 1.2). Une molécule d'ADN linéaire longue de 10 centimètres est ainsi condensée en un chromosome mesurant à peine 10 micromètres.

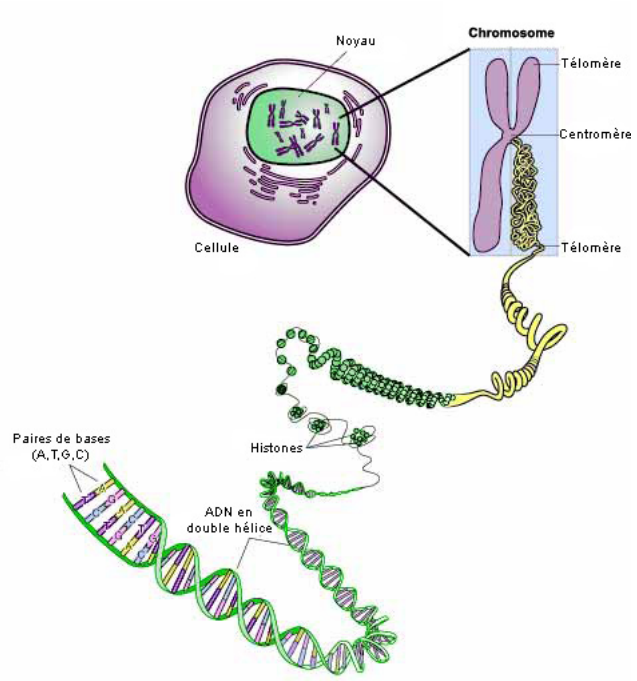

Fig. 1.2: Repliement (condensation) de la molécule d'ADN par spiralisation autour des histones jusqu'à obtention de chromosomes.

L'ADN est présent dans toutes les cellules vivantes et sa structure et ses propriétés chimiques lui permettent de remplir des fonctions importantes. Sa fonction principale est de stocker l'information génétique nécessaire au développement et au fonctionnement d'un organisme, contenue dans l'enchaînement non-aléatoire de nucléotides. Une autre fonction essentielle de l'ADN est la transmission de cette information de génération en génération lors de la reproduction (de manière intégrale ou non). C'est le support de l'hérédité. L'information portée par l'ADN peut être modifiée au cours du temps par des mutations dues principalement à des erreurs lors de la réplication des séquences d'ADN (ajout, délétion ou substitution de nucléotides), ou bien par des recombinaisons

génétiques. Cela aboutit à une diversité des individus et à une évolution possible des espèces grâce à la sélection naturelle.

L'ADN est indispensable à la synthèse des protéines, par l'intermédiaire de l'acide ribonucléique (ARN). Dans la cellule, l'ARN est produit par transcription à partir de l'ADN situé dans le noyau. L'ARN est donc une copie d'une région d'un brin d'ADN et permet la transmission de l'information génétique. L'ARN a de nombreuses similarités avec l'ADN, avec cependant quelques différences importantes : sur le plan de la structure, l'ARN contient un ribose à la place du désoxyribose de l'ADN, ce qui rend l'ARN chimiquement plus instable. La thymine de l'ADN est remplacée par l'uracile (notée U), qui possède les mêmes propriétés d'appariement de base avec l'adénine. Sur le plan fonctionnel, l'ARN est le plus souvent trouvé dans les cellules sous forme simple brin et les molécules d'ARN sont généralement plus courtes que celles d'ADN. L'ARN peut avoir différentes fonctions : il assure soit une fonction structurale, soit une fonction enzymatique (synthèse des protéines, exportation des protéines, etc.), soit une fonction de transport de l'information génétique. L'ARN de type messager (ARNm) est traduit par le ribosome (complexe nucléo-protéique) en séquences d'acides aminés qui forment les protéines. Les protéines sont des éléments essentiels qui remplissent des fonctions diverses au sein de la cellule : les protéines enzymatiques catalysent les réactions chimiques de la cellule, les protéines de structure permettent à la cellule de maintenir son organisation dans l'espace, les protéines de transport assurent le transfert de différentes molécules à l'intérieur et à l'extérieur des cellules, les protéines régulatrices jouent un rôle dans la régulation de l'expression des gènes (facteurs de transcription) ou pour la compaction de l'ADN (histones), etc. En fait, la majorité des fonctions cellulaires est assurée par des protéines.

1.2 Séquençage des génomes, Annotation

Le génome est l'ensemble du matériel génétique d'un individu. Il contient toutes les séquences d'ADN codantes (transcrites en ARN messagers et traduites en protéines, communément appelées gènes ou plus précisément gènes codant pour une protéine) et non-codantes (non transcrites, ou transcrites en ARN mais non traduites). La connaissance de la structure d'un génome passe par son séquençage, qui consiste à déterminer l'ordre d'enchaînement des nucléotides de l'ADN. La taille des génomes étant de plusieurs millions de bases (ou mégabases), il est nécessaire de coupler les approches de biologie moléculaire avec celle de l'informatique pour pouvoir séquencer des génomes entiers. Depuis 1989, d'importants progrès technologiques et informatiques ont permis d'atteindre aujourd'hui une vitesse globale de séquençage de 1 000 nucléotides par seconde.

Le premier véritable séquençage d'un génome est publié en 1972, il s'agit du virus bactériophage MS2. Des bactéries telles que *Escherichia coli* ou *Bacillus subtilis* ont été séquencées en 1997 en raison de leur importance dans le domaine de la recherche fondamentale, ou en raison de leur utilisation industrielle, en particulier dans le domaine agro-alimentaire. Un nombre important de génomes de procaryotes pathogènes ont aussi été séquencés. La priorité a essentiellement été donnée aux pathogènes de l'homme, puisque la moitié des maladies humaines est d'origine bactérienne. Ces dernières années, le séquençage des génomes procaryotes s'est révélé particulièrement productif. Un génome bactérien est séquencé tous les deux mois et une centaine de séquences complètes de génomes bactériens ont été obtenues. Chez les eucaryotes, les génomes de la levure *Saccharomyces cerevisiae* (1997), du ver nématode *Caenorhabditis elegans* (1998), et de la drosophile *Drosophila melanogaster* (2000) ont été les premiers séquencés. Ces

organismes ont été choisis en raison de la petite taille de leur génome, ainsi que pour leur intérêt économique ou leur utilisation dans le domaine de la recherche. Le premier séquençage complet du règne végétal concerne la plante modèle *Arabidopsis thaliana*, il a été terminé en 2000. La complexité et la grande taille des génomes des plantes rendent leur séquençage difficile. En plus d'*Arabidopsis thaliana*, seuls quatre génomes de plantes sont aujourd'hui entièrement séquencés (le riz *Oryza sativa* en 2005, le peuplier *Populus trichocarpa* en 2006, la vigne *Vitis vinifera* en 2007 et le soja *Glycine max.* en 2010). La publication de la première carte du génome humain est annoncée officiellement le 26 juin 2000 et le premier génome complet d'un individu est publié en 2007.

En général, l'objectif n'est pas seulement de séquencer un génome avec un taux d'erreur minimal, mais aussi d'identifier tous les gènes dans cette grande quantité de données. Le processus permettant d'identifier les limites entre les gènes et d'autres caractéristiques sur la séquence d'ADN brute est appelé annotation du génome. L'annotation d'un génome consiste à traiter l'information brute contenue dans la séquence :

- l'**annotation structurale** a pour but de prédire la structure des gènes et leur position sur le chromosome, ainsi que leur organisation (gènes uniques ou en opéron, avec des séquences promotrices, des terminateurs, etc.). Le problème crucial de l'analyse de séquences génomiques est l'identification des séquences codantes. L'identification des unités transcriptionnelles est plutôt facile chez les procaryotes grâce à la possibilité d'identifier relativement facilement les promoteurs des gènes. La fraction codante des génomes procaryotes est globalement très élevée, de l'ordre de 90%. On estime aujourd'hui que le génome d'un organisme procaryote tel qu'une bactérie comporte quelques milliers de gènes. Chez les eucaryotes, le découpage des gènes en introns et exons, et la présence de régions intergéniques, parfois très vastes, compliquent radicalement l'identification des séquences codantes. La plupart du temps, un gène commence par une séquence de nucléotides appelée promoteur, dont le rôle est de permettre l'initiation et surtout la régulation de la transcription de l'ADN en ARN, et se termine par une séquence terminatrice qui marque la fin de la transcription (cf. Figure 1.3).

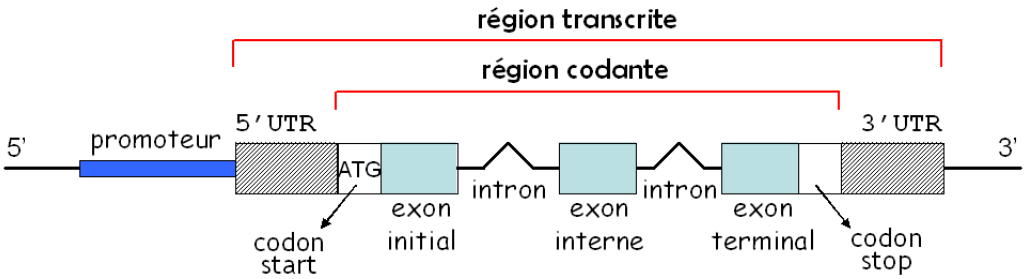

Fig. 1.3: Structure d'un gène eucaryote. UTR = région transcrite non traduite.

Le génome d'un eucaryote comporte entre environ 6 000 gènes pour la levure *Saccharomyces cerevisiae* jusqu'à environ 31 000 gènes pour la daphnie rouge, aussi appelée puce d'eau (*Daphnia pulex*). L'Homme possède environ 23 000 gènes.

- l'**annotation fonctionnelle** a pour but de prédire la fonction potentielle de ces gènes et leurs interactions probables. L'annotation peut être automatique, c'est-à-dire s'appuyer sur des algorithmes recherchant des similarités (de séquences, de structures, de motifs, ...), permettant de prédire la fonction d'un gène. Elle aboutit au transfert « automatique » de l'information figurant dans l'étiquette d'un gène similaire d'un génome déjà annoté au génome en cours d'annotation.

L'annotation automatique est parfois complétée par une annotation manuelle réalisée par des experts qui valident ou invalident la prédiction en fonction de leurs connaissances ou de résultats expérimentaux. Celle-ci peut ainsi éviter le transfert automatique d'erreurs et donc leur propagation.

L'enjeu actuel est de séquencer un grand nombre d'espèces, d'améliorer et de compléter l'annotation des génomes, ainsi que d'étudier les gènes afin d'accéder à une meilleure compréhension de leur fonction. L'exploration de la fonction des gènes peut par exemple permettre d'importants progrès médicaux grâce à l'identification de gènes responsables de maladies génétiques. Chez *Arabidopsis thaliana*, l'identification des gènes impliqués dans la libération des graines et des gènes responsables de la maturation des fruits a déjà des retombées chez les espèces cultivées (Bernot *et al.*, 2001 ; Wilkinson *et al.*, 1997).

1.3 L'expression et le contrôle de l'expression des gènes

L'avancée technologique en matière de séquençage des génomes et d'annotation a permis notamment l'étude et la caractérisation des gènes. Un gène est une séquence d'ADN conçue pour être transcrite en ARN. La molécule d'ARN ainsi produite peut soit être traduite en protéine (elle est dans ce cas appelée ARN messager), soit être directement fonctionnelle (c'est le cas pour les ARN ribosomaux ou les ARN de transfert). Chez les eucaryotes, un gène est constitué d'une alternance de séquences codantes, nommées exons, et de séquences non codantes, les introns, qui seront éliminés de l'ARN messager lors du processus d'épissage, avant la traduction en protéine. Le processus d'épissage des introns permet aussi de supprimer de façon conditionnelle certains exons de l'ARN, donnant la possibilité à partir d'un unique gène de produire plusieurs protéines différentes. On parle alors d'épissage alternatif.
Pour la plupart des organismes eucaryotes, une bonne partie de l'ADN n'est pas codante. Cet ADN non codant, aussi nommé ADN intergénique, est de plus en plus étudié et semble être impliqué dans la régulation de l'expression des gènes, notamment par modification du niveau de condensation de l'ADN. En effet, le conditionnement de l'ADN en chromatine est essentiel pour la régulation de l'activité du génome chez les eucaryotes. La structure organisationnelle de l'ADN dans le noyau constitue en elle-même un mécanisme de répression ou d'activation de la transcription des gènes. Pour activer la transcription d'un gène donné dans une cellule, la chromatine comprise dans la région de contrôle du gène doit être modifiée ou altérée de façon à être permissive à la transcription. Les gènes localisés dans l'euchromatine sont préférentiellement transcrits car la condensation est légère. Ces modifications de l'état transcriptionnel de la chromatine sont généralement en lien direct avec une modification imputant les histones (méthylation, acétylation, ubiquitination ou phosphorilation). Les modifications d'histone, la méthylation d'ADN et d'autres facteurs tels que les petits ARN ou les enzymes de remodelage de la chromatine, définissent des états chromatiniens distincts qui modulent l'accès à l'ADN (Berger, 2007) et sont donc des mécanismes impliqués dans la régulation de l'expression des gènes. L'un des défis actuels en biologie est de comprendre le rôle de chaque gène au sein d'une cellule et de connaître ses conditions d'expression.

2 Évolution des technologies haut-débit

Depuis le début des années 1960, les biologistes moléculaires ont appris à caractériser, isoler et manipuler l'ADN, support de l'information génétique, ainsi que l'ARN et les protéines, molécules structurelles et enzymatiques les plus importantes des cellules. La

majorité des techniques de biologie moléculaire sont fondées sur l'hybridation, technique qui repose sur la complémentarité des bases azotées entre elles. En effet, si l'on place deux simples brins complémentaires d'ADN dans un même milieu, ils vont naturellement s'associer.
En 1977, deux techniques de séquençage des acides nucléiques apparaissent à peu près simultanément : la méthode enzymatique de Frédérick Sanger et l'approche chimique de Walter Gilbert et Allan Maxam. Grâce aux connaissances acquises sur les enzymes, la première est préférée à la seconde, qui est trop toxique. En 1984, la technique d'amplification génétique, ou PCR (Polymerase Chain Reaction), qui permet d'amplifier sélectivement toute séquence d'ADN, devient rapidement un outil puissant et indispensable au séquençage des génomes puisqu'il faut plusieurs millions de copies d'une molécule d'ADN pour pouvoir analyser sa séquence.
Depuis une vingtaine d'années, diverses techniques ont été développées afin d'aborder l'étude de l'expression des gènes. Les premières approches proposées, le Southern blot et le northern blot, permettent d'identifier et de localiser une séquence particulière d'ADN ou d'ARN dans un génome entier. Le Southern blot (respectivement northern) est une technique de transfert des molécules d'ADN (respectivement ARN) sur une membrane, les molécules étant préalablement séparées par taille à l'aide d'une électrophorèse. La détection est réalisée par hybridation avec un ou plusieurs fragments d'ADN marqués, de séquence connue. Ces techniques se limitent à l'analyse d'un petit nombre de gènes à la fois et ne permettent pas d'appréhender la complexité du phénomène de la transcription. Plus récemment, la technique SAGE (Serial Analysis of Genes Expression, Velculescu *et al.*, 1995), permet d'identifier et de quantifier, simultanément, le niveau d'expression de plusieurs milliers de gènes, dans un type cellulaire donné. Cette méthode consiste à réaliser un inventaire des transcrits par séquençage en série de courts fragments d'ADN complémentaire (ADNc). Cette méthode est très sensible, mais aussi longue à mettre en œuvre et coûteuse. Parallèlement à la méthode SAGE, sont apparues, vers la fin des années 1990, les puces ou *microarrays*, moins coûteuses et surtout plus évolutives en termes d'applications. Elles sont rapidement devenues un outil privilégié et la technologie des puces est extrêmement répandue aujourd'hui. Il existe différents types de puces pour étudier tous les niveaux moléculaires : puces à ADN, puces à protéines, etc. Dans la suite, nous nous intéressons uniquement aux puces à ADN (Schena *et al.*, 1995).

2.1 Principe des puces à ADN

Une puce à ADN est une collection de milliers de puits microscopiques sur un support solide miniature tel qu'une lame de microscope (qui peut être du verre, du silicium ou du plastique). Le principe de base des puces à ADN est l'hybridation moléculaire : un mélange complexe d'ADN est marqué puis hybridé avec des ADN complémentaires de séquence connue, fixés sur la surface d'un support solide à des positions déterminées. Plus précisément, un ensemble de fragments d'ADN d'intérêt, simple brin et amplifié par la technique de PCR, est déposé sur la puce dans chaque puits. Ces molécules d'ADN fixées sont appelées "sondes". L'hybridation consiste à placer des molécules "cibles", extraites d'une culture que l'on souhaite analyser et marquées avec un fluorochrome, sur le support où sont fixées les sondes. Les cibles, qui sont aussi des molécules sous forme simple brin, vont naturellement s'apparier aux sondes de séquence complémentaire pour donner de l'ADN double brin. On suppose que la sonde est non restrictive, et en quantité largement suffisante pour accueillir toutes les cibles présentes dans l'échantillon analysé. La quantité d'ADN fluorescent hybridé est donc supposée proportionnelle à la quantité de la molécule cible correspondante dans la cellule de départ. Chaque sonde

de la puce est ensuite excitée par un laser pour récupérer la fluorescence émise via un photo-multiplicateur (PMT) couplé à un microscope confocal. On mesure la quantité de signal dans la longueur d'onde d'émission du fluorochrome. Une image dont le niveau de gris représente l'intensité de la fluorescence lue est alors obtenue. Chaque pixel de l'image scannée représente une mesure de fluorescence (cf. Figure 1.4).
Il existe deux grandes familles de puces à ADN, l'une ne pouvant recevoir qu'un échantillon de cellule par lame et l'autre pouvant recevoir deux échantillons différents (correspondant à deux conditions que l'on souhaite comparer), chacun labellisé avec un fluorochrome de couleur différente. Pour ces dernières, généralement appelées puces deux couleurs, deux images en niveau de gris sont générées (une pour chaque fluorochrome). Traditionnellement, on ne visualise finalement qu'une seule image, obtenue en superposant les deux précédentes et en remplaçant les niveaux de gris par des niveaux de vert pour la première image et des niveaux de rouge pour la seconde. Les sondes peuvent donc être représentées avec des couleurs allant du vert (seulement de l'ADN de la première condition fixé) au rouge (seulement de l'ADN de la seconde condition fixé) en passant par le jaune (ADN des deux conditions fixé en quantité égale).

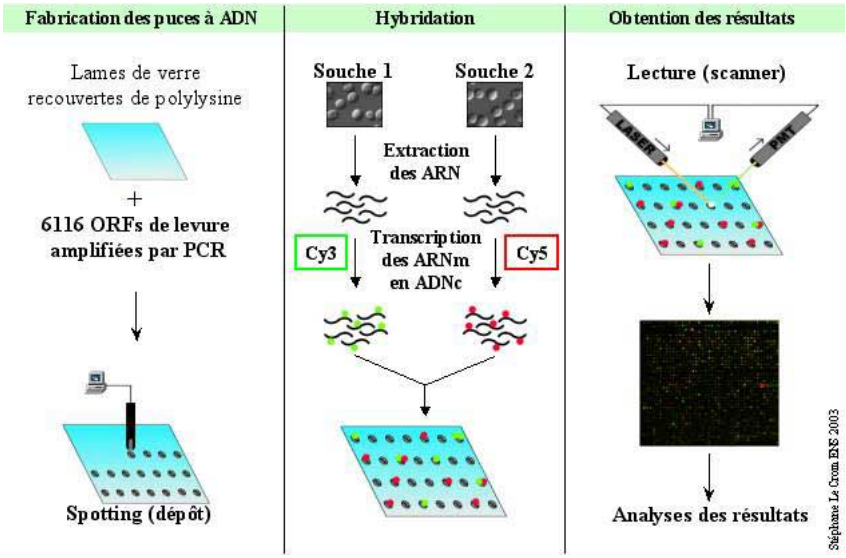

Fig. 1.4: Fabrication des puces à ADN.

Différents types de puces à ADN

Il existe différents types de puces à ADN qui ont des applications différentes en fonction des séquences déposées sur la puce. Nous en présentons brièvement deux.

Les puces SNP (Single Nucleotide Polymorphism) permettent de faire de la génétique à haute résolution. Le polymorphisme nucléotidique est la variation (mutation) d'une seule paire de bases du génome entre individus d'une même espèce. Les SNP représentent 90% de l'ensemble des variations génétiques humaines. Les sondes sont fabriquées à partir d'oligonucléotides synthétisés ayant la séquence du SNP pour chaque mutation connue. Puis les sondes sont hybridées avec l'ADN d'un individu.

Les puces d'expression permettent l'analyse de données d'expression de gènes à l'aide de milliers de sondes représentant les gènes. Les cibles sont alors des ARNm transformés en ADNc par transcription inverse. Les mesures du niveau d'expression de plusieurs milliers de gènes dans un type cellulaire et dans un contexte physiologique et/ou pathologique particulier permettent l'analyse simultanée de l'ensemble des transcrits. La mesure

à grande échelle de l'expression génique est motivée notamment par l'hypothèse que l'état fonctionnel d'un organisme est en grande partie décrit par la quantité de chaque ARN présent dans la cellule à un instant donné.

2.2 Les *tiling arrays*

Depuis une dizaine d'années, les puces à ADN se perfectionnent pour proposer une couverture intégrale du génome, ce sont les *tiling arrays*. La technique reste la même, mais des centaines de milliers de sondes peuvent être fixées sur une même puce. En fonction de la longueur de la sonde et de l'espacement entre sondes, différents degrés de résolution peuvent être obtenus. Les sondes ne représentent plus uniquement des gènes mais couvrent l'intégralité du génome d'un organisme indépendamment de son annotation structurale.
Il existe deux techniques principales de fabrication des *tiling arrays*. La première est la photolithographie : elle implique une synthèse *in situ* où les sondes, d'environ 25bp, sont construites sur la surface de la puce. La seconde technique consiste en l'impression mécanique des sondes sur la puce en utilisant des machines qui placent précisément les sondes déjà synthétisées. Les trois fabricants principaux de *tiling arrays* sont Affymetrix, NimbleGen (Roche) et Agilent.
Bien que la densité des *tiling arrays* ne cesse d'augmenter pour atteindre quasiment deux millions de sondes aujourd'hui, il restera toujours des limites pratiques imposées par le nombre de sondes qui peuvent être synthétisées sur une puce. La couverture complète d'un grand génome eucaryote ($\sim$ 3 Go) peut théoriquement être obtenue mais un très grand nombre de puces serait alors nécessaire et de telles expériences ne sont pas envisageables ne serait-ce que d'un point de vue financier (Mockler et Ecker, 2005). Récemment, la technologie de Next Generation Sequencing (NGS) révolutionne le domaine car elle produit directement des séquences nucléotidiques sans utilisation d'un support. La nouvelle génération de séquenceurs à très haut débit permet de séquencer, en quelques jours, plusieurs gigabases d'ADN composés de courts fragments (35-50 nt).

3 Utilisation des *tiling arrays*

Les *tiling arrays* sont utilisées dans de nombreux types d'applications comme l'étude de l'expression des gènes avec les expériences de transcriptome (Mockler et Ecker, 2005 ; Yamada *et al.*, 2003 ; Hanada *et al.*, 2007). La technologie des puces *tiling arrays* permet l'étude exhaustive de l'activité transcriptionnelle d'un génome. Cet outil permet d'annoter de manière plus complète la fraction transcrite du génome en mettant en évidence de nouvelles unités transcriptionnelles qui avaient échappé aux méthodes d'annotation classiques en raison de leur originalité structurale (petite taille, antisens, gènes à ARN, etc., Aubourg et Rouzé, 2001). De telles puces sont aussi utilisées pour les expériences de ChIP-chip qui permettent d'étudier les interactions protéines/ADN, et en particulier la méthylation d'ADN, les modifications de la chromatine ou les facteurs de transcription (Buck and Lieb, 2004). D'autre part, les expériences CGH (Comparative Genomic Hybridization) permettent la détection d'altérations chromosomiques (Pinkel *et al.*, 1998 ; Snijders *et al.*, 2001). Elles ont été développées pour étudier les délétions, insertions et amplifications des segments d'ADN dans un génome. Différentes approches existent pour déterminer les variations du nombre de copies d'ADN comme la segmentation (Hupé *et al.*, 2004 ; Picard *et al.*, 2005) ou les modèles de Markov cachés (Fridlyand *et al.*, 2004 ; Seifert *et al.*, 2009).

Dans cette thèse, nous nous intéressons uniquement aux expériences de transcriptome et de ChIP-chip qui permettent d'étudier respectivement l'expression des gènes et leurs mécanismes de contrôle.

3.1 Expériences de Transcriptome

Le transcriptome est l'ensemble des ARN messagers (ARNm) produits lors du processus de transcription d'un génome, dans une condition donnée. Les ARN messagers sont une copie des gènes qui sont à l'état actif dans la cellule. La plupart d'entre eux sont traduits en protéines pour participer aux diverses fonctions de la cellule. Bien que les ARNm ne constituent qu'une étape de l'expression des gènes, leur abondance est souvent corrélée à l'activité des protéines codées et leur quantification est plus facile à réaliser que celle des protéines. La quantification systématique des ARNm permet d'avoir une indication relative du taux de transcription des différents gènes dans une condition donnée. Ainsi, la caractérisation et la quantification du transcriptome dans une condition donnée permet d'identifier les gènes actifs dans la cellule et les familles ou réseaux fonctionnels de gènes mis en jeu. Les *tiling arrays* sont utilisées pour mesurer simultanément le niveau d'expression d'un grand nombre d'ARNm dans un but d'identification des gènes ou pour la comparaison de différents états biologiques (cf. Figure 1.5).

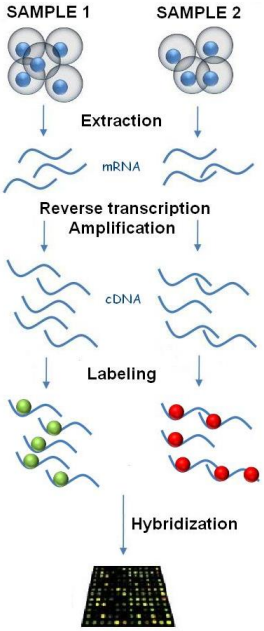

Fig. 1.5: Acquisition des données transcriptome.

Les expériences de transcriptome ont deux objectifs différents : la détection de régions transcrites et l'étude de l'expression des gènes entre plusieurs conditions (analyse différentielle). La plupart des méthodes développées s'intéressent à la détection de régions transcrites. Certaines sont fondées sur les tests statistiques sonde par sonde (par exemple le test de Fisher développé par Halasz *et al.*, 2006), d'autres sont des méthodes de segmentation (Huber *et al.*, 2006 ; Zeller *et al.*, 2008) ou des HMM (Nicolas *et al.*, 2009). L'intégration de l'annotation comme connaissance *a priori* a été proposée dans un cadre supervisé (Du *et al.*, 2006 ; Munch *et al.*, 2006). Étonnamment, peu de méthodes sont consacrées à l'étude des profils d'expression de gènes selon différentes conditions. La méthode gSAM (Ghosh *et al.*, 2007) est une extension de SAM, qui modélise l'expression différentielle pour une région donnée (ensemble de sondes agrégées) par une fonction

constante par morceaux. Dans la méthode TileMap (Ji et Wong, 2005) un test statistique est proposé pour chaque sonde séparément, fondé sur un modèle bayésien empirique hiérarchique ; cette méthode permet la comparaison multiple d'échantillons.

3.2 Expériences de ChIP-chip

L'immunoprécipitation de la chromatine permet d'étudier les interactions entre les protéines et l'ADN ainsi que différents états chromatiniens associés à des états d'activité distincts du génome. Cela participe à l'étude des mécanismes de contrôle de l'expression des gènes de manière générale. Le ChIP-chip est une combinaison de la technique d'immunoprécipitation de la chromatine (ChIP) avec le principe des puces à ADN (chip), ce qui permet une étude à l'échelle du génome des interactions protéines-ADN. Cela permet en particulier d'étudier les modifications d'histone et la méthylation d'ADN, ainsi que de localiser les sites de fixation des facteurs de transcription.
L'originalité et l'intérêt de cette méthode viennent du fait que l'extraction d'ADN se fait *in vivo*, ce qui permet d'avoir une idée plus réaliste des processus à l'oeuvre dans les cellules lors de l'initiation de la transcription. La technique d'immunoprécipitation de la chromatine nécessite l'utilisation d'un anticorps spécifique de la protéine étudiée, et on obtient, après plusieurs étapes d'un protocole précis (précipitation des complexes ADN-protéine-anticorps, purification de l'ADN, séparation du complexe ADN-protéine, etc.), une collection de fragments d'ADN d'assez courte taille et dont on sait qu'ils interagissent avec la protéine sélectionnée par l'anticorps. La partie ChIP est suivie d'une expérience classique sur puce où l'ADN immunoprécipité (fragments liés à la protéine d'intérêt), noté IP, et l'ADN total (noté INPUT) sont marqués avec un fluorochrome puis hybridés sur la puce (cf. Figure 1.6). L'objectif est ensuite de comparer les intensités obtenues pour chacun des deux échantillons afin d'identifier les régions génomiques où la protéine d'intérêt interagit avec l'ADN. Le signal INPUT étant présent partout sur la puce (puisqu'il représente l'ADN total), cela revient à détecter les sondes pour lesquelles il y a un signal IP plus fort que le signal INPUT (régions enrichies).

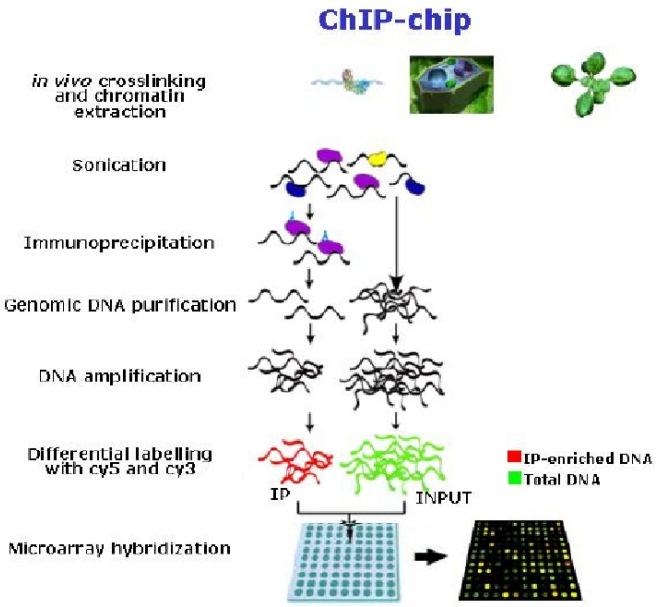

Fig. 1.6: Description de la technique de ChIP-chip.

La finalité des expériences de ChIP-chip est de détecter les régions du génome liées à une protéine et donc enrichies par immunoprécipitation de la chromatine. Diverses méthodes ont été développées pour trouver ces régions. Johnson *et al.* (2006) ont proposé un modèle linéaire fondé sur les caractéristiques de séquence des sondes. Leur algorithme, appelé MAT (Model-based Analysis of Tiling arrays), est dédié à l'analyse des données de puces Affymetrix. Li *et al.* (2005) ont proposé un modèle de Markov caché à deux états, utilisé pour estimer la probabilité d'enrichissement de chaque sonde. Ces deux méthodes font l'hypothèse que la proportion de sondes enrichies à détecter est faible. Cette hypothèse est raisonnable dans le cas des études de facteur de transcription, mais pas pour l'étude de modification d'histone ou de méthylation d'ADN où un fort enrichissement est attendu. Humburg, Bulger et Stone (2008) ont suggéré une procédure d'estimation de paramètres d'un modèle de type HMM pour l'étude de la structure de la chromatine où les régions d'intérêt attendues sont longues et nombreuses. Les données de ChIP-chip peuvent aussi être analysées comme un signal le long du génome dans lequel on recherche des pics, en utilisant le log-ratio entre les intensités des deux échantillons. Les analyses sont alors souvent effectuées en utilisant une fenêtre glissante (Cawley *et al.*, 2004) et des tests statistiques. Keles *et al.* (2004) et He *et al.* (2009) ont proposé respectivement un test de Welch et une méthode de Wilcoxon non paramétrique de la somme des rangs pour comparer le signal entre les deux échantillons.

Habituellement dans une expérience de ChIP-chip, les deux échantillons co-hybridés sont les fragments d'ADN associés à la protéine d'intérêt ou à une marque chromatinienne (IP) et l'ADN génomique total (INPUT) (comme décrit ci-dessus). Dans ce type d'expérience, le résultat est binaire (enrichi ou pas) et ne permet pas une évaluation quantitative du signal. La technique du ChIP-chip permet également d'étudier directement la différence entre deux échantillons d'ADN immunoprécipités (correspondant à deux conditions distinctes), sans hybrider sur la puce l'ADN génomique total. Cette technique est appelée ChIP-chip IP/IP (cf. Figure 1.7). Elle permet de caractériser une différence d'enrichissement entre deux échantillons qui ont alors un rôle symétrique. Johannes *et al.* (2010) ont proposé un modèle de mélange gaussien bidimensionnel avec des contraintes sur les paramètres de moyenne pour étudier la différence d'enrichissement.

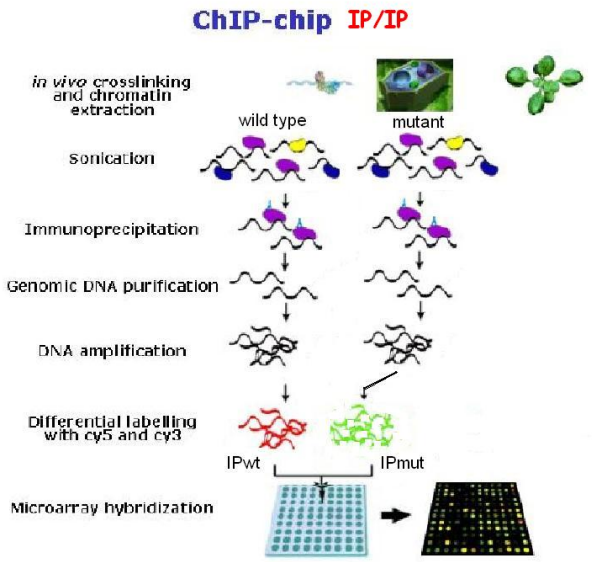

Fig. 1.7: Description de la technique de ChIP-chip IP/IP.

4 Problématique

L'analyse de données de puces *tiling arrays* nécessite le développement de méthodes statistiques adaptées pour la comparaison de deux échantillons issus d'expériences de ChIP-chip (IP/INPUT), de ChIP-chip IP/IP (IP condition 1/IP condition 2), ou de transcriptome (condition 1/condition 2).

Nous nous intéressons à la comparaison de deux échantillons d'un point de vue de classification non supervisée pour caractériser la différence de comportement des sondes. Cela revient à déterminer un statut pour chaque sonde. Pour les données de ChIP-chip, il y a deux statuts possibles pour chaque sonde : enrichi ou normal (cf. Figure 1.8). Pour les données transcriptomiques ou de ChIP-chip IP/IP, les deux échantillons sont symétriques, cela revient à caractériser quatre comportements différents pour chaque sonde : non hybridé, hybridé identiquement dans les deux échantillons ou bien hybridé préférentiellement dans l'un des deux échantillons (cf. Figure 1.9).

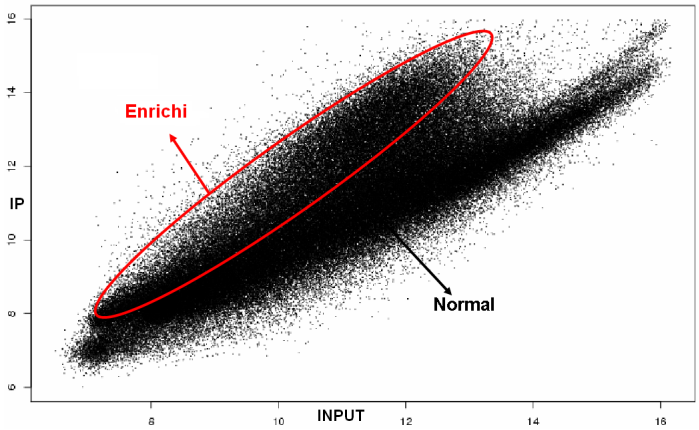

Fig. 1.8: Représentation schématique des deux groupes à définir dans une expérience de ChIP-chip.

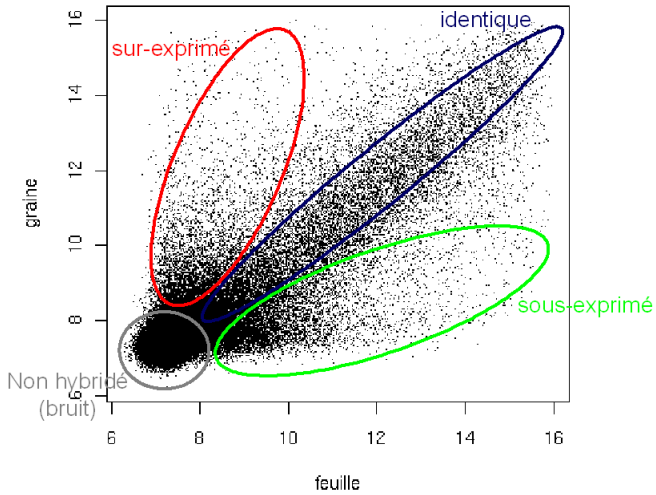

Fig. 1.9: Représentation schématique des quatre groupes à définir dans une expérience de transcriptome.

La plupart des méthodes développées sont fondées sur le log-ratio des intensités des deux conditions (cf. Section 3). Cependant, l'utilisation du log-ratio peut masquer la multimodalité des données en raison de la réduction de dimension. Pour s'affranchir de ce problème, il est préférable de travailler directement avec les deux mesures de chaque sonde. Nous proposons une modélisation qui tient compte de la spécificité du signal issu de puces *tiling arrays*. En effet le signal est bidimensionnel compte tenu des deux intensités à comparer, et longitudinal puisque les sondes sont régulièrement réparties le long du génome. La visualisation de l'intensité du signal indique une dépendance entre des sondes adjacentes couvrant la même région génomique (cf. Figure 1.10). D'autre part, l'annotation structurale nous informe sur la localisation des sondes dans une région intergénique, exonique ou intronique (cf. Figure 1.11). Cette information doit également être prise en compte, en particulier dans les expériences de transcriptome, puisque les sondes situées dans une région exonique ont plus de chance d'être exprimées que celles situées dans des régions intergéniques ou introniques (régions non codantes). Pour parvenir à une telle modélisation, nous utilisons les modèles à variables latentes, qui sont introduits dans le chapitre 2.

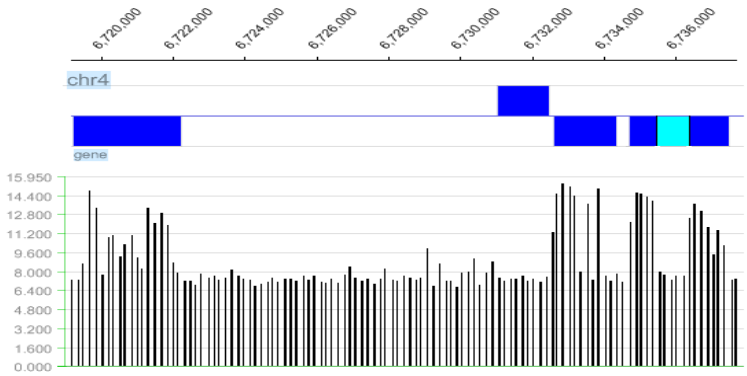

Fig. 1.10: Visualisation de l'intensité du signal : faible dans les régions intergéniques et introniques et fort dans les exons exprimés. Les introns sont représentés en bleu clair et les exons en bleu foncé.

Fig. 1.11: Exemple d'annotation structurale sur une région d'un genome. Carrés jaunes : sondes ; flèches bleues : exons de gènes ; traits fins entre flèches : introns de gènes.

Bibliographie

Aubourg, S. and Rouzé, P. (2001). Genome annotation. *Plant Physiology and Biochemistry* **39**, 181-193.

Berger, S.L. (2007). The complex language of chromatin regulation during transcription. *Nature* **447**, 407-412.

Bernot, A., Choisne, N. and Salanoubat, M. (2001). Séquençage des génomes eucaryotes : Arabidopsis, le quatrième élément. *médecine/sciences* **17**, 829-835.

Buck, M.J. and Lieb, J.D. (2004). Chip-chip : considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83**(3), 349-360.

Cawley, S., Bekiranov, S., Ng, H. *et al.* (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. *Cell* **116**(4), 499-509.

Du, J., Rozowsky, J.S., Korbel, J.O., Zhang, Z.D., Royce, T.E., Schultz, M.H., Snyder, M. and Gerstein, M. (2006). A Supervised Hidden Markov Model Framework for Efficiently Segmenting Tiling Array Data in Transcriptional and ChIP-chip Experiments : Systematically Incorporating Validated Biological Knowlegdge. *Bioinformatics* **22**(24), 3016-3024.

Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G. and Jain, A.N. (2004). Hidden Markov models approach to the analysis of array CGH data. *J Multivariate Analysis* **90**, 132-153.

Ghosh, S., Hirsch, H.A., Sekinger, E.A., Kapranov, P., Struhl, K. and Gingeras, T.R. (2007). Differential analysis for high density tiling microarray data. *BMC Bioinformatics*, **8** :359.

Halasz, G., van Batenburg, M.F., Perusse, J., Hua, S., Lu, X.J., White, K.P. and Bussemaker, H. (2006). Detecting transcriptionally active regions using genomic tiling arrays. *Genome Biology* **7**.

Hanada, K., Zhang, X., Borevitz, J.O., Li, W.H. and Shiu, S.H. (2007). A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. *Genome Research* **17**, 632-640.

He, K., Li, X., Zhou, J., Deng, X.W., Zhao, H. and Luo, J. (2009). NTAP : for NimbleGen tiling array ChIP-chip data analysis. *Bioinformatics* **25**, 1838-1840.

Huber, W., Toedling, J. and Steinmetz, L.M. (2006). Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22**(6), 1963-1970.

Humburg, P., Bulger, D. and Stone, G. (2008). Parameter estimation for robust HMM analysis of ChIP-chip data. *BMC Bioinformatics*, **9** :343.

Hupé, P., Stransky, N., Thiery, J.P., Radvanyi, F. and Barillot, E. (2004). Analysis of array CGH data : from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**(18) :3413-3422.

Ji, H. and Wong, W.H. (2005). TileMap : create chromosomal map of tiling array hybridizations. *Bioinformatics* **21**, 3629-3636.

Johannes, F., Wardenaar, R., Colomé-Tatché M. *et al.* (2010). Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. *Bioinformatics* **26**, 1000-1006.

Johnson, W.E., Li, W., Meyer, C.A., Gottardo, R., Carroll, J.S., Brown, M. and Liu, X.S. (2006). Model-based analysis of tiling-arrays for ChIP-chip. *PNAS* **103**, 12457-12462.

Keles, S., van de Laan, M., Dudoit, S. and Cawley, S.E. (2004). Multiple Testing Methods for ChIP-chip high density oligonucleotide array data. *University of California Berkeley Division of Biostatistics Working Paper Series* **147**.

Li, W., Meyer, A. and Liu, X.S. (2005). A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* **211**, 274-282.

Mockler, T.C. and Ecker, J.R. (2005). Applications of DNA tiling arrays for whole genome analysis. *Genomics* **85**, 1-15.

Munch, K., Gardner, P.P., Arctander, P. and Krogh, A. (2006). A hidden Markov model approach for determining expression from genomic tiling micro arrays. *BMC Bioinformatics*, **7** :239.

Nicolas, P., Leduc, A., Robin, S., Rasmussen, S., Jarmer, H. and Bessières, P. (2009). Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. *Bioinformatics* **25**(18), 2341-2347.

Picard, F., Robin, S., Lavielle, M., Vaisse, C. and Daudin, JJ. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics* **6** :27.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C. and Zhai, Y. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat.Genet.* **20**, 207-211.

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

Seifert, M., Banaei, A., Keilwagen, J., Mette, M.F., Houben, A., Roudier, F., Colot, V., Grosse, I. and Strickert, M. (2009). Array-based Genome comparison of Arabidopsis ecotypes using Hidden Markov models. *Biosignals*, Porto (Portugal).

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S. *et al.* (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat.Genet.* **29**, 263-264.

Velculescu, V.E, Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995). Serial analysis of gene expression. *Science* **270**, 484-487.

Wilkinson, J.Q., Lanahan, M.B., Clark, D.G., Bleecker, A.B., Chang, C., Meyerowitz, E.M. and Klee, H.J. (1997). A dominant mutant receptor from Arabidopsis confers ethylene insensitivity in heterologous plants. *Nature Biotechnology* **15**, 444-447.

Yamada, K., Lim, J., Dale, J.M. *et al.* (2003). Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* **302**, 842-846.

Zeller, G., Henz, S.R., Laubinger, S., Weigel, D. and Rätsch, G. (2008). Transcript normalization and segmentation of tiling array data. *Pacific Symposium on Biocomputing* **12**, 527-538.

Chapitre 2

Modèles à variables latentes

Sommaire

1 Introduction **26**

2 Modèle **27**

2.1 Loi de la variable latente 28

2.2 Loi d'émission 28

3 Inférence **29**

3.1 Présentation de l'algorithme EM 30

3.2 Étape E 31

3.3 Étape M 34

3.4 Initialisation et arrêt de l'algorithme EM 36

3.5 Variantes et extensions de l'algorithme EM 37

4 Sélection de modèles **38**

5 Annexes **41**

5.1 Définitions et propriétés des chaînes de Markov 41

5.2 Démonstrations des formules de l'algorithme Forward/Backward . . 42

Bibliographie **44**

1 Introduction

La classification a pris aujourd'hui une place importante en analyse des données exploratoire et décisionnelle, tant au niveau des domaines d'applications que des développements méthodologiques. Il existe deux types de classification : la classification supervisée et la classification non supervisée. En classification supervisée, les classes sont supposées connues et l'on dispose d'exemples dans chaque classe. Cela convient en particulier au problème de la prise de décision automatisée, en affectant toute nouvelle observation à l'un des groupes préalablement définis. Il s'agit, par exemple, d'établir un diagnostic médical à partir de la description clinique d'un patient, ou de donner une réponse à la demande de prêt bancaire de la part d'un client sur la base de sa situation personnelle. En classification non supervisée, l'appartenance des données aux classes n'est pas connue. C'est justement cette appartenance qu'il s'agit de retrouver à partir des descripteurs disponibles. La classification non supervisée vise donc à déterminer une structure intrinsèque aux données avec les données observées comme seule information disponible. Cette structure dans les données doit satisfaire deux objectifs simultanément : une grande homogénéïté dans chaque classe (les données d'une même classe se ressemblent le plus possible) et une bonne séparation des classes (les données de classes différentes sont les plus différentes possible).
Il existe de nombreuses méthodes dédiées à la classification non supervisée. Les méthodes issues de la statistique sont divisées en deux grandes catégories : les méthodes dites de partitionnement et les modèles de mélange. Les méthodes de partitionnement (méthodes hiérarchiques ou K-means par exemple) font partie des méthodes exploratoires, qui ne font appel à aucune modélisation statistique. Le regroupement des observations en classes s'appuie sur des considérations géométriques (en employant les notions de similarité, dissimilarité et distance entre points) alors que les modèles de mélange, qui se placent dans un cadre probabiliste, sont fondés sur l'analyse de la distribution de probabilité de la population.

Dans ce chapitre, nous abordons le problème de classification non supervisée par cette approche probabiliste qui permet de prendre en compte la variabilité des données. Le support probabiliste a deux avantages majeurs : d'une part, il permet d'avoir accès à des probabilités d'appartenance des individus aux différentes classes. C'est d'ailleurs à partir de ces probabilités que s'établit la classification. D'autre part, le cadre formel de cette approche permet d'apporter des solutions au problème du choix du nombre de classes. Les modèles les plus utilisés sont les modèles de mélange fini de distributions de probabilité. Dans ce contexte, les données sont vues comme étant issues d'un mélange de distributions, où chaque distribution est associée à une classe différente. L'hypothèse sous-jacente des modèles de mélange est que les observations d'une même classe sont issues d'une même distribution de probabilité. La modélisation des observations a bien-sûr une influence sur la classification obtenue.

Dans certaines applications, les observations sont, en plus d'être structurées en classes, organisées spatialement. C'est le cas des données *tiling arrays* dont les sondes présentent une organisation longitudinale le long du génome (cf. Chapitre 1). En effet, deux sondes adjacentes couvrant une même région génomique ont un comportement similaire en termes d'hybridation. Il est donc essentiel de prendre en compte cette information dans la modélisation. L'outil statistique approprié sont les modèles de Markov cachés ou Hidden Markov Models (HMM) qui sont des méthodes usuelles de classification non supervisée pour modéliser des observations ordonnées. Ils fournissent une modélisation capable de

saisir des relations de dépendance et visent à trouver des groupes distincts dans un jeu de données. Introduits par Baum et Petrie (1966), ils peuvent être vus comme un modèle de mélange fini de distributions auquel s'ajoute une hypothèse de dépendance markovienne du processus caché. C'est ce processus caché qui décrit l'appartenance des observations aux classes et que l'on cherche à reconstruire dans un objectif de classification. Les HMM reposent sur l'hypothèse qu'une séquence n'est pas directement générée par une chaîne de Markov mais indirectement par des lois de probabilités attachées aux états de la chaîne de Markov. Ils sont utilisés pour deux raisons principales : la première est la possibilité d'expliquer les variations du processus observé à partir des variations d'un processus sous-jacent caché. La seconde raison est la possibilité de prédire un processus non observé à partir d'un processus observé.
D'un point de vue théorique, les modèles de Markov cachés permettent de bénéficier de l'ensemble des résultats de la statistique mathématique : lois multivariées paramétriques, estimation, choix de modèles.

Dans ce chapitre, nous présentons les modèles à variables latentes dans le contexte des modèles de mélange où les observations sont supposées indépendantes et dans le contexte des modèles de Markov cachés (Section 2). La méthode du maximum de vraisemblance ne donne pas l'expression explicite des estimateurs des paramètres de ces modèles. Nous avons donc recours à l'algorithme EM qui est classiquement utilisé dans ce cas, il est présenté Section 3. La détermination du nombre de classes à partir des données est brièvement abordé Section 4.

2 Modèle

Soient n individus issus d'une population $\mathcal{P}$ structurée en K classes : $\mathcal{P} = \bigcup_{k=1}^{K} C_k$ et $\forall k \neq k', C_k \cap C'_k = \varnothing$. L'échantillon $X = \{X_1, \ldots, X_n\}$ de taille n est aléatoire de réalisations $x_1, \ldots, x_n$. La variable X_t est supposée être issue de la distribution de la classe C_k à laquelle l'individu t appartient. Les variables aléatoires $X_1, \ldots, X_n$ proviennent donc d'un mélange de distributions.
Puisque l'objectif est la classification des données en K classes (le nombre K étant spécifié initialement), l'appartenance des observations aux différentes classes est une information manquante au regard des données observées. On introduit une variable aléatoire latente Z_t, souvent appelée "label", qui est égale à k si l'individu t appartient à la classe k. On note aussi $Z_t = \{Z_{t1}, \ldots, Z_{tK}\}$ un vecteur de booléen où $Z_{tk} = 1$ si l'individu t appartient à la classe k et $Z_{t\ell} = 0$, $\forall\, \ell \neq k$. Ces variables d'appartenance Z_{tk} ne sont pas observées et il s'agit de les reconstruire. Le couple (X, Z) forme ce que l'on appelle les données complètes, où X correspond aux données observées et Z correspond à la classe non observée des individus. La figure 2.1 illustre le problème de classification de données bidimensionnelles issues de trois classes que l'on souhaite retrouver.
Concernant l'identifiabilité du modèle, puisque les Z_t ne sont pas observées, le modèle est invariant pour toute permutation des labels $\{1, ..., K\}$. Le modèle a donc $K!$ définitions équivalentes.

Le modèle nécessite la détermination de la loi de la variable latente Z et de la loi des observations conditionnellement à la variable latente, appelée loi d'émission de $X|Z$.

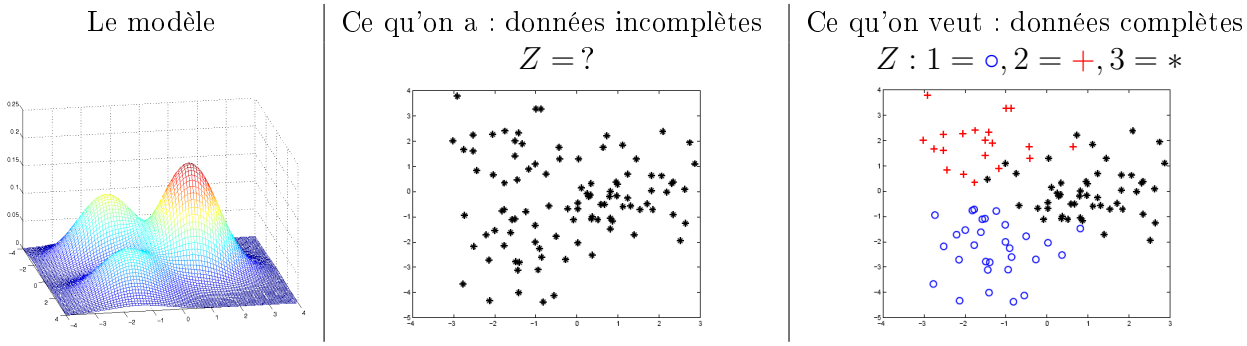

Fig. 2.1: Illustration du problème de classification.

2.1 Loi de la variable latente

Dans les modèles de mélange, les variables $\{Z_1, ..., Z_n\}$ sont supposées indépendantes. Les variables Z_t sont des variables binaires indiquant le label des observations et sont supposées être distribuées selon une loi multinomiale de probabilités $p_1, ..., p_K$, où $p_k = P\{Z_t = k\}$ est la probabilité *a priori* d'appartenance à la population k :

$$Z_t \sim \mathcal{M}(1; p_1, ..., p_K),$$

avec $0 \leq p_k \leq 1$ et $\sum_k p_k = 1$.

Dans un HMM, les variables latentes $\{Z_1, ..., Z_n\}$ sont liées entre elles par des dépendances markoviennes. On suppose que Z est une chaîne de Markov homogène d'ordre 1, c'est-à-dire que toute l'information apportée par le passé est résumée dans l'observation la plus récente : la loi de Z_t conditionnellement à $\{Z_1, Z_2, ..., Z_{t-1}\}$ est égale à la loi de Z_t conditionnellement à Z_{t-1}. La matrice de transition de Z est notée π, de dimension $K \times K$, et dont le terme général $\pi_{k\ell}$ est la probabilité de transition de l'état k à l'état ℓ en une étape : $\pi_{k\ell} = P\{Z_t = \ell | Z_{t-1} = k\}$ pour tout t. La matrice de transition est stochastique : $\forall k, \ell,\ \pi_{k\ell} \geq 0$ et la somme des termes de chaque ligne est égale à 1 : $\forall k,\ \sum_\ell \pi_{k\ell} = 1$.
On a $Z_1 \sim \mathcal{M}(1; m)$, où m est la distribution stationnaire de π.
Les définitions et propriétés des chaînes de Markov sont données en Annexe 5.1. Notons que si toutes les lignes de π sont égales, le modèle est ramené au cas simple des modèles de mélange.

2.2 Loi d'émission

Les variables observées X_t sont supposées indépendantes conditionnellement aux variables latentes Z (principe d'indépendance conditionnelle) :

$$(X_t | Z_t = k) \sim f_k(.),$$

où f_k est la distribution de la classe k. On se place dans un cadre paramétrique, *i.e.* f_k est supposée appartenir à une famille de lois paramétrées : $f_k(.) = f(.; \theta_k)$, où f est caractérisée par le vecteur de paramètres θ_k et θ_k sont les paramètres de la distribution f dans la classe k.
Comme $P\{X_t = x\} = \sum_k P\{X_t = x | Z_t = k\} P\{Z_t = k\}$, la densité de X_t est un mélange

de K densités paramétriques. La distribution marginale de X_t s'écrit donc comme un mélange de distributions :

$$g(x_t;\phi) = \sum_{k=1}^{K} \nu_k f(x_t;\theta_k),$$

où ϕ est l'ensemble des paramètres du modèle. Le coefficient $\nu_k = p_k$ dans le cas des modèles de mélange représente la proportion d'individus appartenant à la classe k, et $\nu_k = m_k^t$ dans le cas des HMM puisque $Z_t \sim \mathcal{M}(1;m^t)$ avec $m^t = m\pi^{t-1}$, et $m_k = P\{Z_1 = k\}$.

Les Figures 2.2 et 2.3 représentent les relations de dépendances conditionnelles dans le cas des modèles de mélange et d'une chaîne de Markov cachée respectivement. L'absence d'arc entre deux sommets signifie que les deux variables aléatoires concernées sont indépendantes conditionnellement aux autres variables.

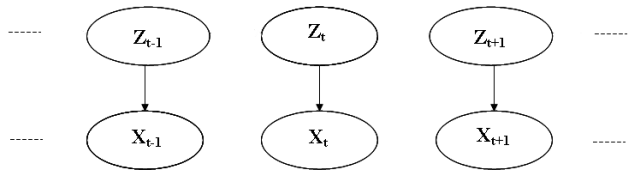

Fig. 2.2: Graphe des dépendances conditionnelles dans un modèle de mélange.

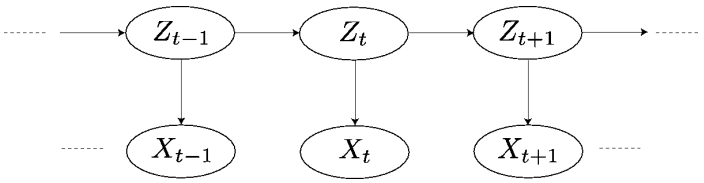

Fig. 2.3: Graphe des dépendances conditionnelles d'une chaîne de Markov cachée d'ordre 1.

La classification des observations est établie à partir de probabilités *a posteriori* d'appartenance à la classe au vu des données observées :

$$\tau_{tk} = P\{Z_t = k|X = x\}.$$

La classification des observations est étudiée plus en détail au chapitre 4.

3 Inférence

Les modèles à variables latentes présument l'existence de variables inobservables directement mais dont on peut mesurer ou observer les effets. L'hypothèse fondamentale est que les covariations entre variables observées s'expliquent par la dépendance de chaque variable observée aux variables latentes. Le problème d'estimation des paramètres peut alors être vu comme un problème de données incomplètes.
Le vecteur des paramètres du modèle à estimer est $\phi = (p, \theta)$ pour les mélanges et $\phi = (m, \pi, \theta)$ pour les HMM. De nombreux travaux ont été développés dans le domaine de l'estimation car les formules explicites des estimateurs des paramètres n'existent pas sous une forme simple et nécessitent des procédures d'estimation itératives. Plusieurs approches peuvent être envisagées pour estimer ϕ, comme par exemple la méthode des

moments, la méthode du maximum de vraisemblance ou les approches bayésiennes. Néanmoins, la méthode du maximum de vraisemblance est la plus utilisée, principalement grâce à l'existence d'une théorie statistique et à l'introduction de l'algorithme EM (Expectation-Maximization algorithm ou Algorithme d'Estimation et Maximisation en français) associé.

La log-vraisemblance des données observées (ou log-vraisemblance observée)

$$\mathcal{L}(X;\phi) = \log P(X;\phi) = \log \sum_Z P(X,Z;\phi)$$

nécessite le calcul de K^n termes, ce qui est impossible en pratique. La solution est d'avoir recours à des algorithmes itératifs de recherche de maximum ou de minimum d'une fonction. L'algorithme le plus utilisé est l'algorithme EM, proposé par Dempster *et al.* (1977).

3.1 Présentation de l'algorithme EM

L'algorithme EM permet d'obtenir les estimateurs du maximum de vraisemblance des paramètres ϕ d'un modèle $P(X,Z|\phi)$ avec Z une variable aléatoire latente. L'idée de cet algorithme est de travailler non pas avec la vraisemblance des données observées mais avec la vraisemblance des données complètes (ou log-vraisemblance complétée), plus facile à manipuler et qui offre une solution simple pour l'estimation des paramètres. L'avantage de cet algorithme est qu'il est facile à implémenter et qu'il fournit en général des formes explicites des estimateurs.

À l'aide de la formule de Bayes, la log-vraisemblance se décompose de la manière suivante :

$$\mathcal{L}(X;\phi) \quad = \quad \mathcal{L}(X,Z;\phi) - \mathcal{L}(Z|X;\phi)$$

En appliquant l'espérance conditionnellement aux données observées X, on obtient :

$$\mathcal{L}(X;\phi) \quad = \quad \mathbb{E}\left[\mathcal{L}(X,Z;\phi)|X\right] - \mathbb{E}\left[\mathcal{L}(Z|X;\phi)|X\right] \ .$$

L'algorithme EM consiste en l'optimisation indirecte de la log-vraisemblance des données observées via l'optimisation itérative de l'espérance conditionnelle de la log-vraisemblance des données complètes. Notons $\phi^{(h)}$ la valeur du paramètre à l'itération h, on obtient :

$$\mathbb{E}\left[\mathcal{L}\left(X_1,...,X_n;\phi\right)|\phi^{(h)}\right] = Q(\phi;\phi^{(h)}) - H(\phi;\phi^{(h)}) \tag{2.1}$$

avec

$$\begin{aligned} Q(\phi;\phi^{(h)}) &= \mathbb{E}\left[\mathcal{L}\left((X_1,Z_1),...,(X_nZ_n);\phi\right)|X,\phi^{(h)}\right] \\ H(\phi;\phi^{(h)}) &= \mathbb{E}\left[\log P(Z|X;\phi)|X,\phi^{(h)}\right] \end{aligned}$$

Chaque itération consiste en deux étapes :
- **Étape E (Expectation) :** Calculer $Q(\phi;\phi^{(h)})$,
- **Étape M (Maximisation) :** Actualiser les paramètres avec $\phi^{(h+1)} = \text{Argmax}_\phi \left\{Q(\phi;\phi^{(h)})\right\}$.

Les étapes E et M sont itérées jusqu'à convergence de l'algorithme.
La propriété fondamentale de l'algorithme EM, établie par Dempster *et al.* (1977), est que la log-vraisemblance des données observées augmente à chaque itération de l'algorithme :

$$\mathcal{L}(X;\phi^{(h+1)}) \geq \mathcal{L}(X;\phi^{(h)}).$$

En effet, l'étape M garantit $Q(\phi;\phi^{(h+1)}) \geq Q(\phi;\phi^{(h)})$, et l'application de l'inégalité de Jensen permet d'obtenir $H(\phi;\phi^{(h+1)}) \leq H(\phi;\phi^{(h)})$. Ainsi, la quantité $H(\phi;\phi^{(h)})$ diminue à chaque itération et la suite définie par

$$\phi^{(h+1)} = \underset{\phi}{\text{Argmax}} \left\{Q(\phi;\phi^{(h)})\right\}$$

fait tendre $\mathcal{L}\left(X_1,...,X_n;\phi\right)$ vers un maximum local si la log-vraisemblance des données observées est majorée. Cependant, il n'y a pas d'assurance de trouver le maximum global.

Dans la suite, nous détaillons chacune des deux étapes de l'algorithme EM dans le cas des modèles de mélange et des HMM.

3.2 Étape E

Selon Archer et Titterington (2002), l'étape E consiste en une restauration probabiliste de toutes les séquences d'états possibles. En effet, l'algorithme EM peut être vu comme un algorithme de restauration-maximisation.
L'étape E dépend de la structure de dépendance des variables latentes, nous traiterons donc séparément les modèles de mélange et les HMM.
Dans cette section, on note $Z_{tk} = 1$ si le processus Z pris en l'instant t est égal à k.

Modèles de mélange

D'après la formule de Bayes, la log-vraisemblance des données complètes s'écrit :

$$\begin{aligned}
\mathcal{L}(X,Z;\phi) &= \mathcal{L}(Z;\phi) + \mathcal{L}(X|Z;\phi) \\
&= \sum_{t=1}^{n}\sum_{k=1}^{K} Z_{tk}\log p_k + \sum_{t=1}^{n}\sum_{k=1}^{K} Z_{tk}\log f(X_t;\theta_k) \\
&= \sum_{t=1}^{n}\sum_{k=1}^{K} Z_{tk}[\log p_k + \log f(X_t;\theta_k)] \ .
\end{aligned}$$

L'espérance de la log-vraisemblance des données complètes sachant X s'écrit alors :

$$\begin{aligned}
Q(\phi,\phi^{(h)}) &= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{k=1}^{K} Z_{tk}\log\left\{p_k^{(h)} f(X_t;\theta_k^{(h)})\right\}|X,\phi^{(h)}\right] \ , \\
&= \sum_{t=1}^{n}\sum_{k=1}^{K} \mathbb{E}\left[Z_{tk}|X,\phi^{(h)}\right]\log\left\{p_k^{(h)} f(X_t;\theta_k^{(h)})\right\} \ , \\
&= \sum_{t=1}^{n}\sum_{k=1}^{K} \tau_{tk}\log\left\{p_k^{(h)} f(X_t;\theta_k^{(h)})\right\}. \qquad (2.2)
\end{aligned}$$

Ainsi l'étape E qui consiste à calculer $Q(\phi,\phi^{(h)})$ se réduit au calcul des τ_{tk}, probabilités *a posteriori* d'appartenance de l'individu t à la classe k. Elles sont calculées par la règle

de Bayes à partir de la valeur courante des paramètres $\phi^{(h)} = (p^{(h)}, \theta^{(h)})$:

$$\tau_{tk}^{(h+1)} = \frac{p_k^{(h)} f(X_t; \theta_k^{(h)})}{\sum_{l=1}^{K} p_l^{(h)} f(X_t; \theta_l^{(h)})}$$

HMM

Dans le contexte des HMM, le vecteur de paramètres est $\phi = (m, \pi, \theta)$, où m est la distribution stationnaire et π est la matrice de transition. La log-vraisemblance des données complètes s'écrit :

$$\mathcal{L}(X, Z; \phi) = \sum_{k=1}^{K} Z_{1k} \log(m_k) + \sum_{t=1}^{n} \sum_{\ell=1}^{K} Z_{t\ell} \log[f(X_t; \theta_\ell)] + \sum_{t \geq 2} \sum_{k=1}^{K} \sum_{\ell=1}^{K} Z_{t-1,k} Z_{t\ell} \log(\pi_{k\ell}).$$

On remarque que seul le dernier terme de l'équation, qui correspond à $\mathcal{L}(Z; \phi)$, diffère par rapport aux modèles de mélange.

L'espérance de la log-vraisemblance des données complètes sachant X s'écrit :

$$\begin{aligned} Q(\phi, \phi^{(h)}) &= \sum_{k=1}^{K} E\left[Z_{1k}|X\right] \log(m_k^{(h)}) + \sum_{t=1}^{n} \sum_{\ell=1}^{K} E\left[Z_{t\ell}|X\right] \log[f(X_t; \theta_\ell^{(h)})] \\ &\quad + \sum_{t \geq 2} \sum_{k=1}^{K} \sum_{\ell=1}^{K} E\left[Z_{t-1,k} Z_{t\ell}|X\right] \log(\pi_{k\ell}^{(h)}). \end{aligned}$$

On remarque que $E\left[Z_{t\ell}|X\right]$ représente la probabilité *a posteriori* que l'individu t appartienne à la classe ℓ, notée $\tau_{t\ell}$.

La structure de dépendance conditionnelle reste simple, ce qui rend le calcul de $P(Z|X)$ possible. L'étape E peut être efficacement implémentée par un algorithme dit "avant-arrière" ou "forward-backward" introduit par Baum *et al.* (1970) puis Devijver (1985).
Cet algorithme est basé sur la décomposition suivante des probabilités $\tau_{t\ell}$ qui traduit l'indépendance conditionnelle entre le passé et le futur du processus à chaque instant t.
On note par la suite $X_a^b = \{X_a, ..., X_b\}$ avec $a < b$.

$$\begin{aligned} \tau_{t\ell} = P(Z_{t\ell}|X_1^n) &= \frac{P(Z_{t\ell}, X_1^n)}{P(X_1^n)} \\ &= \frac{P(X_{t+1}^n|Z_{t\ell}) P(Z_{t\ell}, X_1^t)}{P(X_{t+1}^n|X_1^t) P(X_1^t)} \\ &= \underbrace{\frac{P(X_{t+1}^n|Z_{t\ell})}{P(X_{t+1}^n|X_1^t)}}_{\text{Backward } B_{t\ell}} \times \underbrace{P(Z_{t\ell}|X_1^t)}_{\text{Forward } F_{t\ell}} \end{aligned}$$

Devijver (1985) montre que les quantités $F_{t\ell} = P(Z_{t\ell}|X_1^t)$ peuvent être calculées à l'aide d'une passe avant (c'est-à-dire de 1 à n) tandis que les quantités $B_{t\ell} = \frac{P(X_{t+1}^n|Z_{t\ell})}{P(X_{t+1}^n|X_1^t)}$ (ou directement $\tau_{t\ell}$ car $\tau_{t\ell} = B_{t\ell} \times F_{t\ell}$) peuvent être calculées par une passe arrière (c'est-à-dire de n à 1).

- **Étape Forward**

  Pour $t = 1$,

$$F_{1\ell} = P(Z_{1\ell}|X_1) = \frac{m_\ell f(X_1;\theta_\ell)}{\sum_\ell m_\ell f(X_1;\theta_\ell)} ,$$

  et $\forall\ t \neq 1$,

$$F_{t\ell} \propto \sum_k \pi_{k\ell} F_{t-1,k} f(X_t;\theta_\ell) . \tag{2.3}$$

- **Étape Backward**

  Pour $t = n$,

$$\tau_{n\ell} = F_{n\ell} = P(Z_{n\ell}|X_1^n) ,$$

  et $\forall\ t \neq n$,

$$\tau_{t\ell} = F_{t\ell} \sum_k \frac{\pi_{\ell k}\tau_{t+1,k}}{G_{t+1,k}} , \tag{2.4}$$

  avec $G_{t+1,k} = P(Z_{t+1,k}|X_1^t) = \sum_\ell \pi_{\ell k} F_{t\ell}$.

Les démonstrations des formules 2.3 et 2.4 sont données en Annexe 5.2 dans le cas d'une chaîne de Markov homogène.

Calcul de la vraisemblance observée. On peut remarquer que la récurrence avant permet de calculer la vraisemblance des données observées puisque

$$P(X_1, ..., X_n) = P(X_1) \prod_{t=2}^{n} P(X_t|X_1^{t-1}) = \prod_{t=1}^{n} \left[\sum_{\ell=1}^{K} f(X_t;\theta_\ell) \sum_{k=1}^{K} \pi_{k\ell} F_{t-1,k} \right] .$$

Notons $\alpha_k(t) = P(X_1, ..., X_t, Z_{tk}|\theta)$, la vraisemblance précédente s'écrit alors :

$$P(X_1, ..., X_n) = \sum_{k=1}^{K} \alpha_k(n) .$$

Rabiner (1989) puis Bilmes (1998) ont introduit une procédure récursive pour calculer cette vraisemblance :

1. $\alpha_k(1) = m_k f(X_1;\theta_k)$
2. $\alpha_\ell(t+1) = f(X_{t+1};\theta_\ell) \sum_{k=1}^{K} \pi_{k\ell}\alpha_k(t)$
3. $P(X_1, ..., X_n) = \sum_{k=1}^{K} \alpha_k(n)$

Cependant, ce calcul pose des problèmes numériques. Les valeurs prises par f et α étant généralement petites et la récurrence se faisant sur la multiplication de f par α, on atteint rapidement les limites de la précision numérique de l'ordinateur, et lorsque le nombre d'observations n dépasse 500, on obtient une vraisemblance nulle.
Pour éviter ce problème, il faut pondérer α par une quantité astucieusement choisie. Nous proposons la procédure récursive suivante :

1. $A_{1,k} = \frac{\alpha_k(1)}{\sum_{k=1}^{K} \alpha_k(1)}$
2. $A_{t,k} = \lambda_t \sum_{\ell=1}^{K} A_{t-1,\ell} \pi_{\ell k} f(X_t; \theta_k)$, avec $\lambda_t = 1 \,/\, \sum_{\ell=1}^{K} \sum_{k=1}^{K} A_{t-1,\ell} \pi_{\ell k} f(X_t; \theta_k)$.
3. $P(X_1, ..., X_n) = \sum_{k=1}^{K} \alpha_k(n) = \frac{1}{\prod_{t=1}^{n} \lambda_t} \sum_{k=1}^{K} A_{n,k}$.

Comme $\sum_{k=1}^{K} A_{n,k} = 1$, le calcul de la log-vraisemblance devient :

$$\log P(X_1, ..., X_n) = -\sum_{t=1}^{n} \log(\lambda_t).$$

3.3 Étape M

L'étape de maximisation est identique pour les modèles de mélange et pour les HMM : les paramètres sont obtenus en maximisant l'espérance des données complètes conditionnellement aux données observées.

Dans le cas des modèles de mélange, la maximisation de (2.2) avec les valeurs τ_{tk} obtenues à l'étape précédente s'écrit :

$$\phi^{(h+1)} = \underset{\phi=(p,\theta)}{\text{Argmax}} \sum_{t=1}^{n} \sum_{k=1}^{K} \tau_{tk}^{(h+1)} \log \{p_k f(X_t; \theta_k)\}$$

L'estimation des proportions du mélange est obtenue en maximisant sous la contrainte $\sum_{k=1}^{K} p_k = 1$ l'espérance de la log-vraisemblance des données incomplètes :

$$\hat{p} = \underset{p}{\text{Argmax}} \sum_{t,k} \tau_{tk} \log p_k \ .$$

En utilisant la méthode du multiplicateur de Lagrange, on obtient :

$$p_k^{(h+1)} = \frac{\sum_{t=1}^{n} \tau_{tk}^{(h)}}{n} \ .$$

Dans le cas des HMM, on obtient un estimateur des probabilités de transition de la forme :

$$\begin{aligned} \pi_{k\ell}^{(h+1)} &\propto \sum_{t \geq 2} E(Z_{t-1,k} Z_{t\ell} | X) \\ &\propto \frac{F_{t-1,k} \pi_{k\ell} \tau_{t,\ell}}{G_{t,\ell}} \end{aligned}$$

La distribution stationnaire m est remise à jour avec le vecteur propre normalisé associé à la valeur propre 1 de la matrice de transition π.

Les paramètres θ_k de la distribution sont estimés de la même manière dans un modèle de mélange et dans un HMM.

$$\hat{\theta} = \underset{\theta}{\text{Argmax}} \sum_{t,k} \tau_{tk} \log f(X_t, \theta_k).$$

On suppose que chaque composant du mélange (ou du HMM) appartient à une famille de lois paramétrées et le choix se restreint généralement à des lois classiques appartenant à la famille exponentielle.

Exemple gaussien

Dans le cas continu, le modèle paramétrique le plus utilisé sur $\mathbb{R}^d$ est la loi multinormale $\mathcal{N}(\mu_k, \Sigma_k)$, où μ_k est un vecteur de $\mathbb{R}^d$ désignant la moyenne du composant k et Σ_k est la matrice de variance correspondante, de dimension $K \times K$. La densité d'une distribution gaussienne sur $\mathbb{R}^d$ est définie au point x_t par :

$$f(x_t|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}} [\det(\Sigma_k)]^{-1/2} \exp\left\{ -\frac{1}{2}(x_t - \mu_k)^T \Sigma_k^{-1} (x_t - \mu_k) \right\},$$

où x^T représente la transposée de x.

L'espérance de la log-vraisemblance des données complètes s'écrit :

$$\begin{aligned} Q(\phi, \phi^{(h)}) &= -\frac{1}{2} \sum_{t=1}^{n} \sum_{k=1}^{K} \tau_{tk} \left\{ \log\left[\det(\Sigma_k)\right] + (x_t - \mu_k)^T \Sigma_k^{-1} (x_t - \mu_k) \right\} \\ &\quad + \sum_{t=1}^{n} \sum_{k=1}^{K} \tau_{tk} \log(\pi_k) - \frac{dn}{2} \log(2\pi). \end{aligned}$$

Les estimateurs des paramètres μ et Σ sont donnés par :

$$\hat{\mu}_k = \frac{\sum_{t=1}^{n} \tau_{tk}\ x_t}{\sum_{t=1}^{n} \tau_{tk}}, \qquad \forall\ k = 1, ..., K.$$

$$\hat{\Sigma}_k = \frac{\sum_{t=1}^{n} \tau_{tk} (x_t - \hat{\mu}_k)(x_t - \hat{\mu}_k)^T}{\sum_{t=1}^{n} \tau_{tk}}, \qquad \forall\ k = 1, ..., K.$$

On remarque que ce sont simplement des versions pondérées des estimateurs usuels.

Modélisation de la variance. La densité gaussienne modélise une distribution ellipsoïdale de centre μ_k dont les caractéristiques géométriques (volume, forme, orientation) peuvent être contrôlées grâce à une décomposition spectrale de la matrice de variance Σ_k (Banfield et Raftery, 1993). Chaque matrice de variance des composants du mélange peut s'écrire :

$$\Sigma_k = \lambda_k D_k A_k D_k^T, \tag{2.5}$$

où λ_k représente le volume ($\lambda_k = \det(\Sigma_k)^{1/2}$), D_k représente l'orientation et A_k représente la forme de l'ellipse. La matrice D_k est la matrice des vecteurs propres de Σ_k et A_k est une matrice diagonale telle que $\det(A_k) = 1$ avec les valeurs propres normalisées de Σ_k sur la diagonale dans l'ordre décroissant. Dans les modèles gaussiens multivariés, la variance induit un grand nombre de paramètres et des variances hétéroscédastiques

donnent souvent des résultats instables. Cette décomposition permet donc de prendre en compte au mieux les différentes caractéristiques de la variance. En permettant aux paramètres volumes, formes et orientations de varier ou d'être égaux entre les classes, on obtient quatorze modèles différents et facilement interprétables. Ces quatorze modèles sont détaillés dans Celeux et Govaert (1995) : il y a huit modèles généraux, quatre modèles avec des matrices de variance diagonales et deux modèles avec des formes sphériques ($A_k = I$). Les estimateurs de λ_k, D_k et A_k sont détaillés pour chaque modèle dans Celeux et Govaert (1995).
La géométrie particulière représentée par chaque modèle permet de caractériser simplement non seulement les classes mais aussi leurs similitudes éventuelles, et ainsi d'aider à leur interprétation.

3.4 Initialisation et arrêt de l'algorithme EM

Comme tous les algorithmes itératifs, l'algorithme EM nécessite une initialisation des paramètres et un critère d'arrêt. Dans certains cas, il est possible de partir d'une partition particulière des données ou de valeurs de paramètres prédéfinies. Par contre, si l'on ne dispose d'aucune information *a priori*, les paramètres $\phi^{(0)}$ ou $\tau^{(0)}$ peuvent être initialisés au hasard. Le principal problème de l'algorithme EM est sa sensibilité aux valeurs initiales (Karlis et Xekalaki, 2003). Comme l'algorithme converge vers des maxima locaux de la vraisemblance, les résultats peuvent être différents selon l'initialisation. C'est pourquoi l'étape d'initialisation est cruciale. Différentes procédures d'initialisation ont été proposées (cf. Maitra, 2009 pour une revue détaillée) mais aucune méthode ne semble vraiment meilleure que les autres. Une solution classique consiste à répéter l'algorithme à partir de plusieurs valeurs initiales différentes et à retenir celle maximisant la vraisemblance, mais ceci peut devenir algorithmiquement coûteux. On peut aussi tester plusieurs valeurs initiales sans attendre la convergence de l'algorithme EM mais en l'arrêtant dès que

$$\frac{\mathcal{L}^m - \mathcal{L}^{m-1}}{\mathcal{L}^m - \mathcal{L}^0} \leq 10^{-2} \ ,$$

où $\mathcal{L}^m$ est la log-vraisemblance observée à l'itération m. La valeur seuil 10^{-2} est fixée arbitrairement.
Biernacki, Celeux et Govaert (2003) ont proposé une stratégie d'initialisation efficace en trois étapes :

1. Générer p positions initiales.
2. Exécuter l'algorithme EM un certain nombre de fois à chaque position initiale, avec un nombre fixe d'itérations.
3. Choisir la solution offrant la meilleure vraisemblance parmi les p tests, notée p^*.

Cette stratégie peut être encore améliorée en répétant les trois étapes x fois et en utilisant les $p_1^*, ..., p_x^*$ en tant que positions initiales dans l'étape 1.
Une autre solution est d'initialiser l'algorithme en utilisant les méthodes heuristiques fondées sur des considérations géométriques. L'utilisation de l'approche par classification ascendante hiérarchique comme initialisation de l'algorithme fonctionne bien si les populations sont bien séparées et pour des jeux de données de taille raisonnable. En général, aucune stratégie ne fonctionne bien uniformément dans tous les cas, donc la pratique usuelle est d'essayer différentes stratégies et de choisir la solution de plus grande vraisemblance.

Le critère de convergence peut être choisi soit sur une différence relative de la log-vraisemblance entre deux itérations successives, soit sur les paramètres. L'algorithme

peut aussi être arrêté après un nombre prédéfini d'itérations.

3.5 Variantes et extensions de l'algorithme EM

Depuis le développement de l'algorithme EM, de nombreuses variantes ont vu le jour. Ces variantes permettent de répondre aux difficultés qui peuvent être rencontrées pour le calcul de $Q(\phi, \phi^{(h)})$ à l'étape E, pour la maximisation de $Q(\phi, \phi^{(h)})$ à l'étape M, ou bien pour l'obtention de meilleures performances comme par exemple l'augmentation de la vitesse de convergence.

- L'algorithme CEM est une version classifiante de l'algorithme EM (C pour classification) proposée par Celeux et Govaert (1992). Contrairement à l'algorithme EM qui se positionne dans une optique d'estimation, l'algorithme CEM privilégie la classification en optimisant directement la vraisemblance des données complètes, grâce à l'introduction d'une étape de classification. L'algorithme CEM converge en un nombre fini d'itérations, mais ne converge pas vers l'estimateur du maximum de vraisemblance.
- Lorsqu'il n'y a pas de solution analytique à l'étape M de maximisation, l'algorithme gradient-EM fournit une alternative itérative au calcul d'une solution explicite. De même, l'algorithme Generalized EM (GEM) est une version généralisée de l'algorithme EM dans laquelle la valeur du paramètre $\phi^{(h+1)}$ actualisée à l'étape M ne maximise pas nécessairement Q mais l'augmente simplement tel que :

$$Q(\phi^{(h+1)}|\phi^{(h)}) \geq Q(\phi^{(h)}|\phi^{(h)}).$$

- Celeux et Diebolt (1985) ont introduit l'algorithme SEM en vue de répondre aux limitations de l'EM telles que la dépendance aux valeurs initiales ou la convergence en des points stationnaires stables mais indésirables de la fonction de vraisemblance. Le principe de cette méthode réside dans la maximisation de la log-vraisemblance des données complètes à partir, non pas de son expression analytique, mais grâce à une évaluation numérique de celle-ci *via* l'introduction d'une étape stochastique de classification (étape S) entre les étapes E et M, où l'appartenance des observations aux classes est tirée aléatoirement selon une loi multinomiale de paramètres τ_{tk}.
- Wei et Tanner (1991) ont proposé d'étendre l'algorithme SEM en introduisant l'algorithme Monte Carlo EM (MCEM). Cet algorithme repose sur une approximation de l'espérance des données complètes conditionnellement aux données observées par une approche de Monte Carlo. Il permet de résoudre les problèmes de calcul pouvant intervenir à l'étape E.

On trouvera une revue exhaustive plus détaillée des variantes de l'algorithme EM dans McLachlan et Krishnan (2008).

Mise en pratique

L'utilisation de plus en plus intensive des modèles de mélange a suscité l'apparition de nombreux logiciels ou packages. Mclust (Fraley et Raftery, 2006) est un package R développé en FORTRAN pour les modèles de mélanges gaussiens multivariés. Les quatorze paramétrisations de la matrice de dispersion Σ_k présentées Section 3.3 sont disponibles. Mclust s'appuie sur l'algorithme EM pour l'estimation des paramètres et sur le critère BIC pour la sélection de modèles. L'algorithme EM est initialisé par l'utilisation de l'approche par classification ascendante hiérarchique. Sa flexibilité et ses bonnes performances font de

ce package l'un des plus populaires. MIXMOD (Biernacki *et al.*, 2006) est un logiciel écrit en C++ interfacé avec Scilab ou Matlab entièrement dédié aux modèles de mélange. Il propose des modèles de mélange avec des distributions multinomiales ou des gaussiennes multivariées, ainsi que les différentes paramétrisations de la variance. MIXMOD offre différentes possibilités d'initialisation et plusieurs critères d'arrêt pour l'algorithme EM. Différents critères pour la sélection de modèles ainsi que plusieurs variantes de l'algorithme EM sont aussi inclus dans le logiciel.

4 Sélection de modèles

Une difficulté de la classification non supervisée est le choix du nombre de classes K. Ce nombre peut être fixé par la nature même de la question posée, mais dans la plupart des cas il est inconnu *a priori* et choisir le nombre de classes est souvent une question à laquelle il faut répondre. Remarquons d'abord qu'un choix pragmatique du nombre de classes est de sélectionner une partition dont il sera possible d'interpréter les classes. Choisir le nombre de composants K d'un mélange gaussien constitue un problème majeur qui n'est pas encore complètement résolu. Si le but est de modéliser une densité quelconque, l'influence du nombre de composants sur l'apparence de la densité n'a que peu d'impact, principalement lorsque ce nombre est grand. Cependant, si les mélanges sont utilisés dans une approche de classification, le nombre de composants est alors essentiel car il correspond souvent à une explication physique (réelle ou supposée). Les deux principales méthodes permettant de déterminer le nombre de composants sont fondées sur la vraisemblance : ce sont les critères de sélection et les tests d'hypothèses. Nous nous intéresserons dans la suite uniquement aux critères de sélection.

Le cadre probabiliste des modèles à variables latentes a l'avantage de proposer des solutions au problème du choix du nombre de composants dans le mélange en utilisant des critères statistiques classiques de sélection de modèles. Le choix du nombre de classes peut donc être formulé comme un problème de sélection de modèles et peut être réalisé avec un critère pénalisé de la forme suivante :

$$\mathcal{L}(X; \hat{\phi}_K) - \beta \, pen(K),$$

où $\mathcal{L}(X; \hat{\phi}_K)$ est la log-vraisemblance des données observées prise en son maximum pour un modèle à K populations, β est une constante positive et $pen(K)$ est une fonction appelée pénalité qui représente la pénalité du modèle à K populations. Dans la plupart des cas, la pénalité est fonction croissante du nombre de paramètres à estimer pour le modèle considéré. On choisit le nombre de classes qui maximise le critère pénalisé considéré. Ce type de critère tente donc de réaliser un compromis entre l'adéquation du modèle aux données, mesuré par la log-vraisemblance maximale et la complexité de celui-ci, mesuré par sa dimension. Ces critères ont été davantage étudiés dans le cadre des modèles de mélange que pour les HMM.

Critère AIC

Du point de vue fréquentiste, un "bon" modèle est celui qui réalise un bon compromis « biais-variance ». Dans ce cadre, le critère AIC (Akaike Information Criterion) a été développé par Akaike (1974) avec une pénalité qui ne dépend pas du nombre d'observations n :

$$AIC(K) = \mathcal{L}(X; \hat{\phi}_K) - \nu_K,$$

où ν_K est le nombre de paramètres libres du modèle K. Par exemple, dans le cas d'un modèle de mélange gaussien bidimensionnel avec une matrice de variance non contrainte $\nu_K = 6K - 1$, et pour un modèle HMM, $\nu_K = K^2 + 5K - 1$.

Critère BIC

Le critère BIC (Bayesian Information Criterion) (Schwarz, 1977) se place dans une approche bayésienne et propose une pénalité dépendant de n (cf. Lebarbier et Mary-Huard, 2004 pour plus de détails). Du point de vue du critère BIC, un "bon" modèle est celui qui maximise la vraisemblance intégrée (sur les paramètres) lorsque chaque modèle en compétition est équiprobable *a priori*. Le critère BIC s'écrit :

$$BIC(K) = \mathcal{L}(X; \hat{\phi}_K) - \frac{\log n}{2} \times \nu_K,$$

où ν_K est le nombre de paramètres libres du modèle K.
Ces deux critères sont basés sur des approximations asymptotiques. Le critère BIC est le critère le plus utilisé pour les modèles de mélange, AIC favorisant souvent les modèles trop complexes. D'autres approches ont été considérées pour la sélection de modèle dans le cadre bayésien (cf. Kass et Raftery, 1995 pour un état de l'art complet). Cependant, le critère BIC est souvent préféré pour sa simplicité d'implémentation et pour ses propriétés statistiques. Keribin (2000) a montré que l'utilisation de BIC conduit à un estimateur consistant du nombre de classes sous certaines conditions.

Critère ICL

L'hypothèse de mélange n'étant probablement pas vérifiée dans une population issue de données réelles, on peut s'attendre à ce que les critères classiques sélectionnent un nombre très important de composants pour le mélange afin de réaliser une bonne estimation de la loi inconnue de la population. Si l'on veut que la méthode de choix de modèle retienne un modèle produisant des classes bien séparées tout en respectant au mieux la distribution des données, l'idée est de reporter l'objectif de classification de la méthode d'estimation dans la méthode de choix de modèle. Dans cette perspective, Biernacki, Celeux et Govaert (2000) ont proposé le critère ICL (Integrated Complete Likelihood) qui pénalise la log-vraisemblance complétée (et non observée) par le même terme de complexité que le critère BIC. En pratique, ce critère, à maximiser, évalue le nombre de composants donnant lieu à une classification pertinente des données. Le critère ICL s'écrit :

$$ICL(K) = \mathcal{L}(X, \hat{Z}; \hat{\phi}_K) - \frac{\nu_K}{2} \log(n), \tag{2.6}$$

où la variable latente Z est remplacée par son mode *a posteriori* étant donné $\hat{\phi}_K$. La définition de ICL donnée dans l'équation (2.6) repose sur un partitionnement dur des données. McLachlan et Peel (2000) ont proposé de remplacer $\widehat{Z}$ par la valeur attendue de

Z compte tenu des observations, *ie* $\mathbb{E}[Z_{tk}|X]$.
Hathaway (1986) a montré la décomposition suivante de la log-vraisemblance observée en une log-vraisemblance complétée et un terme entropique (cf. Equation 2.1) :

$$\mathbb{E}\left[\mathcal{L}(X, Z; \phi_K)|X\right] = \mathcal{L}(X; \phi_K) + Ent(K),$$

où l'entropie mesure la séparabilité des populations. Pour des populations très séparées (non imbriquées), $Ent(K)$ est proche de 0.
Ainsi le critère ICL est équivalent au critère BIC avec un terme de pénalité supplémentaire, qui est l'entropie $-\mathbb{E}\left[\log P(Z|X)\right]$. Cette décomposition permet d'interpréter ICL comme une pénalisation du maximum de log-vraisemblance par un terme de complexité du modèle et un terme d'imbrication des classes, ce qui équivaut à un critère BIC pénalisé par ce terme d'imbrication. ICL pénalise donc les modèles produisant des classes trop imbriquées. On remarque que ICL est aussi simple à calculer que BIC, mais son étude théorique est plus complexe (Baudry, 2009).
Le critère ICL a été établi dans le contexte de mélange indépendant, mais Celeux et Durand (1985) ont montré dans une étude de simulation qu'il semble avoir le même comportement dans le contexte HMM.

5 Annexes

5.1 Définitions et propriétés des chaînes de Markov

Introduites en 1889 par Galton et Watson, les chaînes de Markov sont utilisées pour étudier l'évolution au cours du temps de phénomènes aléatoires comme par exemple la propagation d'une épidémie ou l'évolution des cours de titres boursiers, ou bien pour gérer les systèmes de file d'attente.

Un **processus** (aléatoire ou stochastique) rend compte de l'évolution, au cours du temps, d'un phénomène aléatoire. On note Z_t l'état du phénomène au temps t. Z_t est une variable aléatoire.

- La loi de Z_t dépend en général de t.
- Pour deux dates t_1 et t_2 quelconques, Z_{t_1} et Z_{t_2} ne sont pas indépendantes.

Un processus Z_t vérifie la **propriété de Markov** si la loi de Z_t conditionnellement à une série d'observations antérieures $\{Z_1, Z_2, ..., Z_{t-1}\}$ est égale à la loi de Z_t conditionnellement à l'observation la plus récente Z_{t-1}. Cette propriété suppose que toute l'information apportée par le passé est résumée dans l'observation la plus récente.

Soit ε l'espace dans lequel les variables Z_t prennent leurs valeurs.
Une **chaîne de Markov** d'ordre 1 est un processus qui vérifie la propriété de Markov, et pour lequel le temps et l'espace ε évoluent dans ensemble dénombrable.
En notant i_0,i_1,..,i_t des états quelconques de ε, la propriété de Markov s'écrit :

$$P\left\{Z_{t+1} = i_{t+1} | Z_0 = i_0, ..., Z_t = i_t\right\} = P\left\{Z_{t+1} = i_{t+1} | Z_t = i_t\right\}.$$

Une chaîne de Markov est dite **homogène** si les probabilités conditionnelles ci-dessus ne dépendent pas de t, c'est-à-dire si $\forall (k, \ell) \in \varepsilon$:

$$P\left\{Z_{t+1} = \ell | Z_t = k\right\} = P\left\{Z_1 = \ell | Z_0 = k\right\}.$$

La **probabilité de transition** en une étape est :

$$\pi_{k\ell} = P\left\{Z_{t+1} = \ell | Z_t = k\right\} \quad \text{pour tout } t.$$

La **matrice de transition** est la matrice π dont le terme général $\pi_{k\ell}$ est la probabilité de transition de l'état k à l'état ℓ en une étape. C'est une matrice carrée et indépendante du temps.

$$\pi = \begin{pmatrix} \pi_{11} & \dots & \pi_{1K} \\ \vdots & & \vdots \\ \pi_{K1} & \dots & \pi_{KK} \end{pmatrix}$$

La matrice de transition est **stochastique** : $\forall k, \ell,\ \pi_{k\ell} \geq 0$ et la somme des termes de chaque ligne est égale à 1 : $\forall k,\ \sum_\ell \pi_{k\ell} = 1$.

On appelle **distribution stationnaire** la distribution de probabilité correspondant à tout vecteur propre associé à la valeur propre 1 de π :

$$m = m.\pi$$

Un tel vecteur m rend compte d'un comportement stochastique stable du système.

Une chaîne de Markov Z_t d'ordre 1 à K états, homogène dans le temps, est définie par les paramètres suivants :

- les probabilités initiales $m_k = P\{Z_1 = k\}$, pour $k = 1, ..., K$, avec $\sum_k m_k = 1$.
- les probabilités de transition $\pi_{k\ell} = P\{Z_t = \ell | Z_{t-1} = k\}$, pour $k, \ell = 1, ..., K$, avec $\sum_\ell \pi_{k\ell} = 1$.

La connaissance de la distribution de probabilités initiale m_0 et de la matrice de transition π suffit pour décrire complètement le comportement de la chaîne de Markov au cours du temps. En effet, on a la propriété suivante :

$$P\{Z_n = k\} = m_0.\pi^n.$$

Dans le cas d'une chaîne de Markov, le temps de séjour dans l'état k est modélisé implicitement. Rester u fois dans un état consiste à boucler $(u-1)$ fois puis à sortir. La loi d'occupation, ou du temps de séjour dans l'état k, est la loi géométrique de paramètre $1 - \pi_{kk}$, où $\pi_{k\ell}$ est la probabilité de transition de l'état k à l'état ℓ. Le temps moyen de séjour dans l'état k est donc égal à $1/(1 - \pi_{kk})$.

5.2 Démonstrations des formules de l'algorithme Forward/Backward

La récurrence avant (Forward) implique le calcul des probabilités de filtrage $F_{t\ell} = P(Z_{t\ell}|X_1^t)$ pour chaque état ℓ du temps 1 au temps n. Les calculs sont donnés pour une chaîne de Markov homogène.
On note $X_a^b = \{X_a, ..., X_b\}$ avec $a < b$, et $Z_{tk} = 1$ si le processus Z pris en l'instant t est égal à k

- Initialisation pour $t = 1$:

$$F_{1\ell} = P(Z_{1\ell}|X_1) = \frac{m_\ell f(X_1; \theta_\ell)}{\sum_\ell m_\ell f(X_1; \theta_\ell)}.$$

- Pour $t = 2, ..., n$:

$$\begin{aligned} F_{t\ell} = P(Z_{t\ell}|X_1^t) &= \frac{P(Z_{t\ell}, X_1^t)}{P(X_1^t)} \\ &= \frac{\sum_k P(Z_{t-1,k}, Z_{t\ell}, X_1^t)}{P(X_1^t)} \\ &= \frac{\sum_k P(Z_{t-1,k}, Z_{t\ell}, X_t|X_1^{t-1})P(X_1^{t-1})}{P(X_t|X_1^{t-1})P(X_1^{t-1})} \end{aligned}$$

Or, on sait que :

$$\begin{aligned} P(Z_{t-1,k}, Z_{t\ell}, X_t|X_1^{t-1}) &= P(X_t|X_1^{t-1}, Z_{t-1,k}, Z_{t\ell}) \times P(Z_{t-1,k}, Z_{t\ell}|X_1^{t-1}) \\ &= P(X_t|Z_{t\ell}) \times P(Z_{t-1,k}, Z_{t\ell}|X_1^{t-1}) \\ &= P(X_t|Z_{t\ell}) \times P(Z_{t\ell}|X_1^{t-1}, Z_{t-1,k}) \times P(Z_{t-1,k}|X_1^{t-1}) \\ &= \underbrace{P(X_t|Z_{t\ell})}_{\phi_\ell(X_t)} \times \underbrace{P(Z_{t\ell}|Z_{t-1,k})}_{Z \text{ est une chaîne de Markov}} \times \underbrace{P(Z_{t-1,k}|X_1^{t-1})}_{F_{t-1,k}} \end{aligned}$$

Donc en remplacant dans l'expression précédente de $F_{t\ell}$, on obtient la formule de

récurrence suivante :

$$\begin{aligned} F_{t\ell} &= \frac{\sum_k P(X_t|Z_{t\ell})P(Z_{t\ell}|Z_{t-1,k})P(Z_{t-1,k}|X_1^{t-1})}{P(X_t|X_1^{t-1})} \\ &= \frac{f(X_t;\theta_\ell)\sum_k \pi_{k\ell}F_{t-1,k}}{\sum_\ell f(X_t;\theta_\ell)\sum_k \pi_{k\ell}F_{t-1,k}}. \end{aligned}$$

La récurrence arrière (Backward) consiste à calculer soit $B_{t\ell}$ soit directement $\tau_{t\ell}$ (car $\tau_{t\ell} = B_{t\ell} \times F_{t\ell}$) pour chaque état ℓ en reculant de n à 1.

- Initialisation pour $t = n$:

$$\tau_{n\ell} = F_{n\ell} = P(Z_{n\ell}|X_1^n)$$

- Pour $t = n-1, ..., 1$:

$$\begin{aligned} \tau_{t\ell} = \sum_k P(Z_{t+1,k}, Z_{t\ell}|X_1^n) &= \sum_k \underbrace{P(Z_{t\ell}|X_1^n, Z_{t+1,k})}_{Z_{t\ell} \text{ ne dépend pas des } X_{t+1}^n} \underbrace{P(Z_{t+1,k}|X_1^n)}_{\tau_{t+1,k}} \\ &= \sum_k P(Z_{t\ell}|X_1^t, Z_{t+1,k})\tau_{t+1,k} \\ &= \sum_k \frac{P(Z_{t+1,k}, Z_{t\ell}, X_1^t)}{P(Z_{t+1,k}, X_1^t)}\tau_{t+1,k} \end{aligned}$$

Or on a :

$$\begin{aligned} P(Z_{t+1,k}, Z_{t\ell}, X_1^t) &= P(Z_{t+1,k}|Z_{t\ell}, X_1^t)P(Z_{t\ell}|X_1^t)P(X_1^t) \\ &= P(Z_{t\ell}|X_1^t)P(Z_{t+1,k}|Z_{t\ell})P(X_1^t) \end{aligned}$$

et :

$$\begin{aligned} P(Z_{t+1,k}, X_1^t) &= P(Z_{t+1,k}|X_1^t)P(X_1^t) \\ &= \sum_\ell P(Z_{t+1,k}, Z_{t\ell}|X_1^t)P(X_1^t) \\ &= \sum_\ell \underbrace{P(Z_{t+1,k}|X_1^t, Z_{t\ell})}_{Z_{t+1,k} \text{ ind de } X_1^t} P(Z_{t\ell}|X_1^t)P(X_1^t) \\ &= \sum_\ell \pi_{\ell k}F_{t\ell}P(X_1^t) \end{aligned}$$

En notant $G_{t+1,k} = P(Z_{t+1,k}|X_1^t) = \sum_\ell \pi_{\ell k}F_{t\ell}$ et en remplacant dans l'expression précédente de $\tau_{t\ell}$, on obtient la formule de récurrence suivante :

$$\begin{aligned} \tau_{t\ell} &= \sum_k \frac{P(Z_{t\ell}|X_1^t)P(Z_{t+1,k}|Z_{t\ell})P(X_1^t)}{G_{t+1,k}P(X_1^t)}\tau_{t+1,k} \\ &= \sum_k \frac{P(Z_{t\ell}|X_1^t)P(Z_{t+1,k}|Z_{t\ell})}{G_{t+1,k}}\tau_{t+1,k} \\ &= F_{t\ell}\sum_k \frac{\pi_{\ell k}\tau_{t+1,k}}{G_{t+1,k}} . \end{aligned}$$

Bibliographie

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control* **AC-19** 716-723.

Archer, G.E.B. and Titterington, D.M. (2002). Parameter estimation for hidden Markov chains. *Journal of Statistical Planning and Inference* **108** 365-390.

Banfield, J.D. and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803-821.

Baudry, J.P. (2009). Sélection de Modèle pour la Classification Non Supervisée. Choix du Nombre de Classes. *Thèse de Doctorat*, Université Paris-Sud XI.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* **41** 164-171.

Baum, L. and Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics* **37** 1554-1563.

Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based clustering and discriminant analysis with the MIXMOD software. *Computational Statistics and Data Analysis* **51/2** 587-600.

Biernacki, C., Celeux, G. and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis* **41** 561-575.

Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(7) 719-725.

Bilmes, J.A. (1998). A gentle Tutorial of the EM algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. *International Computer Science Institute.*

Celeux, G. and Diebolt, J. (1985). The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comp. Statis. Quaterly* **2** 73-82.

Celeux, G. and Durand, J.B. (2008). Selecting Hidden Markov Model State Number with Cross-Validated Likelihood. *Computational Statistics* 541-564.

Celeux, G. and Govaert, G. (1995). Gaussian Parsimonious Clustering Models. *Pattern Recognition* **28** 781-793.

Celeux, G. and Govaert, G. (1992). A Classification EM Algorithm for Clustering and Two Stochastic versions. *Computational Statistics and Data Analysis* **14** 315-332.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorihtm (with discussion). *J. R. Statis. Soc. B* **39** 1-38.

Devijver, P.A. (1985). Baum's forward-backward algorithm revisited. *Pattern Recognition Letters* **3** 369-373.

Foreman, L.A. (1993). Generalization of the viterbi algorithm. *Journal of Mathematics Applied in Business and Industry* **4** 351-367.

Forney, G.D.Jr (1973). The Viterbi algorithm. *Proceedings of the IEEE.*

Fraley, C. and Raftery, A. E. (2006). MCLUST version 3 for R : Normal mixture modeling and model-based clustering. *Tech. Rep. 504*, University of Washington, Department of Statistics, Seattle, WA.

Hathaway, R.J. (1986). Another interpretation of the em algorithm for mixture distributions. *Statistics & Probability Letters* **4** 53-56.

Karlis, D. and Xekalaki,E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics and Data Analysis*, **41**(3) :577-590.

Kass, E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American statistical Association* **90** 773-795.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A* **62**(1) :49-66.

Lebarbier, E. and Mary-Huard, T. (2011). Classification non supervisée. *Polycopié AgroParisTech.*

Lebarbier, E. and Mary-Huard, T. (2004). Le critère BIC : fondements théoriques et interprétation. *Technical Report 5315*, INRIA.

Maitra, R. (2009). Initializing partition-optimization algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **6** 144-157.

McLachlan, G.J. and Krishnan, T. (2008). The EM Algorithm and Extensions. *Wiley Series in Probability and Statistics.*

McLachlan, G.J. and Peel, D. (2000). Finite Mixture Models. *New York : John Wiley.*

Picard, F. (2007). An introduction to mixture models. *SSB Research Report* **7**.

Rabiner, L.R. (1989). A tutorial on Hidden Markov Models and selected applications in Speech Recognition. *Proceedings of the IEEE.*

Schwarz, G. (1977). Estimating the number of components in a finite mixture model. *Annals of Statistics* **6** 461-464.

Wei, G.C.G, and Tanner, M.A. (1991). A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85** 699-704.

Chapitre 3

Modèles de Markov cachés pour les données *tiling arrays*

Sommaire

1 Modélisation de la loi de la variable latente 48

2 Modélisation de la loi d'émission 51

2.1 Régressions linéaires 51

2.2 Loi gaussienne bidimensionnelle 53

2.3 Mélange de gaussiennes 58

3 Annexes 72

3.1 Estimateurs du modèle gaussien bidimensionnel sous contraintes . . 72

3.2 Estimateurs du modèle gaussien unidimensionnel sous contraintes de colinéarité 74

Bibliographie 76

Le modèle doit répondre à deux exigences essentielles : être raisonnablement simple pour que l'inférence soit réalisable, tout en saisissant au mieux les caractéristiques biologiques. Lors de l'analyse de données *tiling array*, l'objectif est de caractériser le comportement de chaque sonde en définissant un statut pour chacune. Les sondes étant régulièrement réparties le long du génome et suffisamment proches, la prise en compte de la dépendance existant entre deux mesures d'intensités successives est essentielle dans la modélisation. Les modèles HMM permettent d'intégrer les dépendances locales qui peuvent exister entre les sondes (cf. Chapitre 2). D'autre part, l'annotation structurale du génome informe des différentes facultés des sondes à être exprimées (cf. Chapitre 1), et cette information est importante à prendre en compte. Dans ce chapitre, nous présentons une approche de classification non supervisée où toute l'information disponible concernant les sondes est utilisée : les caractéristiques biologiques sont intégrées dans la modélisation de la loi du statut caché des sondes (cf. Section 1). La loi d'émission est choisie en fonction de la question biologique et une modélisation est proposée pour chaque type d'expériences (cf. Section 2). Dans le cadre des expériences de ChIP-chip, les données sont modélisées par un mélange de régressions (Section 2.1). Pour les données de ChIP-chip IP/IP ou de transcriptome, la loi d'émission est une gaussienne bidimensionnelle (Section 2.2). Puis nous proposons de raffiner la modélisation car l'hypothèse de distribution gaussienne n'est pas forcément adaptée au cas de données réelles (Section 2.3) : une modélisation semi-paramétrique où les lois d'émission sont elles-mêmes des mélanges est envisagée.

1 Modélisation de la loi de la variable latente

Nous proposons quatre modèles de classification non supervisée permettant de définir le statut de chaque sonde. Ces modèles diffèrent en fonction de l'information biologique intégrée, allant du plus simple au plus complet (cf. Figure 3.1).

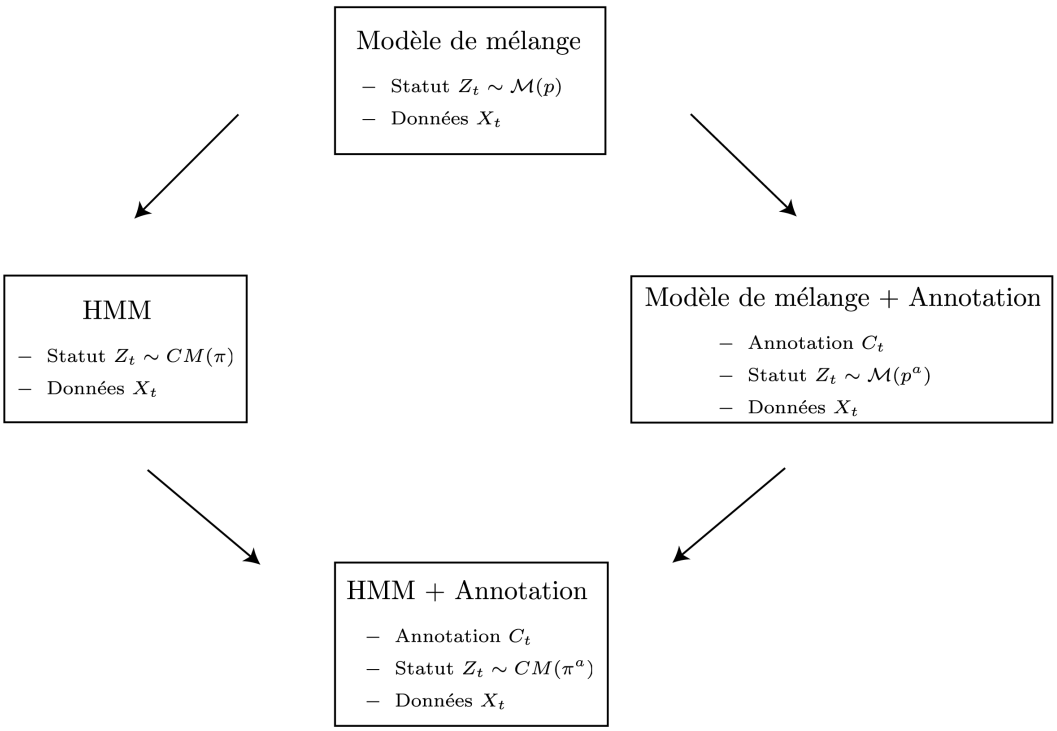

Fig. 3.1: Définition des sous-modèles.

Comme nous l'avons vu dans le chapitre 1, la position des sondes est importante puisqu'en raison de la haute densité des puces *tiling arrays* il existe une dépendance de signal entre les sondes adjacentes. D'autre part, l'annotation structurale informe sur la localisation des sondes dans les régions intergéniques, introniques ou exoniques. Cette information est utile, principalement dans les expériences transcriptomiques, car les sondes annotées exoniques seront préférentiellement hybridées par rapport aux sondes non codantes situées dans l'intergénique ou dans les introns. L'ajustement d'un même modèle d'un bout à l'autre du génome ne saurait refléter l'hétérogénéité qui peut exister entre les régions de nature différente au sein du génome, notamment entre le codant et l'intergénique. La dépendance entre sondes voisines et l'annotation sont donc deux caractéristiques biologiques à intégrer dans le modèle.

On note $Z_t \in \{1, ..., K\}$ la variable latente correspondant au statut caché de la sonde t. Le nombre d'états cachés K est supposé connu, il est déterminé en fonction de la nature des données. On note $Z_t = k$ si la sonde t appartient au groupe k.
L'annotation de la sonde t est notée $C_t \in \{1, ..., A\}$. Les différentes catégories d'annotation à prendre en compte dépendent des données étudiées. De la même manière, $C_t = a$ si la sonde t est dans la catégorie d'annotation a.

Définition des quatre modèles

Le modèle le plus simple, noté $\mathcal{M}_1$, est un modèle de mélange sans ajout d'information, où la variable latente suit une loi multinomiale de paramètres $p_1, ..., p_K$.

$$Z_t \sim \mathcal{M}(p_1, ..., p_K), \quad \text{avec } 0 \leq p_k \leq 1 \quad \text{et} \sum_{k=1}^{K} p_k = 1 \ .$$

Le deuxième modèle proposé tient compte de la dépendance : on suppose que $\{Z_t\}$ est une chaîne de Markov d'ordre 1, de matrice de transition π. Ce modèle est noté $\mathcal{M}_2$. Le statut de la sonde à la position *t-1* donne de l'information pour le statut de la sonde à la position t.

$$P(Z_t = \ell | Z_{t-1} = k) = \pi_{k\ell} \ .$$

Le troisième modèle ($\mathcal{M}_3$) considère des données ne présentant pas de structure spatiale particulière, mais tient compte de l'annotation. Les $\{Z_t\}$ sont alors supposés indépendants et distribués selon une loi multinomiale de paramètres $p_1^a, ..., p_K^a$ où p_k^a correspond à la proportion de sondes du groupe k pour la catégorie d'annotation a.

$$Z_t | C_t = a \sim \mathcal{M}(p_1^a, ..., p_K^a), \quad \text{avec } 0 \leq p_k^a \leq 1 \quad \text{et} \sum_{k=1}^{K} p_k^a = 1 \ .$$

Le quatrième et dernier modèle, noté $\mathcal{M}_4$, prend en compte toute l'information disponible pour chaque sonde : la position de la sonde le long du chromosome et son annotation structurale (exon, intron, intergénique, etc.). Aussi, nous proposons un modèle de Markov caché hétérogène permettant de distinguer différentes catégories d'annotation. La matrice de transition de la chaîne de Markov diffère en fonction de la catégorie d'annotation : elle est notée π^a, de coefficient

$$\pi_{kl}^a = P(Z_t = l | Z_{t-1} = k, C_t = a) \ .$$

Les observations $\{X_t\}$ sont supposées indépendantes conditionnellement aux $\{Z_t\}$. Le modèle est schématisé Figure 3.2. Ainsi, au final, le statut de la sonde t dépend à la fois de son annotation et du statut de la sonde $t - 1$.

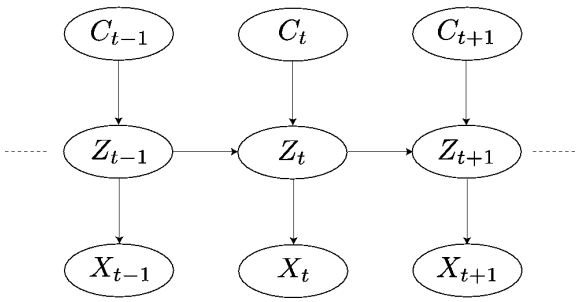

FIG. 3.2: Graphe des dépendances conditionnelles markoviennes d'ordre 1 avec prise en compte de l'annotation.

Inférence

L'inférence des modèles $\mathcal{M}_1$ et $\mathcal{M}_2$ est classique, elle est donnée dans le chapitre 2. Nous détaillons ici uniquement l'inférence des modèles prenant en compte l'annotation. Pour le modèle $\mathcal{M}_3$, les probabilités *a posteriori* sont définies par

$$\tau_{tk} = P(Z_t = k | C_t = a, X_t = x_t) \ .$$

Les estimateurs des probabilités *a posteriori* et des proportions sont :

$$\hat{\tau}_{tk} = \propto \ \hat{p}_k^a f(x_t, \hat{\theta}_k) \ ,$$

$$\hat{p}_k^a = \frac{\sum_{t \in a} \hat{\tau}_{tk}}{\text{card}(t \in a)} \ .$$

Pour le modèle de Markov caché hétérogène $\mathcal{M}_4$, l'espérance conditionnelle de la log-vraisemblance des données complètes s'écrit :

$$\begin{aligned} E\left[\mathcal{L}(X, Z; \phi) | X, C\right] &= \sum_k E(Z_{1,k}|X) \log(m_k^{C_1}) + \sum_{tk\ell a} E(Z_{t-1,k} Z_{t,\ell}|X) \log(\pi_{k\ell}^a) \\ &\quad + \sum_{t\ell} \tau_{t\ell} \log(f(x_t, \theta_k)) \ . \end{aligned}$$

Les estimateurs des probabilités *a posteriori* et de la matrice de transition sont calculés à l'aide de l'algorithme Forward/Backward pour une chaîne de Markov hétérogène :

$$\hat{\tau}_{tk} = F_{t,k} \sum_\ell \frac{\pi_{k\ell}^{C_{t+1}} \tau_{t+1,\ell}}{G_{t+1,\ell}} \ ,$$

avec

$$F_{t,l} = \frac{f(x_t, \theta_\ell) \sum_{k=1}^K \pi_{k\ell}^{C_t} F_{t-1,k}}{\sum_{\ell=1}^K f(x_t, \theta_\ell) \sum_{k=1}^K \pi_{k\ell}^{C_t} F_{t-1,k}} \quad \text{et } G_{t+1,\ell} = \sum_k \pi_{k\ell}^{C_{t+1}} F_{t,k} \ .$$

L'estimateur d'un terme de la matrice de transition est :

$$\hat{\pi}_{k\ell}^{C_t} = \sum_{t \in a} \frac{F_{t-1,k} \pi_{k\ell}^{C_t} \tau_{t\ell}}{G_{t,\ell}} \ ,$$

où l'on remarque que la somme sur t est discontinue : si la sonde t appartient à la catégorie d'annotation a, la sonde $t-1$ n'appartient pas forcément à a.

Choix du modèle

Si le modèle le plus complet est sans doute le plus approprié pour analyser des données *tiling arrays*, tous les sous-modèles peuvent être utiles. En particulier le modèle sans annotation pourra être utilisé si l'on souhaite analyser des données concernant un organisme pas encore séquencé ou dont l'annotation n'est pas encore assez fiable. En effet, l'annotation des génomes est un processus continuellement remis en cause, qui peut contenir des erreurs (cf. Chapitre 1). De même, si l'objectif principal de l'analyse est la détection de nouveaux sites de transcription, il paraît clair que le modèle sans connaissance de l'annotation *a priori* est préférable. Le choix du modèle et les différentes catégories d'annotation à prendre en compte sont discutés dans le chapitre 5 en fonction des données étudiées.

2 Modélisation de la loi d'émission

La loi d'émission est choisie en fonction de la question biologique, la modélisation est différente selon le type d'expériences. La modélisation tient compte de la spécificité du signal issu de puces *tiling arrays* qui est bidimensionnel compte tenu des deux intensités à comparer. Les modèles proposés sont des modèles de mélange bivariés qui permettent la comparaison directe des deux échantillons hybridés sur une puce.
Dans cette section, nous nous intéressons uniquement aux lois d'émission, et par abus de langage nous désignerons par "mélange" la loi des observations, quelle que soit l'hypothèse de dépendance sous-jacente de la variable latente.

2.1 Régressions linéaires

Lors d'une expérience de ChIP-chip, les signaux générés par les deux échantillons sont le signal IP et le signal INPUT. Le signal IP correspond à l'ADN immunoprécipité, c'est-à-dire les fragments d'ADN associés à la protéine d'intérêt ou à une marque chromatinienne ; le signal INPUT correspond à l'ADN génomique total (cf. Chapitre 1). Une observation importante est que les deux signaux issus des données de ChIP-chip ne sont pas équivalents. En effet, l'INPUT correspond à l'ADN génomique total et peut être considéré comme un échantillon de référence. Nous modélisons donc l'IP conditionnellement à l'INPUT. La Figure 3.3 montre que le signal IP ne dépend pas seulement du statut de la sonde, mais aussi du signal INPUT. On distingue deux nuages de points, l'un correspondant aux sondes enrichies (signal IP $>$ signal INPUT) et l'autre aux sondes normales (signal IP proche du signal INPUT). L'objectif est de retrouver ces deux populations afin d'identifier les régions génomiques où la protéine d'intérêt se fixe. En collaboration avec Marie-Laure Martin-Magniette, Tristan Mary-Huard et Stéphane Robin, nous avons développé une méthode appelée ChIPmix, fondée sur un modèle de mélange de régressions linéaires, pour mieux caractériser la relation entre l'IP et l'INPUT. Le modèle $\mathcal{M}_1$ sous hypothèse d'indépendance des observations est publié dans Martin-Magniette *et al.*, 2008 (cf. Annexe A).

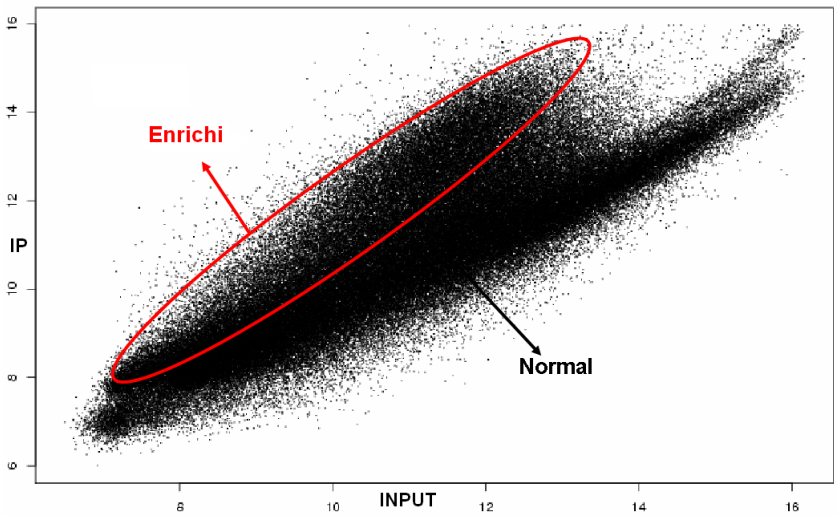

FIG. 3.3: Données de ChIP-chip : IP contre INPUT.

a) Modèle

Soit $X_t = (x_t, Y_t)$ les intensités log-INPUT et log-IP de la sonde t, respectivement, et Z_t son statut (inconnu) qui est binaire. On note $Z_t = 1$ si la sonde t appartient à la population enrichie, et $Z_t = 0$ si elle appartient à la population normale. On suppose que la relation entre l'IP et l'INPUT est linéaire quelle que soit la population, mais avec des paramètres différents. Plus précisément, on a :

$$\begin{array}{rcll} Y_t & \sim & \mathcal{N}(a_0 + b_0 x_t \ , \ \sigma^2) & \text{si } Z_t = 0 \text{ (sonde normale)} \\ Y_t & \sim & \mathcal{N}(a_1 + b_1 x_t \ , \ \sigma^2) & \text{si } Z_t = 1 \text{ (sonde enrichie)} \end{array}$$

Le vecteur de paramètres du modèle est donc $\theta_k = (a_k, b_k, \sigma^2)$ pour $k = 0, 1$, en plus des paramètres de la loi de la variable latente.

Habituellement, la quantité étudiée pour analyser les données de ChIP-chip est le log-ratio (IP/INPUT). La plupart des méthodes cherchent à identifier les sondes enrichies en supposant que la distribution du log-ratio est bimodale. On remarque qu'une analyse fondée sur le log-ratio (IP/INPUT) correspond au cas particulier où $b_0 = b_1 = 1$ dans le modèle de mélange de régressions. La figure 3.4 réalisée avec des données simulées montre que seules des droites de pentes égales et parallèles à la première bissectrice fournissent une distribution des log-ratios parfaitement bimodale. Travailler avec le log-ratio revient à faire l'hypothèse que la pente de la relation linéaire est la même quel que soit le statut de la sonde, ce qui est rarement le cas avec des données réelles.

Le modèle ci-dessus peut être adapté pour gérer simultanément plusieurs réplicats biologiques (MultiChIPmix).
Notons (x_{tr}, Y_{tr}) les intensités log-INPUT et log-IP de la sonde t pour le réplicat r. Pour chaque réplicat biologique r, on a les relations suivantes :

$$\begin{array}{rcll} Y_{tr} & \sim & \mathcal{N}(a_{0r} + b_{0r} x_{tr} \ , \ \sigma_r^2) & \text{si la sonde est normale} \\ Y_{tr} & \sim & \mathcal{N}(a_{1r} + b_{1r} x_{tr} \ , \ \sigma_r^2) & \text{si la sonde est enrichie} \end{array}$$

Les réplicats sont supposés indépendants, ce modèle permet donc de déterminer un unique statut par sonde pour tous les réplicats.

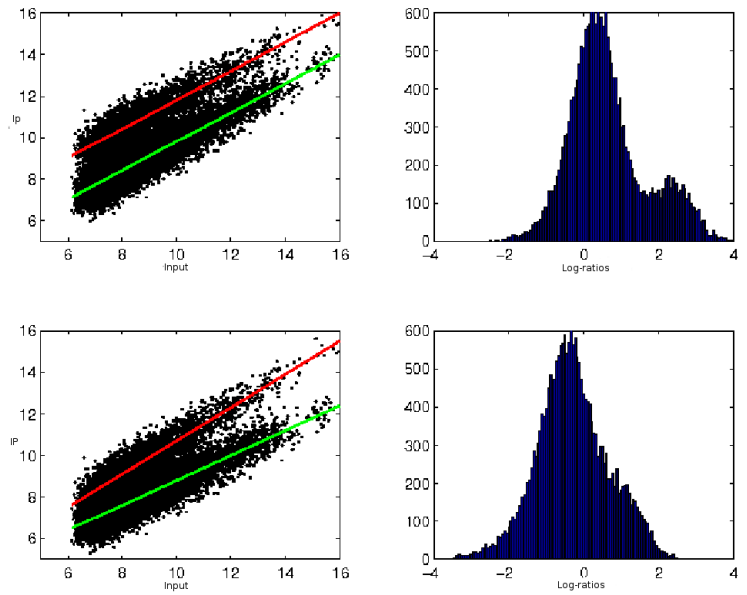

Fig. 3.4: Données simulées. **Haut :** Deux populations avec des relations linéaires de pentes égales. L'histogramme des log-ratios correspondant est bimodal. **Bas :** Deux populations avec des relations linéaires de pentes différentes. L'histogramme des log-ratios correspondant est unimodal.

b) Inférence

Le modèle de mélange de régressions est utilisé pour classer les sondes dans la population enrichie ou normale. Les paramètres du modèle sont les coefficients des deux régressions (a_0, b_0, a_1, b_1) et la variance σ^2 (ainsi que les paramètres de la loi de la variable latente). Ils sont estimés à l'aide de l'algorithme EM (cf. Chapitre 2). Afin d'éviter les problèmes de sensibilité aux valeurs initiales de l'algorithme EM, nous proposons des valeurs initiales issues du premier axe d'une Analyse en Composantes Principales (ACP). À l'étape E, les probabilités *a posteriori* τ_{tk} sont calculées pour chaque sonde, étant donné l'intensité IP et INPUT. À l'étape M, les paramètres de chaque classe sont estimés, en utilisant une régression pondérée, dans laquelle les poids sont donnés par les probabilités *a posteriori*. Les estimateurs de θ_k sont donnés par les formules classiques de régression linéaire :

$$\hat{b}_k = \frac{\sum_t \hat{\tau}_{tk}(x_t - \bar{x}_k)(Y_t - \bar{Y}_k)}{\sum_t \hat{\tau}_{tk}(x_t - \bar{x}_k)^2} \ ,$$

où $\bar{x}_k = \frac{\sum_t \hat{\tau}_{tk} x_t}{\sum_t \hat{\tau}_{tk}}$ et $\bar{Y}_k = \frac{\sum_t \hat{\tau}_{tk} Y_t}{\sum_t \hat{\tau}_{tk}}$.

$$\hat{a}_k = \bar{Y}_k - \hat{b}_k \bar{x}_k \ .$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_t \sum_k \hat{\tau}_{tk} \left[Y_t - (\hat{a}_k + \hat{b}_k x_t) \right]^2 \ .$$

2.2 Loi gaussienne bidimensionnelle

Dans le cadre des expériences transcriptomiques ou de ChIP-chip IP/IP, il s'agit d'étudier la différence entre deux conditions (un mutant et un sauvage par exemple). Les signaux générés par les deux échantillons jouent un rôle symétrique. L'objectif est de distinguer quatre groupes biologiquement interprétables : un groupe où il n'y a pas

d'hybridation (groupe bruit), un groupe où l'hybridation est identique dans les deux échantillons (groupe identique) et deux groupes dans lesquels l'hybridation est plus forte dans un échantillon que dans l'autre (groupes différentiellement hybridés) (cf. Figure 3.5).

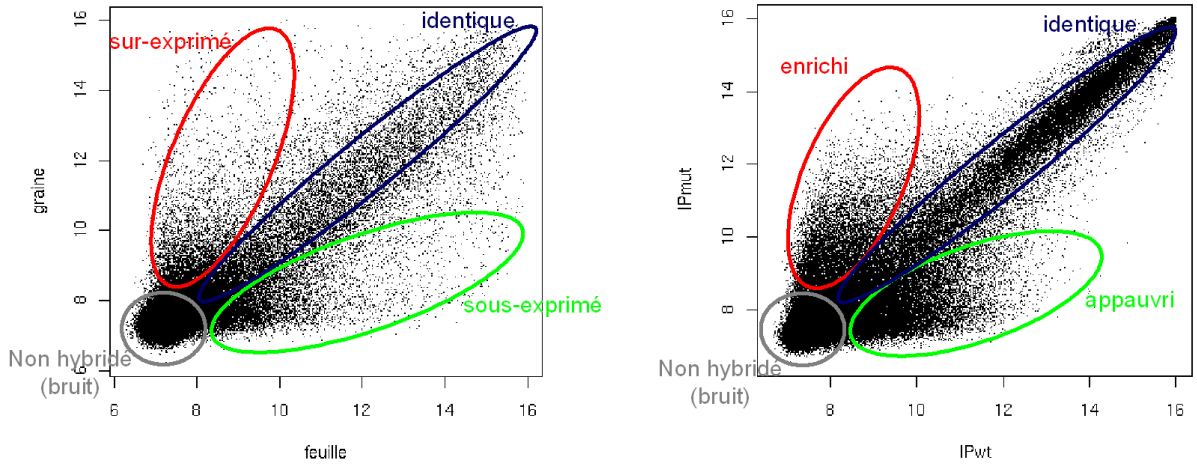

Fig. 3.5: Gauche : Représentation schématique des quatre groupes dans le cas d'une expérience de transcriptome (comparaison de deux conditions d'expression). Droite : Représentation schématique des quatre groupes dans le cas d'une expérience de ChIP-chip IP/IP (étude de la différence de méthylation entre un sauvage et un mutant).

a) Modèle

Soit $X_t = (X_{1t}, X_{2t})$ les log-intensités de la sonde t pour chaque échantillon, et Z_t son statut. Conditionnellement à Z_t on considère que les X_t sont indépendants et qu'ils suivent un mélange gaussien bidimensionnel à $K = 4$ composants :

$$X_t | Z_t = k \sim f(. \,; \mu_k, \Sigma_k) \qquad \forall\, k = 1, ..., K \ ,$$

où f est la fonction de densité du $k$ème composant du mélange. On suppose que $f(.)$ est la densité d'une distribution gaussienne bidimensionnelle de paramètres (μ_k, Σ_k), où μ_k est la moyenne et Σ_k est la matrice de variance-covariance. La densité f est définie par :

$$f(X_t \,; \mu_k, \Sigma_k) = \frac{1}{2\pi}[\det(\Sigma_k)]^{-1/2} \exp\left\{-\frac{1}{2}(x_t - \mu_k)^T \Sigma_k^{-1}(x_t - \mu_k)\right\}.$$

Comme nous l'avons vu dans le chapitre 2 Section 3.3, la densité gaussienne modélise une distribution ellipsoïdale de centre μ_k dont les caractéristiques géométriques (volume, forme, orientation) peuvent être contrôlées grâce à une décomposition spectrale de la matrice de variance Σ_k (Banfield et Raftery, 1993) :

$$\Sigma_k = \lambda_k D_k A_k D_k^T \ , \tag{3.1}$$

où λ_k représente le volume, D_k représente l'orientation et A_k représente la forme.
Il existe déjà de nombreux modèles de classification proposés par Celeux et Govaert (1995). Nous avons appliqué ces modèles sur les jeux de données réelles mais les quatre composants

obtenus ne permettent pas d'identifier correctement les quatre groupes biologiquement interprétables représentés schématiquement Figure 3.5.
Le modèle sélectionné par les critères BIC et ICL est le modèle général $\lambda_k D_k A_k D_k^T$, où toutes les caractéristiques varient en fonction du composant. Les résultats obtenus ne sont pas satisfaisants (cf. Figure 3.6). En effet, un seul composant couvre les groupes différentiellement hybridés et trois composants sont concentrés autour des faibles intensités où se trouve la majorité des sondes.

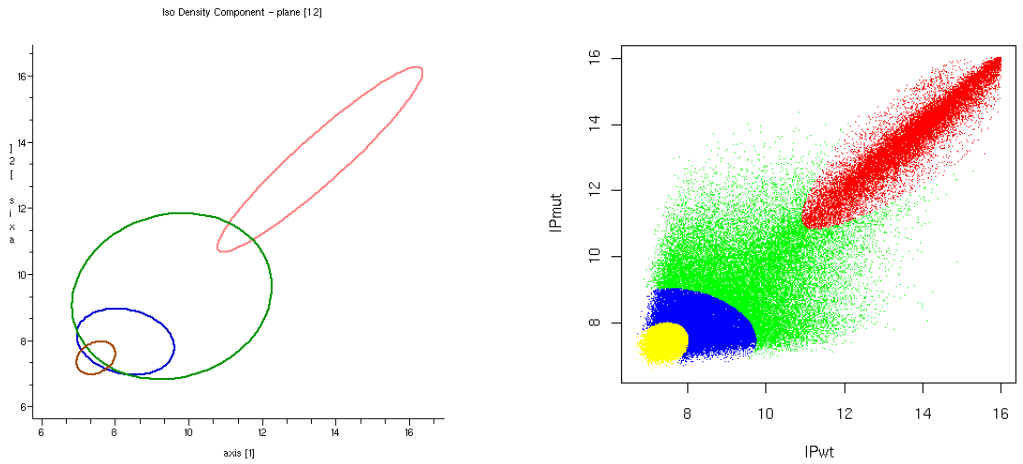

Fig. 3.6: Gauche : Isodensité des quatre gaussiennes pour le modèle $\lambda_k D_k A_k D_k^T$, Droite : Classement des sondes en quatre groupes avec la règle du MAP.

Les modèles qui considèrent un volume constant pour les quatres composants sont un peu meilleurs du point de vue de l'interprétation, mais deux composants sont très chevauchants et on ne retrouve pas le groupe d'intensités faibles représentant le bruit. Beaucoup de sondes sont alors déclarées hybridées à tort (cf. Figure 3.7).

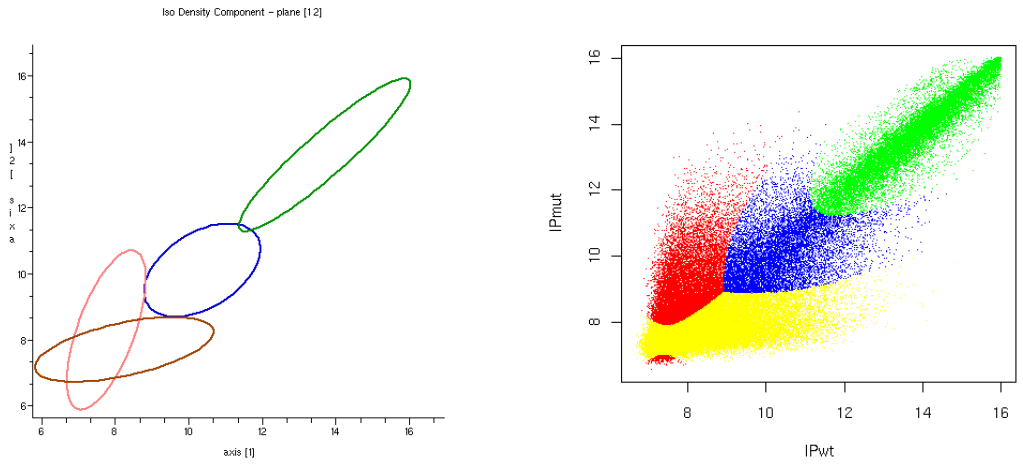

Fig. 3.7: Gauche : Isodensité des 4 gaussiennes pour le modèle $\lambda D_k A_k D_k^T$, Droite : Classement des sondes en 4 groupes avec la règle du MAP.

Afin de modéliser au mieux les données, nous ajoutons donc des contraintes aux modèles proposés par Celeux et Govaert (1995) (cf. Figure 3.8). Les contraintes supplémentaires sont déduites de connaissances biologiques issues des données. En effet, nous savons que le groupe de faibles intensités (groupe bruit) et le groupe où les intensités des sondes ont un comportement similaire dans les deux échantillons (groupe identique) ont la même orientation, qui est proche de la première bissectrice. Nous supposons donc que la matrice d'orientation D est identique pour ces deux groupes. D'autre part, il est connu que le paramètre de variance est le plus difficile à modéliser dans les modèles de mélanges gaussiens et que des variances hétéroscédastiques donnent souvent des résultats instables. Une contrainte de variance est donc aussi imposée. On s'attend à ce que la dispersion autour du grand axe de l'ellipse soit similaire dans les quatre groupes. Étant donné que la première valeur propre de Σ_k est associée au grand axe de l'ellipse et que la deuxième est associée au petit axe, la seconde valeur propre de Σ_k est supposée constante dans les quatre groupes.

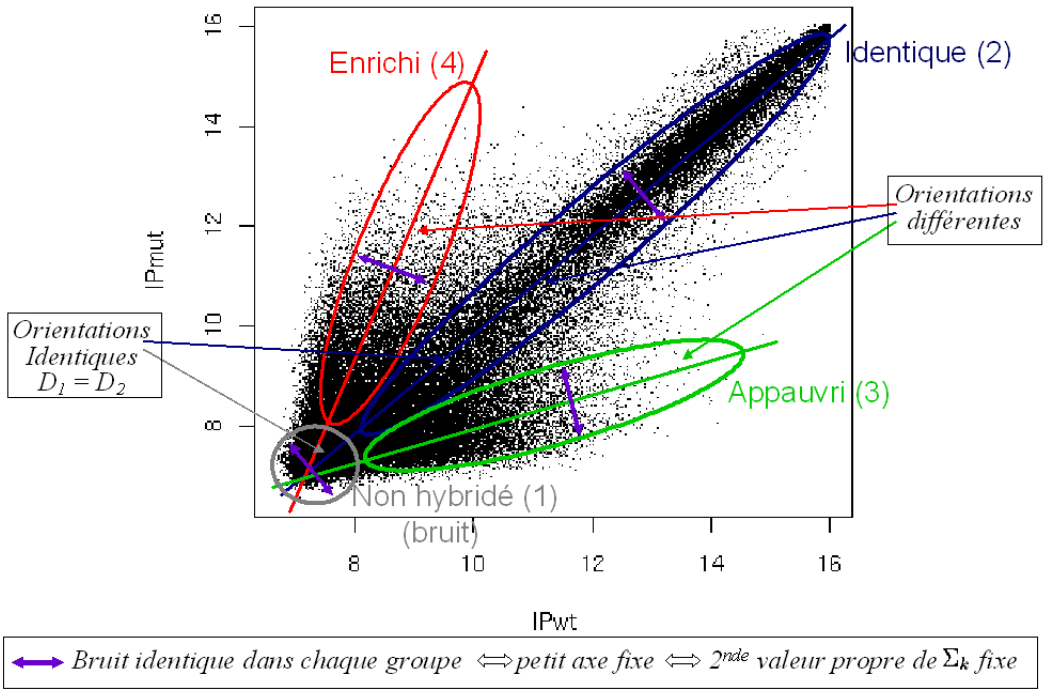

Fig. 3.8: Explication schématique de la modélisation.

En utilisant la décomposition des matrices de variance et sous nos contraintes, le modèle est donc résumé ainsi :

$$\begin{aligned}
\Sigma_k &= D_k \Lambda_k D_k^T, \qquad \text{pour } k = 1, ..., 4 \ ; \\
D_1 &= D_2 = D \ ; \\
\Lambda_k &= \begin{pmatrix} u_{1k} & 0 \\ 0 & u_2 \end{pmatrix}, \qquad \text{avec } u_{1k} > u_2, \ \text{pour } k = 1, ..., 4.
\end{aligned}$$

où les groupes 1 et 2 correspondent aux groupes de même orientation (groupe identique et groupe bruit) et la matrice Λ_k ($\Lambda_k = \lambda_k A_k$) est une matrice diagonale qui contient les valeurs propres de Σ_k.

L'originalité de ce modèle est de proposer la possibilité d'avoir certains composants avec une orientation fixe et d'autres composants avec une orientation libre. De plus il est possible de fixer seulement l'une des deux valeurs propres dans le choix du volume et de la forme pour un même composant du modèle. Cette approche est plus flexible que celle de Celeux et Govaert (1995) où chaque terme de la décomposition est soit égal pour tous les groupes, soit spécifique pour chaque groupe.

Le modèle est décrit pour $K = 4$ groupes, mais les modèles avec $K = 2$ et $K = 3$ sont également envisageables. En effet, selon les expériences, l'un des groupes différentiellement hybridés (ou les deux) peuvent ne pas exister si les échantillons sont suffisamment semblables. La contrainte d'orientation des deux premiers composants reste inchangée car on suppose qu'il existe toujours des sondes non hybridées et identiquement hybridées dans les deux conditions.

Le modèle $\mathcal{M}_1$ le plus simple (sans prise en compte de la dépendance spatiale ni de l'annotation) a fait l'objet d'une publication dans Modulad (Bérard *et al.*, 2009, Annexe B), le modèle complet $\mathcal{M}_4$ est soumis dans SAGMB (cf. Annexe E).

b) Inférence

Les paramètres de moyenne et de variance du modèle sont estimés à l'aide d'un algorithme EM. L'initialisation de l'algorithme EM est réalisée en définissant les quatre groupes par des contraintes géométriques. Le groupe bruit est caractérisé par une proportion initiale de sondes ou par une intensité maximale fixée en fonction de l'expérience biologique, et les trois autres groupes sont déterminés à l'aide de deux droites équidistantes de la première bissectrice.

Dans l'étape M, trouver l'estimateur de Σ_k revient à trouver les estimateurs de D, D_k et Λ_k, où les estimateurs de D et Λ_k sont spécifiques pour satisfaire les contraintes imposées dans le modèle défini Section 2.2.

L'estimateur de μ est :

$$\hat{\mu} = \bar{X}_k = \frac{\sum_{t=1}^{n} \hat{\tau}_{tk} X_t}{\sum_{t=1}^{n} \hat{\tau}_{tk}} .$$

L'estimateur de D_k pour les deux composants d'orientation différente est le même que celui proposé par Celeux et Govaert (1995) pour des composants d'orientations différentes. C'est la matrice des vecteurs propres de W_k, où $W_k = \sum_{t=1}^{n} \tau_{tk}(X_t - \bar{X}_k)(X_t - \bar{X}_k)^T$.

L'estimateur de D est défini dans la proposition suivante.

Proposition 1. *Soit* $W_k = \sum_{t=1}^{n} \tau_{tk}(X_t - \bar{X}_k)(X_t - \bar{X}_k)^T$ *une matrice de la forme* $\begin{pmatrix} w_{1k} & w_{2k} \\ w_{2k} & w_{4k} \end{pmatrix}$ *et* $\Lambda_k = \begin{pmatrix} u_{1k} & 0 \\ 0 & u_2 \end{pmatrix}$,
avec $u_{1k} > u_2$, *pour* $k = 1, ..., 4$. *L'estimateur du maximum de vraisemblance de la matrice d'orientation* D *identique pour les deux premiers composants de même orientation est de la forme* $\begin{pmatrix} \sqrt{\hat{d}} & -\sqrt{1-\hat{d}} \\ \sqrt{1-\hat{d}} & \sqrt{\hat{d}} \end{pmatrix}$, *où* $\hat{d}$ *est le minimum de la fonction :*

$$f(d) = \sum_{k=1}^{2} \left\{ \frac{d^2 w_{1k} + 2w_{2k} d\sqrt{1-d^2} + w_{4k}(1-d^2)}{u_{1k}} + \frac{d^2 w_{4k} + 2w_{2k} d\sqrt{1-d^2} + w_{1k}(1-d^2)}{u_2} \right\}$$

L'estimateur $\hat{d}$ est défini par :

$$\widehat{d^2} - \frac{1}{2} = \pm \frac{N_{1,4}}{2\left[\{N_{1,4}\}^2 + 4\{N_2\}^2\right]^{1/2}}, \quad \text{avec } \hat{d} > 0, \tag{3.2}$$

où $N_{1,4} = \sum_{k=1}^{2} (\hat{w}_{1k} - \hat{w}_{4k})(\hat{u}_2 - \hat{u}_{1k})/\hat{u}_{1k}\hat{u}_2$ *et* $N_2 = \sum_{k=1}^{2} \hat{w}_{2k}(\hat{u}_2 - \hat{u}_{1k})/\hat{u}_{1k}\hat{u}_2$.

L'estimateur de Λ_k est défini dans la proposition suivante.

Proposition 2. *Soit B_k la matrice définie par $B_k = D_k^T W_k D_k$ de la forme $\begin{pmatrix} b_{1k} & b_{3k} \\ b_{4k} & b_{2k} \end{pmatrix}$. L'estimateur du maximum de vraisemblance de Λ_k est de la forme $\begin{pmatrix} \hat{u}_{1k} & 0 \\ 0 & \hat{u}_2 \end{pmatrix}$, où*

$$\begin{cases} \hat{u}_{1k} & = \hat{b}_{1k} / \sum_{t=1}^{n} \hat{\tau}_{tk} \\ \hat{u}_2 & = \sum_{k=1}^{4} \hat{b}_{2k}/n \end{cases} \tag{3.3}$$

Au final, l'estimateur de Σ_k est donc :

$$\hat{\Sigma}_k = \begin{cases} \hat{D}_k^T \hat{\Lambda}_k \hat{D}_k \text{ si } k \geq 2 \\ \hat{D}^T \hat{\Lambda}_k \hat{D} \text{ si } k < 2, \end{cases}$$

avec $\hat{D}$ défini par (3.2), $\hat{D}_k$ est la matrice des vecteurs propres de W_k et $\hat{\Lambda}_k$ défini par (3.3). Les preuves des formules d'estimation sont données dans l'Annexe 3.1.

2.3 Mélange de gaussiennes

Les modèles à variables latentes considèrent généralement des lois d'émission paramétriques pour chaque groupe. Le choix de la distribution pour la loi d'émission est réalisé en fonction du problème, mais il est aussi motivé par la simplicité des calculs. En pratique, les données ne sont pas toujours conformes à ces hypothèses de distribution. La Figure 3.9 représente l'histogramme des données projetées sur les grands axes du groupe identique et des groupes différentiellement hybridés. Les histogrammes sont construits à partir des données pondérées par les probabilités *a posteriori* du groupe correspondant à l'axe sur lequel on projette. On remarque que les distributions empiriques de chaque groupe ne sont clairement pas unimodales et ne correspondent pas à des distributions gaussiennes.

Récemment, Sun *et al.* (2009) ont proposé un modèle de mélange de deux distributions pour analyser des données de ChIP-chip, où la distribution sous l'hypothèse de non enrichissement des sondes est supposée symétrique mais pas forcément gaussienne, ce qui permet une plus grande flexibilité. Chatzis (2010) propose l'utilisation de distributions non elliptiques afin d'obtenir une modélisation plus flexible dans le cas HMM. Il préconise la distribution MNIG (Multivariate Normal Inverse Gaussian) qui a une queue de distribution plus lourde que la gaussienne et qui offre la possibilité de modéliser des distributions asymétriques. Une autre manière de rendre les distributions plus flexibles est de travailler avec des mélanges de distributions, ce qui permet d'envisager une modélisation semi-paramétrique.

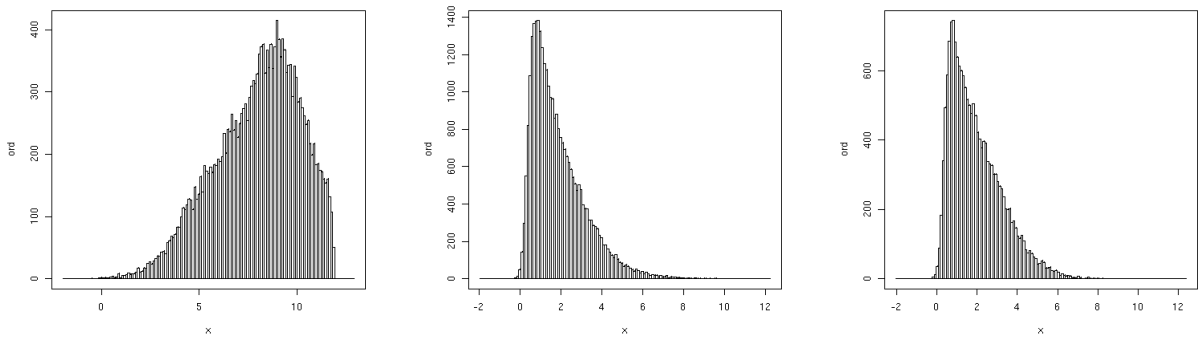

Fig. 3.9: Gauche : Distribution des données projetées sur le grand axe du groupe identique Δ_2 ; Centre : Distribution des données projetées sur le grand axe du groupe différentiellement hybridé dans une condition Δ_3 ; Droite : Distribution des données projetées sur le grand axe du groupe différentiellement hybridé dans l'autre condition Δ_4.

Dans cette section, nous proposons de nous affranchir de l'hypothèse de distribution gaussienne en considérant la loi d'émission comme étant un mélange de distributions gaussiennes (cf Section a)). Deux approches sont envisagées : l'une avec contraintes de colinéarité des composants dans les groupes, et l'autre plus générale, sans contraintes sur les composants. Cette seconde approche a été développée en collaboration avec Stevenn Volant [1]. L'objectif est d'obtenir une meilleure estimation de la densité de chaque groupe et donc de mieux définir la frontière de classification entre les groupes. L'inférence est réalisée avec l'algorithme EM (cf Section b)). La Section c) s'intéresse au choix du nombre de groupes dans les mélanges de mélange. Dans la Section d), nous proposons une initialisation de l'algorithme EM fournissant un agencement des composants dans chaque groupe. Cette initialisation est requise pour le calcul des estimateurs. Enfin, la Section e) fait l'objet d'une étude de simulation pour évaluer les performances des critères de sélection définis Section c) et pour comparer les critères d'appariement introduits Section d).

a) Modèle

Soient $\{X_t\}_{1,...,n}$ les données observées, où $X_t \in \mathbb{R}^d$. On note $\{Z_t\}$ la variable latente, prenant ses valeurs dans $\{1, ..., K\}$. Les observations $\{X_t\}$ sont indépendantes conditionnellement à Z, de loi d'émission ψ_k $(k = 1, ..., K)$:

$$X_t | Z_t = k \sim \psi_k \ . \tag{3.4}$$

La loi d'émission ψ_k est considérée comme étant elle-même un mélange de L_k distributions paramétriques :

$$\psi_k = \sum_{\ell=1}^{L_k} \eta_{k\ell} f(.; \theta_{k\ell}), \tag{3.5}$$

où $\eta_{k\ell}$ est la proportion du ℓ-ième composant du groupe k ($\forall \ell \in \{1, ..., L_k\}$, $0 < \eta_{k\ell} < 1$ et $\sum_\ell \eta_{k\ell} = 1$) et L_k est le nombre de composants du groupe k. On note L le nombre de composants du modèle et $\sum_{k=1}^K L_k = L$.

[1] UMR AgroParisTech/INRA MIA 518

Notons que ce modèle peut être écrit de manière équivalente en introduisant une seconde variable latente $\{W_t\}$ prenant ses valeurs dans $\{1, ..., L\}$, et avec une distribution paramétrique standard (gaussienne) comme loi d'émission :

$$\forall\, t,\ X_t | W_{tk\ell} = 1 \sim f(.; \theta_{k\ell})\ .$$

Dans le contexte des HMM, les deux variables latentes Z et W sont deux chaînes de Markov emboîtées. La variable latente $\{Z_t\}$ est une chaîne de Markov de distribution stationnaire m qui prend ses valeurs dans l'ensemble $\{1, ..., K\}$ tandis que $\{W_t\}$ est une chaîne de Markov qui prend ses valeurs dans $\{1, ..., L\}$. La matrice de transition de W, notée $\Omega = \{\omega_{k,\ell;k',\ell'}\}$ avec $(k, k') \in \{1, ..., K\}^2$ et $(\ell, \ell') \in \{1, ..., L_k\}^2$ est contrainte telle que :

$$\omega_{k,\ell;k',\ell'} = \pi_{k,k'} \eta_{k'\ell'}\ , \tag{3.6}$$

où π est la matrice de transition de Z.Le vecteur de paramètres du modèle est : $\Theta = (\Pi, m, \{\eta_{k\ell}\}_{k,\ell}, \{\theta_{k\ell}\}_{k,\ell})$.
Ce modèle n'est actuellement pas défini dans le cas où l'annotation est prise en compte dans la loi de la variable latente.

Dans la suite, deux approches différentes sont envisagées : l'une avec contraintes de colinéarité des composants dans les groupes, et l'autre plus générale, sans contraintes sur les composants.

Modèle avec contraintes de colinéarité. Le premier modèle proposé est similaire à celui présenté Section 2.2, mais la distribution de chaque groupe est un mélange de gaussiennes au lieu d'une unique distribution gaussienne. Ce modèle est développé pour un nombre fixe de groupes $K = 4$, où les quatre groupes sont ceux présentés Section 2.2 : groupe bruit, groupe identique, et deux groupes différentiellement hybridés (notés respectivement groupes 1, 2, 3 et 4). Nous considérons trois axes Δ_2, Δ_3 et Δ_4 correspondant respectivement au grand axe des groupes 2, 3 et 4, et concourants au barycentre du groupe bruit. Les composants gaussiens du groupe k sont contraints à être colinéaires le long de l'axe Δ_k pour $k = 2, 3, 4$ (cf Figure 3.10).

Soit $X_t = (X_{1t}, X_{2t})$ les log-intensités de la sonde t.

- Le groupe bruit est un groupe particulier, considéré circulaire et représenté par une gaussienne sphérique :

$$X_t | Z_t = 1 \sim \mathcal{N}\left(\begin{pmatrix}\mu_1^1\\ \mu_1^2\end{pmatrix}, \sigma^2 I_2\right)\ .$$

- Les trois autres groupes sont chacun représentés par un mélange de gaussiennes. La projection des données sur les axes Δ_k nous permet de travailler avec des mélanges unidimensionnels. Soient (U_{tk}, V_{tk}) les coordonnées de (X_{1t}, X_{2t}) dans le repère $(\Delta_k, \Delta_k^\perp)$. On considère un mélange gaussien unidimensionnel le long de chaque axe Δ_k et une distribution unique pour tous les composants selon $\Delta_k^\perp$:

$$(V_{tk} | Z_t = k) \sim \mathcal{N}(0, \sigma^2) \ \text{ et } \ (U_{tk} | Z_t = k) \sim \psi_k\ ,$$

  où ψ_k est défini comme dans l'équation (3.5), avec $\eta_{k\ell}$ est la proportion du ℓ-ième composant du mélange dans le groupe k et $f(.; \theta_{k\ell}) \sim \mathcal{N}(\mu_{kl}, \sigma_{kl}^2)$.

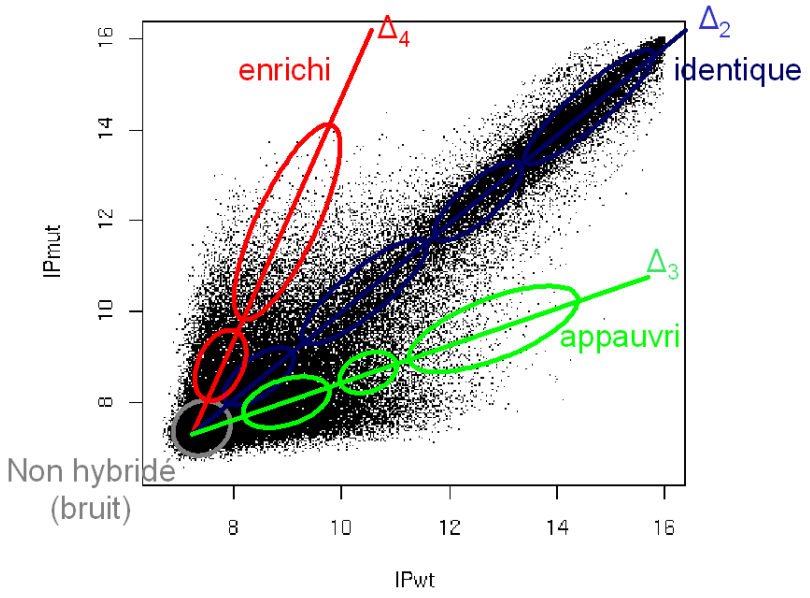

Fig. 3.10: Représentation schématique des mélanges de gaussiennes dans chaque groupe, le long des trois axes.

Modèle sans contraintes de colinéarité. Le second modèle proposé est un modèle sans contraintes particulières sur les composants. Chaque composant est distribué selon une gaussienne bidimensionnelle sphérique. La matrice de variance est supposée identique pour tous les composants (cf Figure 3.11). Ce modèle est plus général que le précédent, le nombre de groupes K n'est pas nécessairement fixé égal à 4.

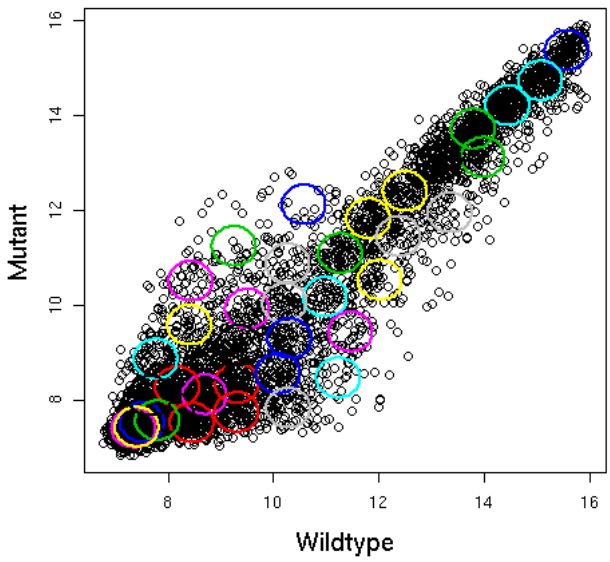

Fig. 3.11: Représentation schématique des composants gaussiens sphériques sans contraintes.

b) Inférence

L'inférence peut être réalisée avec l'algorithme EM. À l'étape M, les estimateurs des paramètres sont obtenus en maximisant :

$$\mathbb{E}\left[\log P(X,Z;\Theta)|X\right] = \mathbb{E}\left[\log P(Z;\Theta)|X\right] + \mathbb{E}\left[\log P(X|Z;\Theta)|X\right] \ . \tag{3.7}$$

La maximisation du premier terme de l'équation (3.7) est classique et résulte en l'estimation des paramètres de la distribution de la variable latente (cf. Chapitre 2). Comme les lois d'émission sont des mélanges, la maximisation du second terme

$$\mathbb{E}\left[\log P(X|Z;\Theta)|X\right] = \sum_t \sum_k \tau_{tk} \log \left[\sum_\ell \eta_{k\ell} f(X_t;\theta_{k\ell})\right] \tag{3.8}$$

nécessite des calculs supplémentaires. Dans l'équation (3.8), τ_{tk} est la probabilité *a posteriori* qu'une observation t appartienne au groupe k, définie par $P(Z_t = k|X)$ et est estimée à l'étape E.
Notons que pour k fixé, l'équation (3.8) correspond à une version pondérée de la vraisemblance d'un modèle de mélange indépendant. On peut donc appliquer une seconde fois la décomposition de l'algorithme EM. Pour ce faire, nous introduisons la variable latente $\{W_t\}_t$ qui fait référence au composant ℓ du groupe k telle que :

$$(W_{tk}|Z_t = k) \sim \mathcal{M}(\eta_k)$$

où η_k est le vecteur $(\eta_{k1}, ..., \eta_{kL_k})$.
On a :

$$\mathbb{E}\left[\log P(X|Z;\Theta)|X\right] \ = \ \mathbb{E}\left[\log P(X,W|Z;\Theta)|X\right] - \mathbb{E}\left[\log P(W|X,Z;\Theta)|X\right] \ . \tag{3.9}$$

Par analogie avec l'algorithme EM classique (avec une seule variable latente), la propriété fondamentale établie par Dempster *et al.* (1977) (cf. Chapitre 2, Section 3.1) peut être appliquée à l'équation (3.9). Maximiser $\mathbb{E}\left[\log P(X|Z;\Theta)|X\right]$ en Θ revient donc à maximiser uniquement le premier terme de l'équation $\mathbb{E}\left[\log P(X,W|Z;\Theta)|X\right]$, et on a :

$$\begin{aligned}\mathbb{E}\left[\log P(X,W|Z;\Theta)|X\right] &= \mathbb{E}\left[\log P(X|W,Z;\Theta)|X\right] + \mathbb{E}\left[\log P(W|Z;\Theta)|X\right] \\ &= \sum_t \tau_{tk} \sum_\ell \delta_{tk\ell} \log f(X_t;\theta_{k\ell}) + \sum_t \tau_{tk} \sum_\ell \delta_{tk\ell} \log \eta_{k\ell}\end{aligned} \tag{3.10}$$

où $\delta_{tk\ell} = \mathbb{E}\left[W_{tk\ell} = 1|Z_t = k, X_t\right]$.

La méthode d'inférence d'un modèle de mélange paramétrique standard peut à présent être appliquée pour estimer les paramètres de l'équation (3.10). À l'étape E on a :

$$\widehat{\delta}_{tk\ell} = \frac{\widehat{\eta}_{k\ell} f(X_t;\widehat{\theta}_{k\ell})}{\sum_{\ell=1}^{L_k} \widehat{\eta}_{k\ell} f(X_t;\widehat{\theta}_{k\ell})}, \tag{3.11}$$

et à l'étape M :

$$\widehat{\eta}_{k\ell} = \frac{\sum_t \widehat{\tau}_{tk}\widehat{\delta}_{tk\ell}}{\sum_t \widehat{\tau}_{tk}} \quad \text{et} \quad \widehat{\theta}_{k\ell} = \underset{\theta_{k\ell}}{\text{Argmax}} \sum_t \widehat{\tau}_{tk} \sum_\ell \widehat{\delta}_{tk\ell} \log f(X_t;\theta_{k\ell}) \ . \tag{3.12}$$

Modèle avec contraintes de colinéarité. Les estimateurs des paramètres spécifiques au modèle avec contraintes de colinéarité sont donnés dans la proposition ci-dessous.

Proposition 3. *On note les quatre groupes de 1 à 4, où 1 est le groupe bruit. L'estimateur de $\mu_{k\ell}$ (pour $k = 2, 3, 4$) est :*

$$\hat{\mu}_{k\ell} = \frac{\sum_t \hat{\tau}_{tk}\hat{\delta}_{tk\ell}U_{tk}}{\sum_t \hat{\tau}_{tk}\hat{\delta}_{tk\ell}} .$$

On a

$$\hat{\mu}_1^1 = \frac{\sum_t \hat{\tau}_{t1}X_{1t}}{\sum_t \hat{\tau}_{t1}} \qquad et \qquad \hat{\mu}_1^2 = \frac{\sum_t \hat{\tau}_{t1}X_{2t}}{\sum_t \hat{\tau}_{t1}} .$$

L'estimateur de $\sigma^2_{k\ell}$ (pour $k = 2, 3, 4$) est :

$$\hat{\sigma}^2_{k\ell} = \frac{\sum_t \hat{\tau}_{tk}\hat{\delta}_{tk\ell}(U_{tk} - \hat{\mu}_{k\ell})^2}{\sum_t \hat{\tau}_{tk}\hat{\delta}_{tk\ell}} .$$

L'estimateur de σ^2 est :

$$\hat{\sigma}^2 = \frac{\sum_t \hat{\tau}_{t1}(X_{1t} - \hat{\mu}_1^1)^2 + \sum_t \hat{\tau}_{t1}(X_{2t} - \hat{\mu}_1^2)^2 + \sum_{k=2}^4 \sum_{\ell=1}^{L_k} \sum_t \hat{\tau}_{tk}\hat{\delta}_{tk\ell}V_{tk}^2}{\sum_t \hat{\tau}_{t1} + \sum_t \hat{\tau}_{t1} + \sum_{k=2}^4 \sum_{\ell=1}^{L_k} \sum_t \hat{\tau}_{tk}\hat{\delta}_{tk\ell}} .$$

Soit Δ_k la droite de vecteur directeur $\vec{d_k} = (d_k, \sqrt{1 - d_k^2})$ représentant le grand axe du groupe k. L'estimateur de Δ_k (pour $k = 2, 3, 4$) est donné par :

$$\widehat{d}_k = Argmin(\widehat{Q}_2 + \widehat{Q}_3) ,$$

où

$$\widehat{Q}_2 = \sum_{k=2}^4 \sum_{\ell=1}^{L_k} \sum_t \hat{\tau}_{tk}\hat{\delta}_{tk\ell} \log f(U_{tk}; \hat{\mu}_{k\ell}, \hat{\sigma}^2_{k\ell}) ,$$

$$\widehat{Q}_3 = \sum_{k=2}^4 \sum_{\ell=1}^{L_k} \sum_t \hat{\tau}_{tk}\hat{\delta}_{tk\ell} \log f(V_{tk}; 0, \hat{\sigma}^2) + \sum_t \hat{\tau}_{t1} \log f(X_{1t}; \hat{\mu}_1^1, \hat{\sigma}^2) + \sum_t \hat{\tau}_{t1} \log f(X_{2t}; \hat{\mu}_1^2, \hat{\sigma}^2) .$$

Les détails des calculs des estimateurs sont donnés en Annexe 3.2.
Ce modèle est illustré sur des données réelles, et les ajustements des densités sur les histogrammes de la Figure 3.9 sont présentés dans le chapitre 5, Section 3.3. Le choix du nombre de composants par groupe ainsi que la contrainte de variance résiduelle σ^2 identique dans tous les groupes sont aussi discutés dans le chapitre 5, Section 3.3.

Modèle sans contraintes de colinéarité. Les estimateurs des paramètres du modèle sans contraintes de colinéarité sont ceux définis dans les équations (3.11) et (3.12).

c) Sélection de modèles

L'objectif de cette section est l'estimation du nombre de groupes K ou du nombre de composants L qui est souvent inconnu en pratique. Nous introduisons trois critères de vraisemblance pénalisée pour estimer le nombre de groupes, adaptés au contexte des HMM lorsque la loi d'émission est un mélange de distributions (définie équation (3.5)).

Lorsqu'un groupe est caractérisé par un mélange de composants, l'approximation de type BIC pour estimer le nombre de groupes est :

$$BIC(K,L) = \log P(X;\widehat{\Theta}_{K,L}) - \frac{\nu_{K,L}}{2}\log(n), \tag{3.13}$$

où $\nu_{K,L}$ est le nombre de paramètres libres du modèle à K groupes et L composants.
Concernant le critère ICL, deux approches sont envisageables. La plus simple est de le définir à partir de la chaîne de Markov W.

$$ICL_W(K,L) = \log P(X,Z,W;\widehat{\Theta}_{K,L}) - \mathbb{E}_{Z|X}\left[\mathcal{H}(W|Z,X)\right] - \mathcal{H}(Z|X) - \frac{\nu_{K,L}}{2}\log(n) \tag{3.14}$$

Cependant, la formule du critère ICL_W définie dans l'équation (3.14) est fondée uniquement sur la variable latente W se référant aux composants et non aux groupes. L'objectif principal étant l'estimation du nombre de groupes, nous définissons un autre critère, noté $ICL_Z(K)$, fondé sur la vraisemblance complète intégrée $P(X,Z|\Theta)$:

$$ICL_Z(K,L) = \log P(X,Z;\widehat{\Theta}_{K,L}) - \mathcal{H}(Z|X) - \frac{\nu_{K,L}}{2}\log(n)\ . \tag{3.15}$$

La différence entre ICL_W et ICL_Z est équivalente à la différence entre les deux entropies $\mathcal{H}_W$ et $\mathcal{H}_Z$. Le nombre de groupes estimé par ICL_W a tendance à être plus grand que celui estimé par ICL_Z.
Dans le cas indépendant, BIC et ICL_W sont constants quel que soit le nombre de groupes (le nombre de paramètres libres ne dépend pas de L et la vraisemblance observée reste toujours la même). D'autre part, le critère ICL_Z augmente avec le nombre de groupes. Ainsi, aucun de ces critères ne peut être utilisé pour estimer le nombre de groupes dans le cas indépendant. En revanche, dans le contexte des HMM, la vraisemblance varie en fonction du nombre de groupes. De plus, comme $\nu_{K,L}$ dépend de K et L, le nombre de paramètres libres d'un HMM à K états diffère de celui d'un HMM à $K-1$ états. Les trois critères peuvent donc être appliqués pour estimer le nombre de groupes.

d) Initialisation de l'algorithme EM

L'inférence du modèle, présentée Section b), requiert la connaissance de l'agencement des composants dans chaque groupe.

Modèle avec contraintes de colinéarité. Pour le modèle avec contraintes de colinéarité, cet agencement est connu *via* les contraintes imposées dans le modèle. L'initialisation de l'algorithme EM est déterminée à l'aide des résultats obtenus avec le modèle gaussien bidimensionnel présenté Section 2.2.

Modèle sans contraintes de colinéarité. Pour le modèle sans contraintes, cette étape d'initialisation est fondamentale. En effet, les composants ne sont pas définis par rapport à un groupe. L'idée naturelle de tester tous les regroupements possibles des composants n'est pas réalisable car elle conduit à une complexité algorithmique trop

élevée. Ce problème de combinaison de composants a suscité un certain intérêt. Dans le cas indépendant, Tantrum et Murua (2003) ont proposé une méthode hiérarchique de combinaison de composants gaussiens fondée sur un critère de vraisemblance. Deux composants sont fusionnés si l'influence de leur combinaison sur la vraisemblance est la plus faible, en supposant que le mélange des deux composants est ajusté selon une unique distribution gaussienne dont les paramètres sont obtenus à partir de ces deux composants. Li (2005) a proposé une méthode non-hiérarchique fondée sur l'approche des k-means. Cette méthode suppose implicitement que les composants sont de forme sphérique et est donc sous optimale pour d'autres formes de distribution. Baudry *et al.* (2010) ont proposé une méthode hiérarchique permettant d'obtenir la composition des groupes à partir de la minimisation d'un critère d'entropie. Dans ce cas, la densité résultant de chaque groupe est définie par un mélange de distributions. Cette approche conduit à une meilleure estimation de la densité de chaque groupe. Contrairement aux autres méthodes, Baudry *et al.* (2010) ne se concentrent pas uniquement sur la classification, mais aussi sur l'estimation de la densité. Pour plus de détails sur les méthodes d'appariement des composants dans le cas gaussien indépendant, voir Hennig (2010).

La méthode proposée par Baudry *et al.* (2010) est définie dans le cas d'observations indépendantes, nous proposons d'étendre cette méthode sous hypothèse de dépendance markovienne. Nous présentons donc une initialisation de l'algorithme EM fournissant un agencement des K groupes localement optimal, à partir d'un procédé hiérarchique. Cette initialisation consiste à fusionner les L composants en K groupes. À chaque étape du processus hiérarchique, les composants i et j les plus pertinents sont sélectionnés pour être appariés. Cela permet d'obtenir un optimum local de la log-vraisemblance.

Dans la suite, on se place à l'itération G de la classification hiérarchique ascendante : le modèle a été réduit de L à G groupes et l'objectif est maintenant d'obtenir la meilleure combinaison en $G-1$ groupes.
Le terme "composant" se réfère indifféremment à un ensemble d'observations de distributions f (aucun appariement) ou ψ (au moins un appariement).

La définition du modèle dans le contexte HMM (Section a)) permet d'envisager trois critères d'appariement différents fondés respectivement sur la vraisemblance de X, de X, W et de X, Z. Les trois critères d'appariement sont définis par :

$$\nabla^1_{ij} = \mathbb{E}\left[\log P(X; G'_{i\cup j})|X\right], \ \nabla^2_{ij} = \mathbb{E}\left[\log P(X, W; G'_{i\cup j})|X\right], \ \nabla^3_{ij} = \mathbb{E}\left[\log P(X, Z; G'_{i\cup j})|X\right]. \tag{3.16}$$

où $G'_{i\cup j}$ est le modèle à $G-1$ groupes obtenu en combinant les deux composants i et j du modèle à G groupes.
Les deux composants i et j appariés sont ceux maximisant l'un des critères $\nabla_{k\ell}$:

$$(i,j) = \underset{k,\ell\in\{1,\dots,G\}^2}{\text{Argmax}} \ \nabla_{k\ell}\,. \tag{3.17}$$

Notons que la maximisation de ∇ peut être définie de manière similaire à l'aide des critères de sélection de modèles :

$$\begin{aligned}
\underset{i,j\in\{1,\dots,G\}^2}{\text{Argmax}} \ \nabla^1_{ij} &= \underset{i,j\in\{1,\dots,G\}^2}{\text{Argmin}} \left[BIC(G) - BIC(G'_{i\cup j})\right], \\
\underset{i,j\in\{1,\dots,G\}^2}{\text{Argmax}} \ \nabla^2_{ij} &= \underset{i,j\in\{1,\dots,G\}^2}{\text{Argmin}} \left[ICL_W(G) - ICL_W(G'_{i\cup j})\right], \\
\underset{i,j\in\{1,\dots,G\}^2}{\text{Argmax}} \ \nabla^3_{ij} &= \underset{i,j\in\{1,\dots,G\}^2}{\text{Argmin}} \left[ICL_Z(G) - ICL_Z(G'_{i\cup j})\right].
\end{aligned}$$

Lorsqu'un nouveau composant i' est formé à partir de la réunion des deux composants i et j, un modèle à $G-1$ groupes est obtenu, pour lequel la densité du groupe i' est le mélange des distributions des composants i et j.
Dans le modèle à $G-1$ groupes, les estimateurs des paramètres ne correspondent pas à ceux du maximum de vraisemblance. Ceci s'explique par les contraintes appliquées à la matrice de transition permettant de rester dans le cadre des HMM à chaque étape. Afin d'approcher un maximum local, quelques itérations de l'algorithme EM sont effectuées pour mettre à jour les paramètres.

L'algorithme ci-dessous résume la procédure de combinaison des composants d'un HMM, en utilisant le critère d'appariement ∇^1 et le critère de sélection de modèle ICL_Z défini dans la section c). Les critères ∇^1 et ICL_Z semblent être les plus appropriés d'après l'étude de simulation Section e).

Algorithme.

1. *Inférer un HMM à L composants.*
2. *Pour $G = L, L-1, ..., 1$*
   - *Sélectionner les deux composants i et j à combiner tels que :*

$$(i,j) = \underset{k,\ell \in \{1,...,G\}^2}{\text{Argmax}} \ \nabla^1_{k\ell},$$

   - *On obtient un modèle à $G-1$ groupes où la densité du composant i' est ajustée par le mélange des distributions des composants i et j.*
   - *Mettre à jour les paramètres avec quelques itérations de l'algorithme EM*
3. *Sélectionner le nombre de groupes $\widehat{K}$:*

$$\hat{K} \ = \ \underset{\ell \in \{L,...,1\}}{\text{Argmax}} \, ICL_Z(\ell)$$

Notons que cette procédure nécessite un temps de calcul important de par l'implémentation des critères d'appariement impliquant le calcul de la vraisemblance observée, ainsi qu'à cause du nombre de modèles considérés dans la procédure hiérarchique qui vaut $\sum_{k=K}^{L} k(k-1)/2$. Nous nous intéressons actuellement à la diminution de ce temps de calcul. Une solution consiste à réduire l'espace des modèles à explorer en utilisant un critère d'élagage qui permet de considérer uniquement les composants les plus pertinents (ce point est discuté plus en détail en Conclusion et perspectives).

e) Étude de simulation.

Dans ce paragraphe, nous présentons une étude de simulation qui met en évidence les performances des différents critères de sélection BIC, ICL_W et ICL_Z. D'autre part, les performances des critères d'appariement de composants ∇^1, ∇^2 et ∇^3 sont comparées, et cette étude consiste aussi en la détermination de la meilleure combinaison des critères de sélection et d'appariement. Un autre objectif de l'étude est d'illustrer l'avantage de la prise en compte de la dépendance dans les données en comparant notre méthode à celle proposée par Baudry *et al.* (2010) dans le cas indépendant.

Le plan de simulation correspond à celui étudié par Baudry *et al.* (2010) dans lequel une dépendance markovienne a été ajoutée. Pour cela, une variable $\{Z_t\}_t$ est simulée selon une chaîne de Markov qui prend ses valeurs dans $\{1, ..., 4\}$. La loi d'émission est une gaussienne bidimensionnelle pour les deux premiers états et un mélange de deux gaussiennes bidimensionnelles pour les deux autres. Il y a donc six composants au total et seulement quatre groupes. Afin de mesurer l'influence de la dépendance markovienne, nous considérons quatre matrices de transition différentes telles que $\forall k \in \{1, ..., 4\}$, $P(Z_t = k|Z_{t-1} = k) = a$, avec $a \in \{0.8, 0.7, 0.6, 0.5\}$, et $P(Z_t = k'|Z_{t-1} = k) = (1 - a)/3$.
Pour contrôler la dispersion des données, un paramètre b est introduit dans la matrice de variance-covariance Σ_ℓ de la ℓ-ième distribution gaussienne tel que $\Sigma_\ell = b \begin{pmatrix} \sigma_{\ell 1} & \sigma_{\ell 12} \\ \sigma_{\ell 12} & \sigma_{\ell 2} \end{pmatrix}$, où $b \in \{1, 3, 5, 7\}$. Les groupes sont bien séparés si $b = 1$ (cas de Baudry *et al.*, 2010), et $b = 7$ correspond à un fort recouvrement des groupes. La Figure 3.12 présente un exemple de jeu de données simulées pour différentes valeurs de b.

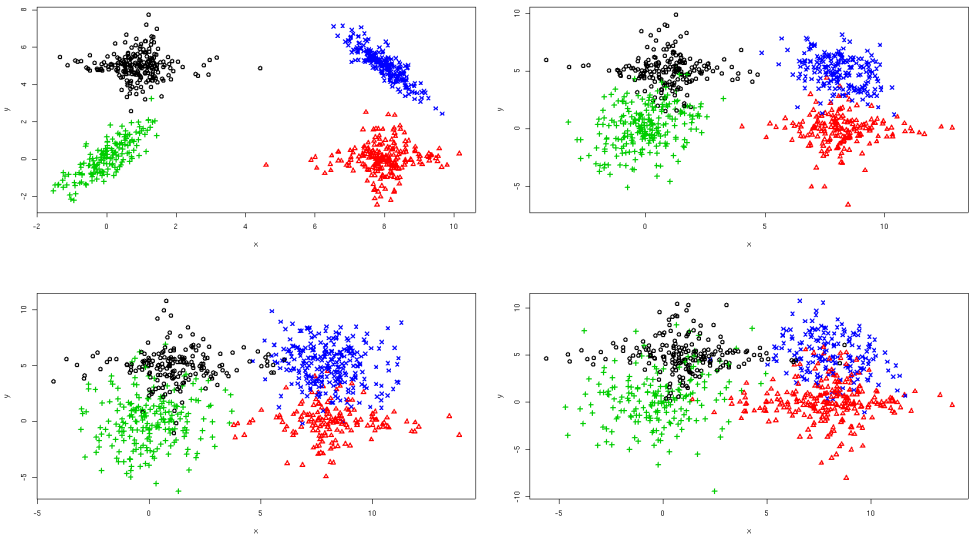

Fig. 3.12: Exemple de données simulées. Haut gauche : $b = 1$, cas simple, qui correspond à celui étudié par Baudry *et al.* (2010) ; Haut droit : $b = 3$; Bas gauche : $b = 5$; Bas droit : $b = 7$. Chaque groupe est représenté par une combinaison symbole/couleur.

Pour chacune des 16 configurations possibles du couple (a, b), 100 jeux de données de taille $n = 800$ ont été simulés. Le package Mclust (Fraley et Raftery, 1999) a été utilisé pour estimer les paramètres du modèle de mélange dans l'approche de Baudry *et al.* (2010). L'inférence du HMM à L états de notre méthode a été réalisée sous l'hypothèse de distributions sphériques pour la loi d'émission, c'est-à-dire la matrice de variance-covariance s'écrit $\begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$. Cette matrice est commune à tous les composants.

Nous présentons les résultats obtenus pour chaque condition de simulation.
Premièrement, nous étudions l'intérêt de la prise en compte de la dépendance en comparant la méthode de Baudry *et al.* (2010) et notre approche. Pour comparer la méthode proposée avec chacun des trois critères et la méthode de Baudry *et al.* (2010), on calcule le MSE (Mean Square Error ou Erreur Quadratique Moyenne) des probabilités *a posteriori*. Le MSE évalue la différence entre la probabilité *a posteriori* estimée $\widehat{\tau}^{(m)}$ d'une méthode m avec la probabilité théorique $\tau^{(th)}$:

$$MSE^{(m)} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{n}\sum_{t=1}^{n}||\widehat{\tau}_t^{(m)} - \tau_t^{(th)}||_2 \ , \tag{3.18}$$

où C est le nombre de jeux de données, $C = 100$.
Plus le MSE est petit, plus la méthode est performante.
Nous évaluons ensuite les performances des critères d'appariement ∇^1, ∇^2 et ∇^3 en termes de classification à l'aide du MSE. D'autre part, étant donné que notre but est de classer les données en un nombre de groupes donné, un indicateur intéressant est le taux de bonne classification. Ce taux permet de mesurer la justesse de la classification estimée. Nous calculons ce taux pour chacune des conditions de simulation décrites précédemment. La classification est obtenue en utilisant la règle du MAP (Maximum A Posteriori) sur les probabilités *a posteriori*.
Enfin, nous nous intéressons à la combinaison du critère de sélection et du critère d'appariement qui donne la meilleure estimation du nombre de groupes.

- **Intérêt de la prise en compte de la dépendance.**
  La Figure 3.13 présente les valeurs du MSE pour chaque condition de simulation. On remarque que la méthode proposée par Baudry *et al.* (2010) fournit de bons résultats uniquement lorsque les composants sont bien séparés ($b = 1$, à gauche de chaque graphique). Ceci peut-être expliqué par le fait que leur critère d'appariement dépend principalement des caractéristiques géométriques des composants. Dans des configurations plus complexes, c'est-à-dire lorsque les composants sont imbriqués, la prise en compte de la dépendance apporte un gain sur l'estimation des probabilités *a posteriori*. En effet, lorsque le paramètre b augmente, la valeur du MSE de l'approche de Baudry *et al.* (2010) augmente, alors que la méthode proposée ici dans le contexte des HMM semble plus stable quel que soit le critère utilisé.

- **Étude des critères d'appariement.**
  L'objectif est de trouver le critère d'appariement permettant la meilleure combinaison des composants d'un HMM. En étudiant le MSE, on remarque que les trois critères d'appariement ne donnent pas les mêmes résultats (cf. Figure 3.13). Le critère ∇^1 fournit les meilleurs résultats pour la plupart des valeurs des paramètres a et b. Ceci n'est pas étonnant car l'algorithme hiérarchique avec le critère ∇^1 et l'algorithme EM maximisent tous les deux la vraisemblance observée. Ainsi, on s'attend à ce que le critère ∇^1 donne les meilleures combinaisons de composants, comparées aux deux autres critères proposés. Lorsque les groupes sont bien séparés, les critères ∇^3 et ∇^1 produisent des estimations similaires, quel que soit le niveau de dépendance. Dans les cas de forte dépendance, les résultats générés par ∇^2 sont analogues à ceux des deux critères précédents. Cependant, lorsque le niveau de dépendance est faible, ∇^2 donne des estimations assez éloignées des vraies, même si les groupes sont facilement distinguables ($b = 1, 3$). Les critères ∇^2 et ∇^3 sont directement liés à la dépendance dans les données, c'est pourquoi, quand le niveau de dépendance

est faible, ils sont moins pertinents pour estimer les probabilités *a posteriori* que ∇^1.

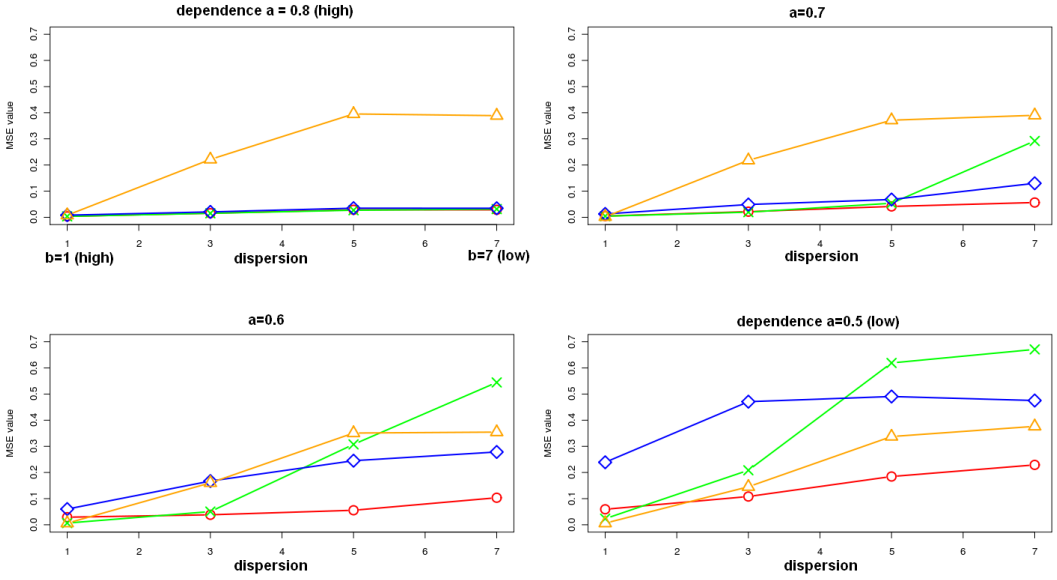

Fig. 3.13: Valeurs du MSE pour chaque condition de simulation. Les graphiques représentent des niveaux de dépendance différents : forte dépendance (haut gauche) à faible (Bas droit). "Δ" : Méthode de Baudry *et al.* (2010), "$\circ$" : ∇^1, "$\Diamond$" : ∇^2, "$\times$" : ∇^3.

La Figure 3.14 décrit la variation de l'écart-type du MSE pour chaque condition de simulation. Le critère ∇^1 fournit une fois de plus les meilleurs résultats parmi les quatre méthodes, particulièrement dans le cas d'un niveau de dépendance élevé.

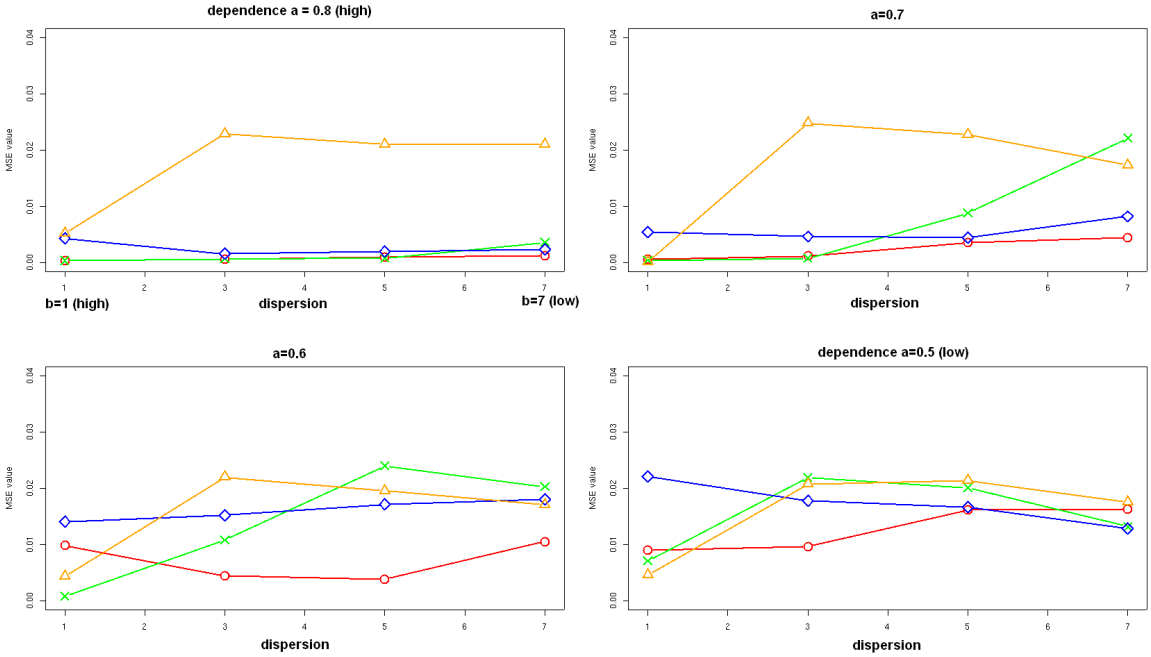

Fig. 3.14: Écart-type du MSE pour chaque condition de simulation. Les graphiques représentent des niveaux de dépendance différents : forte dépendance (haut gauche) à faible (Bas droit). "Δ" : Méthode de Baudry *et al.* (2010), "$\circ$" : ∇^1, "$\Diamond$" : ∇^2, "$\times$" : ∇^3.

Le tableau 3.1 présente les différents taux de classification et leur écart-type.

		moyenne (sd)		
Critères d'appariement		∇^1	∇^3	∇^2
a=0.8	b=1	**0.998 (0.0002)**	**0.998 (0.0002)**	0.994 (0.001)
	b=3	**0.976 (0.0006)**	**0.976 (0.0006)**	0.972 (0.001)
	b=5	**0.951 (0.0009)**	**0.951 (0.0009)**	0.947 (0.001)
	b=7	**0.950 (0.0011)**	0.948 (0.0021)	0.945 (0.018)
a=0.7	b=1	**0.996 (0.0003)**	**0.996 (0.0002)**	0.991 (0.0036)
	b=3	**0.966 (0.0008)**	**0.966 (0.0007)**	0.947 (0.0031)
	b=5	**0.928 (0.0027)**	0.924 (0.0046)	0.909 (0.0033)
	b=7	**0.889 (0.0035)**	0.781 (0.0112)	0.859 (0.0051)
a=0.6	b=1	0.979 (0.0072)	**0.995 (0.0004)**	0.959 (0.0095)
	b=3	**0.948 (0.0035)**	0.945 (0.0055)	0.864 (0.0093)
	b=5	**0.905 (0.0026)**	0.783 (0.0122)	0.781 (0.0107)
	b=7	**0.846 (0.0059)**	0.636 (0.0103)	0.731 (0.0109)
a=0.5	b=1	0.963 (0.0055)	**0.986 (0.0036)**	0.855 (0.0130)
	b=3	**0.899 (0.0057)**	0.861 (0.0109)	0.664 (0.0104)
	b=5	**0.817 (0.0091)**	0.619 (0.0103)	0.608 (0.0104)
	b=7	**0.757 (0.0094)**	0.564 (0.0065)	0.574 (0.0085)

TAB. 3.1: Taux de bonne classification et leur écart-type (sd). Haut : Forte dépendance. Bas : faible dépendance. Les valeurs en gras correspondent aux meilleurs taux de classification calculés pour chacune des trois méthodes.

Dans la plupart des conditions de simulation, le critère ∇^1 fournit les meilleurs taux de classification. Apparier les composants avec ∇^3 permet d'obtenir des résultats proches (voire meilleurs) que ceux obtenus par ∇^1 quand b est égal à 1 ou 3. Cependant, dans des cas plus complexes, ∇^1 est nettement meilleur. Les résultats générés par le critère ∇^2 sont les moins satisfaisants (cf. Figure 3.13).

- **Étude de la combinaison des critères de sélection et d'appariement.**
  Nous nous intéressons à l'estimation du nombre de groupes en comparant les neuf combinaisons possibles des trois critères. Ce nombre est égal à 4 pour les jeux de données simulées. Le tableau 3.2 décrit le pourcentage d'estimations correctes du nombre de groupes, calculé pour chaque niveau de dépendance et pour chaque valeur de b.
  La première remarque est que si le critère ICL_W est utilisé en tant que critère de sélection, l'estimation du nombre de groupes est mauvaise quels que soient les critères d'appariement. En effet, le critère ICL_W a tendance à surestimer le nombre de groupes car il est fondé sur la variable latente W qui est liée aux L composants initiaux. Il semble plus judicieux d'estimer le nombre de groupes avec les critères BIC et ICL_Z qui ne dépendent pas de W. Le tableau 3.2 indique que la meilleure estimation du nombre de groupes est obtenue avec ICL_Z quel que soit le critère d'appariement.

Critères d'appariement		∇^1			∇^3			∇^2		
Critères de sélection		BIC	ICL_Z	ICL_W	BIC	ICL_Z	ICL_W	BIC	ICL_Z	ICL_W
a=0.8	b=1	0.86	**0.91**	0.33	0.92	**0.96**	0.76	0.89	**0.94**	0.31
	b=3	0.93	**0.98**	0.05	0.96	**0.98**	0.35	0.93	**0.98**	0.08
	b=5	0.93	**0.98**	0.05	0.97	**0.98**	0.21	0.89	**0.98**	0.07
	b=7	0.95	**0.99**	0.12	**0.97**	**0.97**	0.26	0.91	**0.98**	0.1
a=0.7	b=1	0.90	**0.97**	0.42	0.91	**0.92**	0.78	0.86	**0.97**	0.44
	b=3	0.93	**0.98**	0.13	0.93	**0.96**	0.28	0.92	**0.99**	0.10
	b=5	0.91	**0.98**	0.04	0.95	**0.98**	0.28	0.91	**0.97**	0.02
	b=7	0.89	**0.99**	0.07	0.86	**0.98**	0.16	0.93	**0.98**	0.09
a=0.6	b=1	0.82	**0.93**	0.48	0.90	**0.95**	0.80	0.82	**0.93**	0.33
	b=3	0.85	**0.94**	0.13	0.97	**0.98**	0.33	0.81	**0.90**	0.12
	b=5	0.90	**0.98**	0.07	0.86	**0.99**	0.24	0.85	**0.88**	0.09
	b=7	0.94	**0.95**	0.09	0.91	**0.97**	0.26	0.89	**0.95**	0.06
a=0.5	b=1	0.79	**0.90**	0.43	0.91	**0.94**	0.75	0.76	**0.90**	0.30
	b=3	0.86	**0.91**	0.15	0.91	**0.96**	0.43	0.86	**0.77**	0.12
	b=5	0.91	**0.94**	0.14	0.92	**0.95**	0.36	0.92	**0.80**	0.10
	b=7	**0.93**	0.92	0.16	0.96	**0.98**	0.42	0.89	**0.78**	0.10

TAB. 3.2: Pourcentage d'estimations correctes du nombre de groupes pour chaque condition de simulation. Haut : forte dépendance. Bas : faible dépendance. Les valeurs en gras correspondent aux meilleurs taux de classification calculés pour chacune des trois méthodes.

L'étude de simulation a mis en évidence un réel gain de la prise en compte de la dépendance, particulièrement lorsque les composants sont imbriqués. De plus, on a montré que le critère d'appariement ∇^1 donne les meilleurs résultats parmi les trois critères proposés. En effet, l'estimation des probabilités *a posteriori* est proche des vraies valeurs et ces estimations sont stables en termes de MSE. Cette remarque a été confirmée par l'étude du taux de bonne classification.
Concernant l'estimation du nombre de groupes, l'étude de simulation suggère de sélectionner le nombre de groupes avec ICL_Z. Pour conclure, nous proposons d'utiliser ∇^1 comme critère d'appariement et d'estimer le nombre de groupes avec ICL_Z.

3 Annexes

3.1 Estimateurs du modèle gaussien bidimensionnel sous contraintes

Nous rappelons que le modèle est défini ainsi (cf. Section 2.2) :

$$\begin{cases} \Sigma_k &= D_k \Lambda_k D_k^T, \qquad \text{pour } k = 1, ..., 4; \\ D_1 &= D_2 = D \ ; \\ \Lambda_k &= \begin{pmatrix} u_{1k} & 0 \\ 0 & u_2 \end{pmatrix}, \qquad \text{avec } u_{1k} > u_2, \text{ pour } k = 1, ..., 4. \end{cases}$$

où les groupes 1 et 2 correspondent aux groupes de même orientation et la matrice Λ_k ($\Lambda_k = \lambda_k A_k$) est une matrice diagonale qui contient les valeurs propres de Σ_k.

L'estimation des paramètres est spécifique pour satisfaire les contraintes imposées sur les matrices de variance.
Posons $W_k = \sum_{t=1}^n \tau_{tk}(x_t - \bar{x}_k)(x_t - \bar{x}_k)^T$ une matrice de la forme $\begin{pmatrix} w_{1k} & w_{2k} \\ w_{2k} & w_{4k} \end{pmatrix}$, avec $\bar{x}_k = \frac{\sum_{t=1}^n \hat{\tau}_{tk} x_t}{\sum_{t=1}^n \hat{\tau}_{tk}}$. On note $n_k = \sum_{t=1}^{n} \hat{\tau}_{tk}$, $n = \sum_{k=1}^{K} n_k$ et $tr(A)$ la trace de la matrice A.

L'expression générale de l'espérance de la log-vraisemblance des données complètes conditionnellement à X du modèle, notée Q, est :

$$\begin{aligned} Q \ = \ & \sum_{k=1}^{K} n_k \log(\pi_k) - n \log(2\pi) + \frac{1}{2} \sum_{k=1}^{K} n_k \log \left[\det\{(D_k \Lambda_k D_k^T)^{-1}\} \right] \\ & - \frac{1}{2} \sum_{k=1}^{K} tr\{(D_k \Lambda_k D_k^T)^{-1} W_k\} - \frac{1}{2} \sum_{k=1}^{K} n_k (\bar{x_k} - \mu_k)^T (D_k \Lambda_k D_k^T)^{-1} (\bar{x_k} - \mu_k). \end{aligned}$$

Maximiser Q revient à minimiser $F = \sum_{k=1}^{K} tr((D_k \Lambda_k D_k^T)^{-1} W_k) + \sum_{k=1}^{K} n_k \log(\det(D_k \Lambda_k D_k^T))$.
Après simplification :

$$F = \sum_{k=1}^{K} tr(D_k^T W_k D_k \Lambda_k^{-1}) + \sum_{k=1}^{K} n_k \log(\det(\Lambda_k))$$

En écrivant la contrainte sur les matrices d'orientation de manière plus générale avec un ℓ fixé, on a :

$$\begin{aligned} \Sigma_k &= D_k^T \Lambda_k D_k \ si \ k \geq \ell \\ \Sigma_k &= D^T \Lambda_k D \ si \ k < \ell \end{aligned}$$

Au final, on doit minimiser :

$$F = \sum_{k=1}^{\ell} tr(D^T W_k D \Lambda_k^{-1}) + \sum_{k=\ell+1}^{K} tr(D_k^T W_k D_k \Lambda_k^{-1}) + \sum_{k=1}^{K} n_k \log \left\{ \det(\Lambda_k) \right\},$$

avec $\ell = 2$ et $K = 4$ dans notre cas.

Dans l'étape M, trouver l'estimateur de Σ_k revient à trouver les estimateurs de D, D_k et Λ_k qui sont donc calculés en minimisant F. On remarque que seul Λ_k est présent dans les trois termes de F.

Estimateur de D

Minimiser F en D revient à minimiser $f(D) = \sum_{k=1}^{2} tr(D\Lambda_k^{-1}D^T W_k)$. On peut réécrire $f(D)$ sous la forme suivante :

$$f(D) = \sum_{k=1}^{2} \left(\frac{d_1' W_k d_1}{u_{1k}} + \frac{d_2' W_k d_2}{u_2} \right),$$

où d_1' est le premier vecteur de la matrice D et d_2' le second.
Puisque D est une matrice orthogonale et normée, elle est de la forme $\begin{pmatrix} \sqrt{d} & -\sqrt{1-d} \\ \sqrt{1-d} & \sqrt{d} \end{pmatrix}$. En développant $f(D)$ et en dérivant par rapport à d, on obtient un pôlynome de degré 4 en d qui se résout facilement.
L'estimateur de $\hat{d}$ est défini par :

$$\widehat{d^2} - \frac{1}{2} \;=\; \pm \frac{N_{1,4}}{2\left[\{N_{1,4}\}^2 + 4\{N_2\}^2\right]^{1/2}}\ , \quad \text{avec } \hat{d} > 0,$$

où $N_{1,4} = \sum_{k=1}^{2} (\hat{w}_{1k} - \hat{w}_{4k})(\hat{u}_2 - \hat{u}_{1k})/\hat{u}_{1k}\hat{u}_2$ et $N_2 = \sum_{k=1}^{2} (\hat{w}_{2k})(\hat{u}_2 - \hat{u}_{1k})/\hat{u}_{1k}\hat{u}_2$.

Estimateur de D_k

L'estimateur de D_k pour les composants $k = 3, 4$ d'orientation différente est le même que celui proposé par Celeux et Govaert (1995) pour des composants d'orientations différentes. La matrice W_k est carrée, symétrique et réelle. Elle est donc diagonalisable dans une base orthonormée. On pose $W_k = L_k \Omega_k L_k^T$, où Ω_k est la matrice diagonale avec les valeurs propres de W_k dans l'ordre décroissant. On obtient $F = \sum_{k=l}^{K} tr(D_k^T L_k \Omega_k L_k^T D_k \Lambda_k^{-1})$, avec Λ_k^{-1} matrice diagonale avec les valeurs propres dans l'ordre croissant. On déduit que $D_k^T L_k = Id$, c'est-à-dire $D_k = L_k$, où L_k est la matrice des vecteurs propres de W_k.

Estimateur de Λ_k

Soit B_k la matrice définie par $B_k = D_k^T W_k D_k$ de la forme $\begin{pmatrix} b_{1k} & b_{3k} \\ b_{4k} & b_{2k} \end{pmatrix}$. En développant la trace et le déterminant, on peut réécrire F sous la forme :

$$F = \sum_{k=1}^{4} (b_{1k} u_{1k}^{-1} + b_{2k} u_2^{-1}) + \sum_{k=1}^{4} n_k \left\{ \log(u_{1k}) + \log(u_2) \right\},$$

et minimiser F en Λ_k revient à minimiser F en u_{1k} et u_2. On doit résoudre le système suivant :

$$\begin{cases} \frac{\partial F}{\partial u_{1k}} & = \frac{-b_{1k}}{u_{1k}^2} + \frac{n_k}{u_{1k}} = 0 \\ \frac{\partial F}{\partial u_2} & = \frac{-\sum_{k=1}^{K} b_{2k}}{u_2^2} + \frac{\sum_{k=1}^{K} n_k}{u_2} = 0 \end{cases}$$

L'estimateur du maximum de vraisemblance de Λ_k est de la forme $\begin{pmatrix} \hat{u}_{1k} & 0 \\ 0 & \hat{u}_2 \end{pmatrix}$, où

$$\begin{cases} \hat{u}_{1k} & = \hat{b}_{1k}/n_k \\ \hat{u}_2 & = \sum_{k=1}^{4} \hat{b}_{2k}/n\ . \end{cases}$$

3.2 Estimateurs du modèle gaussien unidimensionnel sous contraintes de colinéarité

Cette annexe présente les calculs des estimateurs de la procédure d'initialisation de l'algorithme EM avec contraintes de colinéarité (cf. Proposition 3).
Les quatre groupes sont déterminés à l'aide du modèle de mélange gaussien bidimensionnel présenté Section 2.2 : groupe bruit, groupe identique, et deux groupes différentiellement hybridés. Le groupe bruit est considéré circulaire et représenté par une gaussienne sphérique. Les trois autres groupes sont chacun représentés par un mélange de gaussiennes. Les estimateurs du modèle sont calculés en maximisant l'équation (3.10) qui correspond à :

$$\sum_t \sum_k \sum_\ell \mathbb{E}\left[Z_{t,k} W_{tk\ell} \mid X\right] \log(\psi_k) = \sum_t \sum_k \sum_\ell \mathbb{E}\left[Z_{t,k} W_{tk\ell} \mid X\right] \log\left\{\eta_{k\ell} f(X_t; \theta_{k\ell})\right\} \ .$$

En utilisant les coordonnées (U_{tk}, V_{tk}) projetés de (X_{1t}, X_{2t}) dans le repère $(\Delta_k, \Delta_k^\perp)$ et en notant les quatre groupes de 1 à 4, où 1 est le groupe bruit, on obtient :

$$\begin{aligned}\sum_{tk\ell} \left[Z_{t,k} W_{tk\ell} \mid X\right] \log\left\{\eta_{k\ell} f(X_t; \theta_{k\ell})\right\} &= \sum_t \left[Z_{t,1} \mid X\right] \log\left\{f(X_{1t}; \mu_1^1, \sigma^2) f(X_{2t}; \mu_1^2, \sigma^2)\right\} \\ &\quad + \sum_{k=2}^{4} \sum_{t\ell} \left[Z_{t,k} W_{tk\ell} \mid X\right] \log\left\{\eta_{k\ell} f(U_{tk}; \mu_{k\ell}, \sigma_{k\ell}^2) f(V_{tk}; 0, \sigma^2)\right\} .\end{aligned}$$

On remarque que $\mathbb{E}\left[Z_{t,1} \mid X\right] = \tau_{t1}$ et $\mathbb{E}\left[Z_{t,k} W_{tk\ell} \mid X\right] = \tau_{tk}\, \mathbb{E}\left[W_{tk\ell} \mid Z_{t,k}, X\right]$ et on note

$$\delta_{tk\ell} = \mathbb{E}\left[W_{tk\ell} \mid Z_{t,k}, X\right] .$$

On a $\sum_l \delta_{tk\ell} = 1$.

Pour simplifier, notons :

$$Q_1 = \sum_{t=1}^{n} \sum_{k=2}^{4} \sum_{\ell=1}^{L_k} \tau_{tk} \delta_{tk\ell} \log \eta_{k\ell} \ ,$$

$$Q_2 = \sum_{t=1}^{n} \sum_{k=2}^{4} \sum_{\ell=1}^{L_k} \tau_{tk} \delta_{tk\ell} \log f(U_{tk}; \mu_{k\ell}, \sigma_{k\ell}^2) \ ,$$

$$Q_3 = \sum_{t=1}^{n} \sum_{k=2}^{4} \sum_{\ell=1}^{L_k} \tau_{tk} \delta_{tk\ell} \log f(V_{tk}; 0, \sigma^2) + \sum_t \tau_{t1} \log f(X_{1t}; \mu_1^1, \sigma^2) + \sum_t \tau_{t1} \log f(X_{2t}; \mu_1^2, \sigma^2) \ .$$

Estimateur de $\mu_{k\ell}$ (pour $k = 2, 3, 4$)

$$\begin{aligned} Q_2 &= \sum_{t=1}^{n} \sum_{k=2}^{4} \sum_{\ell=1}^{L_k} \tau_{tk} \delta_{tk\ell} \log f(U_{tk}; \mu_{k\ell}, \sigma_{k\ell}^2) \\ &= \sum_{t=1}^{n} \sum_{k=2}^{4} \sum_{\ell=1}^{L_k} \tau_{tk} \delta_{tk\ell} \log \left[\frac{1}{\sqrt{2\pi}\sigma_{k\ell}} \exp\left(\frac{-1}{2\sigma_{k\ell}^2} (U_{tk} - \mu_{k\ell})^2\right)\right] \\ &= \sum_{t=1}^{n} \sum_{k=2}^{4} \sum_{\ell=1}^{L_k} \tau_{tk} \delta_{tk\ell} \left[-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_{k\ell}^2) - \frac{1}{2\sigma_{k\ell}^2} (U_{tk} - \mu_{k\ell})^2\right] \end{aligned}$$

En annulant la dérivée de Q_2 par rapport à $\mu_{k\ell}$, on a :

$$\frac{\partial Q_2}{\partial \mu_{k\ell}} = \frac{1}{\hat{\sigma}^2_{k\ell}} \sum_t \hat{\tau}_{tk} \hat{\delta}_{tk\ell} (U_{tk} - \hat{\mu}_{k\ell}) = 0$$

Donc

$$\hat{\mu}_{k\ell} = \frac{\sum_t \hat{\tau}_{tk} \hat{\delta}_{tk\ell} U_{tk}}{\sum_t \hat{\tau}_{tk} \hat{\delta}_{tk\ell}} \ .$$

Estimateur de $\hat{\sigma}^2_{k\ell}$ (pour $k = 2, 3, 4$) En annulant la dérivée de Q_2 par rapport à $\sigma^2_{k\ell}$, on obtient :

$$\frac{\partial Q_2}{\partial \sigma^2_{k\ell}} = \frac{-1}{2\hat{\sigma}^2_{k\ell}} \sum_t \hat{\tau}_{tk} \hat{\delta}_{tk\ell} + \frac{1}{2\hat{\sigma}^4_{k\ell}} \sum_t \hat{\tau}_{tk} \hat{\delta}_{tk\ell} (U_{tk} - \hat{\mu}_{k\ell})^2 = 0.$$

D'où

$$\hat{\sigma}^2_{k\ell} = \frac{\sum_t \hat{\tau}_{tk} \hat{\delta}_{tk\ell} (U_{tk} - \hat{\mu}_{k\ell})^2}{\sum_t \hat{\tau}_{tk} \hat{\delta}_{tk\ell}} \ .$$

On remarque que l'on peut facilement calculer les estimateurs des variances avec des contraintes différentes.

- Si on suppose qu'on a la même variance pour tous les composants d'un même groupe, on obtient :

$$\hat{\sigma}^2_k = \frac{\displaystyle\sum_{t=1}^{n} \sum_{\ell=1}^{L_k} \hat{\tau}_{tk} \hat{\delta}_{tk\ell} (U_{tk} - \hat{\mu}_{k\ell})^2}{\displaystyle\sum_{t=1}^{n} \sum_{\ell=1}^{L_k} \hat{\tau}_{tk} \hat{\delta}_{tk\ell}} \ .$$

- Si on suppose qu'on a la même variance pour tous les composants de tous les groupes, on obtient :

$$\hat{\sigma}^2_U = \frac{1}{n} \sum_{t=1}^{n} \sum_{k=2}^{4} \sum_{\ell=1}^{L_k} \hat{\tau}_{tk} \hat{\delta}_{tk\ell} (U_{tk} - \hat{\mu}_{k\ell})^2 \ .$$

Estimateur de $\hat{\sigma}^2$

$$\begin{aligned} Q_3 &= \sum_t \tau_{t1} \left[-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{(X_{1t} - \mu^1_1)^2}{2\sigma^2} \right] \\ &+ \sum_t \tau_{t1} \left[-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{(X_{2t} - \mu^2_1)^2}{2\sigma^2} \right] \\ &+ \sum_{t=1}^{n} \sum_{k=2}^{4} \sum_{\ell=1}^{L_k} \tau_{tk} \delta_{tk\ell} \left[-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{V^2_{tk}}{2\sigma^2} \right] \end{aligned}$$

$$\begin{aligned} \frac{\partial Q_3}{\partial \sigma^2} &= -\frac{1}{2\sigma^2} \sum_t \tau_{t1} + \frac{1}{2\sigma^4} \sum_t \tau_{t1} (X_{1t} - \mu^1_1)^2 - \frac{1}{2\sigma^2} \sigma^2 \sum_t \tau_{t1} + \frac{1}{2\sigma^4} \sum_t \tau_{t1} (X_{2t} - \mu^2_1)^2 \\ &- \frac{1}{2\sigma^2} \sum_{t=1}^{n} \sum_{k=2}^{4} \sum_{\ell=1}^{L_k} \tau_{tk} \delta_{tk\ell} + \frac{1}{2\sigma^4} \sum_{t=1}^{n} \sum_{k=2}^{4} \sum_{\ell=1}^{L_k} \tau_{tk} \delta_{tk\ell} V^2_{tk} \end{aligned}$$

En annulant la dérivée, on obtient :

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{n} \hat{\tau}_{t1}(X_{1t} - \hat{\mu}_1^1)^2 + \sum_{t=1}^{n} \hat{\tau}_{t1}(X_{2t} - \hat{\mu}_1^2)^2 + \sum_{t=1}^{n}\sum_{k=2}^{4}\sum_{\ell=1}^{L_k} \hat{\tau}_{tk}\hat{\delta}_{tk\ell}V_{tk}^2}{\sum_{t=1}^{n} \hat{\tau}_{t1} + \sum_{t=1}^{n} \hat{\tau}_{t1} + \sum_{t=1}^{n}\sum_{k=2}^{4}\sum_{\ell=1}^{L_k} \hat{\tau}_{tk}\hat{\delta}_{tk\ell}} .$$

Estimation de Δ_k (pour $k = 2, 3, 4$) Δ_k est la droite représentant le grand axe du groupe k. C'est la droite de vecteur directeur $\vec{d_k} = (d_k, \sqrt{1 - d_k^2})$ passant par le barycentre du groupe k noté $\bar{X}_k$ ($\bar{X}_k = \sum_t \hat{\tau}_{tk} X_t / \sum_t \hat{\tau}_{tk}$) et par le barycentre du groupe bruit noté $\bar{X}_1$.
On estime Δ_k en minimisant $Q_2 + Q_3$, où U_k et V_k sont écrits en fonction de d_k :

$$\begin{pmatrix} U_k \\ V_k \end{pmatrix} = A_k^{-1} \begin{pmatrix} \bar{X}_{1k} \\ \bar{X}_{2k} \end{pmatrix} - A_k^{-1} \begin{pmatrix} \bar{X}_{11} \\ \bar{X}_{21} \end{pmatrix} ,$$

avec $A_k^{-1} = \begin{pmatrix} \sqrt{1 - d_k^2} & d_k \\ -d_k & \sqrt{1 - d_k^2} \end{pmatrix}$.

Au final,

$$\widehat{d_k} = \text{Argmin}(\widehat{Q}_2 + \widehat{Q}_3).$$

En pratique, $\widehat{d_k}$ est obtenu à l'aide de la fonction *optimize* du logiciel R.

Bibliographie

Banfield, J.D. and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803-821.

Baudry, J.P., Raftery, A.E., Celeux, G., Lo, K. and Gottardo, R. (2008). Combining Mixture Components for Clustering. *JCGS*.

Bérard, C., Martin-Magniette, M.-L., To, A., Roudier, F., Colot, V. and Robin, S. (2009). Mélanges gaussiens bidimensionnels pour la comparaison de deux échantillons de chromatine immunoprécipitée. *La revue de MODULAD*, **40** 53-68.

Celeux, G. and Govaert, G. (1995). Gaussian Parsimonious Clustering Models. *Pattern Recognition* **28** 781-793.

Chatzis, S.P. (2010). Hidden Markov Models with Nonelliptically Contoured State Densities. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(12) 2297-2304.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorihtm (with discussion). *J. R. Statis. Soc. B* **39** 1-38.

Fraley, C. and Raftery, A.E. (1999). Mclust : Software for Model-based Cluster Analysis. *Journal of Classfication* **16** 297-306.

Hennig, C. (2010). Methods for merging Gaussian mixture components. *Adv Data Anal Classif* 3-34.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing* **85** 1501-1510.

Li, J. (2005). Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics* 547-568.

Martin-Magniette, M.L., Mary-Huard, T., Bérard, C. and Robin, S. (2008). ChIPmix : mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics* **24** :i181-i186.

Schwarz, G. (1978). Estimating the number of components in a finite mixture model. *Annals of Statistics* **6** 461-464.

Sun, W., Buck, M.J., Patel, M. and Davis, I.J. (2009). Improved ChIP-chip analysis by a mixture model approach. *BMC Bioinformatics* **10** :173.

Tantrum, J. and Murua, A. (2003). Assessment and pruning of hierarchical model based clustering. *Pattern Recognition, Clustering, ACM Press.*

Turner, T.R. (2000). Estimating the Propagation Rate of a Viral Infection of Potato Plants via Mixtures of Regressions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **49**(3) 371-384.

Chapitre 4

Classification

Sommaire

1 Règle du MAP 81

1.1 Seuil de classification 81

1.2 Association de groupes 82

2 Contrôle des faux-positifs 83

3 Classification de régions 86

3.1 Probabilités *a posteriori* pour une région 86

3.2 Règle de classification 88

Bibliographie 89

Dans le chapitre précédent, nous nous sommes intéressés à la modélisation de la distribution des intensités des sondes à l'aide de modèles à variables latentes. D'un point de vue de classification non supervisée, le but ultime est de classer chaque sonde dans un groupe afin de définir son statut.
Une fois les paramètres du modèle estimés, la probabilité *a posteriori* pour une sonde d'appartenir au groupe k est calculée :

$$\tau_{tk} = P\left\{Z_t = k | X = x\right\}.$$

La Figure 4.1 présente un exemple de classification à deux groupes où les sondes sont colorées en fonction des probabilités *a posteriori* estimées à partir du modèle de mélange de régressions (cf. Chapitre 3 Section 2.1). Les probabilités *a posteriori* indiquent le degré de confiance du classement des observations et c'est à partir de ces probabilités que s'établit la classification. On remarque qu'il existe une région aux frontières des groupes où la classification est difficile à déterminer.

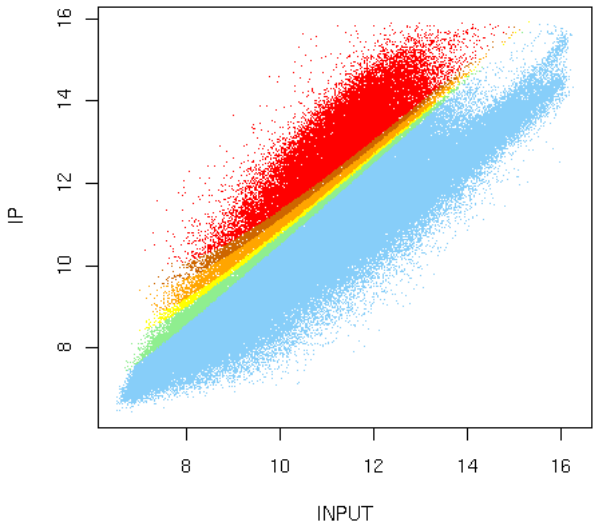

Fig. 4.1: Les sondes sont colorées en fonction de la valeur de la probabilité *a posteriori* d'appartenir au groupe du haut : rouge correspond à une probabilité > 0.9, marron > 0.8, orange > 0.6, jaune > 0.5, vert > 0.3, et bleu correspond à une probabilité < 0.3.

Dans la Section 1, nous présentons la règle du *Maximum A Posteriori* qui est couramment utilisée pour établir la classification, ainsi que deux variantes. Puis, nous proposons un contrôle de faux-positifs qui permet d'éviter un trop grand nombre d'erreurs de classification dans le cas où les observations sont indépendantes et où il n'y a que deux groupes (Section 2). Enfin, nous généralisons la formule des probabilités *a posteriori* pour une observation à un ensemble d'observations, et nous proposons une règle de classification pour classer directement un ensemble d'observations constituant une région d'intérêt (Section 3).

1 Règle du MAP

Pour classer les observations à partir des probabilités *a posteriori*, la méthode la plus utilisée est la règle du *Maximum A Posteriori* (MAP) qui assigne chaque observation à la population pour laquelle la probabilité *a posteriori* est la plus grande.

$$\widehat{Z}_t = \underset{k}{\text{Argmax}}\, P\left\{Z_t = k | X = x\right\} = \underset{k}{\text{Argmax}}\, \tau_{tk} \ .$$

Dans le cas d'une classification à deux groupes, la règle du MAP revient à comparer les probabilités *a posteriori* à une valeur seuil de 0.5.

La règle du MAP est très simple à mettre en œuvre, mais elle présente un inconvénient majeur : le classement est trop arbitraire. En effet, une sonde située à la frontière de deux groupes a des probabilités *a posteriori* très proches pour ces deux groupes ($\tau_{tk} \simeq \tau_{tk'}$), et le classement par la règle du MAP n'est alors pas très fiable.

1.1 Seuil de classification

Dans certaines applications, ne pas définir de classement pour une observation est préférable à une mauvaise prédiction. L'intérêt des probabilités *a posteriori* est de permettre la caractérisation des classements incertains. Afin d'éviter les erreurs de classification, un seuil s peut être fixé arbitrairement sur les probabilités *a posteriori* pour ne classer que les observations dont l'une des probabilités *a posteriori* est forte. Ce seuil de classification arbitraire délimite une marge de non classement autour de chacun des groupes (cf. Figure 4.2).

$$\widehat{Z}_t = \begin{cases} k \text{ si } \underset{\ell}{\text{Argmax}}\ \tau_{t\ell} = k \ \text{ et } \tau_{tk} > s \\ 0 \ \text{ sinon} \end{cases}$$

où $\widehat{Z}_t = 0$ correspond à une sonde t non classée.

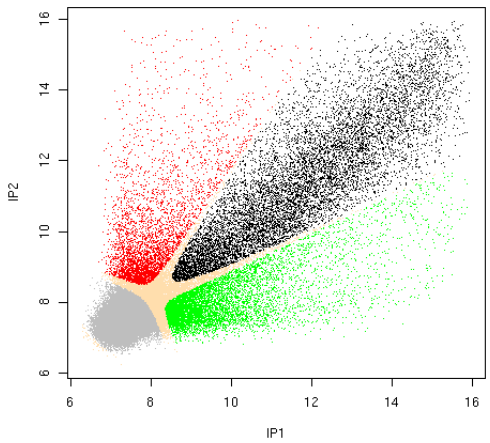

Fig. 4.2: Classement des sondes en quatre groupes avec un seuil de classification égal à 0.7. La zone de non classement est en beige.

1.2 Association de groupes

Une autre manière de réduire les classements incertains est d'associer différents groupes qui, en fonction de la question posée, peuvent être regroupés pour établir une classification spécifique. Les différentes classifications sont faciles à réaliser puisqu'il suffit de sommer les probabilités *a posteriori* des groupes prédéfinis.
Ainsi, pour les expériences de ChIP-chip IP/IP ou de transcriptome pour lesquelles nous proposons un modèle à quatre composants (Chapitre 3 Section 2.2), la classification peut être modifiée en fonction de l'intérêt biologique. Si l'objectif principal est de distinguer les différences d'hybridation entre les deux conditions, une classification en trois groupes est satisfaisante. Il suffit pour cela de sommer les probabilités *a posteriori* du groupe identiquement hybridé et celles du groupe bruit (pas d'hybridation). Une autre possibilité est de classer en deux groupes seulement, l'un correspondant à une hybridation identique dans les deux échantillons, et l'autre correspondant à une hybridation différente, sans distinguer les groupes différentiellement hybridés (cf. Figure 4.3 réalisée avec le modèle HMM $\mathcal{M}_2$ défini Chapitre 3 Section 1). Si l'objectif biologique est plutôt de trouver des régions hybridées, une classification en deux groupes peut être choisie en séparant le groupe bruit des trois autres.

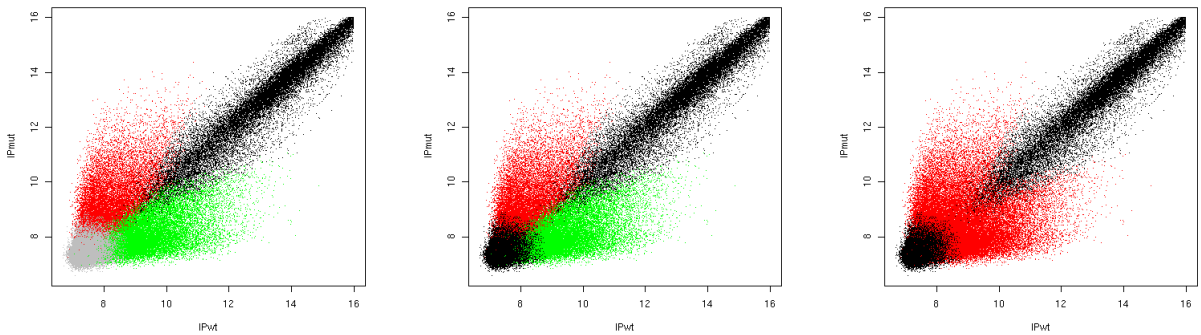

Fig. 4.3: Classement des sondes en 4 groupes (gauche), 3 groupes (centre), 2 groupes (droite).

Sur l'exemple donné Figure 4.4 (réalisée avec le modèle de mélange gaussien bidimensionnel $\mathcal{M}_1$ défini Chapitre 3 Section 1, pour une meilleure visualisation des frontières), on remarque que la zone de non classement est réduite dans le cas d'une classification à trois ou deux groupes (respectivement 11.9% et 9.3% de sondes non classées) par rapport à une classification à quatre groupes (12.5% de sondes non classées).

Ces différentes variantes de la règle du MAP restent pragmatiques et peu formelles, et les difficultés de classification aux frontières des groupes ne sont pas résolues. Dans la section suivante, nous proposons une procédure permettant de contrôler les erreurs de classification dans le cas d'une classification à deux groupes et pour des observations indépendantes.

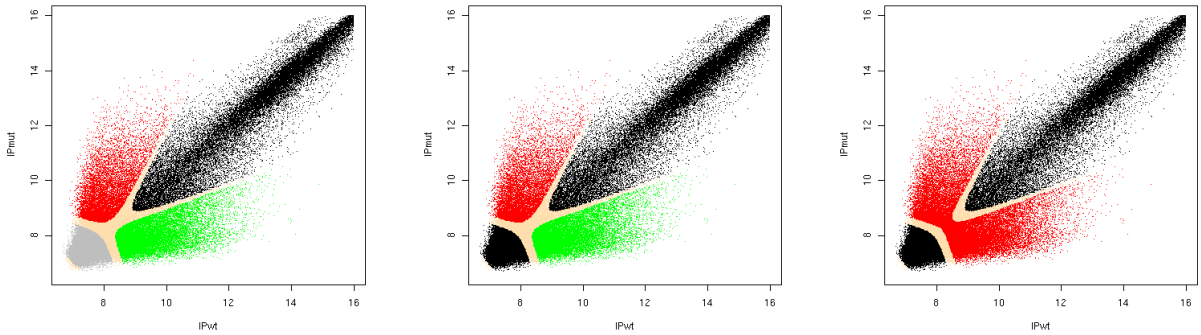

Fig. 4.4: Classement des sondes avec un seuil de classification égal à 0.7 (zone en beige). Gauche : classification en 4 groupes avec 12.5% de sondes non classées. Centre : classification en 3 groupes avec 11.9% de sondes non classées. Droite : classification en 2 groupes avec 9.3% de sondes non classées.

2 Contrôle des faux-positifs

On se place dans le cadre d'une classification à deux groupes ($K = 2$), sous l'hypothèse d'observations indépendantes. Nous reprenons le modèle de mélange de régressions défini Chapitre 3 Section 2.1 pour l'analyse de données de ChIP-chip.

Dans le cas d'une classification à deux groupes, les deux groupes jouent généralement un rôle différent : un seul des deux groupes est un groupe d'intérêt. Afin d'éviter de trop nombreuses erreurs de classification dans le groupe d'intérêt, nous proposons un contrôle de faux-positifs, où un faux-positif est une observation classée à tort dans le groupe d'intérêt. Dans l'exemple d'une expérience de ChIP-chip où l'objectif est de classer les sondes en deux groupes (un groupe enrichi, noté 1, et un groupe normal, noté 0), l'identification des sondes enrichies est particulièrement importante d'un point de vue biologique. Le contrôle de faux-positifs s'intéresse donc uniquement à la probabilité d'affecter une sonde à tort dans le groupe enrichi.

L'identification des sondes enrichies à partir des probabilités *a posteriori* revient à trouver un seuil s tel que si $\tau_{t1} > s$, la sonde t est déclarée enrichie. La règle du MAP fixe le seuil $s = 1/2$, ce qui signifie implicitement que les erreurs de classification dans la population enrichie et dans la population normale ont le même coût. Le contrôle de faux-positifs permet de contrôler les erreurs de classification en déterminant un seuil s de manière non arbitraire.
Nous définissons la règle de classification suivante :

$$\widehat{Z}_t = \mathbb{1}_{\{\tau_{t1} > s\}} \ .$$

Dans la théorie des tests d'hypothèses, le contrôle de faux-positifs est réalisé en contrôlant la probabilité de rejeter à tort l'hypothèse nulle. De manière analogue au contrôle du risque de première espèce, le taux de faux-positifs dans le cadre des modèles de mélange est contrôlé par la probabilité d'affecter une sonde à tort dans la population enrichie. Le risque à contrôler est :

$$P\{\widehat{Z}_t = 1 \mid Z_t = 0,\ X_t = x_t\} \ .$$

En pratique, on fixe un niveau α tel que :

$$P\{\widehat{Z}_t = 1 \mid Z_t = 0,\ X_t = x_t\} \leq \alpha \ , \tag{4.1}$$

et on cherche le seuil s correspondant. Le seuil s dépend à la fois de α et du log-INPUT x_t et doit donc être calculé pour chaque sonde.

Proposition 4. *Notons $X_t = (x_t, Y_t)$ les intensités log-INPUT et log-IP de la sonde t, respectivement, et Z_t son statut. La distribution marginale de Y_t, pour un niveau d'INPUT x_t donné, est un mélange de deux régressions :*

$$(1-p)f(Y_t; x_t, \theta_0) + pf(Y_t; x_t, \theta_1),$$

où p représente la proportion de sondes enrichies, et $f(\cdot; x, \theta_k)$ est la fonction de densité de probabilité d'une distribution gaussienne de moyenne $\mu_k = a_k + b_k x$ et de variance σ^2 pour $k = 0, 1$.
Le seuil s qui garantit $P\{\widehat{Z}_t = 1 \mid Z_t = 0,\ X_t = x_t\} \leq \alpha$ est :

$$s = \frac{\exp^{\lambda}}{1 + \exp^{\lambda}},$$

où

$$\lambda = \left(\frac{\mu_1 - \mu_0}{\sigma}\right)\left(u_{1-\alpha} - \frac{\mu_1 - \mu_0}{2\sigma}\right) - \log\left(\frac{1-p}{p}\right),$$

où $u_{1-\alpha}$ est le quantile de $\mathcal{N}(0,1)$ au niveau α.

Preuve 1. *Nous rappelons que la sonde t, d'IP Y_t et d'INPUT x_t appartient au groupe enrichi si et seulement si $\tau_{t1} > s$, avec :*

$$\tau_{t1} = P_{x_t}\{Z_t = 1 \ \mid \ y_t\} = \frac{pf(y_t; x_t, \theta_1)}{(1-p)f(y_t; x_t, \theta_0) + pf(y_t; x_t, \theta_1)},$$

où p représente la proportion de sondes enrichies, θ_1 est le vecteur de paramètres (a_1, b_1, σ) et $f(y_t; x_t, \theta_1)$ est la densité de la loi $\mathcal{N}(\mu_1 = a_1 + b_1 x, \sigma^2)$, et où θ_0 est le vecteur de paramètres (a_0, b_0, σ) et $f(y_t; x_t, \theta_0)$ est la densité de la loi $\mathcal{N}(\mu_0 = a_0 + b_0 x, \sigma^2)$.

L'équation (4.1) peut être réécrite de la manière suivante :

$$P\{pf(y_t; x_t, \theta_1)(1-s) - s(1-p)f(y_t; x_t, \theta_0) > 0 \ \mid \ Z_t = 0,\ X_t = x_t\} \leq \alpha. \tag{4.2}$$

Le seuil s est solution de cette équation, qui nous amène à étudier le signe d'un polynôme du second degré en y_t.

En remplaçant les fonctions de densité de probabilité $f(y_t; x_t, \theta_0)$ et $f(y_t; x_t, \theta_1)$ par leur expression, on obtient que l'équation (4.2) est équivalente à :

$$P\left[2\frac{(\mu_1 - \mu_0)}{\sigma^2} y_t + (\mu_0^2 - \mu_1^2)\frac{1}{\sigma^2} - 2\log\left(\frac{s(1-p)}{(1-s)p}\right) > 0 \mid Z_t = 0,\ X_t = x_t\right] \leq \alpha.$$

Posons

$$\gamma = (\mu_0^2 - \mu_1^2)\frac{1}{\sigma^2} - 2\log\left(\frac{s(1-p)}{(1-s)p}\right).$$

En centrant et réduisant, on obtient :

$$P\left[\frac{y_t - \mu_0}{\sigma} > \frac{-\gamma\sigma}{2(\mu_1 - \mu_0)} - \frac{\mu_0}{\sigma} \mid Z_t = 0,\ X_t = x_t\right] \leq \alpha.$$

Puisque le statut de la sonde t est normal ($Z_t = 0$), la distribution de Y_t est une gaussienne de moyenne $\mu_0 = a_0 + b_0 x_t$ et de variance σ^2. On déduit donc que la solution est :

$$\frac{-\gamma\sigma}{2(\mu_1 - \mu_0)} = u_{1-\alpha} + \frac{\mu_0}{\sigma},$$

où $u_{1-\alpha}$ est le quantile de $\mathcal{N}(0,1)$ au niveau α.
En utilisant la définition de γ, on a :

$$2\sigma \log\left(\frac{s}{1-s}\right) = 2u_{1-\alpha}(\mu_1 - \mu_0) - \frac{1}{\sigma}(\mu_1 - \mu_0)^2 - 2\sigma \log\left(\frac{1-p}{p}\right) \cdot$$

L'expression du seuil s est donc donnée par :

$$s = \frac{\exp^{\lambda}}{1 + \exp^{\lambda}},$$

où

$$\lambda = \left(\frac{\mu_1 - \mu_0}{\sigma}\right)\left(u_{1-\alpha} - \frac{\mu_1 - \mu_0}{2\sigma}\right) - \log\left(\frac{1-p}{p}\right) \cdot \blacksquare$$

Ce contrôle se concentre sur les erreurs de classement dans le groupe enrichi. La même stratégie peut être appliquée pour contrôler la probabilité d'assigner une sonde à tort dans le groupe normal. Notons tout de même que les quantités intervenant dans l'expression de λ sont toutes estimées.

Perspectives du contrôle de faux-positifs

Le contrôle de faux-positifs proposé ci-dessus n'est défini que pour $K = 2$ et dans le cas des modèles de mélange. Dans le contexte des HMM, il existe une procédure de contrôle de faux-positifs pour les tests multiples (False Discovery Rate, Sun et Cai, 2009) lorsqu'il n'y a que deux groupes. La règle de classification est la suivante :

$$\tau_{t1} > s \;\Rightarrow\; \widehat{Z}_t = 1 \Leftrightarrow \quad \text{sonde } t \text{ déclarée enrichie},$$

où s est un seuil arbitraire qui doit être fixé. On note $\mathcal{P}(x)$ l'ensemble des sondes t tel que $\tau_t > s$ et on a :

$$\widehat{\mathrm{FDR}}(s) = \frac{\sum_{t \in \mathcal{P}(x)} (1 - \hat{\tau}_t)}{\mathrm{card}(\mathcal{P}(x))}$$

En pratique, on fixe $\widehat{\mathrm{FDR}} \leq \alpha$ et on trouve le seuil s dépendant de α.

Une perspective intéressante serait d'étendre le contrôle de faux-positifs au cas où l'on a plus de deux groupes ($K > 2$), ce qui est difficile à cause de la définition intrinsèque d'un faux-positif. Une solution simple consiste à se ramener à deux populations d'intérêt en sommant les probabilités *a posteriori* des différents groupes associés (cf. Section 1.2).

3 Classification de régions

En classification non supervisée, les observations sont en général classées individuellement. Cependant, dans certains cas, l'observation n'est pas une unité pertinente et il est préférable de classer conjointement un ensemble d'observations. Pour les données *tiling arrays*, les sondes ne sont pas des unités biologiquement pertinentes. En effet, dans les analyses transcriptomiques en particulier, on s'intéresse plutôt au statut d'une région comme par exemple un gène, couvrant généralement plusieurs sondes. Néanmoins, la plupart des méthodes appliquées aux *tiling arrays* pour des problèmes de classification de données génomiques fournissent un résultat par sonde. La classification de régions biologiquement intéressantes n'est pas un procédé habituel. Les quelques méthodes développées pour donner un résultat par région sont fondées sur une approche par fenêtre glissante où les intensités des sondes sont fusionnées *a priori* (Johnson *et al.*, 2006). La méthode proposée par Li *et al.* (2005) est de définir une région *a posteriori* à partir de critères fixés : elle doit contenir au moins deux sondes ayant une valeur d'enrichissement positive dans un échantillon de ChIP, et au moins une sonde avec une valeur d'enrichissement inférieure à -15 dans l'échantillon de référence. Mais ces méthodes ne considèrent pas les régions qui ont une organisation particulière, comme par exemple les régions non connexes couvertes par plusieurs sondes non-adjacentes telles que les gènes avec leurs exons et leurs introns.

3.1 Probabilités *a posteriori* pour une région

Nous généralisons le calcul des probabilités *a posteriori* pour une observation à un ensemble d'observations ayant une structure arbitraire. La généralisation des probabilités *a posteriori* pour une observation est le critère naturel de classification d'une région. Il permet de restaurer localement le chemin caché, et une région est déclarée appartenant au groupe k si le chemin caché reste dans l'état k durant toute la région. On définit la probabilité *a posteriori* $P_{gk,X}$ d'une région g d'appartenir au groupe k comme la probabilité que toutes les sondes de la région appartiennent au groupe k :

$$P_{gk,X} = P(\forall t \in g, Z_t = k \mid X, C) \ , \tag{4.3}$$

où X correspond aux observations, et C à l'annotation.
La région est dite *homogène* puisque toutes les sondes ont le même statut.
Notons que les régions à classer doivent être définies *a priori*.

Une région est définie comme un ensemble de sondes, pas forcément connexe, mais qui peut être décomposée en sous-régions de sondes adjacentes. En référence à la structure des gènes et sans perte de généralité, nous appelerons 'gènes' ces régions, 'exons' les sous-régions et 'introns' les espaces entre les sous-régions.
Dans les gènes eucaryotes, les exons correspondent aux régions codantes qui sont épissées pour former le transcrit d'ARNm, après élimination des introns, qui ne sont pas exprimés. Les probabilités *a posteriori* d'un gène sont donc obtenues en considérant uniquement les sondes exoniques couvrant le gène.

Nous calculons la probabilité $P_{gk,X}$ définie à l'équation (4.3) pour un gène g avec Q exons (et $Q-1$ introns). On note e_q la position de la première sonde de l'exon q et i_q la position de la première sonde de l'intron q ; ainsi $i_q - 1$ correspond à la dernière sonde de l'exon $q-1$. On note par convention i_Q la position de la première sonde après la fin du gène. Le schéma explicatif du calcul pour un gène avec Q exons est donné Figure 4.5.

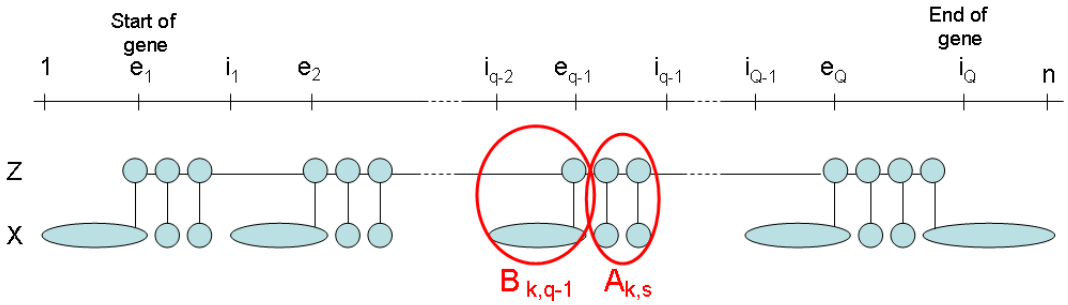

Fig. 4.5: Représentation schématique d'un gène avec Q exons

Remarquons que sous l'hypothèse de dépendance markovienne, les probabilités conditionnelles de la région ne sont pas égales au produit des probabilités de chaque sonde couvrant la région. Le calcul donné Proposition 5 est défini dans le cas de la prise en compte de l'annotation et sous hypothèse de dépendance markovienne (correspondant au Modèle $\mathcal{M}_4$ du chapitre 3 Section 1).

Proposition 5. *On note $X_u^v = \{X_t\}_{u \leq t \leq v}$. La probabilité a posteriori d'une région g d'appartenir au groupe k s'écrit :*

$$\begin{aligned} P_{gk,X} &= P(\forall t \in g, Z_t = k | X, C) \\ &= P(Z_{e_1} = k | X_1^{e_1}, C) \times \left(\prod_{t=e_1+1}^{i_1-1} A_{k,t} \right) \times \prod_{q=2}^{Q-1} \left(B_{k,q} \times \prod_{t=e_q+1}^{i_q-1} A_{k,t} \right) \\ &\quad \times B_{k,Q} \times \left(\prod_{t=e_Q+1}^{i_Q-2} A_{k,t} \right) \times P(Z_{i_Q-1} = k | Z_{i_Q-2} = k, X_{i_Q-1}^n) \ , \end{aligned}$$

où $A_{k,s} = P(Z_s = k \mid Z_{s-1} = k, X_s, C)$ et
$B_{k,q} = P(Z_{e_q} = k \mid Z_{i_{q-1}-1} = k, X_{i_{q-1}}^{e_q}, C)$, avec $C = \{C_t\}$.

Idée de la preuve.

$$\begin{aligned} P_{gk,X} &= P(\forall\ e \text{ sonde exon}, Z_e = k \mid X, C) \\ &= \underbrace{P(Z_{e_1} = k \mid X_1^{e_1}, C)}_{\text{sonde 1 jusqu'à la première sonde de l'exon 1}} \times \underbrace{\prod_{s=e_1+1}^{i_1-1} P(Z_s = k \mid Z_{s-1} = k, X_s, C)}_{\text{sondes de l'exon 1}} \\ &\quad \times \prod_{q=2}^{Q-1} \left[\underbrace{P(Z_{e_q} = k \mid Z_{i_{q-1}-1} = k, X_{i_{q-1}}^{e_q}, C)}_{\text{première sonde de l'exon } q} \times \underbrace{\prod_{s=e_q+1}^{i_q-1} P(Z_s = k \mid Z_{s-1} = k, X_s, C)}_{\text{sondes de l'exon } q} \right] \\ &\quad \times \underbrace{P(Z_{e_Q} = k \mid Z_{i_{Q-1}-1} = k, X_{i_{Q-1}}^{e_Q}, C)}_{\text{première sonde de l'exon } Q} \times \underbrace{\prod_{s=e_Q+1}^{i_Q-2} P(Z_s = k \mid Z_{s-1} = k, X_s, C)}_{\text{sondes de l'exon } Q} \\ &\quad \times \underbrace{P(Z_{i_Q-1} = k \mid Z_{i_Q-2} = k, X_{i_Q-1}^n)}_{\text{dernière sonde de l'exon } Q \text{ jusqu'à la sonde } n} \ . \end{aligned}$$

Tous ces termes peuvent être calculés avec la procédure Forward de l'algorithme Forward/Backward. ■

Un gène est une région particulière puisqu'il est composé d'introns et d'exons. Le calcul est simplifié dans le cas d'une région connexe. D'autre part, la définition des régions faite *a priori* rend possible l'étude de l'épissage alternatif en modifiant la liste des exons associés à un gène. Cela permet aussi d'exclure le dernier exon pour lequel le niveau d'expression pourrait être inférieur en raison du protocole de marquage utilisé lors des expériences *tiling arrays* (Nicolas *et al.*, 2009).

3.2 Règle de classification

Nous présentons à présent une procédure de classification de gènes en deux étapes. Nous définissons d'abord un critère d'homogénéité, puis une règle de classification fondée sur les probabilités *a posteriori*.

1. La probabilité pour un gène d'être homogène quel que soit le statut est $\sum_{k=1}^{K} P_{gk,X}$.

   Nous vérifions d'abord si le gène peut être considéré comme homogène en examinant un ratio similaire à un facteur de Bayes :

   $$\sum_{k=1}^{K} P_{gk,X} \ / \ \sum_{k=1}^{K} P_{gk} \ , \tag{4.4}$$

   où P_{gk} est la version non conditionnelle de $P_{gk,X}$.

   $$\begin{aligned} P_{gk} &= P(\forall t \in g, Z_t = k|C) \\ &= m_k^E \times (\pi_{kk}^E)^{\sum_{q=1}^{Q}(i_q - e_q) - 1} \times \prod_{q=1}^{Q-1} [(\pi^I)^{e_{q+1} - i_q}]_{kk} \ , \end{aligned}$$

   où les exposants E et I se réfèrent aux catégories exoniques et introniques, respectivement.

2. Le calcul de $P_{gk,X}$ dépend de la longueur du gène. En effet, comme le calcul implique un produit de probabilités avec autant de termes que de sondes exoniques dans le gène, $P_{gk,X}$ tend vers zéro pour les gènes longs. Afin de contrôler l'effet de la dépendance de $P_{gk,X}$ à la longueur des gènes, nous appliquons une correction linéaire concernant la longueur des exons et le nombre d'exons dans le gène sur le log-ratio défini Équation (4.4). Ce log-ratio corrigé est défini comme un critère *unistatut* qui est un outil d'aide à la décision. Il doit être positif et le plus élevé possible pour considérer la région homogène. L'avantage de ce critère de classification est d'offrir la possibilité de ne pas classer le gène s'il est trop hétérogène en termes de statuts différents des sondes couvrant le gène.

3. Si le gène est déclaré homogène, la deuxième étape consiste à classer le gène dans l'un des K groupes en fonction des probabilités *a posteriori*. On note que la somme des $P_{gk,X}$ pour $k \in \{1, ...K\}$ n'est pas égale à 1, puisque nous considérons uniquement le cas d'un gène homogène. Il existe bien sûr beaucoup d'autres possibilités où toutes les sondes d'un même gène n'ont pas forcément le même statut. Cette hypothèse simplificatrice d'homogénéité est discutable, mais peut facilement être modifiée car les calculs sont génériques.

Le gène g est finalement attribué au groupe k pour lequel la probabilité $P_{gk,X}/\sum_\ell P_{g\ell,X}$ est la plus élevée :

$$\hat{Z}_g = \underset{k}{\text{Argmax}} \ \frac{P_{gk,X}}{\sum_{\ell=1}^{K} P_{g\ell,X}} \ .$$

Ces probabilités sont renormalisées pour que la somme sur k soit égale à 1.

La classification par région est illustrée sur quelques exemples de gènes dans le chapitre 5 Section 4.3.

Bibliographie

Johnson, W.E., Li, W., Meyer, C.A., Gottardo, R., Carroll, J.S., Brown, M. and Liu, X.S. (2006). Model-based analysis of tiling-arrays for ChIP-chip. *PNAS* **103**, 12457-12462.

Li, W., Meyer, A. and Liu, X.S. (2005). A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* **211**, 274-282.

Nicolas, P., Leduc, A., Robin, S., Rasmussen, S., Jarmer, H. and Bessières, P. (2009). Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. *Bioinformatics* **25**(18), 2341-2347.

Sun, W. and Cai, T.T. (2009) Large-scale multiple testing under dependence. *J.R. Statist.Soc. B*, **71** Part2 393-424.

Chapitre 5

Applications

Sommaire

1 Contexte et outils 92

1.1 Plante modèle *Arabidopsis thaliana* 92

1.2 Caractéristiques de la puce 93

1.3 Normalisation des données 95

1.4 Visualisation dans FLAGdb++ 96

2 Analyse de données de ChIP-chip 97

2.1 Étude de H3K9me3 98

2.2 Étude de l'épigénome d'*Arabidopsis thaliana* avec MultiChIPmix . . 101

2.3 Package ChIPmix 102

3 Analyse de données de ChIP-chip IP/IP - Étude de H3K9me2 . 103

3.1 Analyse avec le modèle gaussien bidimensionnel $\mathcal{M}_2$ 104

3.2 Comparaison avec la méthode de Johannes *et al.* (2010) 106

3.3 Analyse avec le modèle de mélanges de mélange 108

4 Analyse de données transcriptome - Étude des données graine vs feuille 113

4.1 Comparaison de modèles 113

4.2 Analyse avec le modèle gaussien bidimensionnel $\mathcal{M}_4$ 116

4.3 Classification par gène 117

4.4 Détection de nouveaux transcrits 118

4.5 Package TAHMMAnnot 120

5 Étude de simulations - Comparaison de méthodes 121

Bibliographie 124

Les méthodes développées dans le cadre de cette thèse sont génériques et permettent l'exploitation de données de puces *tiling arrays*. Elles ont été appliquées à trois types d'expériences issus du projet ANR/Génoplante Tiling Array Genome (TAG), qui a débuté en 2007, coordonné par l'Unité de Recherche en Génomique Végétale (URGV) spécialisée dans la Génomique des Plantes (Unité Mixte de Recherche INRA/Université Evry-Val d'Essonne ERL[1] CNRS). L'objectif de ce projet était de concevoir, valider et exploiter une puce *tiling array* unique couvrant le génome entier d'*Arabidopsis thaliana*. L'idée était de créer un outil pour toute la communauté travaillant sur *Arabidopsis* afin d'étudier l'activité du génome de cette plante modèle. Les applications de cette puce ont été mises au point pour divers types d'expérimentations (méthylation de l'ADN, ChIP-chip, RIP-chip, détection et analyse de transcrits codants et non-codants, etc.). Les données analysées dans cette thèse concernent les expériences de ChIP-chip IP/INPUT, ChIP-chip IP/IP et transcriptome.

Nous décrivons dans la Section 1 les caractéristiques générales de la puce, nous discutons brièvement de la normalisation des données et nous présentons le logiciel de visualisation des résultats développé à l'URGV. Les Sections 2, 3 et 4 présentent respectivement les résultats obtenus et les interprétations biologiques des trois types d'expériences analysées. Les méthodes développées sont automatisées sous forme de packages R afin d'être directement utilisables par les biologistes (cf. Sections 2.3 et 4.5). Enfin une étude de simulation permettant de comparer différentes méthodes d'exploitation de puces *tiling arrays* est proposée Section 5.

1 Contexte et outils

1.1 Plante modèle *Arabidopsis thaliana*

Arabidopsis thaliana, de son nom commun "Arabette des Dames" ou "Fausse Arabette", est une plante de la famille des Brassicacées (Crucifères) à laquelle appartiennent de nombreuses espèces cultivées utilisées dans l'alimentation (chou, navet, radis, moutarde, etc). C'est une plante herbacée mesurant 25 à 35 cm de haut à l'âge adulte (cf. Figure 5.1). Cette plante est un organisme modèle pour la recherche génétique dans le monde végétal. Les raisons de ce choix sont nombreuses : son génome est l'un des plus petits du monde végétal avec ses 5 chromosomes regroupant environ 120 millions de paires de bases, soit environ vingt fois plus petit que celui de plantes cultivées comme l'orge ou le maïs. Son cycle de reproduction, principalement par autofécondation, est court (6 à 8 semaines) et peut s'accomplir entièrement *in vitro*. C'est une plante de petite taille (en laboratoire on peut cultiver un millier de pieds sur un mètre carré) et très prolifique (plusieurs milliers de graines à chaque génération). Plusieurs centaines de mutations sont connues et beaucoup sont aisément détectables à l'oeil nu (couleur, poils, fleurs, etc.). De plus, l'absence d'intérêts économiques sur cette espèce facilite la diffusion des informations entre laboratoires. En 2000 ce fut le premier génome végétal séquencé, avec environ 25 000 gènes identifiés et localisés. Les recherches continuent pour établir la fonction de chacun d'eux.

[1] Équipe de Recherche Labellisée

Fig. 5.1: *Arabidopsis thaliana.*

1.2 Caractéristiques de la puce

La puce utilisée est une puce *tiling array* deux couleurs à oligos courts issue de la technologie NimbleGen. Cette puce permet d'étudier le génome nucléaire d'*Arabidopsis thaliana*, composé de cinq chromosomes et des génomes mitochondrial et chloroplastique à l'aide d'environ 720 000 sondes réparties le long du génome. Les sondes sont synthétisées *in situ* par un procédé de photolithographie. Pour déterminer leur emplacement, le génome est segmenté en fenêtres de 165 nt à l'intérieur desquelles sont dessinées des sondes de longueur variable (entre 50 et 75 nt, la moyenne étant environ de 55 nt). Ainsi deux sondes consécutives ne doivent pas être séparées par plus de 230 nt. Une analyse descriptive indique que l'écart moyen est de 112 nt pour les cinq chromosomes. Cependant, certaines régions de faible complexité, le plus souvent péricentromériques, riches en micro-satellites et en éléments répétés, ne sont pas encore séquencées. La société NimbleGen n'a donc pas pu fabriquer de sondes dans ces régions mais le nombre d'espaces supérieurs à 230 nt reste tout de même petit par rapport au nombre total de sondes.
Les chromosomes chloroplastiques et mitochondriaux sont représentés différemment des chromosomes nucléaires, avec une taille moyenne des sondes plus petite et un espacement moyen entre sondes plus grand. Ceci est dû à leur plus petite taille et au fait qu'ils ont un plus grand nombre de régions non séquencées.

Afin de maîtriser au mieux l'étape d'hybridation lors de la préparation de la puce, il est préférable que la température d'association des deux brins d'ADN soit la même pour toutes les sondes. Cette température dépend fortement du taux de GC dans la séquence nucléotidique et est mesurée grâce au Tm. On s'attend à ce qu'un oligo à Tm faible ($< 76^\circ$) hybride moins bien, et qu'à l'inverse, un oligo à Tm élevé ($> 76^\circ$) hybride de façon forte mais moins spécifique. La particularité des puces NimbleGen est d'assurer la plus grande homogénéité d'hybridation possible entre les sondes et les cibles en choisissant des sondes ayant toutes un Tm proche de 76°. C'est la raison pour laquelle les sondes NimbleGen sont de longueur variable, cela permet de choisir celles qui ont le Tm le plus proche possible de 76°C. Une analyse descriptive montre que la moyenne du Tm sur l'ensemble des sondes est 74.7°C, la médiane est 75.7°C. Néanmoins, à cause des contraintes fortes sur la longueur et l'espacement des sondes imposées sur le design de la puce, le Tm varie beaucoup, entre 60 et 90°C (cf. Figure 5.2).

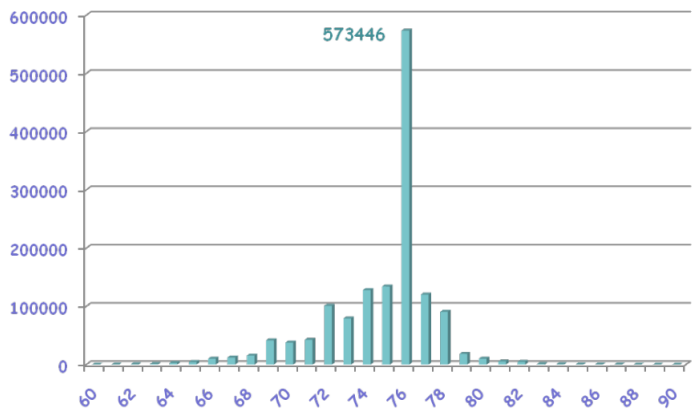

Fig. 5.2: Histogramme des Tm.

La spécificité d'une sonde représente le nombre de fois où l'on trouve la sonde dans le génome. Cette spécificité peut influer sur la valeur du signal d'hybridation. En effet, si une sonde correspond à une séquence répétée plusieurs fois dans le génome, le signal d'hybridation sera plus fort uniquement à cause du fait qu'un plus grand nombre de protéines peut venir se fixer dessus. La spécificité de la sonde a été calculée à l'aide du logiciel SPADS (Thareau *et al.*, 2003) développé à l'URGV. Le calcul de spécificité se base sur des alignements locaux cherchant à détecter les régions paralogues (issues d'un événement de duplication) les plus proches et à estimer le pourcentage de conservation (pourcentage d'identité) entre la sonde et la région paralogue la plus proche. Quatre niveaux de conservation sont définis : le niveau 1 correspond à un pourcentage d'identité inférieur à 40%, le niveau 2 correspond à un pourcentage d'identité compris entre 40% et 70%, le niveau 3 correspond à un pourcentage d'identité supérieur à 70% et le niveau 4 correspond à un pourcentage d'identité égal à 100%. Les sondes ayant un pourcentage d'identité égal à 100% sont appelées sondes *repeat*, cela signifie qu'il y a au moins une autre région du génome exactement identique. Le logiciel SPADS a permis d'établir qu'environ 8% des sondes de la puce sont des sondes *repeat*.

Bien que les sondes soient dessinées sans tenir compte de l'annotation structurale, il est possible de caractériser ultérieurement la répartition des sondes le long du génome en fonction de cette annotation. Les sondes peuvent couvrir intégralement des régions intergéniques, exoniques ou introniques, mais aussi être à cheval sur deux régions. La description de tous les cas possibles est schématisée Figure 5.3. La majorité des sondes (67%) couvre de l'intergénique, 14% des sondes couvrent un exon et 4% couvrent un intron. Les 15% restantes sont pour moitié des sondes non spécifiques et enfin 7% des sondes couvrent des régions chevauchant deux annotations différentes (cf. Figure 5.3). Cette information est nécessaire pour prendre en compte l'annotation dans les modèles statistiques.

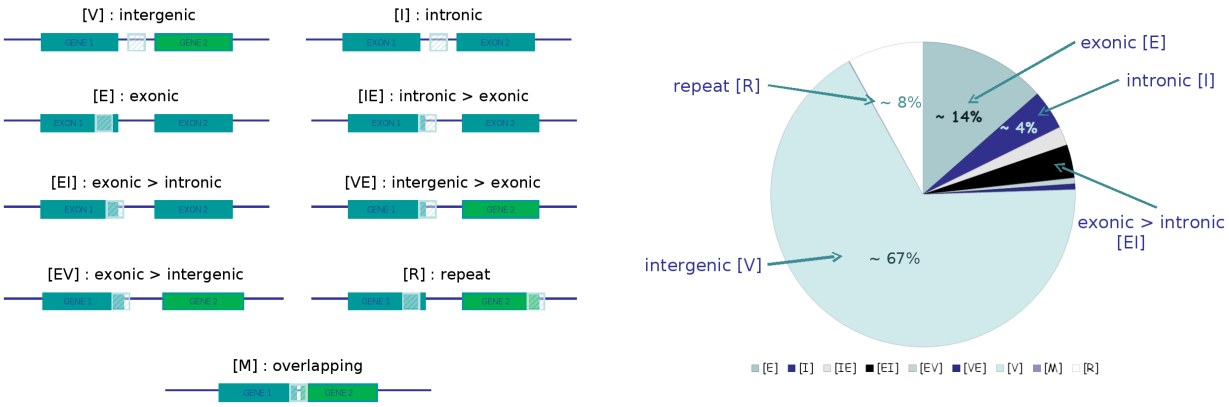

Fig. 5.3: Gauche : Description des différentes possibilités d'annotation pour une sonde. La sonde est représentée par un carré bleu rayé, les deux rectangles (bleu et/ou vert) représentent soit deux gènes adjacents, soit deux exons successifs d'un même gène. Droite : Répartition des sondes en fonction de leur annotation.

1.3 Normalisation des données

Les biais techniques sont nombreux et affectent de manière non négligeable la mesure de l'intensité des sondes. Le but de la normalisation des données est de les quantifier puis de les corriger. La normalisation des données constitue un objet d'étude en soi. Smyth, Yang et Speed (2003) proposent une étude complète des questions de normalisation des données de puces à ADN.
Pour les puces où deux échantillons sont cohybridés, le biais principal est le biais de marquage lié à une différence d'incorporation des deux fluorochromes. En effet, l'échantillon marqué avec le fluorochrome vert a des intensités globalement plus élevées que celui marqué en rouge. Pour éliminer cet effet, on utilise un dispositif en dye-swap (Boulicaut et Gandrillon, 2004). Le principe du dye-swap est de faire une répétition technique en inversant les marquages, chaque traitement est ainsi marqué par les deux fluorochromes. Les valeurs des intensités des signaux sont ensuite moyennées sur les deux puces du dye-swap.

Toutes les expériences analysées dans ce chapitre ont été effectuées sur des puces NimbleGen deux-couleurs décrites Section 1.2. Pour chaque jeu de données, deux réplicats biologiques sont disponibles et les hybridations sont réalisées en dye-swap. Les intensités des signaux obtenues pour chaque traitement sont moyennées sur le dye swap pour corriger les biais de marquage gène spécifique. De plus, nous utilisons les logarithmes en base 2 des intensités des signaux car cela permet de stabiliser la variance et d'avoir une échelle de valeurs plus petite.
Le jeu de données de ChIP-chip a fait l'objet d'une étape de normalisation complémentaire car les sondes de la puce étaient réparties sur trois lames différentes, ce qui a engendré des biais techniques supplémentaires. Comme les signaux IP et INPUT diffèrent sensiblement, les données ont été normalisées avec le modèle d'analyse de la variance suivant (adapté du modèle décrit dans Kerr *et al.*, 2002) :
Notons Y_{plfts} le log_2 du signal de la sonde s sur la puce p et la lame l, de traitement t et marqué avec le fluorochrome f. On s'intéresse uniquement aux effets qui ne concernent

pas la sonde s.

$$Y_{plfts} = \mu + \alpha_p + \beta_l + (\alpha\beta)_{pl} + \gamma_f + (\alpha\gamma)_{pf} + E_{plfts}. \tag{5.1}$$

où

- $\alpha_p + \beta_l + (\alpha\beta)_{pl}$ correspond à l'effet support (lame + puce + interactions),
- γ_f correspond à l'effet fluorochrome,
- $(\alpha\gamma)_{pf}$ correspond à l'effet puce×fluorochrome,
- les résidus E_{plfts} sont supposés indépendants, identiquement distribués et centrés.

Les biais sont quantifiés en estimant les paramètres du modèle, puis soustraits des données brutes. Dans la suite, les analyses sont effectuées sur les données normalisées $\widehat{E}_{plfts}$.

1.4 Visualisation dans FLAGdb++

FLAGdb++[1] (Dèrozier *et al.*, 2011) est une base de données intégrative dédiée à la génomique structurale et fonctionnelle des plantes, développée à l'URGV. Actuellement, elle gère les génomes d' *Arabidopsis thaliana*, d'*Oryza sativa* (riz), de *Vitis vinifera* (vigne) et de *Populus trichocarpa* (peuplier). FLAGdb++ est utilisée par les biologistes pour étudier la fonction des gènes végétaux en les considérant dans un contexte large : un environnement chromosomique, une famille multigénique et/ou un réseau fonctionnel. Les outils d'exploration sont développés de manière générique afin de pouvoir s'appliquer à différents génomes et pouvoir stocker, organiser, visualiser et exploiter de nombreux types de données génomiques telles que les annotations structurales et fonctionnelles, les séquences transcrites (ESTs, ADNc, $\cdots$), les motifs protéiques, les séquences répétées, etc. FLAGdb++ s'articule en deux parties : une base de données permettant le stockage et une interface graphique (développée en Java) permettant une visualisation et une interrogation aisées de l'information.

Étant donné le nombre important de données issues d'une expérience de *tiling array*, l'exploitation et la visualisation des résultats deviennent difficiles. Pourtant, la visualisation des résultats fait partie intégrante de l'analyse et constitue la première étape dans l'inférence d'hypothèses biologiques. Durant ma thèse j'ai travaillé en collaboration avec Sandra Dèrozier et Sébastien Aubourg (Équipe bioinformatique pour la génomique prédictive, URGV) afin de mettre au point un outil de visualisation des résultats statistiques obtenus. Dans un premier temps, un module Java a été intégré à FLAGdb++ afin de permettre la visualisation de données générées à partir de puces *tiling arrays* (cf. Figure 5.4). La visualisation concerne les sondes et leurs caractéristiques (séquence, Tm, unicité dans le génome, etc.) ainsi que leur position relative par rapport aux annotations structurales du génome. Un module de visualisation statistique a ensuite été ajouté. Chaque sonde peut être colorée en fonction des résultats statistiques obtenus pour l'expérience analysée. La couleur des sondes dépend des probabilités *a posteriori* estimées par le modèle statistique (cf. Figure 5.5). Le nombre de groupes K de la classification ainsi que les couleurs associées aux K groupes sont définis par l'utilisateur. Le rectangle représentant la sonde est divisé en K rectangles de taille proportionnelle aux K probabilités *a posteriori* et colorés en fonction du groupe correspondant.

[1] http ://urgv.evry.inra.fr/FLAGdb++

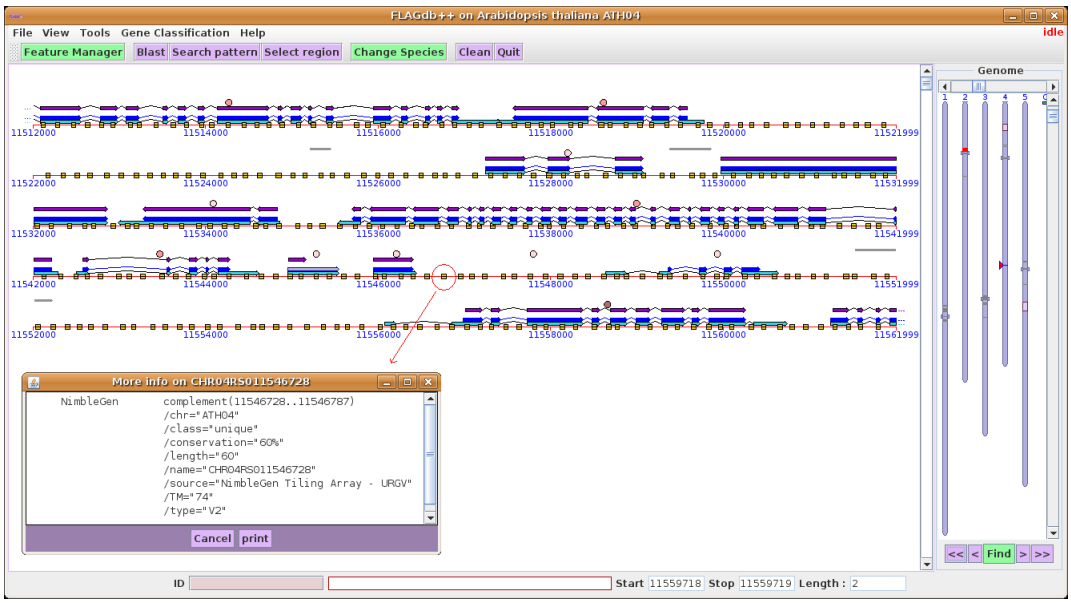

Fig. 5.4: Fenêtre de FLAGdb++ où l'on visualise les informations d'annotation structurale, ainsi que les sondes (carrés jaunes).

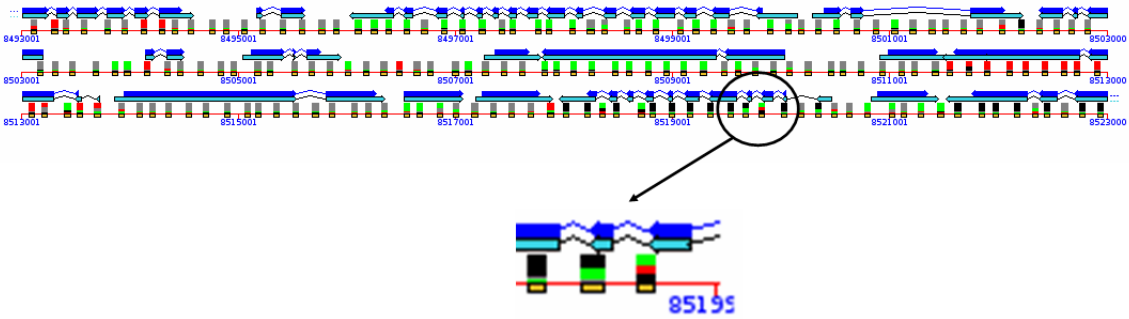

Fig. 5.5: Visualisation des résultats statistiques sous forme de couleur sur chaque sonde. La couleur des sondes dépend des probabilités *a posteriori*.

2 Analyse de données de ChIP-chip

Les expériences de ChIP-chip ont été les premières données produites avec la puce *tiling array* d'Arabidopsis présentée Section 1. Avant exploitation, la puce, ainsi que la méthode d'analyse ont été dans un premier temps validées sur un jeu de données spécifiquement choisi. Ce jeu de données concerne l'étude du comportement de l'histone H3 tri-méthylée au niveau de la lysine 9 (H3K9me3) chez *Arabidopsis thaliana*. Il a déjà été étudié par Turck *et al.* (2007) sur une puce couvrant uniquement le chromosome 4 d'Arabidopsis et les résultats obtenus ont été validés biologiquement. Le même jeu de données a donc été analysé avec la méthode ChIPmix, modèle de mélange de régressions présenté Chapitre 3 Section 2.1 (Martin-Magniette *et al.*, 2008) et des comparaisons avec d'autres méthodes ainsi qu'avec les résultats de Turck *et al.* (2007) ont permis de valider la méthode (cf. Section 2.1). De nombreuses expériences de ChIP-chip ont ensuite été analysées afin d'étudier l'épigénome d'*Arabidposis thaliana* (cf. Section 2.2). Le package R associé à la méthode ChIPmix de mélange de régressions est présenté Section 2.3.

2.1 Étude de H3K9me3

L'objectif biologique est d'identifier les régions du génome qui sont des cibles de H3K9me3. Les deux échantillons hybridés représentent l'ADN immunoprécipité (IP) et l'ADN génomique total (échantillon de référence INPUT), et il faut détecter les sondes enrichies, c'est-à-dire les sondes pour lesquelles la valeur du signal IP est significativement plus grande que la valeur de l'INPUT.

a) Analyse avec ChIPmix

Le jeu de données H3K9me3 a été analysé avec la méthode ChIPmix, modèle de mélange de régressions présenté Chapitre 3 Section 2.1 (Martin-Magniette *et al.*, 2008, cf. AnnexeA). Les modèles HMM et annotation présentés Chapitre 3 Section 1 n'ayant été développés qu'ultérieurement, cette analyse a été faite sous hypothèse d'indépendance des sondes. La classification des données en deux groupes (enrichi et normal) est obtenue avec le contrôle des faux-positifs présenté Chapitre 4 Section 2, en fixant $\alpha = 0.01$.

Nous présentons en particulier les résultats obtenus pour l'analyse du chromosome 4. Les estimations des paramètres du modèle sont données Table 5.1. On remarque que les paramètres estimés $\hat{b}$ des pentes ne sont pas égaux à 1, ce qui signifie que l'hypothèse sous-jacente des méthodes fondées sur le log-ratio n'est pas vérifiée dans cet exemple (cf. Chapitre 3 Section 2.1).
Pour évaluer la reproductibilité des analyses, nous avons comparé les résultats pour deux réplicats biologiques : les estimations obtenues sont semblables pour les deux réplicats.

		constante a	pente b	variance σ^2	probabilité π
REP1	Groupe Normal	1.47	0.82	0.42	0.74
	Groupe Enrichi	-0.47	1.17	0.42	0.26
REP2	Groupe Normal	2.29	0.75	0.38	0.76
	Groupe Enrichi	0.47	1.07	0.38	0.24

Tab. 5.1: Tableau des estimateurs pour le chromosome 4, pour les deux réplicats biologiques.

Après classification avec contrôle des faux-positifs, environ 17% des sondes sont déclarées enrichies pour le réplicat biologique 1, et 15.5% pour le réplicat 2. Environ 75% des sondes déclarées enrichies dans le premier réplicat le sont également dans le deuxième. La population de sondes déclarées enrichies est représentée en rouge sur la Figure 5.6.

Après visualisation des résultats, on remarque que malgré l'absence de prise en compte de la dépendance dans le modèle, les sondes enrichies sont regroupées sous forme de plage. De plus, les régions correspondant aux plages de sondes enrichies sont riches en gènes, ce qui est cohérent avec le fait que H3K9me3 est une marque euchromatiniennne. Cette marque euchromatinienne a déjà fait l'objet d'une étude dans Turck *et al.* (2007), en utilisant une puce *tiling array* couvrant uniquement le chromosome 4. Les régions identifiées ont été validées biologiquement. Plus de 75% des sondes du chromosome 4 déclarées enrichies par ChIPmix couvrent des régions génomiques identifiées par Turck *et al.* (2007).

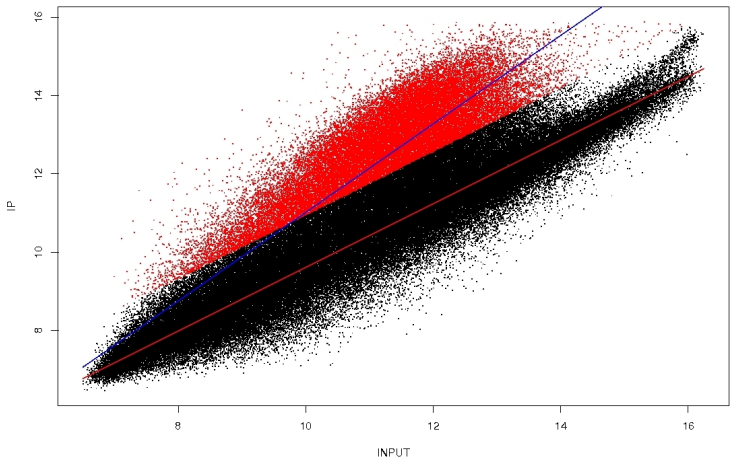

Fig. 5.6: Résultats de la méthode de régression pour le chr4 (Rep1), avec les droites de régression estimées. En rouge : sondes déclarées enrichies après classification.

Les résultats sont similaires pour les quatre autres chromosomes du génome nucléaire. On obtient entre 15% et 17% de sondes enrichies selon les chromosomes. Les génomes chloroplastique et mitochondrial ont un comportement particulier puisqu'il n'existe pas de régions cibles de H3K9me3 en dehors du génome nucléaire. Le cas du génome mitochondrial est complexe car certaines régions ont été dupliquées dans le génome nucléaire sur le chromosome 2, nous ne l'étudierons donc pas en détail. Le génome chloroplastique constitue un bon exemple test où l'on s'attend à n'avoir qu'une seule population de sondes normales et aucune sonde enrichie. Le modèle sélectionné par BIC pour le génome chloroplastique est un modèle à deux droites de régression (cf. Figure 5.7), mais seulement 15 sondes sont déclarées enrichies. Le critère BIC vaut 180 pour le modèle à deux droites de régression contre 188 pour le modèle à une seule population. Les résultats sont similaires sur le deuxième réplicat biologique.

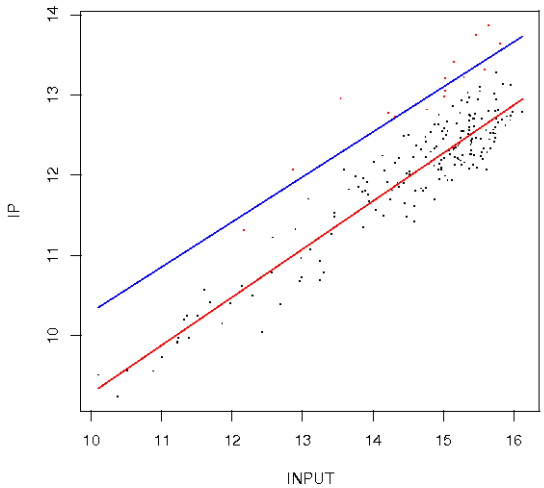

Fig. 5.7: Résultats de la méthode de régression pour le chromosome chloroplastique.

b) Comparaison de méthodes

Nous avons comparé les résultats fournis par ChIPmix avec ceux de deux autres méthodes (choisies parmi celles publiées lors de la réalisation de ce travail) : la méthode ChIPOTle (Buck, Nobel et Lieb, 2005) et le logiciel de NimbleGen. La société NimbleGen fournit avec la puce une liste de sondes déclarées enrichies par une méthode de détection de pics. Cette méthode applique un algorithme de permutation pour trouver les pics statistiquement significatifs, en utilisant le log-ratio des intensités. ChIPOTle est une méthode dédiée à la détection de pics, utilisant une approche par fenêtre glissante sur le log-ratio. Deux paramètres, influant sur le nombre de pics détectés, doivent être fixés : nous avons choisi la taille de la fenêtre à 500 et le pas à 100.
D'autre part, les résultats des trois méthodes ont été comparés aux régions validées par Turck *et al.* (2007) sur le chromosome 4, considérées comme référence.

Les sondes déclarées enrichies par ChIPmix incluent presque toutes celles trouvées enrichies par le logiciel NimbleGen, mais couvrent des régions génomiques plus larges. Par ailleurs ChIPmix identifie d'autres régions génomiques non détectées par le logiciel NimbleGen (voir Figure 5.8), et qui sont validées par une comparaison avec les résultats de Turck *et al.* (2007). ChIPmix détecte 30 477 sondes enrichies, dont 24 575 en commun avec Turck *et al.* (2007). ChIPOTle détecte 24 357 sondes (20 866 communes avec Turck *et al.* (2007)) et NimbleGen détecte 19 837 sondes (16 600 communes avec Turck *et al.* (2007)) (voir Figure 5.9). Parmi les trois méthodes, ChIPmix fournit les résultats les plus proches de la publication de référence.
En ce qui concerne le génome chloroplastique, ChIPOTle détecte aussi des sondes enrichies, supérieures en nombre à celles détectées par ChIPmix. Le logiciel NimbleGen fournit uniquement les analyses des chromosomes nucléaires.

Fig. 5.8: Comparaison des résultats avec $SignalMap^{TM}$ sur une région du chromosome 4 : la 1ère ligne donne l'annotation du génome (les plages en rose sont des gènes) ; la 2ème ligne correspond aux résultats de la méthode de détection de pics de NimbleGen (les pics sont en couleur par dessus les sondes, colorés selon la valeur du FDR) ; la 3ème ligne correspond aux sondes déclarées enrichies par ChIPmix (en rouge) avec un contrôle de faux positifs de seuil $\alpha = 0.01$; la 4ème ligne correspond aux résultats de ChIPotle (fenêtre=500, pas=100).

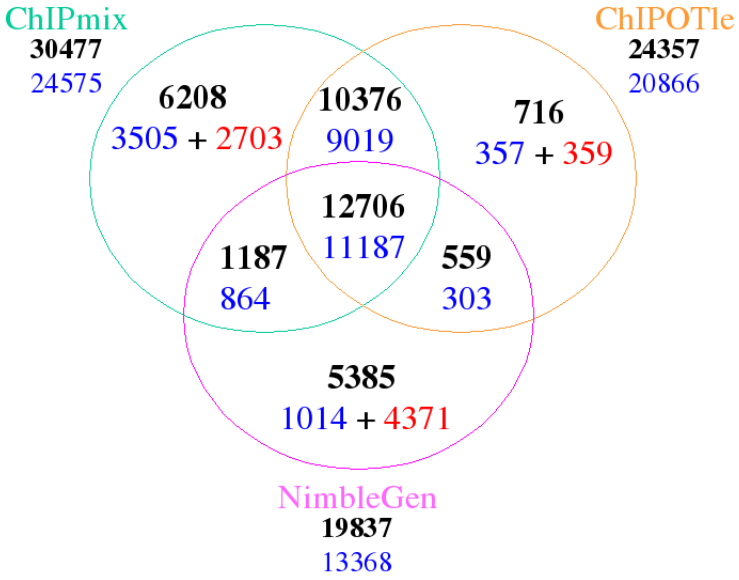

Fig. 5.9: Diagramme de Venn résumant les résultats des trois méthodes. Le nombre de sondes enrichies est en noir, le nombre de sondes validées par Turck *et al.* (2007) est en bleu, le nombre de sondes détectées uniquement par la méthode est en rouge.

2.2 Étude de l'épigénome d'*Arabidopsis thaliana* avec MultiChIPmix

Une fois la puce et la méthode validées sur le jeu de données H3K9me3, elles ont été utilisées dans des projets biologiques pour analyser de nombreuses expériences de ChIP-chip, dans l'objectif d'étudier l'épigénome d'*Arabidopsis thaliana*. Les modifications d'histone et la méthylation d'ADN définissent différents états chromatiniens qui engendrent des états transcriptionnels distincts. Ainsi, la composition précise de la chromatine le long du génome, qui définit l'épigénome, contribue à la régulation de la transcription du génome. Les principes gouvernant l'organisation de l'épigénome restent assez peu compris, c'est pourquoi ils intéressent les biologistes. En réponse aux besoins des biologistes pour étudier simultanément plusieurs réplicats biologiques, les données ont été analysées en utilisant la méthode MultiChIPmix, adaptée de Martin-Magniette *et al.* (2008), présentée Chapitre 3 Section 2.1. La classification est réalisée en contrôlant un critère analogue au FDR (Mary-Huard *et al.*, 2010), qui est une extension du contrôle de faux-positifs présenté Chapitre 4 Section 2 pour le contrôle de tests multiples. Les sondes pour lesquelles le risque ajusté par la procédure de Benjamini-Hochberg est inférieur à 0.01 sont déclarées enrichies.
Les sondes voisines enrichies sont ensuite regroupées en domaines, en exigeant une séquence minimale de 400 paires de bases (*pb*) et en permettant un écart maximal de 200 *pb*. Ainsi, les sondes isolées enrichies ne sont pas considérées dans les analyses.

Pour caractériser les états chromatiniens le long du génome d'*Arabidopsis thaliana*, Roudier *et al.* (2011) ont généré des cartes épigénomiques pour huit modifications d'histone (H3K4me2 et 3, H3K27me1 et 2, H3K36me3, H3K56ac, H4K20me1 et H2Bub) en utilisant une puce *tiling array* couvrant le chromosome 4 d'Arabidopsis avec une résolution d'environ 900 *pb*. Quatre cartes épigénomiques obtenues précédemment dans la littérature pour H3K9me2 et 3, H3K27me3 et la méthylation de l'ADN (5mC) ont aussi été considérées. D'autre part, des profils épigénomiques ont été effectués pour sept des marques

précédentes (H3K4me2 et 3, H3K27me1, H3K27me3, H3K36me3, H2Bub et 5mC) en utilisant la puce *tiling array* présentée Section 1 couvrant le génome entier d'Arabidopsis avec une résolution de 165 *pb*. Les analyses des huit marques chromatiniennes sur la puce couvrant le chromosome 4 ainsi que les analyses des sept marques sur la puce couvrant le génome entier ont été réalisées avec MultiChIPmix.
Le profilage épigénomique d'*Arabidopsis* a fourni des indications sur la relation entre l'activité transcriptionnelle et les marques chromatiniennes. Par exemple, H3K4me3 et H3K36me2 sont détectées respectivement aux extrémités 3' et 5' des gènes activement transcrits, tandis H3K27me3 est largement corrélée avec la répression des gènes (Oh, Park et van Nocker, 2008). En revanche, la méthylation de l'ADN (5mC) a une double localisation. Elle est présente principalement dans les régions hétérochromatines, sur les éléments transposables et répétés, où elle est associée à H3K9me2 et H3K27me1, mais aussi dans environ 30% des gènes, dont plusieurs sont caractérisés par des niveaux d'expression modérée (Bernatavichute *et al.*, 2008). La distribution d'H3K36me3, H3K9me3 et H2Bub sur les régions codantes des gènes exprimés suggère que ces modifications sont liées à l'élongation de la transcription.
Les analyses intégratives indiquent que ces 12 modifications de la chromatine, qui couvrent à elles seules environ 90% du génome, sont présentes à n'importe quelle position dans un nombre limité de combinaisons. De plus, Roudier *et al.* (2011) ont montré que l'épigénome d'*Arabidopsis* peut être partitionné en 4 types chromatiniens principaux, avec des propriétés fonctionnelles distinctes. Ces types chromatiniens forment des domaines courts et très entrecoupés le long du chromosome.
Cette première étude de l'organisation de la chromatine chez *Arabidopsis* suggère des principes d'organisation simples et fournit les principes de bases de la chromatine qui façonnent l'activité transcriptionnelle chez les plantes.

Moghaddam *et al.* (2011) se sont intéressés à l'héritage des modifications d'histone dans les hybrides intra-spécifiques d'*Arabidopsis thaliana*. Peu d'informations ont à ce jour été établies sur la stabilité ou la dynamique des modifications de la chromatine en réponse à une hybridation intra-espèces. Des expériences de ChIP-chip et de CGH ont été réalisées pour étudier la stabilité des modifications d'histones en réponse à l'hybridation intra-espèces de divers écotypes consanguins d'*Arabidopsis* (Col-0, C24 et Cvi). Les modifications d'histone H3K4me2 et H3K27me3 ont été comparées à l'échelle du génome entre les écotypes parentaux Col-0, C24 et Cvi et leur descendance hybride. En combinant l'hybridation de génomique comparative et le profilage épigénomique, Moghaddam *et al.* (2011) ont montré que les patrons de distribution de H3K4me2 et H3K27me3 sont globalement similaires dans divers écotypes d'*Arabidopsis thaliana*, et demeurent largement inchangés dans leur descendance F1. H3K4me2 et H3K27me3 sont plutôt stables en réponse à l'hybridation intra-espèces, avec un héritage additif dans la descendance hybride. Cependant, les analyses ont révélé qu'il existe de faibles variations de modification de la chromatine parmi les écotypes d'*Arabidopsis*, les polymorphismes de séquence se trouvant principalement dans des éléments transposables. La gamme de ces variations est plus élevée pour H3K27me3 (généralement une marque répressive) que pour H3K4me2 (généralement une marque active).

2.3 Package ChIPmix

Lorsque la méthode est utilisée comme un outil d'analyse dans de nombreuses études, il est nécessaire de proposer aux utilisateurs un outil automatique et efficace. La méthode de mélange de régressions (cf. Chapitre 3 Section 2.1) et le calcul du contrôle de

faux-positifs (cf. Chapitre 4 Section 2) sont implémentés sous R (fonction ChIPmix.R). L'initialisation de l'algorithme d'estimation des paramètres peut se faire de manière aléatoire, ou en utilisant deux droites issues du premier axe d'une Analyse en Composantes Principales (ACP) appliquée au jeu de données. Les deux droites sont choisies respectivement avec une pente 5% plus forte ou plus faible que celle estimée par l'ACP. Le programme sélectionne le modèle à un composant (régression linéaire simple) ou deux composants (mélange de deux régressions) qui minimise le critère BIC (Schwarz, 1977, cf. Chapitre 2 Section 4). En effet, si la protéine d'intérêt n'a aucune cible, c'est-à-dire qu'aucune sonde n'est enrichie, une régression linéaire simple est suffisante pour ajuster les données. Dans notre cas, le nombre de paramètres libres du modèle est égal à $4k - 1$, où k est le nombre de composants du modèle.

La fonction ChIPmix.R nécessite un fichier d'entrée comportant 3 colonnes : l'identifiant de la sonde, l'intensité INPUT et l'intensité IP. En sortie, le programme retourne l'estimation numérique des paramètres du modèle de mélange (l'ordonnée à l'origine et la pente pour chaque droite, les variances et les probabilités *a posteriori* de la proportion de sondes dans chaque population), un graphe avec les 2 droites de régression et les sondes enrichies colorées en rouge, et un fichier texte comportant 6 colonnes : l'identifiant de la sonde, l'intensité INPUT, l'intensité IP, la probabilité *a posteriori* d'être enrichi, le seuil calculé par le contrôle de faux positif et le statut (0 si normal, 1 si enrichi).
Dans le package, des codes perl et R permettent de mettre en forme les données à partir des fichiers *pair* fournis par NimbleGen, ainsi que de réaliser la normalisation ANOVA présentée Section 1.3.

La méthode ChIPmix peut être appliquée à différentes questions biologiques (modifications d'histone, méthylation d'ADN, promoteurs, etc.). Elle fonctionne pour différents organismes et s'adapte aux différentes situations, quelle que soit la proportion de sondes enrichies. Actuellement, elle est utilisée sur la plateforme de l'URGV pour analyser la méthylation d'ADN chez le bovin et elle est mise à disposition des biologistes de l'équipe épigénétique et épigénomique végétales de l'ENS (CNRS UMR 8197) avec qui nous avons collaboré.

La fonction ChIPmix.R est disponible sur le site `http://www.agroparistech.fr/mia/doku.php?id=productions:logiciels`. Les packages "ChIPmixHMM" (permettant de prendre en compte la dépendance dans les données), "MultiChIPmix" (permettant l'analyse simultanée de réplicats biologiques) et "MultiChIPmixHMM" (dépendance des données et analyse simultanée de réplicats biologiques) sont disponibles sur demande. Les packages avec HMM nécessitent des fichiers ordonnés en entrée et n'effectuent pas de contrôle de faux-positifs.

3 Analyse de données de ChIP-chip IP/IP - Étude de H3K9me2

La technique du ChIP-chip IP/IP permet d'étudier directement la différence entre deux échantillons d'ADN immunoprécipités (correspondant à deux conditions distinctes), sans hybrider sur la puce l'ADN génomique total. Elle permet de caractériser une différence d'enrichissement entre deux échantillons qui ont alors un rôle symétrique.
Dans le jeu de données étudié, les deux échantillons co-hybridés sur la puce visent à

étudier le comportement de l'histone H3 di-méthylée au niveau de la lysine 9 (H3K9me2) chez *Arabidopsis thaliana*. Le but est d'étudier les différences de méthylation entre un échantillon sauvage et un échantillon mutant (mutant nrpd1ab, qui est un double mutant polIV et polV) en définissant quatre groupes de sondes : pas d'hybridation (groupe bruit), méthylation identique (groupe identique), perte de méthylation chez le mutant (groupe appauvri), gain de méthylation chez le mutant (groupe enrichi).
Cette méthylation d'histone est faiblement présente dans le génome, et le mutant est connu pour avoir une perte de méthylation par rapport au sauvage (Bernatavichute *et al.*, 2008). Les proportions attendues sont environ 15% des sondes dans le groupe appauvri et 5% des sondes dans le groupe enrichi.

Nous avons dans un premier temps analysé ces données avec le modèle gaussien bidimensionnel défini Chapitre 3 Section 2.2 sous hypothèse de dépendance des observations (cf. Section 3.1). Nous avons ensuite comparé les résultats à ceux obtenus par Johannes *et al.* (2010), qui ont proposé un mélange gaussien bidimensionnel avec des contraintes sur les paramètres de moyenne (cf. Section 3.2). Enfin, nous avons ré-analysé ces données avec le modèle de mélange de distributions pour la loi d'émission présentée Chapitre 3 Section 2.3 afin d'illustrer les améliorations apportées par cette modélisation plus flexible (cf. Section 3.3).

3.1 Analyse avec le modèle gaussien bidimensionnel $\mathcal{M}_2$

Le jeu de données H3K9me2 est analysé avec le modèle gaussien bidimensionnel défini Chapitre 3 Section 2.2. La loi de la variable latente est un HMM sans prise en compte de l'annotation, modèle noté $\mathcal{M}_2$, défini Chapitre 3 Section 1. En effet, les modifications d'histone affectent de larges régions adjacentes de l'histone (Humburg, Bulger et Stone, 2008) donc les sondes enrichies sont détectées principalement sur les éléments transposables mais aussi dans leur environnement. Comme la méthylation n'affecte pas une catégorie d'annotation spécifique, les informations d'annotation structurale ne sont pas utiles pour détecter les sondes différentiellement enrichies. La classification est réalisée avec la règle du MAP.

Nous présentons les résultats obtenus pour le chromosome 4, les résultats des autres chromosomes étant similaires. Les paramètres contraints de la loi d'émission gaussienne du modèle $\mathcal{M}_2$ sont donnés dans le tableau 5.2. Les matrices d'orientation D estimées pour les groupes bruit et identique sont égales, et très proches de la matrice d'orientation attendue pour une direction sur la première bissectrice. La deuxième valeur propre de la matrice de variance Σ est identique dans les quatre groupes. On note que la matrice de variance estimée du groupe bruit est proche d'une distribution gaussienne sphérique.

La matrice de transition estimée est

$$\hat{\pi} = \begin{pmatrix} 0.87 & 0.01 & 0.07 & 0.05 \\ 0.01 & 0.93 & 0.04 & 0.02 \\ 0.14 & 0.04 & 0.77 & 0.05 \\ 0.16 & 0.04 & 0.07 & 0.73 \end{pmatrix}.$$

On remarque que les probabilités de transition sont élevées sur la diagonale, ce qui signifie que la sonde $t+1$ a tendance à rester dans le même groupe que la sonde t.

	bruit	identique	appauvri	enrichi
$\hat{\mu}$	(7.6 ;7.6)	(12.9 ;12.7)	(9.3 ;8.1)	(8.3 ;9.3)
$\hat{D}$	$\begin{pmatrix} \mathbf{0.71} & \mathbf{-0.70} \\ \mathbf{0.70} & \mathbf{0.71} \end{pmatrix}$	$\begin{pmatrix} \mathbf{0.71} & \mathbf{-0.70} \\ \mathbf{0.70} & \mathbf{0.71} \end{pmatrix}$	$\begin{pmatrix} -0.92 & 0.39 \\ -0.39 & -0.92 \end{pmatrix}$	$\begin{pmatrix} 0.44 & -0.90 \\ 0.90 & 0.44 \end{pmatrix}$
$\hat{\Lambda}$	$\begin{pmatrix} 0.13 & 0 \\ 0 & \mathbf{0.16} \end{pmatrix}$	$\begin{pmatrix} 6.34 & 0 \\ 0 & \mathbf{0.16} \end{pmatrix}$	$\begin{pmatrix} 1.28 & 0 \\ 0 & \mathbf{0.16} \end{pmatrix}$	$\begin{pmatrix} 1.37 & 0 \\ 0 & \mathbf{0.16} \end{pmatrix}$
$\hat{\Sigma}$	$\begin{pmatrix} 0.15 & -0.01 \\ -0.01 & 0.15 \end{pmatrix}$	$\begin{pmatrix} 3.28 & 3.09 \\ 3.09 & 3.22 \end{pmatrix}$	$\begin{pmatrix} 1.11 & 0.4 \\ 0.4 & 0.33 \end{pmatrix}$	$\begin{pmatrix} 0.39 & 0.47 \\ 0.47 & 1.14 \end{pmatrix}$

Tab. 5.2: Estimation des paramètres de la loi d'émission gaussienne du modèle HMM $\mathcal{M}_2$.

Le graphe obtenu après classification est donné Figure 5.10. Les proportions de sondes dans chacun des groupes sont 43% pour le groupe bruit, 23% pour le groupe identique, 21% pour le groupe appauvri et 13% pour le groupe enrichi.

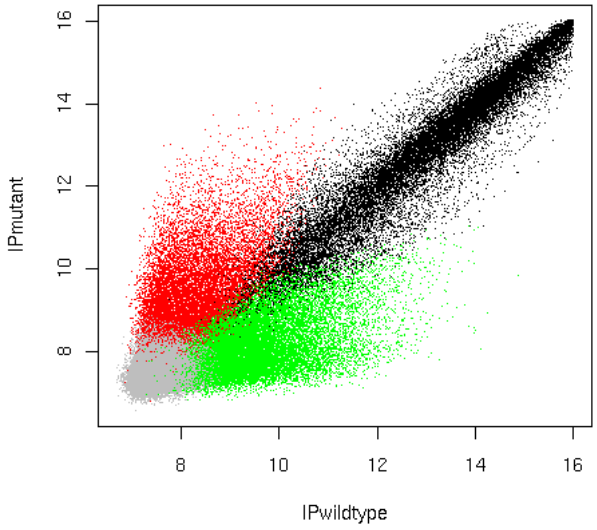

Fig. 5.10: Graphe obtenu avec le modèle HMM $\mathcal{M}_2$ après classification par la règle du MAP. Le groupe bruit est en gris, le groupe identique est en noir, le groupe appauvri est en vert et le groupe enrichi est en rouge.

La modification d'histone étudiée est une marque hétérochromatinienne. La plupart des régions couvertes par H3K9me2 sont contigües et couvrent plusieurs mégabases dans les régions péricentromériques ou dans l'hétérochromatine interstitielle (comme le knob sur le chromosome 4), mais il y a aussi des petites régions (îlots d'hétérochromatine) localisées dans l'euchromatine et qui couvrent principalement des éléments transposables (Bernatavichute *et al.*, 2008). Les résultats obtenus corroborent ces informations : 91.3% des sondes dans l'hétérochromatine sont méthylées alors que seulement 49.5% le sont dans l'euchromatine. Dans l'hétérochromatine, 82% des sondes ont un comportement identique

entre le sauvage et le mutant alors que seulement 9.5% des sondes sont identiques dans l'euchromatine. De plus, 56% des sondes méthylées couvrent un élément transposable ou une région environnante de 500 paires de bases (pb).

Les probabilités de transition du modèle HMM fournissent des indications sur la longueur des régions de chaque groupe par le biais du temps moyen de séjour théorique (distribution géométrique de la longueur des régions enrichies). La taille moyenne des sites de fixation est de 14.3 sondes (correspondant à 3 289 pb) pour le groupe identique, 4.5 sondes (1 035 pb) pour le groupe appauvri, 3.7 sondes (851 pb) pour le groupe enrichi et 7.7 sondes (1 771 pb) pour le groupe bruit. Les sites de fixation étant souvent des éléments transposables, ces estimations montrent que les éléments transposables enrichis ou appauvris sont trois fois plus petits que les éléments transposables qui ont un comportement identique entre le mutant et le sauvage. Ceci suggère que le mutant affecte principalement les éléments transposables de l'euchromatine qui sont généralement 2 à 3 fois plus petits que ceux de l'hétérochromatine.

L'élément transposable META1 (situé entre les positions 5 326 458 et 5 331 580 sur le chromosome 4) est connu pour avoir une perte de méthylation chez le mutant. La région régulatrice de META1 est située au début de l'élément transposable puisqu'on y trouve des petits ARN impliqués dans le processus de méthylation. Notre méthode déclare la première moitié des sondes couvrant META1 dans le groupe de perte de méthylation chez le mutant. Les autres sondes sont déclarées identiquement méthylées entre les deux échantillons (cf. Figure 5.11).

Fig. 5.11: Représentation de l'élément transposable META1 à l'aide du logiciel FLAGdb++ avec les résultats du modèle $\mathcal{M}_2$ (positions 5326458-5331580, chr4). Les sondes sont d'abord déclarées avec une perte de méthylation chez le mutant (vert), puis identiquement méthylées (noir).

3.2 Comparaison avec la méthode de Johannes *et al.* (2010)

Dans cette section, nous comparons les résultats présentés ci-dessus avec le modèle $\mathcal{M}_2$ à ceux obtenus par Johannes *et al.* (2010). Cette comparaison est réalisée sur le jeu de données réelles H3K9me2. Une étude de comparaison de méthodes est aussi effectuée Section 5 sur des données simulées.

Johannes *et al.* (2010) ont proposé deux modèles de mélange avec quatre gaussiennes bidimensionnelles contraintes sur les paramètres de moyenne. Pour simplifier les notations, on note 1, 2, 3, 4 les groupes correspondant respectivement aux groupes bruit, identique, appauvri et enrichi. Le premier modèle, appelé modèle "full-switching", fixe les contraintes suivantes sur les paramètres de moyenne : $\vec{\mu}_1 = (\mu_1, \mu_1)$, $\vec{\mu}_2 = (\mu_2, \mu_2)$, $\vec{\mu}_3 = (\mu_2, \mu_1)$, $\vec{\mu}_4 = (\mu_1, \mu_2)$ et les matrices de variance des groupes 3 et 4 sont égales ($\Sigma_3 = \Sigma_4$). Le second modèle, appelé "flexible-switching", correspond au modèle full-switching avec des contraintes moins restrictives sur les paramètres : $\vec{\mu}_3 = (\mu_4, \mu_3)$ et $\vec{\mu}_4 = (\mu_3, \mu_4)$.
Nous avons comparé le modèle HMM $\mathcal{M}_2$ avec ces deux modèles sur les données H3K9me2.

Le modèle full-switching conduit à une plus petite proportion de sondes différentiellement enrichies (7.8% appauvries chez le mutant et 1.2% enrichies chez le mutant, cf.

Figure 5.12) que le HMM (respectivement 22% et 14%). L'élément transposable META1, déclaré appauvri avec $\mathcal{M}_2$ (voir ci-dessus), est déclaré identiquement méthylé avec le modèle full-switching. La classification du modèle flexible-switching ne semble pas être appropriée pour les sondes ayant des intensités similaires entre 8 et 10 et qui devraient être déclarées identiquement méthylées (cf. Figure 5.12).
En comparant les résultats du modèle HMM $\mathcal{M}_2$ avec les résultats obtenus avec notre modèle le plus simple $\mathcal{M}_1$ (modèle de mélange, sans dépendance), nous notons que la classification correcte des sondes différentiellement enrichies est une conséquence des contraintes imposées sur les matrices de variance dans le modèle. L'avantage supplémentaire du HMM est de fournir une classification qui ne se fonde pas uniquement sur des considérations géométriques (position des X_t observés), mais aussi sur l'environnement de la sonde t (grâce à l'hypothèse de dépendance). Cependant, le modèle HMM $\mathcal{M}_2$ a tendance à déclarer différentiellement enrichies les sondes ayant des intensités similaires entre 8 et 10.
En conclusion, il semble que l'hypothèse d'indépendance, les contraintes symétriques sur les paramètres de moyennes et les variances égales pour les groupes 3 et 4 conduisent à un modèle trop simple pour analyser ces données. Une comparaison des méthodes est aussi réalisée sur des données simulées dans la Section 5.

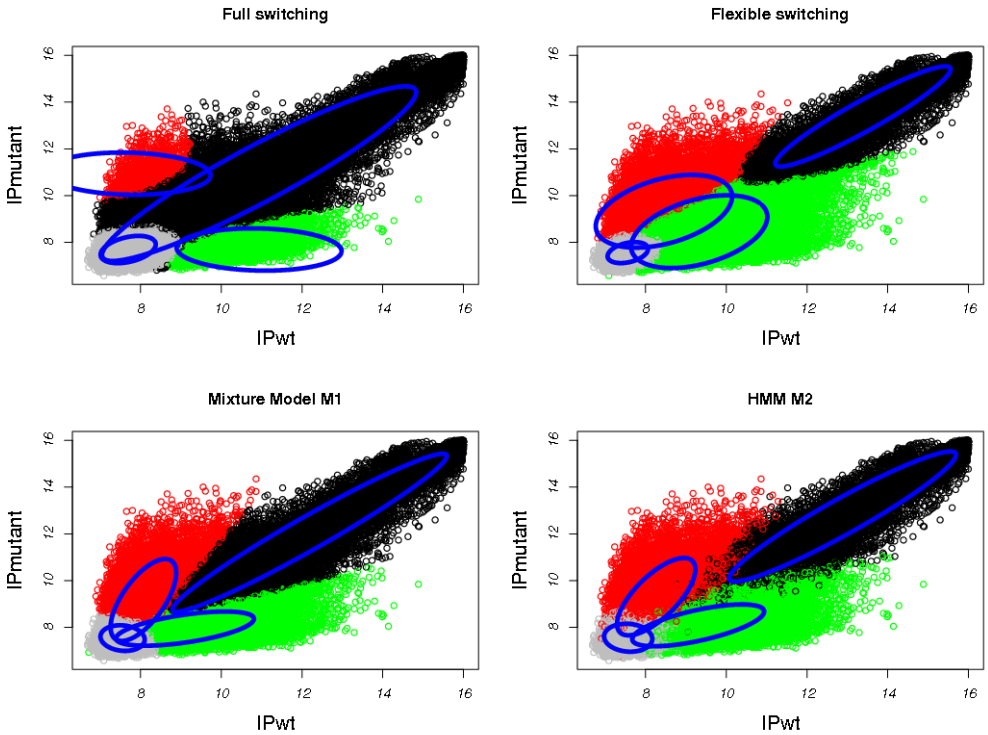

Fig. 5.12: Comparaison de la classification entre les deux modèles de Johannes *et al.* (2010) et le modèle de mélange et le HMM sur le jeu de données H3K9me2. Haut-gauche : modèle full-switching, Haut-droit : modèle flexible-switching, Bas-gauche : modèle de mélange $\mathcal{M}_1$, Bas-droit : HMM $\mathcal{M}_2$.

3.3 Analyse avec le modèle de mélanges de mélange

Les deux approches proposées Chapitre 3 Section 2.3, où la loi d'émission est un mélange de gaussiennes, ont aussi été appliquées sur le jeu de données H3K9me2.

a) Méthode avec contraintes de colinéarité

Pour la méthode avec contraintes de colinéarité (présentée Chapitre 3 Section 2.3), le nombre de groupes K est fixé à 4 et nous avons étudié plusieurs modèles en fonction du nombre de composants par groupe. Le choix des modèles est restreint au cas où l'on a un nombre de composants identique pour chaque groupe, entre deux composants par groupe (modèle noté 1222, où les chiffres correspondant respectivement au nombre de composants pour le groupe bruit, identique, appauvri et enrichi) et sept composants par groupe (noté 1777). Ces différents modèles ont été comparés à l'aide des critères de sélection présentés Chapitre 3 Section c) et les résultats sont donnés dans le tableau 5.3. Les critères BIC et ICL_Z sélectionnent le modèle à cinq composants par groupe.

	1222	1333	1444	1555	1666	1777
nbparam	69	123	207	303	417	549
-2*LogVrais	478477	477368	477373	477116	**477071**	477103
BIC	478895	477891	478000	**477849**	477908	478045
ICL_Z	508216	506181	506159	**505915**	505993	506176

Tab. 5.3: Comparaison des modèles en fonction du nombre de composants par groupe.

Les résultats obtenus après classification avec ce modèle à cinq composants par groupe sont exactement identiques à ceux obtenus par le modèle $\mathcal{M}_2$ présentés dans la Section 3.1. Après discussion avec les biologistes, l'hypothèse de variance résiduelle σ^2 identique pour tous les groupes ne semblait pas très appropriée pour ce type de données et nous avons donc modifié cette hypothèse en fonction des connaissances biologiques. Il s'avère que la variance résiduelle des groupes différentiellement hybridés peut être largement supérieure à celle des groupes bruit et identique, ce qui est vérifié sur la Figure 5.13 des distributions des données projetées sur le petit axe de chaque groupe.

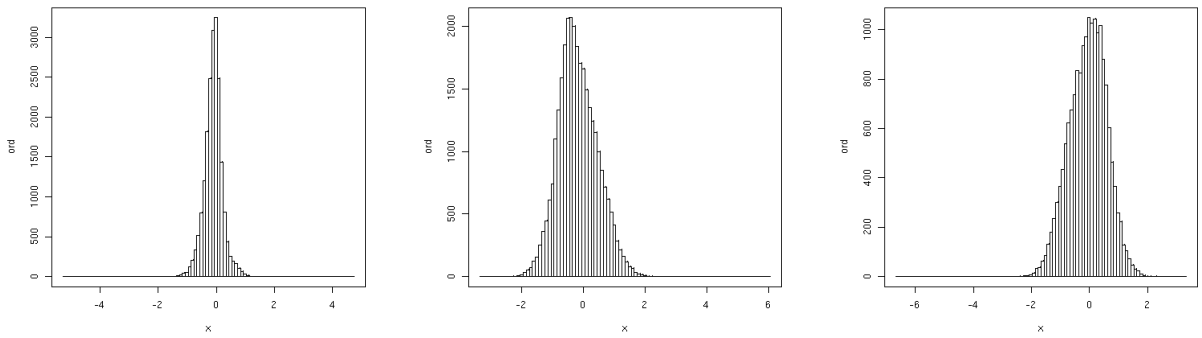

Fig. 5.13: Gauche : Distribution des données projetées sur le petit axe du groupe identique ; Centre : Distribution des données projetées sur le petit axe du groupe appauvri ; Droite : Distribution des données projetées sur le petit axe du groupe enrichi.

Nous avons donc introduit deux variances résiduelles différentes dans le modèle : σ_1^2 pour les groupes bruit et identique (notés 1 et 2) et σ_2^2 pour les groupes différentiellement hybridés (notés 3 et 4). De manière analogue au calcul de l'estimateur de σ^2 (Chapitre 3 Annexe 3.2), les estimateurs de σ_1^2 et σ_2^2 sont donnés par :

$$\hat{\sigma}_1^2 = \frac{\sum_t \hat{\tau}_{t1}(X_{1t} - \hat{\mu}_1^1)^2 + \sum_t \hat{\tau}_{t1}(X_{2t} - \hat{\mu}_1^2)^2 + \sum_{\ell=1}^{L_2} \sum_t \hat{\tau}_{t2}\hat{\delta}_{t2\ell}V_{t2}^2}{\sum_t \hat{\tau}_{t1} + \sum_t \hat{\tau}_{t1} + \sum_{\ell=1}^{L_2} \sum_t \hat{\tau}_{t2}\hat{\delta}_{t2\ell}} ,$$

et

$$\hat{\sigma}_2^2 = \frac{\sum_{k=3}^{4} \sum_{\ell=1}^{L_k} \sum_t \hat{\tau}_{tk}\hat{\delta}_{tk\ell}V_{tk}^2}{\sum_{k=3}^{4} \sum_{\ell=1}^{L_k} \sum_t \hat{\tau}_{tk}\hat{\delta}_{tk\ell}} .$$

Les critères BIC et ICL_Z sélectionnent cette fois-ci le modèle à six composants par groupe, noté 1666-$\sigma_1^2\sigma_2^2$ (cf. Table 5.4).

	1222-$\sigma_1^2\sigma_2^2$	1333-$\sigma_1^2\sigma_2^2$	1444-$\sigma_1^2\sigma_2^2$	1555-$\sigma_1^2\sigma_2^2$	1666-$\sigma_1^2\sigma_2^2$	1777-$\sigma_1^2\sigma_2^2$
nbparam	70	124	208	304	418	550
-2*LogVrais	455195	453149	452465	452223	452056	**451997**
BIC	455625	453683	453104	452967	**452904**	452950
ICL_Z	486285	482988	481662	481257	**481028**	481041

Tab. 5.4: Comparaison des modèles avec une variance résiduelle différente, en fonction du nombre de composants par groupe.

Les ajustements des densités estimées pour chaque groupe sont représentés sur les histogrammes des données projetées sur les grands axes de chaque groupe (cf. Figure 5.14).

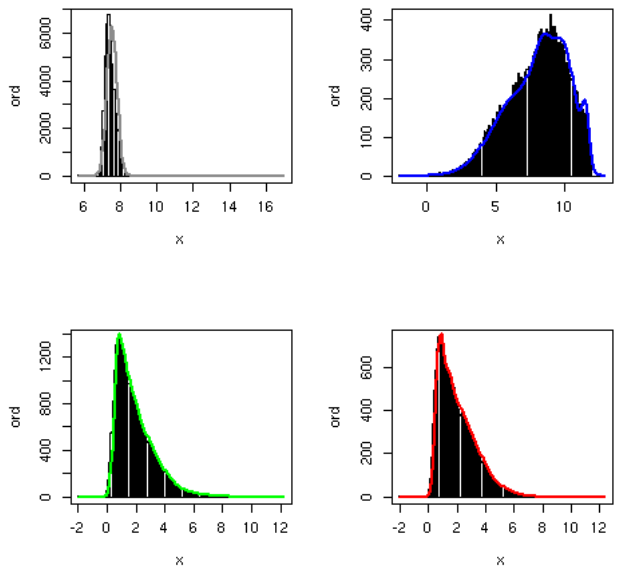

Fig. 5.14: Distribution des données et ajustement de la densité estimée pour chaque groupe (bruit en gris, identique en bleu, appauvri en vert et enrichi en rouge). Pour les groupes identique, appauvri et enrichi, les données sont projetées sur le grand axe du groupe correspondant.

Le graphe obtenu après classification avec le modèle 1666-$\sigma_1^2\sigma_2^2$ est présenté Figure 5.15. Les proportions de sondes dans chacun des groupes sont 40% pour le groupe bruit, 18% pour le groupe identique, 27% pour le groupe appauvri et 15% pour le groupe enrichi. En comparaison avec la Figure 5.10 obtenue par le modèle $\mathcal{M}_2$, on remarque que les frontières du groupe identique sont réduites. Il y a un plus grand nombre de sondes d'intensités élevées classées dans les groupes différentiellement hybridés. Par contre, les sondes d'intensités similaires entre 8 et 10 ont tendance à être déclarées différentiellement hybridées. En comparant les classifications à l'aide d'une table de contingence (cf. Table 5.5), on trouve environ 90% des sondes ayant le même statut entre les deux modèles. La plus grosse différence de classement intervient pour des sondes classées dans les groupes bruit et identique avec $\mathcal{M}_2$, et qui sont classées dans les groupes différentiellement hybridés avec 1666-$\sigma_1^2\sigma_2^2$. La Figure 5.16 illustre les différences de classements entre les deux modèles.

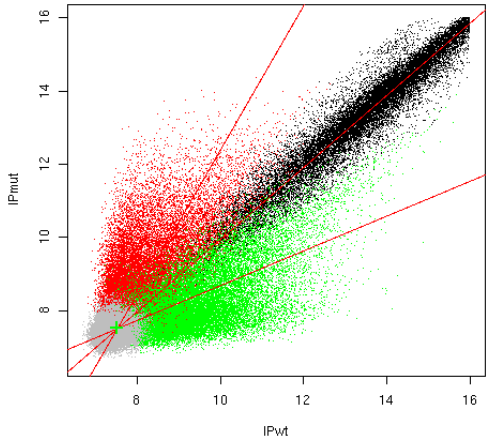

Fig. 5.15: Graphe obtenu avec le modèle sous contraintes de colinéarité avec une variance résiduelle différente et avec six composants par groupe (1666-$\sigma_1^2\sigma_2^2$), après classification par la règle du MAP. Le groupe bruit est en gris, le groupe identique est en noir, le groupe appauvri est en vert et le groupe enrichi est en rouge.

1666-$\sigma_1^2\sigma_2^2$ \ $\mathcal{M}_2$	Bruit	Ident.	Appauvri	Enrichi
Bruit	44308	6	329	113
Ident.	1	19906	3	0
Appauvri	2466	3558	23191	1381
Enrichi	1068	1930	545	12894

Tab. 5.5: Table de contingence entre le modèle HMM $\mathcal{M}_2$ et le modèle sous contraintes de colinéarité avec une variance résiduelle différente et avec six composants par groupe.

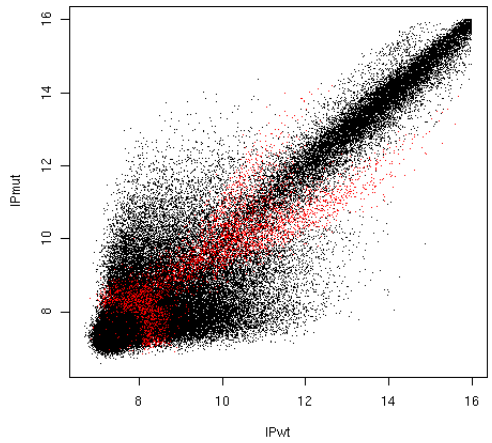

FIG. 5.16: Différence de classement entre les modèles $\mathcal{M}_2$ et 1666-$\sigma_1^2\sigma_2^2$. Les sondes différemment classées sont en rouge.

L'élement transposable META1 présenté Section 3.1 est schématisé Figure 5.17 avec les résultats du modèle 1666-$\sigma_1^2\sigma_2^2$. On remarque que cette fois-ci, la majorité des sondes est déclarée avec une perte de méthylation chez le mutant.

FIG. 5.17: Représentation de l'élément transposable META1 à l'aide des résultats du modèle 1666-$\sigma_1^2\sigma_2^2$. Les sondes sont majoritairement déclarées avec une perte de méthylation chez le mutant (vert).

b) Méthode sans contraintes de colinéarité

Nous avons aussi illustré la méthode sans contraintes (présentée Chapitre 3 Section 2.3) sur le jeu de données H3K9me2. En raison du temps de calcul élevé (qui est discuté en Conclusion et perspectives), nous appliquons notre méthode sur un échantillon spécifique de 5000 sondes du chromosome 4 d'*Arabidopsis thaliana*, entre les positions 5237473 et 6071023.

Nous choisissons arbitrairement le nombre initial de composants $L = 40$ et nous ajustons un HMM avec 40 composants sphériques (cf. Figure 5.18, Gauche) avant de dérouler l'algorithme hiérarchique avec ∇^1 comme critère d'appariement et ICL_Z comme critère de sélection.
La Figure 5.18 (Droite) représente la classification finale, où le nombre de groupes sélectionné par ICL_Z est 8. Cette classification à huit groupes est facilement interprétable et cohérente avec la connaissance biologique. Le groupe bruit est représenté en gris. Les quatre groupes sur la diagonale correspondent aux sondes identiquement hybridées dans les deux conditions. Ces sondes sont faiblement méthylées (groupe noir) ou fortement méthylées (groupe cyan). Le groupe enrichi est en rouge, tandis que les deux groupes verts (du côté droit de la diagonale) correspondent aux sondes appauvries en méthylation. Les proportions des groupes enrichi et appauvri sont respectivement 6.5% et 15.5%, ce

qui est très proche des proportions attendues par les biologistes (5% et 15%, cf. Section 3). Les frontières de classification sont mieux définies, en particulier au niveau des sondes d'intensités entre 8 et 10 qui sont majoritairement classées dans le groupe identique. D'autre part, notre méthode identifie 75% des sondes de l'élément transposable META1 comme étant appauvries (cf. Figure 5.19).

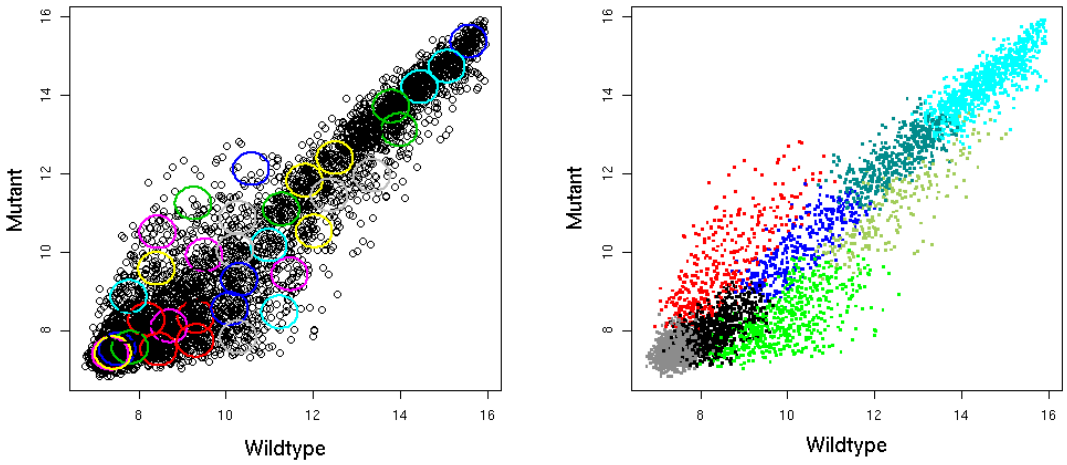

Fig. 5.18: Gauche : Représentation du HMM initial avec 40 composants. Droite : Classification obtenue après appariement des composants, chaque couleur représente un groupe spécifique.

Fig. 5.19: Représentation de l'élément transposable META1 à l'aide du logiciel FLAGdb++. Chaque sonde est colorée en fonction des probabilités *a posteriori* d'appartenir à chacun des groupes : groupe appauvri en vert

Les densités estimées par la méthode sans contraintes de colinéarité sont représentées pour chaque groupe sur la Figure 5.20. Les histogrammes ont été construits en projetant les données, pondérées par leur probabilité *a posteriori*, sur l'axe des abscisses. Les distributions empiriques ne sont clairement pas unimodales et considérer un mélange de distributions permet un meilleur ajustement par rapport à une unique gaussienne.
Notons toutefois que si l'on fixe arbitrairement le nombre de groupes à $K = 4$ (au lieu de 8 sélectionné par le critère ICL_Z), les quatre groupes obtenus par la méthode sans contraintes de colinéarité ne sont plus biologiquement interprétables.

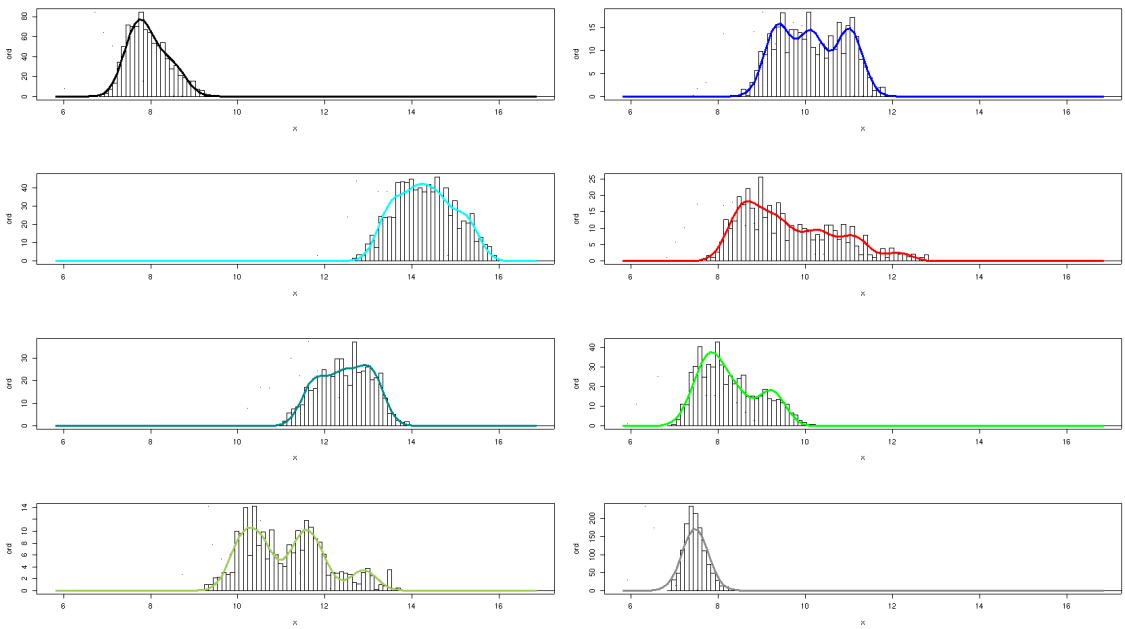

Fig. 5.20: Représentation de l'ajustement des densités pour chaque groupe. Les couleurs des densités estimées correspondent aux couleurs des groupes de la Figure 5.18.

4 Analyse de données transcriptome - Étude des données graine vs feuille

Lors d'une expérience de transcriptome, deux objectifs sont généralement considérés : l'étude de la différence d'expression des gènes entre deux conditions et la détection de régions transcrites.
Le jeu de données analysé permet d'étudier la différence d'expression des gènes entre les graines dix jours après pollinisation (10dap) et les feuilles d'*Arabidopsis thaliana*. L'objectif principal est de caractériser les conditions d'expression des gènes en classant les sondes en quatre groupes : un groupe où il n'y a pas d'hybridation (bruit), un groupe d'expression identique entre la feuille et la graine (identique), un groupe où l'expression est plus faible chez la graine (sous-exprimé) et un groupe où l'expression est plus forte chez la graine (sur-exprimé). Un autre objectif est de mettre en évidence de nouvelles régions transcrites.

Dans un premier temps, nous avons comparé les quatre sous-modèles présentés Chapitre 3 Section 1 (cf. Section 4.1), et nous présentons les résultats obtenus avec le modèle le plus pertinent (Section 4.2). Nous donnons ensuite quelques exemples de classification de gènes (cf. Section 4.3). La détection de nouvelles régions transcrites fait l'objet de la Section 4.4.

4.1 Comparaison de modèles

Nous comparons les quatre sous-modèles présentés Chapitre 3 Section 1 en termes de vraisemblance. Nous rappelons que le modèle le plus simple, sans annotation ni dépendance spatiale, est noté $\mathcal{M}_1$. Le modèle $\mathcal{M}_2$ correspond au HMM, $\mathcal{M}_3$ est le modèle avec annotation mais sans dépendance spatiale, et $\mathcal{M}_4$ correspond au modèle HMM avec annotation utilisant toute l'information disponible. La loi d'émission est une gaussienne

bidimensionnelle (modèle présenté Chapitre 3 Section 2.2), où les contraintes sur les matrices de variance sont valables pour les quatre sous-modèles.
Pour le modèle $\mathcal{M}_4$, nous avons considéré deux possibilités pour le nombre de catégories d'annotation :

- Soit $a = 7$ catégories d'annotation, correspondant à celles définies Section 1.2 : exonique, intronique, intergénique et quatre catégories supplémentaires qui représentent les sondes chevauchant un exon et un intron ou bien un exon et de l'intergénique (avec deux catégories différentes selon que la sonde couvre majoritairement l'exon, l'intron ou l'intergénique).
- Soit $a = 3$ catégories d'annotation : exonique, intronique et intergénique, où seuls les exons sont supposés être exprimés. Les sondes chevauchant plusieurs catégories d'annotation sont intégrées dans la catégorie pour laquelle le recouvrement est majoritaire.

Le tableau 5.6 présente l'ajustement des quatre modèles pour le chromosome 4. En comparant les modèles $\mathcal{M}_1$ et $\mathcal{M}_3$ ou les modèles $\mathcal{M}_2$ et $\mathcal{M}_4$, on remarque que la combinaison du HMM avec les informations d'annotation conduit à une réelle amélioration en termes de vraisemblance. Dans les deux cas, la minimisation du critère BIC est obtenue en ajoutant l'annotation au modèle. La connaissance de l'annotation est une information très utile dans un objectif de classification à cause de la différence intrinsèque des sondes exoniques ou intergéniques. La minimisation des critères BIC et ICL est atteinte pour le modèle complet $\mathcal{M}_4$, ce qui suggère que toutes les informations disponibles doivent être prises en compte. Les résultats obtenus pour le modèle $\mathcal{M}_4$ diffèrent très peu en fonction du nombre de catégories d'annotation. Le nombre de paramètres du modèle augmente logiquement avec la valeur de a, mais les valeurs des critères BIC et ICL sont similaires pour $a = 3$ et $a = 7$. Une des raisons est peut-être qu'il y a peu de sondes dans les quatre catégories chevauchantes ajoutées, en raison de la résolution de la puce *tiling array* utilisée.

	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$, $a = 3$	$\mathcal{M}_4$, $a = 3$	$\mathcal{M}_4$, $a = 7$
nombre de paramètres	19	31	25	61	121
-2 log-vraisemblance	406249	371309	373283	356617	**355949**
BIC	406469	371668	373573	**357323**	357350
ICL	436197	412706	399986	398272	**398186**

Tab. 5.6: Ajustement des 4 modèles. $\mathcal{M}_1$ = mélange, $\mathcal{M}_2$ = HMM, $\mathcal{M}_3$= mélange + annotation, $\mathcal{M}_4$ = HMM + annotation. La valeur de a correspond au nombre de catégories d'annotation.

En termes d'interprétation biologique, on remarque que le modèle HMM a tendance à lisser la classification des sondes. Lorsqu'une sonde isolée est déclarée exprimée dans une région intergénique en raison d'une intensité anormalement élevée, le lissage permet de corriger le statut de cette sonde en un statut non exprimé. Cependant, ce lissage n'est pas intéressant pour une sonde intronique isolée entre deux sondes exoniques exprimées (cf. Figure 5.21). L'ajout de l'annotation dans le modèle permet de corriger ce lissage. D'autre part, le modèle $\mathcal{M}_4$ (HMM et annotation) donne clairement des régions plus homogènes, qui sont plus faciles à interpréter d'un point de vue biologique (cf. Figure 5.22).

Dans la suite, nous utiliserons donc le modèle $\mathcal{M}_4$ pour analyser ces données d'expression. Le nombre de catégories d'annotation $a = 3$ est préféré pour faciliter les interprétations biologiques.

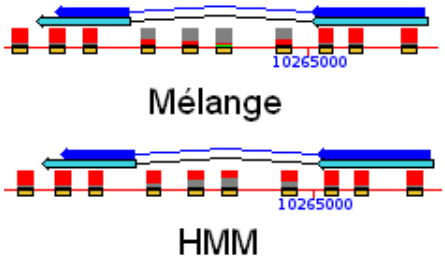

Fig. 5.21: Exemple d'un gène déclaré sur-exprimé, où les sondes introniques ont tendance à être lissées avec le modèle HMM (rouge=sur-exprimé, gris=non-exprimé).

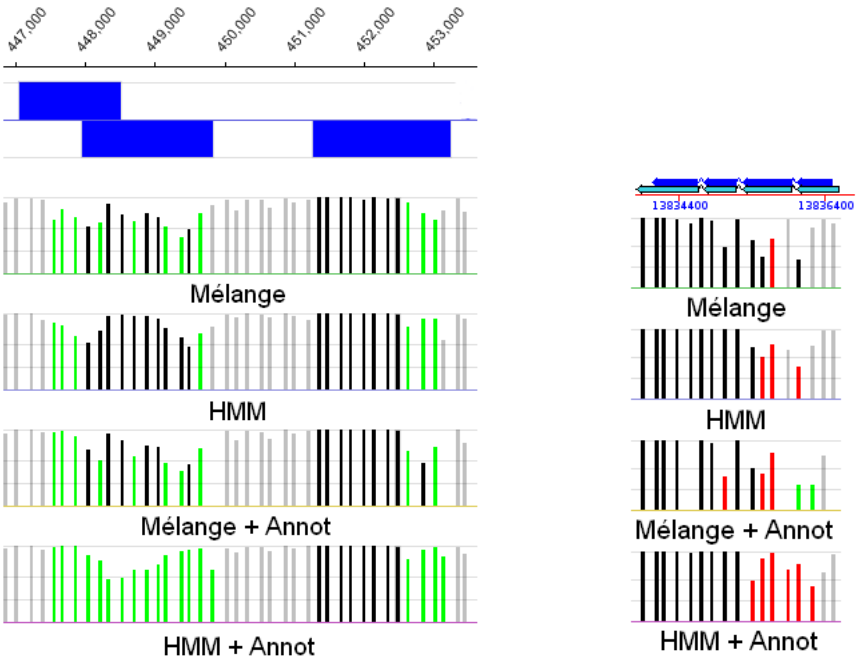

Fig. 5.22: Comparaison des quatre modèles sur deux régions où les gènes sont exprimés. Les régions déclarées exprimées par le modèle HMM+Annotation sont plus homogènes. Sur la région de droite, on remarque que tous les exons d'un même gène ne sont pas identifiés dans un même groupe.

La prise en compte de la structure spatiale des sondes nécessite d'avoir des sondes régulièrement réparties le long du chromosome. De nombreux articles suggèrent une réinitialisation de la chaîne de Markov lorsque la distance entre deux sondes voisines est trop grande, généralement à cause de régions non séquencées sur le génome (Ji et Wong, 2005 ; Li *et al.*, 2005). Lorsque l'annotation est intégrée dans le modèle HMM, un problème analogue est mis en évidence. En effet, la matrice de transition est différente pour chaque catégorie d'annotation et par convention, la matrice de transition utilisée correspond à celle de l'annotation de la sonde t. Si l'annotation change entre la sonde t et la sonde $t+1$, la probabilité de transition n'est plus valable, et ces changements sont fréquents le long du chromosome, en particulier entre les exons et les introns. Nous travaillons donc actuellement sur un modèle où une réinitialisation de la chaîne de Markov serait intégrée au modèle $\mathcal{M}_4$ à chaque changement d'annotation (à l'aide de la distribution stationnaire correspondante).

4.2 Analyse avec le modèle gaussien bidimensionnel $\mathcal{M}_4$

Comme nous l'avons vu dans la section précédente, les données ont été analysées avec le modèle $\mathcal{M}_4$ prenant en compte la dépendance spatiale des sondes et la connaissance de l'annotation. Trois catégories d'annotation fondamentales dans les expériences de transcriptome sont considérées : intergénique, intronique et exonique.
Nous présentons les résultats obtenus pour le chromosome 4, les résultats obtenus pour les autres chromosomes étant similaires. Les proportions estimées du modèle sont présentées dans le tableau 5.7. Ces résultats prouvent que les proportions dans chaque groupe dépendent fortement de la catégorie d'annotation. Raisonnablement, il y a une majorité de sondes intergéniques ou introniques non-exprimées (une discussion au sujet des sondes exprimées dans l'intergénique et l'intonique est proposée Section 4.4). Au contraire, 78% des sondes sont exprimées dans la catégorie exonique.

	Intergénique	Intronique	Exonique
Bruit	**84**	**60**	22
Ident.	1	7	**41**
Sous-exp	9	24	**23**
Sur-exp	6	9	**14**

Tab. 5.7: Proportions de sondes dans les quatre groupes pour chaque type d'annotation (en %)

Les matrices de transition des catégories intergénique et intronique sont similaires (cf. Table 5.8 et Table 5.9) : quel que soit le statut de la sonde t, la sonde $t+1$ a entre 70% et 95% de chance d'être dans le groupe bruit. Ceci est différent pour les sondes exoniques où la matrice de transition a de fortes probabilités sur la diagonale, ce qui signifie que la sonde $t+1$ a une forte probabilité (80 % à 90%) d'avoir le même statut que la sonde t (cf. Table 5.10).

	Bruit	Ident.	Sous-exp	Sur-exp
Bruit	**87**	1	7	5
Ident.	**95**	3	1	1
Sous-exp	**77**	1	**19**	3
Sur-exp	**75**	2	5	**18**

Tab. 5.8: Matrice de transition de la catégorie intergénique.

	Bruit	Ident.	Sous-exp	Sur-exp
Bruit	**87**	2	8	3
Ident.	**89**	0	1	10
Sous-exp	**55**	2	43	0
Sur-exp	**96**	1	0	3

Tab. 5.9: Matrice de transition de la catégorie intronique.

	Bruit	Ident.	Sous-exp	Sur-exp
Bruit	**83**	14	3	0
Ident.	2	**90**	6	2
Sous-exp	7	5	**87**	1
Sur-exp	8	6	1	**85**

Tab. 5.10: Matrice de transition de la catégorie exonique.

La Figure 5.23 montre que les sondes déclarées dans un même groupe sont regroupées en régions génomiques. Un gène est couvert par une majorité de sondes ayant le même statut excepté les sondes introniques qui sont raisonnablement déclarées non-exprimées. Tous ces résultats semblent cohérents et confirment les connaissances biologiques sur ces données d'expression.

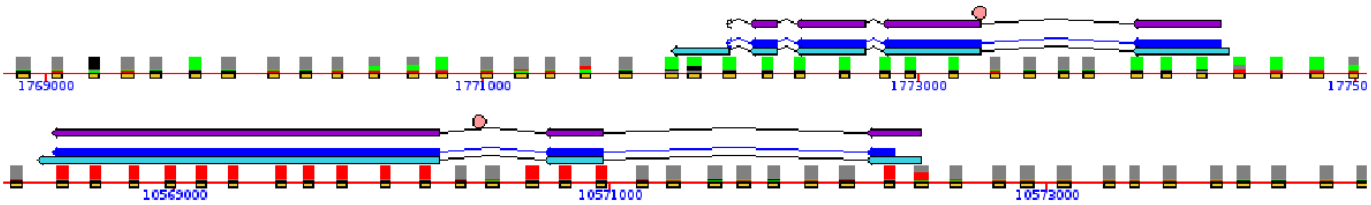

Fig. 5.23: Visualisation des résultats statistiques avec le logiciel FLAGdb++. Exemple de deux gènes, dont l'un est couvert par des sondes sous-exprimées (en vert) et l'autre est couvert par des sondes sur-exprimées (en rouge). Les flèches en bleu correspondent aux gènes, le traits fins entre les flèches correspondent aux introns.

4.3 Classification par gène

En utilisant la procédure de classification de régions donnée dans le Chapitre 4 Section 3, nous trouvons 81% de gènes ayant une valeur de critère unistatut supérieure à zéro (correspondant à 1 736 gènes). Parmi ces 1 736 gènes, 920 sont déclarés identiquement exprimés entre la graine et la feuille, 318 sont déclarés sous-exprimés chez la graine et 181 sont déclarés sur-exprimés chez la graine. À l'aide des sorties graphiques de Genevestigator, qui est une base de données de résultats d'analyse du transcriptome (Zimmermann *et al.*, 2004), huit gènes ont été clairement identifiés comme préférentiellement transcrits dans les graines. Parmi ces huit gènes, sept ont une valeur de critère unistatut supérieure à 0 et sont déclarés sur-exprimés dans les graines avec notre calcul.

Les exemples de trois gènes sont présentés dans le tableau 5.11. Pour chaque gène nous disposons de son identifiant, du nombre de sondes, du nombre de sondes exoniques, de l'annotation de chaque sonde, du statut défini pour chaque sonde, du critère unistatut et enfin des 4 probabilités *a posteriori* du gène.

Le gène AT4G00210 a une valeur unistatut positive et on peut définir son statut comme étant sur-exprimé. Cela corrobore avec les statuts obtenus par sonde. Le gène AT4G00390 a une valeur unistatut égale à -0.3, ce qui signifie que ce gène n'est pas homogène et il n'est donc pas classé. Le gène AT4G00110 a une valeur unistatut positive et on peut définir son statut comme étant identiquement exprimé dans les deux échantillons. Dans ce cas, la classification du gène donne un résultat nettement plus facile à interpréter que la classification obtenue pour chaque sonde.

ID	AT4G00210	AT4G00390	AT4G00110
nb probes	8	9	11
nb exonic probes	5	8	11
C_t	E.E.E.I.I.I.E.E	V.E.E.E.E.E.E.E.E	E.E.E.E.E.E.E.E.E.E.E
$\hat{Z}_t$	4.4.4.1.1.1.4.4	1.2.2.2.3.3.1.1.1	2.2.2.2.2.2.4.4.4.4.1
critère unistatut	**5.3**	**-0.3**	**3**
$\frac{Q_{gk,X}}{\sum_l Q_{gl,X}}$	0 ; 0 ; 0 ; **1**	0 ; 0.7 ; 0.1 ; 0.2	0 ; **1** ; 0 ; 0 ;

Tab. 5.11: Exemples de 2 gènes. Pour l'annotation, E=exon, I=intron, V=intergénique. Pour les statuts, 1=bruit, 2=identique, 3=sous-exprimé, 4=sur-exprimé.

4.4 Détection de nouveaux transcrits

Dans les sections précédentes, l'objectif principal était l'identification de sondes ou de gènes différentiellement exprimés. Mais un autre défi majeur de l'utilisation des puces *tiling arrays* est la détection de nouveaux transcrits (Jarvis et Robertson, 2011). En effet, la technologie des puces *tiling arrays* dont les sondes ont la particularité de couvrir l'intégralité du génome avec une haute résolution et indépendamment de l'annotation structurale rend possible l'étude exhaustive de l'activité transcriptionnelle d'un génome. Les données transcriptomiques permettent de mettre en évidence de nouvelles unités transcriptionnelles qui avaient échappé aux méthodes d'annotation classiques en raison de leur originalité structurale (petite taille, antisens, gènes à ARN, etc.). Bien que notre modèle soit construit sur la comparaison de deux échantillons, il permet aussi la détection de nouveaux sites de transcription. Une analyse exploratoire a permis de détecter des plages de sondes exprimées dans l'intergénique qui suggèrent clairement de nouveaux gènes. Par ailleurs, les résultats obtenus montrent étonnamment beaucoup de transcription dans les introns en 5'UTR (40% des sondes introniques sont déclarées exprimées dans le tableau 5.7, cf. Figure 5.24). Cela semble cohérent avec l'article récent de Cenik *et al.* (2010) qui suppose un rôle fonctionnel des introns courts en 5'UTR.

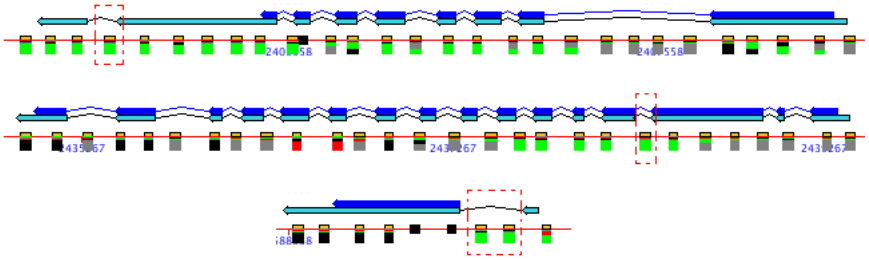

Fig. 5.24: Exemple de trois introns déclarés sous-exprimés avec une probabilité *a posteriori* supérieure à 0.75.

Si l'objectif principal de l'étude est la détection de nouvelles régions transcriptionnelles afin d'améliorer l'annotation existante, il s'avère logique de ne pas utiliser la connaissance de l'annotation *a priori*. Dans ce cas, le modèle HMM sans annotation $\mathcal{M}_2$ semble donc plus approprié. Notons toutefois que le modèle $\mathcal{M}_4$ avec prise en compte de l'annotation détecte aussi des régions exprimées dans l'intergénique, mais en moins grand nombre. Ces régions, détectées malgré la prise en compte de l'information intergénique sur les

sondes, sont supposées être vraiment authentiques, c'est pourquoi nous focalisons notre analyse dessus.

La difficulté est de généraliser cette étude au génome entier pour détecter des régions intéressantes de manière automatique. Il existe des tests statistiques fondés sur les statistiques de balayage (scan) pour évaluer la signification de clusters atypiques en utilisant une fenêtre de balayage (Glaz, Pozdnyakov et Wallenstein, 2009). Cependant, dans le cas général, l'expression exacte de la distribution de la statistique de scan n'est pas connue et des approximations basées sur des heuristiques sont proposées.

En collaboration avec Michel Koskas [1], nous avons mis au point un algorithme permettant une analyse exploratoire du génome. Cette approche n'assure aucune garantie statistique mais est algorithmiquement efficace, ce qui est essentiel pour l'analyse de données génomiques. Cet algorithme permet de détecter des régions intéressantes de manière automatique : dans le cas des expériences transcriptomiques, il a été utilisé pour rechercher des plages de sondes exprimées dans l'intergénique. Le programme est écrit en C++ et correspond à une recherche de somme partielle maximale ainsi que de sa localisation. Les sondes exprimées dans l'intergénique sont notées positivement et toutes les autres sondes sont notées négativement. L'algorithme est itératif : il détecte en premier la plus grande plage de sondes exprimées (celle de plus grande somme cumulée), puis de manière itérative il trouve toutes les autres jusqu'à obtenir des plages de longueur 1 (correspondant à une unique sonde).
Ce programme permet d'obtenir les positions des plages sur le chromosome d'une manière rapide, fiable (non subjective) et précise. La complexité en temps (pour trouver une plage) et en occupation mémoire sont toutes deux linéaires.
La difficulté est de choisir le coefficient positif ou négatif des sondes puisque cela modifie les sommes partielles cumulées et change donc radicalement la solution des plages proposées. En effet, plus il est petit plus les plages de sondes exprimées pourront être parsemées de sondes non exprimées. Dans cette analyse, nous avons arbitrairement choisi ce coefficient égal à 1.

Afin de réduire l'analyse, nous nous sommes intéressés uniquement aux régions exprimées couvrant plus de cinq sondes consécutives et détectées dans les deux réplicats biologiques, ce qui conduit à 143 régions sur tout le génome. Nous avons dans un premier temps comparé ces régions à l'annotation fournie par TAIR10[1]. Sur les 143 régions, 82 correspondent à des *other*RNA, *sn*RNA, *sno*RNA, *r*RNA, *t*RNA, *npc*RNA, pseudogène, gènes éléments transposables ou gènes codant pour une protéine, présents dans l'annotation officielle. L'annotation TAIR10 a donc validé 57% des régions détectées. Les 61 régions restantes (non annotées officiellement) ont été analysées individuellement, en explorant tous les indices de transcription répertoriés : les EST (Expressed Sequence Tags), où l'on repère ceux qui disposent d'un intron ou d'une queue polyA permettant de les orienter de manière non ambigüe, les MPSS mRNA qui indiquent une fin de transcription, les MIR (prédictions de microRNA) et les gènes prédits par le logiciel Eugène (Schiex *et al.*, 2001). On cherche aussi des similarités de régions chez d'autres organismes. On vérifie les régions qui ont un gène exprimé sur le brin complémentaire, signalant éventuellement une région régulatrice. D'autre part, on croise les régions avec les premières données RNAseq (feuilles).

[1]UMR AgroParisTech/INRA MIA 518

[1]Base de données de biologie génétique et moléculaire de la plante modèle *Arabidopsis thaliana*, `http://www.arabidopsis.org/`

Les Figures 5.25 et 5.26 présentent deux exemples de régions exprimées dans l'intergénique parmi les 61, avec les indices de transcription détectés. La première région correspond à une extension de gène, validée par un MPSS mRNA, un gène Eugène prédit et 23 EST dont 4 ont un intron confirmant l'orientation de la transcription. La seconde région correspond à une région identiquement exprimée, validée par un MPSS mRNA, 23 EST qui ont un intron confirmant l'orientation de la transcription. De plus, un gène sous-exprimé sur le brin complémentaire est adjacent à la région.

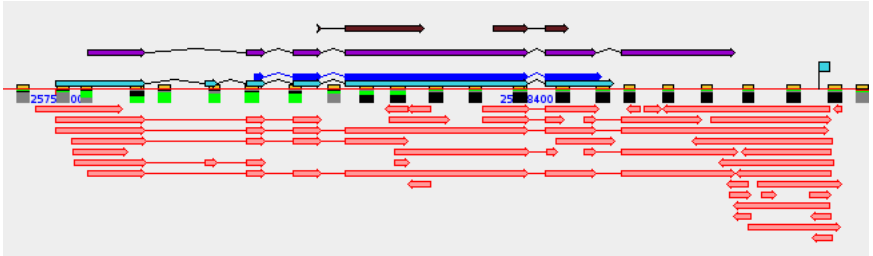

Fig. 5.25: Exemple d'une région correspondant à une extension de gène, validée par un MPSS mRNA (drapeau bleu), un gène Eugène prédit (flèche violette) et 23 EST (flèches roses) dont 4 ont un intron confirmant l'orientation de la transcription.

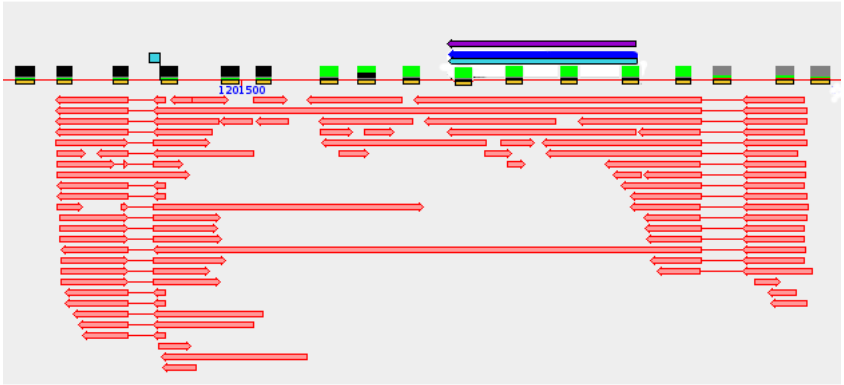

Fig. 5.26: Exemple d'une région correspondant à une région identiquement exprimée (sondes en noir), validée par un MPSS mRNA (drapeau bleu), 23 EST (flèches roses) qui ont un intron confirmant l'orientation de la transcription. Un gène sous-exprimé sur le brin complémentaire est adjacent à la région (sondes en vert).

4.5 Package TAHMMAnnot

Le package "TAHMMAnnot" propose les quatre différents modèles présentés Chapitre 3 Section 1 avec la loi d'émission présentée Chapitre 3 Section 2.2. Il est disponible sur le CRAN (`http://cran.r-project.org/`).

Le nombre de groupes K et le nombre de catégories d'annotation a sont fixés par l'utilisateur.
L'initialisation de l'algorithme d'estimation des paramètres peut se faire de manière aléatoire (obligatoire si $K \neq 3$ ou 4), ou en fixant des contraintes. Pour $K = 4$, l'utilisateur peut faire varier les proportions initiales de sondes dans les groupes bruit et identique, et les groupes sont définis par des contraintes géométriques à l'aide de deux droites parallèles et équidistantes de la bissectrice des axes. Pour $K = 3$, l'initialisation est fondée

sur l'identification de deux droites autour de la bissectrice des axes. L'utilisateur fixe une intensité maximum pour le groupe bruit, une coordonnée d'intersection des droites et un angle d'écartement entre les droites.

Les fichiers d'entrée comportent trois colonnes : l'identifiant de la sonde, l'intensité de la première condition et l'intensité de la seconde condition. Les modèles avec HMM nécessitent des fichiers ordonnés en entrée, et les modèles avec annotation nécessitent une quatrième colonne correspondant à la catégorie d'annotation de chaque sonde.
En sortie, le programme retourne l'estimation numérique des paramètres du modèle de mélange (proportion de sondes dans chaque groupe, moyennes et variances de chaque gaussienne, ainsi que la matrice de transition et la distribution stationnaire dans le cas des modèles HMM), un graphe avec les sondes colorées selon leur groupe d'appartenance, et un fichier texte comportant l'identifiant de la sonde, l'intensité de la première condition (Int1), l'intensité de la seconde condition (Int2), le log-ratio (Int1/Int2), les probabilités *a posteriori* d'appartenir à chacun des groupes, le statut de la sonde. La classification peut être réalisée à l'aide d'un seuil (fixé par l'utilisateur) de non classement des sondes incertaines.
Dans le package, des codes perl et R permettent de mettre en forme les données à partir des fichiers *pair* fournis par NimbleGen, ainsi que de réaliser au choix une normalisation lowess ou la normalisation ANOVA présentée Section 1.3.

Le package est utilisé dans l'équipe "Epigénétique et épigénomique végétale" (CNRS UMR 8197) dirigée par Vincent Colot à l'Institut de Biologie de l'Ecole Normale Supérieure, pour l'analyse de données génomiques de ChIP-chip IP/IP et de transcriptome, ainsi que sur la plateforme transcriptome de l'URGV (équipe génomique fonctionnelle d'Arabidopsis).
Les expériences transcriptomiques étudient l'expression différentielle entre des souches sauvages et des souches mutantes d'*Arabidposis thaliana*. En particulier, l'analyse du mutant *ddm1* donne des résultats intéressants avec une population identifiée de sondes surexprimées coïncidant clairement avec des éléments transposables.
Les expériences de ChIP-chip IP/IP s'intéressent à la comparaison de méthylation d'ADN ou de différentes marques d'histone entre des souches sauvages et des souches mutantes d'*Arabidposis thaliana*, ainsi qu'à la comparaison de lignées différentes d'epiRIL (qui ont le même génotype, mais un épigénotype différent). De nombreux jeux de données ont un nombre relativement faible de sondes différentiellement enrichies, avec une population appauvrie souvent sous-représentée. Dans ce cas, l'identification de la population appauvrie avec TAHMMAnnot n'est généralement pas satisfaisante. Le modèle avec $K = 3$ populations est plus adapté, mais ne permet pas de détecter les quelques sondes appauvries qui sont toutefois intéressantes d'un point de vue biologique.

5 Étude de simulations - Comparaison de méthodes

Afin de valider notre méthode, nous avons effectué une étude de simulations et comparé notre approche avec trois méthodes : ChIPOTle (Buck, Nobel et Lieb, 2005), la méthode de Johannes *et al.* (2010) et celle de Seifert *et al.* (2009). Comme il n'y a pas d'annotation dans les jeux de données simulés, nous avons appliqué le modèle $\mathcal{M}_2$ correspondant au HMM avec des contraintes sur les matrices de variance. Cette étude de simulation n'est pas spécifique aux expériences de ChIP-chip ou de transcriptome. Elle est réalisée dans un contexte général de comparaison de deux conditions mais le vocabulaire utilisé est celui des données d'expression.

Plan de simulation

Nous avons généré deux jeux de données de taille $n = 90\,000$ pour lesquels l'intérêt est l'identification des sondes différentiellement exprimées. Les jeux de données sont simulés avec une variable latente Z étant une chaîne de Markov d'ordre 1 qui prend ses valeurs dans $\{1, ..., 4\}$. La matrice de transition π et la distribution stationnaire m ont été ajustées sur les données réelles décrites ci-dessous.

Pour s'affranchir de l'hypothèse gaussienne, les observations X ont été échantillonnées selon une distribution empirique dans chacun des quatre groupes d'un jeu de données réelles, afin d'être semblables à des données *tiling array* réalistes. Le rééchantillonnage est réalisé en utilisant les probabilités *a posteriori* en tant que poids pour chaque sonde. Deux jeux de données réelles ont été choisis avec différentes proportions de sondes différentiellement exprimées. Le premier jeu de données, présenté à la Section 3, concerne l'étude d'une marque chromatinienne (H3K9me2) chez *Arabidopsis thaliana* pour un sauvage et un mutant. Dans ce jeu de données, environ 30% de sondes sont censées être différentiellement exprimées. Le second jeu de données est un jeu de données de ChIP-chip en libre accès, issu de Penterman *et al.* (2007), qui compare le profil de méthylation d'un type sauvage d'*Arabidopsis* à celui d'un triple mutant. Cette étude conduit à de faibles proportions de sondes différentiellement exprimées.

Nous avons analysé les jeux de données simulés avec le modèle $\mathcal{M}_2$ et avec trois autres méthodes.

- ChIPOTle est une méthode dédiée à la détection de pics dans les expériences classiques de ChIP-chip. Par conséquent, elle ne fournit que deux populations. Elle utilise une approche par fenêtre glissante fondée sur le log-ratio. La taille de la fenêtre et le pas sont les deux paramètres à régler. Avec les paramètres par défaut, ChIPOTle ne détecte qu'un seul pic. Nous prenons les paramètres taille de fenêtre=200 et pas=50, ce qui semble être une bonne combinaison donnant un nombre raisonnable de pics pour chaque jeu de données simulées. Nous avons utilisé la valeur absolue du log-ratio pour imiter les situations habituellement analysées.
- Seifert *et al.* (2009) ont proposé un HMM à trois états modélisant les log-ratios des deux intensités. Cette méthode requiert l'intégration de connaissances *a priori* en utilisant les distributions *a priori*. Le choix des priors n'est pas facile et ceux donnés par défaut ne fournissent pas trois populations. C'est pourquoi nous les avons modifiés. Nous posons startDistribution = (0.1,0.7,0.2), means = (-1,0.0,1), stds = (0.3,1,0.5), scaleOfAprioriMeans = (0.1,1,75) et shapeOfStandardDeviations = (20, 1 ; 100).
- Johannes *et al.* (2010) ont proposé deux modèles de mélange de quatre gaussiennes bidimensionnelles avec des contraintes sur les paramètres de moyennes pour l'analyse simultanée de deux échantillons. Ces deux modèles sont décrits plus précisément dans la Section 3.2.

Dans les jeux de données simulés, le but est toujours d'identifier les quatre groupes. La classification est réalisée avec la règle du MAP pour notre modèle $\mathcal{M}_2$ et le modèle de Johannes *et al.* (2010). Toutefois, ChIPOTle et la méthode de Seifert *et al.* (2009) ne fournissent pas quatre groupes : les groupes bruit et identique sont fusionnés. La méthode de Seifert *et al.* (2009) applique l'algorithme de Viterbi pour déterminer la séquence d'états la plus probable et donner une classification des sondes en trois groupes. Pour ChIPOTle, les sondes différentiellement exprimées sont déduites des pics détectés et il y a seulement deux groupes. Pour comparer notre méthode avec celles de Seifert *et al.* (2009) et ChIPOTle, nous avons additionné les probabilités *a posteriori* des groupes bruit et identique pour obtenir une classification en trois groupes, et aussi les probabilités *a*

posteriori des groupes sur-exprimés et sous-exprimés pour obtenir une classification en deux groupes.

Les méthodes sont comparées en utilisant les résultats de classification en termes de sensibilité, de spécificité et False Discovery Rate (FDR) pour un groupe donné k. La sensibilité est définie comme $\frac{TP_k}{TP_k+FN_k}$, la spécificité est définie comme $\frac{TN_k}{TN_k+FP_k}$, et le FDR est défini comme $\frac{FP_k}{TP_k+FP_k}$, où :

$$TP_k = \sum_t \mathbb{1}_{(\hat{Z}_t=k)}\mathbb{1}_{(Z_t=k)} \qquad FN_k = \sum_t \mathbb{1}_{(\hat{Z}_t=k)}\mathbb{1}_{(Z_t\neq k)}$$
$$FP_k = \sum_t \mathbb{1}_{(\hat{Z}_t\neq k)}\mathbb{1}_{(Z_t=k)} \qquad TN_k = \sum_t \mathbb{1}_{(\hat{Z}_t\neq k)}\mathbb{1}_{(Z_t\neq k)}$$

Nous nous concentrons sur les sondes déclarées différentiellement exprimées. La sensibilité et la spécificité doivent être grandes alors que le FDR est souhaité petit.

Résultats

Les résultats sont présentés dans les tableaux 5.12 et 5.13.
Les pics détectés par la méthode ChIPOTle représentent seulement entre 31% et 35% de sondes différentiellement exprimées pour les deux jeux de données. Le modèle "flexible-switching" de Johannes *et al.* (2010) fournit de meilleurs résultats que le modèle "full-switching", nous nous concentrons donc sur le modèle flexible-switching. Dans le premier jeu de données, le modèle flexible-switching de Johannes *et al.* (2010) et la méthode de Seifert *et al.* (2009) ont un comportement similaire. Ils ont trouvé respectivement 85% ou 82% de sondes sous-exprimées et 63% ou 72% de sondes sur-exprimées.

Dans le second jeu de données où peu de sondes sont différentiellement exprimées, les deux méthodes se comportent différemment. La méthode de Seifert *et al.* (2009) semble avoir des difficultés à trouver le groupe sur-exprimé (seulement 55% de sondes détectées). Au contraire, le modèle flexible-switching de Johannes *et al.* (2010) trouve 100% de sondes différentiellement exprimées, mais détecte du même coup un grand nombre de faux-positifs (FDR = 99%). Environ 40 000 sondes avec des intensités similaires entre 8 et 10 sont déclarées différentiellement exprimées alors qu'elles devraient être déclarées dans les groupes identique ou bruit. Dans les deux jeux de données simulés, notre modèle $\mathcal{M}_2$ identifie plus de 83% des sondes différentiellement exprimées, avec moins de faux-positifs (FDR entre 2% et 13%). Parmi les quatre méthodes, le modèle $\mathcal{M}_2$ a fourni les meilleurs triplets de sensibilité, spécificité et FDR, quel que soit le nombre de sondes différentiellement exprimées dans le jeu de données.
Afin de s'affranchir de l'hypothèse de dépendance markovienne, nous avons également simulé deux autres jeux de données où le chemin caché Z a été échantillonné dans un processus de saut à quatre états, avec des transitions markoviennes entre état et des temps de séjour suivant la loi Binomiale Négative (au lieu de géométrique). Des résultats similaires sont obtenus. Les triplés de sensibilité, spécificité et FDR obtenus avec le modèle $\mathcal{M}_2$ sont respectivement (92,99,4) et (91,99,5) pour les deux groupes différentiellement exprimés dans le premier jeu de données, (99,100,3) et (85,100,16) dans le deuxième jeu de données. Cela garantit que notre modèle n'est pas trop dépendant de l'hypothèse markovienne.

	bruit	identique	sous-exp	sur-exp
Z	38135	19527	19413	12925
ChIPOTle	99, 35, 27		35, 99, 6	
$\mathcal{M}_2$ with 2 groups	97, 94, 4		94, 94, 5	
Seifert *et al.*	92, 79, 11		82, 95, 18	72, 98, 13
$\mathcal{M}_2$ with 3 groups	98, 93, 4		92, 99, 6	90, 99, 6
Johannes *et al.*	94, 90, 12	82, 100, 0.1	85, 89, 31	63, 99, 11
TAHMMAnnot $\mathcal{M}_2$	96, 96, 5	99, 100, 1	92, 98, 6	90, 99, 6

Tab. 5.12: Jeu de données H3K9me2 avec une forte proportion de sondes différentiellement exprimées. La notation correspond dans l'ordre à sensibilité, spécificité et FDR, en %.

	bruit	identique	sous-exp	sur-exp
Z	45300	43401	782	517
ChIPOTle	100, 31, 1		31, 100, 0	
$\mathcal{M}_2$ with 2 groups	100, 91, 0.1		91, 100, 6	
Seifert *et al.*	98, 99, 0		100, 99, 48	55, 98, 84
$\mathcal{M}_2$ with 3 groups	100, 91, 0.1		97, 100, 2	82, 100, 12
Johannes *et al.*	23, 96, 13	74, 79, 12	100, 100, 31	100, 56, 99
TAHMMAnnot $\mathcal{M}_2$	85, 84, 16	83, 85, 16	97, 100, 2	83, 100, 13

Tab. 5.13: Jeu de données issu de Penterman *et al.* (2007) avec une faible proportion de sondes différentiellement exprimées. La notation correspond dans l'ordre à sensibilité, spécificité et FDR, en %.

Bibliographie

Bérard, C., Martin-Magniette, M-L. and Robin, S. (2011) Mixture model approach to compare two samples of tiling array data : ChIP-chip and Transcriptome. *Statistical Applications in Genetics and Molecular Biology* **10**, Iss. 1, Article 50.

Bernatavichute, Y.V., Zhang, X., Cokus, S., Pellegrini, M. and Jacobsen, S.E. (2008). Genome-Wide Association of Histone H3 Lysine Nine Methylation with CHG DNA Methylation in Arabidopsis thaliana. *PLoS ONE* **3**(9) :e3156.

Boulicaut, J.F. and Gandrillon, O. (2004). Informatique pour l'analyse du transcriptome. Chapitre Techniques statistiques pour l'analyse du transcriptome. *Hermès.*

Buck, M.J., Nobel, A.B. and Lieb, J.D. (2005). ChIPOTle : a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol.* **6**(11).

Cenik, C., Derti, A., Mellor, J.C., Berriz, G.F. and Roth, F.P. (2010). Genome-wide functional analysis of human 5'untranslated region introns. *Genome Biology* **11** :R29.

Dèrozier, S., Samson, F., Tamby, J.P., Guichard, C., Brunaud, V., Grevet, P., Gagnot, S., Label, P., Leplé, J.C., Lecharny, A. and Aubourg, S. (2011). Exploration of plant genomes in the FLAGdb++ environment. *Plant Methods* **7**(1) :8.

Glaz, J., Pozdnyakov, V. and Wallenstein, S. (2009). Scan Statistics Methods and Applications. *Statistics for Industry and Technology, Birkhauser.*

Humburg, P., Bulger, D. and Stone, G. (2008). Parameter estimation for robust HMM analysis of ChIP-chip data. *BMC Bioinformatics* **9** :343.

Jarvis, K. and Robertson, M. (2011). The noncoding universe. *BMC Biology* **9** :52.

Ji, H. and Wong, W.H. (2005). TileMap : create chromosomal map of tiling array hybridizations. *Bioinformatics* **21**, 3629-3636.

Johannes, F., Wardenaar, R., Colomé-Tatché M. *et al.* (2010). Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. *Bioinformatics* **26** 1000-1006.

Kerr, M.K., Afshari, C.A., Bennett, L., Bushel, P., Martinez, J., Walker, N.J. and Churchill, G.A. (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* **12** 203-217.

Li, W., Meyer, A. and Liu, X.S. (2005). A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* **211**, 274-282.

Martin-Magniette, M.L., Mary-Huard, T., Bérard, C. and Robin, S. (2008). ChIPmix : mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics* **24** :i181-i186.

Mary-Huard, T., Martin-Magniette, M.L., Bérard, C. and Robin, S. (2010). Statistical methodology for the analysis of multi-sample ChIP-chip experiments. *International Biometric Conference* Florianopolis (Brésil).

Moghaddam, A.M.B., Roudier, F., Seifert, M., Bérard, C., Martin-Magniette, M.L., Ashtiyani, R.K., Houben, A., Colot, V. and Mette, M.F. (2011). Additive inheritance of histone modifications in Arabidopsis thaliana intra-specific hybrids. *The Plant Journal.*

Oh, S., Park, S. and van Nocker, S. (2008). Genic and global functions for Paf1C in chromatin modification and gene expression in Arabidopsis. *PLoS Genet* **4** :e1000077.

Penterman, J., Zilberman, D., Huh, J.H., Ballinger, T., Henikoff, S. and Fischer R.L. (2007). DNA demethylation in the Arabidopsis genome. *PNAS* **104** 6752-6757.

Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T. *et al.* (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO Journal* **30** 1928-1938.

Schiex, T., Moisan, A. and Rouzé, P. (2001). EuGene : An Eucaryotic Gene Finder that combines several sources of evidence. *Computational Biology*, Eds. O. Gascuel and M-F. Sagot, LNCS 2066, 111-125.

Schwarz, G. (1978). Estimating the number of components in a finite mixture model. *Annals of Statistics* **6** 461-464.

Seifert, M., Banaei, A., Keilwagen, J., Mette, M.F., Houben, A., Roudier, F., Colot, V., Grosse, I. and Strickert, M. (2009). Array-based Genome comparison of Arabidopsis ecotypes using Hidden Markov models. *Biosignals.*

Smyth, G.K., Yang, Y.H. and Speed, T.P. (2003). Statistical issues in cDNA microarray data analysis. *Methods in Molecular Biology* **224** :111-136.

Thareau, V., Déhais, P., Serizet, C., Hilson, P., Rouzé, P. and Aubourg, S. (2003). Automatic design of gene-specific sequence tags for genome-wide functional studies. *Bioinformatics* **19(17)** :2191-8.

Turck, F., Roudier, F., Farrona, S., Martin-Magniette, M-L., Guillaume, E., Buisine, N., Gagnot, S., Martienssen, R., Coupland, G. and Colot, V. (2007). Arabidopsis TFL2/LHP1 Specifically Associates with Genes Marked by Trimethylation of Histone H3 Lysine 27. *PLoS Genet.v* **3(6)**.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L. and Gruissem, W. (2004). GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiology* **136**(1), 2621-2632.

Conclusion et perspectives

Dans cette thèse, nous nous sommes intéressés à la modélisation de données génomiques issues de technologies haut-débit. L'approche adoptée est celle des modèles à variables latentes, en particulier des modèles de mélange et des modèles de Markov cachés, qui sont des méthodes usuelles de classification non supervisée. Pour obtenir une analyse pertinente des données, il est essentiel que le modèle saisisse au mieux les caractéristiques biologiques et tienne compte de la spécificité des données. Dans le cas de puces *tiling arrays* deux couleurs, le signal est bidimensionnel, les observations apparaissent de manière séquentielle puisque les sondes sont régulièrement réparties le long du génome, et l'annotation structurale du génome informe de la prédisposition des sondes à être exprimées. Toutes ces caractéristiques biologiques sont intégrées dans la modélisation proposée. D'autre part, les modèles sont adaptés en fonction de la question biologique et une modélisation est proposée pour chaque type d'expériences. Nous proposons un mélange de régressions pour la comparaison de deux échantillons dont l'un peut être considéré comme un échantillon de référence, et un mélange gaussien bidimensionnel lorsque les deux échantillons jouent des rôles symétriques. Enfin, une modélisation semi-paramétrique autorisant des mélanges de distributions plus flexibles qu'une distribution gaussienne est envisagée pour mieux s'ajuster aux données réelles. La seconde contribution de ce travail est apportée en termes de classification des observations. Dans le cas d'observations indépendantes et pour une classification à deux groupes, nous proposons un contrôle de faux-positifs qui permet de contrôler les erreurs de classification. Puis, sous hypothèse de dépendance markovienne, nous nous intéressons à la classification d'un ensemble d'observations constituant une région d'intérêt en généralisant la formule des probabilités *a posteriori*.

Comme nous en avons déjà discuté Chapitre 4 Section 2, une perspective intéressante serait d'étendre le contrôle de faux-positifs quand il y a plus que deux groupes dans la classification, et aussi pour les résultats de classification par région.
D'autre part, le travail de modélisation semi-paramétrique avec des mélanges de mélanges de distributions (Chapitre 3 Section 2.3) est très prometteur au regard de l'estimation de la densité des groupes et de la classification obtenues sur quelques exemples d'applications (cf. Chapitre 5 Section 3.3). L'intégration de l'annotation au modèle améliorerait encore certainement les résultats. Cependant, cette méthode, fondée sur la combinaison hiérarchique de composants à l'aide de critères de vraisemblance pénalisée, n'est applicable en temps raisonnable que pour des jeux de données de l'ordre du millier d'observations, ce qui est environ cent fois moins que le nombre d'observations issues d'une expérience de *tiling array*. Cela est dû au fait que les critères impliquent le calcul de la vraisemblance observée nécessitant l'implémentation de l'étape forward de l'algorithme forward-backward, dont la complexité est linéaire en le nombre d'observations. En collaboration avec Stevenn Volant, nous travaillons actuellement sur l'implémentation d'une étape d'élagage permettant de réduire l'espace des modèles à explorer dans la procédure hiérarchique. Ceci peut être fait en utilisant un critère d'élagage qui permet de

considérer uniquement les composants à associer les plus pertinents. Les critères usuels tels que la distance euclidienne ne sont pas adpatés au contexte des HMM car ils ne tiennent pas compte de la dépendance dans les données. Il faut définir un critère qui prend en compte à la fois la distance entre les composants et la dépendance markovienne des observations.

Les méthodes proposées dans cette thèse ont été développées dans l'objectif de répondre à des questions biologiques concrètes, ce qui a suscité de nombreuses collaborations avec les biologistes et bioinformaticiens, avant, pendant et après la modélisation. Cette interaction est essentielle et contribue à une meilleure modélisation, adaptée aux différents types de données. La validation et l'exploitation des méthodes requièrent également des compétences multi-disciplinaires. La finalité de mon travail est de proposer aux biologistes un outil automatique efficace permettant d'analyser leurs données et de fournir des résultats biologiquement interprétables. Ceci a été réalisé sous forme de packages R et grâce à l'outil de visualisation des résultats statistiques intégré à la base de données FLAGdb++.

Dans une perspective d'amélioration des outils mis à disposition des utilisateurs biologistes, la gestion simultanée de plusieurs réplicats biologiques est importante. Cette adaptation a déjà été réalisée pour la méthode ChIPmix de mélange de régressions (avec MultiChIPmix, cf. Chapitre 3 Section 2.1) et peut également être prise en compte pour le mélange gaussien bidimensionnel.
D'autre part, le modèle gaussien bidimensionnel pourrait être généralisé à la comparaison de $d > 2$ échantillons. Le nombre de paramètres, linéaire en d et quadratique en le nombre de groupes K, reste encore très faible par rapport au nombre d'observations. Toutefois, la difficulté majeure repose sur les contraintes géométriques dans le modèle (Chapitre 3 Section 2.2) qui devront être redéfinies en dimension supérieure.
Par ailleurs, nous souhaitons nous intéresser au problème, intrinsèque au modèle de mélange, de la caractérisation d'un groupe qui ne contient que peu d'observations. En effet, dans certaines expériences, une proportion très faible de sondes est attendue préférentiellement hybridée dans une condition et le groupe correspondant n'est alors pas identifié convenablement. Le modèle considérant un groupe de moins ($K = 3$ au lieu de $K = 4$) n'est pas satisfaisant car c'est justement cette population qui intéresse le biologiste. Un calcul de distance de Mahalanobis ou un test d'appartenance des sondes aux groupes pourra être envisagé.

Les technologies de séquençage à haut-débit sont en constante évolution depuis l'apparition des puces à ADN. La densité des *tiling arrays* ne cesse d'augmenter pour atteindre quasiment deux millions de sondes aujourd'hui. La quantité de données grandissante a impliqué un fort investissement en termes d'analyse. Alors qu'en 2004 de nombreux articles avaient souligné la nécessité de mettre au point des procédures statistiques efficaces (Buck et Lieb, 2004), il existe aujourd'hui beaucoup de méthodes pour analyser les données *tiling arrays*. Les méthodes sont variées, mais les modèles à variables latentes semblent de plus en plus faire consensus.
Parallèlement, depuis deux ou trois ans, les technologies de séquençage ont connu une véritable révolution grâce au développement d'outils pour le séquençage parallèle massif des molécules d'ADN, qui fournissent une résolution de l'ordre du nucléotide. La technologie appelée Next Generation Sequencing (NGS) produit plusieurs gigabases de séquences nucléotidiques en quelques jours seulement, et sans utilisation d'un support. Cette technique a été étendue à l'analyse du transcriptome (RNA-Seq) et des interactions

protéines-ADN (ChIP-Seq). Avec un accès direct à la séquence, le RNA-Seq permet d'analyser le transcriptome de manière plus fine, avec une très grande sensibilité de détection et une gamme dynamique d'expression plus large que pour les *tiling arrays*. Cependant, comme toute nouvelle technologie, la technologie NGS reste pour l'instant coûteuse (séquencer un échantillon coûte environ $1000, soit dix fois plus qu'une puce) et souffre de biais techniques incontrôlés (Oshlack, Robinson et Young, 2010). Elle soulève de nouvelles questions comme l'assemblage des génomes ou le *read mapping*, ainsi que comment choisir une profondeur de séquençage suffisante en fonction de l'analyse (c'est-à-dire combien de fois l'échantillon doit être séquencé). De plus, alors que le signal d'une puce *tiling array* est défini comme une valeur d'intensité pour chaque sonde, le signal des données NGS est défini comme un comptage du nombre de *reads* (courts fragments d'ADN) chevauchant chaque paire de base. Ce changement de la nature des données nécessite le développement de stratégies d'analyse appropriées et pose des problèmes algorithmiques. Les modèles proposés dans cette thèse ne sont plus adaptés, le choix de la loi d'émission devra par exemple être modifié et les algorithmes devront être plus efficaces pour analyser ces données à très haute densité. Néanmoins, l'expérience acquise sur les puces *tiling arrays* facilitera la transition vers l'analyse de données NGS car la démarche générale de modélisation reste la même.

Annexe A

Article publié dans *Bioinformatics*

BIOINFORMATICS

Vol. 24 ECCB 2008, pages i181–i186
doi:10.1093/bioinformatics/btn280

ChIPmix: mixture model of regressions for two-color ChIP–chip analysis

Marie-Laure Martin-Magniette[1,2,*,†], Tristan Mary-Huard[1,†], Caroline Bérard[1,2] and Stéphane Robin[1]

[1]UMR AgroParisTech/INRA MIA 518, 16 rue Claude Bernard, 75231 Paris Cedex 05 and [2]URGV UMR INRA/CNRS/UEVE, 2 rue Gaston Crémieux, CP5708, 91057, Evry Cedex, France

ABSTRACT

Motivation: Chromatin immunoprecipitation (ChIP) combined with DNA microarray is a high-throughput technology to investigate DNA–protein binding or chromatin/histone modifications. ChIP–chip data require adapted statistical method in order to identify enriched regions. All methods already proposed are based on the analysis of the log ratio (Ip/Input). Nevertheless, the assumption that the log ratio is a pertinent quantity to assess the probe status is not always verified and it leads to a poor data interpretation.

Results: Instead of working on the log ratio, we directly work with the Ip and Input signals of each probe by modeling the distribution of the Ip signal conditional to the Input signal. We propose a method named ChIPmix based on a linear regression mixture model to identify actual binding targets of the protein under study. Moreover, we are able to control the proportion of false positives. The efficiency of ChIPmix is illustrated on several datasets obtained from different organisms and hybridized either on tiling or promoter arrays. This validation shows that ChIPmix is convenient for any two-color array whatever its density and provides promising results.

Availability: The ChIPmix method is implemented in R and is available at http://www.agroparistech.fr/mia/outil_A.html

Contact: marie_laure.martin@agroparistech.fr

1 INTRODUCTION

Chromatin immunoprecipitation (ChIP) is a well-established procedure used to investigate proteins associated with DNA. ChIP on chip involves analysis of DNA recovered from ChIP experiments by hybridization to microarray. In a two-color ChIP–chip experiment, two samples are compared: DNA fragments crosslinked to a protein of interest (IP) and genomic DNA (Input). The two samples are differentially labeled and then co-hybridized on a single array. The goal is then to identify actual binding targets of the IP, i.e. probes whose IP signal is significantly larger than the Input signal.

Many authors have already pointed out the need for efficient statistical procedures to detect enriched probes (Buck and Lieb, 2004; Keles, 2007). Recently, two strategies have been widely applied for the detection of enriched DNA regions. The first strategy takes advantage of the spatial structure of the data. Since probes are positioned all along the genome, if one region is enriched we expect several adjacent probes to obtain high ratio measurements, resulting in a 'peak' of intensity. Spatial methods such as sliding windows (Cawley *et al.*, 2004; Keles, 2007) or Hidden Markov Models (Ji and Wong, 2005; Li *et al.*, 2005) have been proposed to detect these peaks. Alternatively, the second strategy is to consider that the whole population of probes can be divided into two components: the population of IP-enriched genomic fragments, and the population of genomic DNA that is not IP enriched. Different statistical methods have been proposed to distinguish between the two populations by considering the distribution of the ratios (or their associated rank). Assuming that a non-negligible proportion of the fragments are enriched, the log ratio distribution is bimodal, the highest mode corresponding to the enriched population. A probe is then declared enriched when its ratio exceeds a selected cutoff, which is fixed according to the data distribution (Buck and Lieb, 2004).

Importantly, both strategies assume that the log ratio measurement is a pertinent statistical quantity to assess the probe status (enriched or not). This assumption is correct if the distribution of the ratio mostly depends on the status (normal/enriched) of the probe. Figure 1A shows the ideal situation described in Buck and Lieb (2004), where the distribution is bimodal. In many applications, the distribution of the log ratios is closer to Figure 1B, and the performance of log ratio-based methods may be poor. At least two technical reasons may explain the difference between the ideal and real cases. First, there are some technical difficulties to obtain the IP sample: it requests the use of a very specific antibody and a careful experimental process to avoid a high level of contamination. The second reason comes from the possible cross-hybridization phenomena.

From observation of Figure 1, we argue that it is worth working directly with the two measurements of each probe (Input and IP) rather than with the log ratio. In Figure 1C, we observe that the relationship between the two measurements is almost linear. Working on log ratio amounts to stating that the slope of the linear relationship is the same whatever the status of the probe. In many cases the slopes are different: Figure 2 (synthetic data) shows that even a slight difference between the two slopes may turn the distribution of the log ratios into unimodal rather than bimodal, as observed for the NimbleGen slide in Figure 1.

In this work, we propose a new statistical method that we call ChIPmix, based on a mixture model of regressions. This framework allows us to well characterize the IP–Input relationship, and to provide a statistical procedure to control the proportion of probes wrongly classified as enriched. The article is organized as follows. The statistical model and the procedure for false positive control are described in Section 2. In Section 3, we consider several large datasets obtained from different organisms and hybridized on different array types (tiling or promoter). We show that the method

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from bioinformatics.oxfordjournals.org by guest on September 22, 2011

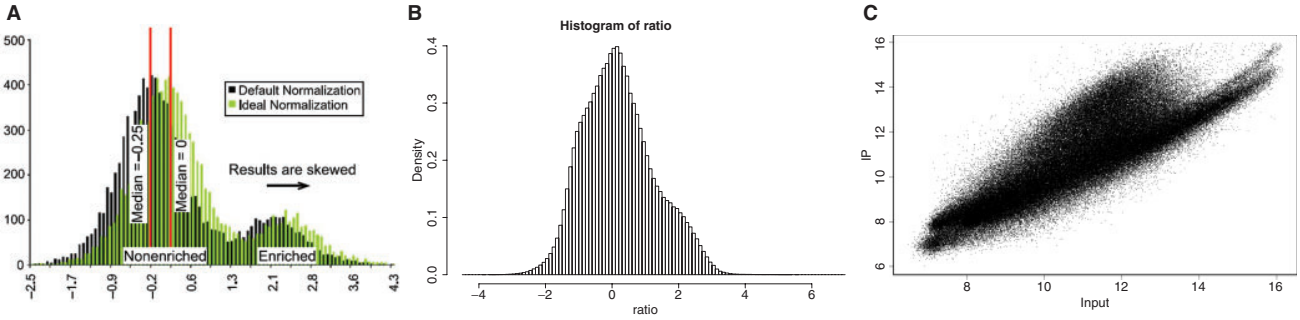

Fig. 1. **(A)** Ideal log ratio distribution with two distinct peaks. **(B)** Log ratio distribution on a real example (NimbleGen array). **(C)** Associated plot of IP versus Input (NimbleGen array).

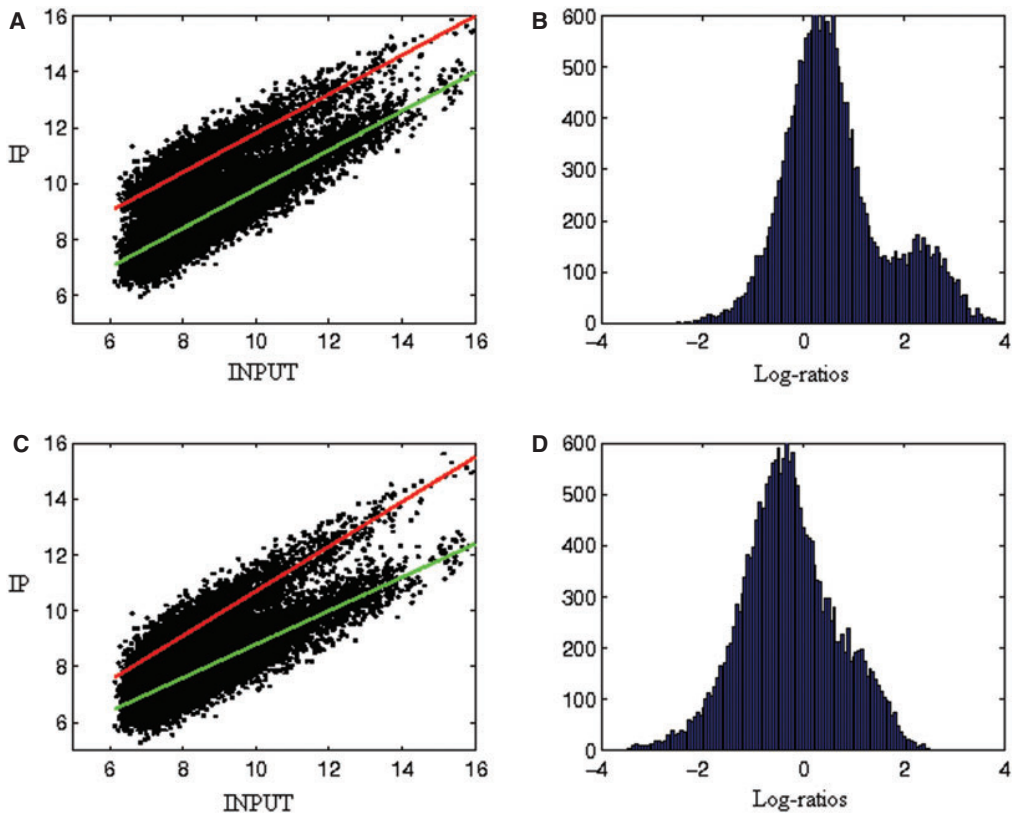

Fig. 2. Synthetic data. **(A,B)** Two populations with linear relationship and equal slopes. The corresponding log ratio histogram is bimodal. **(C,D)** Two populations with linear relationship but different slopes. The corresponding log ratio histogram is unimodal.

outperforms competing methods in terms of sensitivity. The main conclusions and some possible extensions are discussed in Section 4.

2 STATISTICAL FRAMEWORK

2.1 Model and inference

Let (x_i, Y_i) be the log-Input and log-IP intensities of probe i, respectively. The (unknown) status of the probe is characterized through a label Z_i which is 1 if the probe is enriched and 0 if it is normal (not enriched). We assume the Input–IP relationship to be linear whatever the population, but with different slope and intercept. More precisely, we have:

$$\begin{aligned} Y_i &= a_0 + b_0 x_i + \epsilon_i \qquad \text{if } Z_i = 0 \text{ (normal)} \\ &= a_1 + b_1 x_i + \epsilon_i \qquad \text{if } Z_i = 1 \text{ (enriched)} \end{aligned}$$

where ϵ_i is a Gaussian random variable with mean 0 and variance σ^2. Such a model is named a mixture model of regressions.

Downloaded from bioinformatics.oxfordjournals.org by guest on September 22, 2011

The marginal distribution of Y_i for a given level of Input x_i is

$$(1-\pi)\phi_0(Y_i|x_i)+\pi\phi_1(Y_i|x_i), \quad (1)$$

where π is the proportion of enriched probes, and $\phi_j(\cdot|x)$ stands for the probability density function (PDF) of a Gaussian distribution with mean a_j+b_jx and variance σ^2.

The mixture model is used to classify probes as normal or enriched. To do this, we calculate the probability of a probe to be enriched given its Input and IP intensities. This probability is called the *posterior* probability and is defined from Equation (1) by

$$\tau_i=\Pr\{Z_i=1\,|\,x_i,Y_i\}=\frac{\pi\phi_1(Y_i|x_i)}{(1-\pi)\phi_0(Y_i|x_i)+\pi\phi_1(Y_i|x_i)}. \quad (2)$$

The mixture parameters (proportion, intercepts, slopes and variance) are estimated using the EM algorithm. The EM algorithm is dedicated to the class of incomplete data models where the status of the observations is unknown. In the E step, the posterior probability for each observation to belong to each class is calculated. In the M step, the parameters of each class are estimated using a weighted regression, in which the weights are given by the posterior probabilities. This algorithm is implemented in the mixreg function of the mixreg R package (Turner, 2000). Figure 4A shows the application of Chipmix on the NimbleGen high-density array data, presented in Section 3.3.

In the mixreg function, the initial values of the parameters must be given by the users otherwise they are chosen randomly. Nevertheless, the EM algorithm is well-known to be sensitive to the initial values (Bohning and Seidel, 2003; Karlis and Xekalaki, 2003) and to solve this difficulty, we propose initial values derived from the first axis of the Principal Component Analysis (PCA) of the whole dataset (see the ChIPmix R function for details).

The mixture model with two linear regressions is adapted if the protein under study has some targets. When the protein has no target, all probes belong to the normal class. In this case, a simple linear regression is sufficient to fit the data. For each dataset the two models (one or two classes) are fitted and the best model is selected according to the BIC criterion (Schwarz, 1978).

2.2 False discovery control

Posterior probabilities are used to classify probes into the normal or enriched class, using the following classification rule

$$\tau_i > s \quad \Rightarrow \quad \widehat{Z}_i=1 \text{ classified as enriched,}$$

where s is an arbitrary threshold that has to be fixed. In the context of mixture models, s is usually fixed to 1/2 (*Maximum A Posteriori* rule) which implicitly means that misclassifications in population 0 or in population 1 have the same cost.

In ChIP–chip experiments, where false positives are of concern, it is important to control the false positive proportion and to fix s accordingly. In the hypothesis test theory, the false discovery control is performed by controlling the probability to reject wrongly the null hypothesis. We propose an analogous concept in the mixture model framework. Our aim is to control the probability for a probe to be wrongly assigned to the enriched class. Therefore, we want $\Pr\{\tau_i>s\,|\,x_i,Z_i=0\}$ to be equal to a predefined level α. In practice, we fix α and we find the threshold s depending on α and x_i (see Appendix).

3 RESULTS

We present three applications of ChIPmix to assess the performance of the method whatever the specificity and density of the array (tiling or promoter array). The first two applications validate the method on already published data. The third dataset is used to compare ChIPmix with existing methods.

3.1 Promoter DNA methylation in the human genome

Weber *et al.* (2007) measured DNA methylation using a NimbleGen microarray representing 15 609 promoter regions of the human genome. Each promoter region is covered by 15 probes and is classified into a category according to its CpG rate. We focus on the analysis of the class ICP (intermediate CpG promoter). Weber *et al.* based their classification on the mean log ratio value for the 15 probes per promoter region. If this value was larger than 0.4 (threshold based on bisulfite sequencing), the promoter region was declared hypermethylated. Among the 2056 promoter regions under study, 460 were declared hypermethylated.

We applied ChIPmix to these data without averaging the 15 values per promoter region. The estimated proportion π of enriched probes was 0.794. This is in keeping with a large proportion of targets expected in such experiments. The estimated regression slopes were $\widehat{b}_0=0.613$ for the normal class and $\widehat{b}_1=1.162$ for the enriched one, which shows that the Input–Ip relations substantially differ between the two status. At the level $\alpha=0.01$, a total of 1706 promoter regions were found to have at least one probe enriched. Except for one region, all the promoter regions of the Weber's list have at least enriched probe, and 403 have 5 or more enriched probes. Besides, ChIPmix identified 38 promoter regions with 9 probes or more classified as enriched that were not detected in Weber *et al.* (2007).

3.2 Histone modification in *Arabidopsis thaliana*

Turck *et al.* (2007) studied several histone modifications of *A.thaliana* using a custom genomic tiling array of Chromosome 4. To declare a tile enriched, they developed a two-step method based on a Gaussian mixture model and a total of 2775 tiles were found to be marked by histone H3 trimethylated at lysine 27 (H3K27me3) according to their analysis.

We analyzed the same dataset using ChIPmix. The estimated proportion and slopes were $\widehat{\pi}=0.361$, $\widehat{b}_0=0.907$ and $\widehat{b}_1=1.167$. The tiles classified as enriched at risk $\alpha=0.01$ include all the tiles found by Turck *et al.* (2007) plus 2346 others: 1404 tiles extend the genomic region already found marked by H3K27me3 and 942 tiles form 62 new genomic regions. The difference between the two slopes enables us to better discriminate the two classes for high Input intensities. This may explain the higher number of enriched probes detected by ChIPmix.

3.3 NimbleGen high-density array (Histone modification H3K9me3)

In this last example, we considered Chip–chip data produced on a two-color NimbleGen array of 1 132 140 probes. Each chromosome of the model plant *A.thaliana* is covered by about 200 000 probes. Such very high-density arrays are more and more popular, so we need to assess the efficiency of ChIPmix on such a very large dataset.

Downloaded from bioinformatics.oxfordjournals.org by guest on September 22, 2011

From a biological point of view, the same IP and Input samples were already hybridized on a custom genomic tilling array covering the Chromosome 4 (Turck *et al.*, 2007). Regions identified in Turck *et al.* (2007) were biologically validated and are used as true positives. In addition, the chloroplastic genome can be used as a negative control, since no histone modification target is expected in this region. We did not use the mitochondrial genome as a negative control since some regions have been duplicated in the nuclear genome. From a statistical point of view, since ChIPmix does not take the spatial structure into account, it is important to compare it with methods using this information. We compared our results with those provided by the NimbleGen software and ChIPOTle method (Buck *et al.*, 2005). NimbleGen software uses a permutation-based algorithm to find statistically significant peaks, using scaled log ratio data, and ChIPOTle method uses a sliding window approach.

Two biological replicates were available, for which hybridizations were performed in dye-swap. We performed a normalization step to remove technical biases as well as dye bias. Since the Input and IP samples differed substantially, array-by-array normalization such as lowess could not be applied. We quantified biases by an ANOVA model (Kerr *et al.*, 2002), and removed them from the raw data. The IP and Input signals for each biological replicate were averaged on the dye-swap to remove the gene-specific dye bias. Analyses per chromosome were performed on the normalized data.

For a risk $\alpha = 0.01$, a total of 30 477 probes were detected in the first replicate and 27 553 in the second. The intersection contains more than two-thirds of the probes declared enriched in at least one replicate (23 546 probes). Although ChIPmix does not take the spatial structure of the genome into account, enriched probes are clustered in genomic regions (Fig. 3). These regions are rich in genes and corroborate the results of Turck *et al.* (2007), who have shown that H3K9me3 is actually a euchromatin mark. Moreover, more than 80% of the probes classified as enriched in this experiment cover genomic regions already found in Turck *et al.* (2007).

For the chloroplastic genome, the BIC criterion selected a two component regression model. For the first biological replicate, two 2-probe clusters and one 3-probe cluster were declared enriched. On the same replicate, ChIPOTle (window = 500 and step = 100) found five 2-probe clusters, two 3-probe cluster and one 6-probe cluster. With other parameters (window = 200 and step = 50), the number of detected peaks increased. Results are similar for the second biological replicate, and one cluster was declared enriched with ChIPmix in both biological replicates (two with ChIPOTle). NimbleGen did not provide the analysis of the chloroplastic genome.

We also compared ChIPmix to the results given by NimbleGen and ChIPOTle on Chromosome 4, studied in Turck *et al.* (2007). The probes declared enriched by ChIPmix include almost all of those found enriched by the NimbleGen software, but cover much larger genomic regions. Moreover ChIPmix identifies other genomic regions not found by the NimbleGen software (Fig. 3), that are validated by a comparison with results of Turck *et al.* (2007). ChIPmix detects 30 477 enriched probes, including 24 575 in common with Turck *et al.* (2007). ChIPOTle detects 24 357 probes [20 866 common with Turck *et al.* (2007)] and NimbleGen detects 19 837 probes [16 600 common with Turck *et al.* (2007)] (Fig. 4B). Among the three methods ChIPmix provides the closest results to the reference publication.

4 DISCUSSION

We propose a statistical method based on mixture of regression to classify probes in ChIP–chip experiments. Our approach accounts for different relations between IP and Input intensity in the two classes of probes (enriched and normal). The ChIPmix method outperforms the standard approaches based on the log ratio.

We presented various applications each dedicated to one specific biological question (histone modification and DNA methylation on different organisms). ChIPmix can also be applied to the detection of transcription factor binding sites (TFBS, results not shown).

Fig. 3. Genomic region of Chromosome 4 of *A.thaliana* visualized with SignalMap^{TM}. In the first line annotation is given, the boxes are the genes, the second line shows the genomic regions found by the NimbleGen software. Thick bars are not enriched and the others bars are colored according to a FDR value and are all enriched. The third line gives the probes declared enriched by ChIPmix with $\alpha = 0.01$. The fourth line gives the results of ChIPOTle (window = 500, step = 100).

Downloaded from bioinformatics.oxfordjournals.org by guest on September 22, 2011

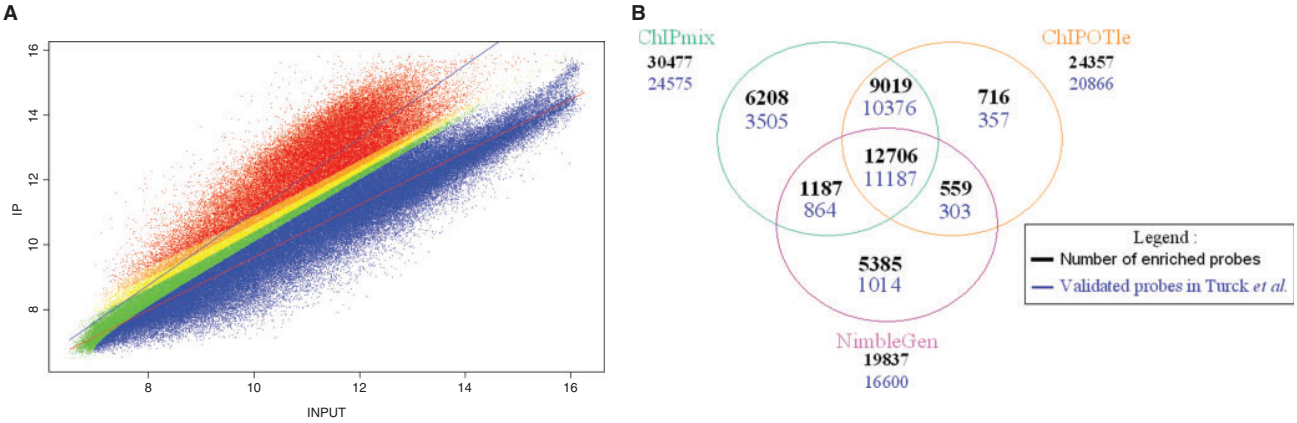

Fig. 4. **(A)** IP intensities versus Input Intensities colored according to the posterior probabilities $\hat{\tau}_i$. Colors change every 20% (blue: $\hat{\tau}_i < 20\%$, red: $\hat{\tau}_i > 80\%$). The two lines are the two estimated linear regressions of the mixture. **(B)** Venn diagram summarizing the results of the three methods.

The method is valid when the proportion of positive probes is expected to be large (e.g. histone modification), or small (e.g. TFBS). Through the examples we have shown that ChIPmix is convenient for any two-color chip whatever its density (array size from thousands to hundreds of thousands of probes) and the nature of the probe (tiling and promoter arrays).

ChIPmix does not account for the spatial structure of the data. While this could be seen as a drawback, we showed that enriched probes are clustered into genomic regions in the presented applications. Moreover, this may become perfectly relevant for specific experiments as well as RIP-chip, which investigates interactions between protein and RNA (Schmitz-Linneweber *et al.*, 2005) or ChIP–chip experiments performed on array where promoter are represented by only one probe (see project SAP at `www.psb.ugent.be/SAP/`).

The only parameter of the ChIPmix method is the risk α, which can be easily interpreted. In contrast, two parameters have to be tuned in the ChIPOTle method (window size and step). The tuning of this two parameters depends on both the experimental protocol and the array type. The results are very sensitive to this tuning.

The proposed strategy can be extended in different ways. The ChIPmix extension to the unequal variance case is straightforward. However, the equal variance case provides an efficient framework for the false discovery control. If the equality of variance is not assumed, the calculations given in appendix do not hold anymore, and the solving of the equation system becomes much more complex.

The proposed regression models allow us to correct the IP intensity with respect to the Input one. Other elements may influence the level of IP signal. Weber *et al.* (2007) show that the CpG rate has to be taken into account to classify probes. The specificity of the probes (number of hits) may also alter the IP intensity. All this information can be considered as covariates and added in the model. This will lead to a mixture of multiple regression for which the statistical framework is almost the same as the one we propose.

The proportion of false negative results can be controlled in the same way as the false discovery described in Section 2.2. This allows us to evaluate the sensitivity of the classification at each Input level. Moreover, the two criteria (false negative and false discovery) can be combined to derive a threshold s that optimizes some trade-off between them.

ACKNOWLEDGEMENTS

The authors want to thank Vincent Colot, Alain Lecharny and Michel Caboche from the URGV unit for helpful discussions and advice.

Funding: This work was supported by the TAG ANR/Genoplante project.

REFERENCES

Bohning,D. and Seidel,W. (2003) Editorial: recent developments in mixture models. *Comput. Stat. Data Anal.*, **41**, 349–357.

Buck,M.J. and Lieb,J.D. (2004) Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349–360.

Buck,M.J. *et al.* (2005) ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol.*, **6**, R97.

Cawley,S. *et al.* (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.

Ji,H. and Wong,W.H. (2005) Tilemap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, **21**, 3629–3636.

Karlis,D. and Xekalaki,E. (2003) Choosing initial values for the EM algorithm for finite mixtures. *Comput. Stat. Data Anal.*, **41**, 577–590.

Keles,S. (2007) Mixture modeling for genome-wide localization of transcription factors. *Biometrics*, **63**, 10–21.

Kerr,M.K. *et al.* (2002) Statistical analysis of a gene expression microarray experiment with replication. *Stat. Sin.*, **12**, 203–217.

Li,W. *et al.* (2005) A hidden markov model for analyzing chip-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, **21**, i274–i282.

Schmitz-Linneweber,C. *et al.* (2005) RNA immunoprecipitation and microarray analysis show a chloroplast pentatricopeptide repeat protein to be associated with the 5′ region of mRNAs whose translation it activates. *Plant Cell*, **17**, 2791–2804.

Weber,M. *et al.* (2007) Distribution, silencing potential and evolutionary impact of promoter dna methylation in the human genome. *Nat. Genet.*, **39**, 457–466.

Schwarz,G. (1978) Estimating the dimension of a model. **6**, 461–464.

Turck,F. *et al.* (2007) Arabidopsis tfl2/lhp1 specifically associates with genes marked by trimethylation of histone h3 lysine 27. *PLoS Genet.*, **3**, e86.

Turner,T.R. (2000) Estimating the rate of spread of a viral infection of potato plants via mixtures of regressions. *Appl. Stat.*, **49**, 371–384.

Downloaded from bioinformatics.oxfordjournals.org by guest on September 22, 2011

APPENDIX

We propose to control the probability for a normal probe i to be wrongly assigned to the enriched class:

$$\Pr\{\tau_i > s \,|\, x_i, Z_i = 0\} = \alpha. \tag{3}$$

In practice, we fix α and we find the threshold s depending on α and x_i. Using definition 2, $\Pr\{\tau_i > s \,|\, x_i, Z_i = 0\}$ can be rewritten as

$$\Pr\{(1-\pi)\phi_0(Y_i|x_i)(1-s) - s\pi\phi_1(Y_i|x_i) > 0 \,|\, x_i, Z_i = 0\}.$$

Replacing the probability density functions $\phi_0(Y_i|x_i)$ and $\phi_1(Y_i|x_i)$ with their expression, we get Equation (3) equivalent to

$$\Pr\left(2\frac{(a_0-a_1)+(b_0-b_1)x_i}{\sigma^2}Y_i + \gamma(s,x_i) > 0 \,|\, x_i, Z_i = 0\right) = \alpha, \tag{4}$$

where

$$\gamma(s,x_i) = \frac{(a_0+b_0x_i)^2-(a_1+b_1x_i)^2}{\sigma^2} - 2\log\{s(1-\pi)\} + 2\log\{(1-s)\pi\}.$$

Since the status of probe i is normal ($Z_i = 0$), the distribution of Y_i is a Gaussian with mean $a_0 + b_0x_i$ and variance σ^2, and we deduce that Equation (4) is equivalent to solve

$$\gamma(s,x_i) = \frac{2(a_0-a_1+(b_0-b_1)x_i)}{\sigma}\left\{u_{1-\alpha} + \frac{(a_0+b_0x_i)}{\sigma}\right\},$$

where $u_{1-\alpha}$ is the $(1-\alpha)$ quantile of Gaussian with mean 0 and variance 1.

Using the definition of $\gamma(s,x_i)$, the expression of threshold s is given by

$$s = \frac{e^{\lambda}}{1+e^{\lambda}},$$

where

$$\lambda = \left(\frac{a_1-a_0+(b_1-b_0)x_i}{\sigma}\right)\left(u_{1-\alpha} - \frac{a_1-a_0+(b_1-b_0)x_i}{2\sigma}\right) - \log\left(\frac{1-\pi}{\pi}\right).$$

Downloaded from bioinformatics.oxfordjournals.org by guest on September 22, 2011

Annexe B

Article publié dans *La Revue de Modulad*

Mélanges gaussiens bidimensionnels pour la comparaison de deux échantillons de chromatine immunoprécipitée

Caroline Bérard[1], Marie-Laure Martin-Magniette[1,2], Alexandra To[3], François Roudier[3], Vincent Colot[3] et Stéphane Robin[1].

[1] UMR AgroParisTech/INRA MIA 518, 16 rue Claude Bernard, PARIS Cedex 05.
[2] UMR INRA 1165 - CNRS 8114 - UEVE URGV, 2 rue Gaston Crémieux, EVRY.
[3] UMR CNRS 8186, Département de Biologie, 46 rue d'Ulm, PARIS Cedex 05.

caroline.berard@agroparistech.fr , marie_laure.martin@agroparistech.fr
to@biologie.ens.fr , roudier@biologie.ens.fr
colot@biologie.ens.fr , stephane.robin@agroparistech.fr

Résumé L'immunoprécipitation de la chromatine (ChIP) permet d'étudier les interactions entre les protéines et l'ADN ainsi que différents états chromatiniens. Le ChIP-chip est une technique combinant l'immunoprécipitation de la chromatine avec le principe des puces à ADN, ce qui permet une étude à l'échelle du génome. Nous nous intéressons ici à l'analyse des différences entre deux échantillons d'ADN immunoprécipité. Biologiquement, on s'attend à distinguer quatre groupes différents : un groupe d'ADN non-immunoprécipité, un groupe d'ADN immunoprécipité identiquement dans les deux échantillons et deux groupes dans lesquels l'ADN est immunoprécipité en quantités différentes. Nous modélisons ces données par un mélange de gaussiennes bidimensionnelles à quatre composants. Les matrices de variance sont contraintes afin d'intégrer des connaissances biologiques. Les paramètres sont estimés par l'algorithme EM. Nous appliquons cette méthode pour étudier la différence de méthylation d'une histone entre l'écotype sauvage de la plante modèle *Arabidopsis thaliana* et un mutant.
Mots-clés : Mélange gaussien, décomposition spectrale, algorithme EM, ChIP-chip.

Résumé Chromatin immunoprecipitation (ChIP) enables to investigate interactions between proteins and DNA and also various chromatin states. ChIP-chip is a well-established procedure combining chromatin immunoprecipitation with DNA microarrays, which allows a study of the whole genome. We are interested in the analyze of the differences between two immunoprecipitated DNA samples. From a biological point of view, we expect to distinguish four different groups : a group of non-immunoprecipited DNA, a group of immunoprecipited DNA in both samples, and then two groups in which DNA is differently immunoprecipited. We propose to model these data with a mixture of two-dimensional Gaussians with four components. Biological knowledges are included as constraints on the variance matrices. The parameters are estimated by the EM algorithm. This method is applied to NimbleGen data in order to study the histone methylation difference between the wild ecotype of the model plant *Arabidopsis thaliana* and a mutant.
Keywords : Gaussian mixture, eigenvalue decomposition, EM algorithm, ChIP-chip.

1 Introduction

La connaissance des mécanismes de régulation des gènes est essentielle pour comprendre certains concepts biologiques importants. On sait par exemple que le développement d'un organisme dépend grandement de l'harmonisation de l'expression de ses gènes. Après le séquençage entier des génomes à grande échelle, le défi consiste donc aujourd'hui à comprendre le fonctionnement des gènes, c'est-à-dire à déterminer leur fonction et leur patron d'expression.

Dans le noyau des cellules eucaryotes, l'ADN est fractionné en chromosomes et il est condensé sous forme de chromatine. La chromatine est un complexe ADN-protéines qui joue un rôle essentiel dans le contrôle de l'activité des gènes. Les protéines présentes sont principalement des histones. La condensation de l'ADN en chromatine s'organise de manière séquentielle et ordonnée. En premier lieu, 147 paires de bases d'ADN s'enroulent autour d'un octamère d'histones pour former un nucléosome. Dans un second niveau d'organisation, les nucléosomes se compactent et forment une hélice. Cette hélice est finalement condensée en euchromatine (condensation légère) ou en hétérochromatine (condensation plus prononcée) constituant un troisième niveau d'organisation. Les gènes localisés dans l'euchromatine peuvent être plus facilement transcrits car la condensation est légère. Cette structure d'organisation du génome dans le noyau constitue en elle-même un mécanisme de répression ou d'activation de la transcription des gènes. En effet, pour activer la transcription d'un gène donné dans une cellule, la chromatine comprise dans la région de contrôle du gène doit être modifiée ou altérée de façon à être permissive à la transcription. Les modifications post-traductionnelles d'histone (comme la méthylation, l'acétylation, l'ubiquitination ou la phosphorilation) sont des mécanismes impliqués dans la régulation de l'expression des gènes (Turck *et al.* [18]).

L'immunoprécipitation de la chromatine (ChIP) permet d'étudier les interactions entre les protéines et l'ADN ainsi que différents états chromatiniens associés à des états d'activité distincts du génome. Le ChIP-chip est une technique combinant l'immunoprécipitation de la chromatine avec le principe des puces à ADN (Amaratunga et Cabrera [1]), ce qui permet une étude à l'échelle du génome. Habituellement dans une expérience de ChIP-chip, les deux échantillons co-hybridés sont les fragments d'ADN associés à la protéine d'intérêt ou à une marque chromatinienne (IP) et l'ADN génomique total (INPUT). Le but est ensuite de détecter les sondes de la puce pour lesquelles il y a un signal IP afin d'identifier les régions génomiques où la protéine d'intérêt se fixe.

Buck et Lieb [7] ont montré la nécessité de développer de nouvelles méthodes statistiques pour détecter les sondes enrichies dans les expériences de ChIP-chip. Récemment, deux stratégies ont été largement appliquées : la première tient compte de la structure spatiale des données (Cawley *et al.* [9], Keles [14]) et la seconde considère que la totalité des sondes peut être divisée en deux populations : les sondes enrichies et les non-enrichies (Buck et Lieb [7], Turck *et al.* [18], Martin-Magniette *et al.* [15]). Différentes méthodes statistiques ont été proposées pour distinguer ces deux populations : toutes sont fondées sur la distribution du log-ratio $\log(IP/INPUT)$ (Buck et Lieb [7], Turck *et al.* [18]), exceptée la méthode proposée par Martin-Magniette *et al.* [15] qui utilise un mélange de régressions pour modéliser la loi de l'IP conditionnellement à l'INPUT.

La technique du ChIP-chip permet également d'étudier directement la différence entre deux échantillons d'ADN immunoprécipités, sans hybrider sur la puce l'ADN génomique total (INPUT). À notre connaissance il n'existe pas de méthode pour analyser ce type de données (IP/IP) dans la littérature. Les méthodes de segmentation initialement développées pour l'analyse des données CGH (Hupé *et al.* [13], Olshen *et al.* [16], Picard *et al.* [17]) pourraient être utilisées, mais les régions génomiques non immunoprécipitées et les régions immunoprécipitées identiquement dans les deux échantillons seraient indistinguables. De plus ces méthodes sont assez coûteuses en temps de calcul pour des puces tiling-array qui ont un grand nombre de sondes.

L'objectif de notre travail est de proposer une modélisation conjointe des signaux IP obtenus par un modèle de mélange de gaussiennes bi-dimensionnelles. La description des données est détaillée section 2. Les mélanges gaussiens bidimensionnels modélisés à l'aide d'une décomposition de la matrice de variance sont étudiés section 3. Les connaissances biologiques sont prises en compte sous forme de contraintes sur les paramètres du modèle et sur le nombre de composants. Cette modélisation est détaillée dans la section 4. Une application de la méthode sur des données issues de la technologie NimbleGen est présentée dans la section 5.

2 Description des données

Les données analysées concernent la plante modèle *Arabidopsis thaliana*. Les deux échantillons co-hybridés sur la puce visent à étudier le comportement de l'histone H3 diméthylée au niveau de la lysine 9 (H3K9me2). On compare un échantillon sauvage et un échantillon mutant (mutant nrpdlalb).
L'expérience est faite en dye-swap (Boulicaut et Gandrillon [6]) : le principe est de faire une répétition technique en inversant les marquages. Chaque traitement est ainsi marqué par les deux fluorochromes, ce qui permet de contrôler le biais dû au marquage (biais technique). Les intensités des signaux sont ensuite moyennées sur le dye-swap.
La puce à ADN utilisée est une puce tiling-array à oligos courts issue de la technologie NimbleGen. Cette puce permet d'étudier le génome nucléaire d'*Arabidopsis thaliana*, composé de cinq chromosomes et des génomes mitochondrial et chloroplastique. La puce est constituée d'environ 700 000 sondes.

Lorsque l'on étudie des données de ChIP-chip IP/IP, on s'attend à distinguer quatre groupes (cf Figure 1) :

- Un groupe d'intensité faible qui correspond aux séquences d'ADN qui ne sont pas immunoprécipitées (bruit).
- Un groupe où les séquences d'ADN sont immunoprécipitées en même quantité chez le sauvage et chez le mutant. Cela correspond aux endroits sur le génome où l'histone est méthylée identiquement dans les deux échantillons. Ce groupe sera défini dans la suite comme groupe normal.
- Deux groupes où les séquences d'ADN sont immunoprécipitées en quantités différentes

chez le sauvage et chez le mutant. Le taux de méthylation de l'histone H3K9me2 peut être plus faible chez le mutant (groupe appauvri), ou bien au contraire, plus élevé (groupe enrichi).

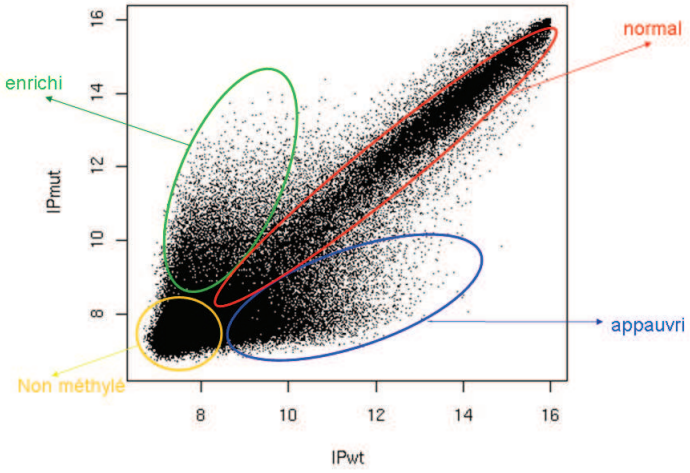

FIG. 1 – Comparaison de deux échantillons de chromatine immunoprécipitée (sauvage vs mutant) : Identification schématique des différents groupes.

3 Modèle de mélanges gaussiens bidimensionnels

Dans cette section, nous rappelons brièvement l'approche de classification par les mélanges gaussiens et reprenons la modélisation des modèles de mélanges gaussiens à l'aide d'une décomposition des matrices de variance, puis nous appliquons certains modèles définis dans Biernacki *et al.* [4] à nos données.

3.1 Approche par classification

Si le but de l'analyse est la classification, le label de chaque donnée est manquant au regard de l'échantillon observé. Notons Z_{ik}, ce label pour l'individu i, qui est une variable aléatoire égale à 1 si le point x_i appartient à la population k et 0 sinon. Les variables $\{Z_1, ...Z_n\}$ (avec $Z_i = \{Z_{i1}, ..., Z_{iK}\}$) sont supposées indépendantes et suivent une loi multinomiale de probabilités $\pi_1, ..., \pi_K$, qui sont les proportions des K classes dans le mélange. Si nous notons Y le vecteur des données complètes (X, Z) où seul X est observé, alors cette reformulation montre clairement que les modèles de mélange peuvent être vus comme un cas particulier des modèles à structure cachée comme par exemple les modèles de Markov cachés (Cappé *et al.* [8], Ephraim et Merhav [12]), la différence étant que les variables $\{Z_1, ...Z_n\}$ sont supposées ici indépendantes.

Dans notre travail, la variable observée $X_i = (X_{1i}, X_{2i})$ est le signal log-IP de chaque échantillon pour la sonde i et nous supposons que les observations proviennent d'un mélange de densités gaussiennes. La densité du couple s'écrit :

$$f(X_i, \psi) = \sum_{k=1}^{K} \pi_k \phi(X_i|\mu_k, \Sigma_k),$$

où π_k est la proportion du k-ième composant du mélange ($0 < \pi_k < 1 \ \forall k = 1, ..., K$ et $\sum_{k=1}^{K} \pi_k = 1$), $\psi = (\pi_1, ..., \pi_{K-1}, \mu_1, ..., \mu_K, \Sigma_1, ..., \Sigma_K)$ est le vecteur des paramètres du mélange et $\phi(.|\mu_k, \Sigma_k)$ est la densité d'une distribution gaussienne bidimensionnelle de moyenne μ_k et de variance Σ_k définis au point x_i par :

$$\phi(x_i|\mu_k, \Sigma_k) = \frac{1}{2\pi}[\det(\Sigma_k)]^{-1/2} \exp\left\{-\frac{1}{2}(x_i - \mu_k)'\Sigma_k^{-1}(x_i - \mu_k)\right\},$$

où M' représente la transposée de M.

Nous calculons les probabilités conditionnelles que la sonde i appartienne à chacun des groupes sachant l'ensemble des observations. Nous rappelons que par définition, la probabilité conditionnelle que la sonde i appartienne au groupe k sachant l'ensemble des observations est définie par :

$$\tau_{ik} = \frac{\hat{\pi}_k \phi(X_i|\hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{l=1}^{K} \hat{\pi}_l \phi(X_i|\hat{\mu}_l, \hat{\Sigma}_l)}.$$

Nous pouvons ensuite classer la sonde i en l'attribuant au groupe pour lequel la probabilité conditionnelle est la plus grande (règle du Maximum A Posteriori).

3.2 Paramétrisation spectrale des matrices de variance

La densité gaussienne modélise une distribution ellipsoïdale de centre μ_k dont les caractéristiques géométriques (volume, forme, orientation) sont définies à l'aide d'une décomposition spectrale de la matrice de variance Σ_k. Pour cela, nous reprenons une paramétrisation proposée par Banfield et Raftery [2] qui permet de proposer de nombreux modèles de classification. Cette paramétrisation considère la décomposition spectrale des matrices de variance :

$$\Sigma_k = \lambda_k D_k A_k D_k' \ , \tag{1}$$

où λ_k représente le volume ($\lambda_k = \det(\Sigma_k)^{1/2}$), D_k représente l'orientation et A_k représente la forme de l'ellipse. La matrice D_k est la matrice des vecteurs propres de Σ_k et A_k est une matrice diagonale telle que $\det(A_k) = 1$ avec les valeurs propres normalisées de Σ_k sur la diagonale dans l'ordre décroissant. En permettant aux paramètres volumes, formes et orientations de varier ou d'être égaux entre les classes, on obtient 14 modèles de mélanges gaussiens différents et facilement interprétables. Les 14 modèles sont détaillés dans Celeux et Govaert [10] : il y a 8 modèles généraux, 4 modèles avec des matrices de variance diagonales et 2 modèles avec des formes sphériques ($A_k = I$).

3.3 Application de 4 modèles de classification aux données

Les 14 modèles de classification (Celeux et Govaert [10]) sont implémentés dans le logiciel MIXMOD [4]. À la vue des données IP/IP (cf Figure 1), nous considérons uniquement les modèles à quatre composants d'orientations différentes, c'est-à-dire les modèles $\lambda D_k A D'_k$, $\lambda_k D_k A D'_k$, $\lambda D_k A_k D'_k$ et $\lambda_k D_k A_k D'_k$. En reprenant les conventions de Celeux et Govaert [10], nous notons λ (respectivement D, A) lorsque le volume (respectivement l'orientation, la forme) est égal pour tous les composants, et λ_k (respectivement D_k, A_k) lorsque le volume (respectivement l'orientation, la forme) est différent pour tous les composants.
On peut choisir le meilleur modèle à l'aide du critère BIC ou du critère ICL. Le critère BIC (Bayesian Information Criterion, Schwarz (1978)) est très utilisé pour les modèles à structure cachée, en particulier les modèles de mélange. Soit $x = (x_1, ..., x_n)$ un n-échantillon où $x_i = (x_{i1}, x_{i2})$ est le signal log-IP observé pour un individu i, le critère BIC du modèle m vaut :

$$BIC_m = -2 \log \left\{ f(x|\hat{\psi}_m) \right\} + \nu_m \log(n),$$

où $\hat{\psi}_m$ est l'estimateur des paramètres pour le modèle m et ν_m est le nombre de paramètres du modèle m. Le critère ICL (Integrated Complete-data Likelihood, Biernacki *et al.* [5]) prend en compte la capacité d'un modèle de mélange à révéler une structure en classes dans les données. Il correspond au critère BIC pénalisé par un terme d'entropie qui mesure le degré d'imbrication des composants :

$$ICL_m = BIC_m + H_m,$$

où H_m correpond à l'entropie du modèle m, avec :

$$H_m = -2 \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\tau_{ik}).$$

Les deux critères sélectionnent le modèle $\lambda_k D_k A_k D'_k$. Ce modèle est celui qui a le plus de paramètres à estimer (23 paramètres pour un mélange de 4 gaussiennes bidimensionnelles), ce qui n'est pas un problème étant donné le très grand nombre de données (environ 150 000 observations par jeu de données).
Les résultats obtenus avec le modèle $\lambda_k D_k A_k D'_k$ ne nous satisfont pas (cf Figure 2). En effet, un seul composant couvre les groupes enrichi et appauvri et trois composants sont presque concentriques autour du groupe d'ADN non immunoprécipité (bruit). Ceci est dû au fait que la densité de points est beaucoup plus importante au niveau du groupe d'ADN non immunoprécipité qui regroupe environ 50% des données.

Les modèles non choisis par les critères BIC et ICL et qui considèrent un volume, λ, constant pour les quatres composants sont un peu meilleurs du point de vue de l'interprétation, mais deux composants sont très chevauchants et on ne retrouve pas le groupe d'ADN non immunoprécipité. Beaucoup de sondes sont alors classées dans le groupe appauvri à tort (cf Figure 3).

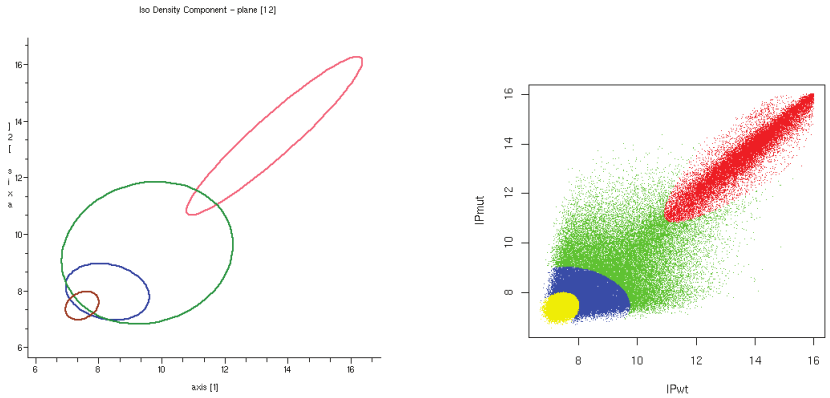

FIG. 2 – Droite : Isodensité des 4 gaussiennes pour le modèle $\lambda_k D_k A_k D'_k$, Gauche : Classement des sondes en 4 groupes avec la règle du MAP

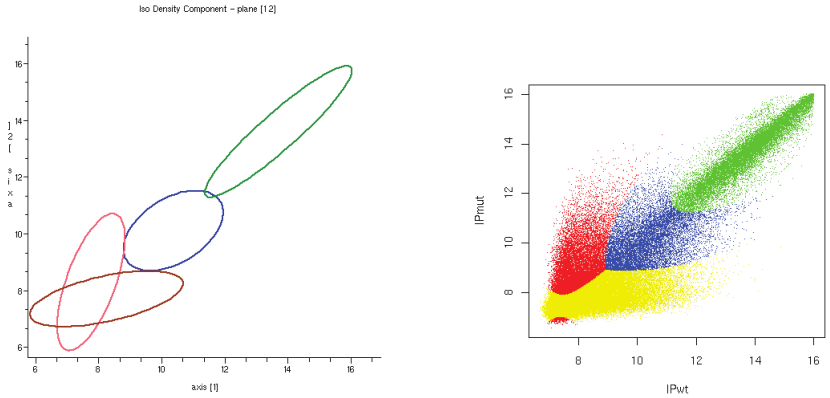

FIG. 3 – Droite : Isodensité des 4 gaussiennes pour le modèle $\lambda D_k A_k D'_k$, Gauche : Classement des sondes en 4 groupes avec la règle du MAP

4 Modélisation des données avec prise en compte des connaissances biologiques

4.1 Modélisation

Afin de modéliser au mieux les données, nous ajoutons des contraintes aux modèles détaillés section 3.3. Les contraintes supplémentaires sont déduites de connaissances biologiques que nous avons sur les données. En effet, nous avons vu dans la section 2 que l'on s'attend à identifier 4 groupes différents lorsqu'on analyse des données de ChIP-chip IP/IP. Le nombre de composants du modèle de mélange est donc fixé à K=4. De plus, nous avons certaines connaissances sur les 4 groupes que l'on souhaite identifier : le groupe d'ADN non immunoprécipité et le groupe normal ont la même orientation proche de la première bissectrice. D'autre part, on suppose que le bruit est égal dans chaque groupe, ce qui revient à fixer la deuxième valeur propre de Σ_k. En effet, la première valeur propre est associée au grand axe de l'ellipse et la deuxième est associée au petit axe de l'ellipse.

Cette dernière hypothèse est utile pour contraindre la modélisation car des variances hétéroscédastiques donnent souvent des résultats très instables et ne permettent pas de retrouver les 4 groupes de la Figure 1.
Nous reprenons la paramétrisation définie section 3.2 :

$$\Sigma_k = \lambda_k D_k A_k D_k'.$$

Afin d'avoir le même bruit dans chaque groupe, on contraint la seconde valeur propre de Σ_k à être constante dans les 4 groupes. Les deux groupes qui ont la même orientation auront la même matrice D. En utilisant la décomposition des matrices de variance et sous nos contraintes, on obtient donc :

$$\begin{cases} \Sigma_k &= \lambda_k D_k A_k D_k' = D_k \Lambda_k D_k', \text{ pour } k = 1, .., 4, \text{ avec } \Lambda_k = \lambda_k A_k \\ D_1 &= D_2 = D \\ \Lambda_k &= \begin{pmatrix} u_{1k} & 0 \\ 0 & u_2 \end{pmatrix}, \text{ avec } u_{1k} > u_2, \text{ pour } k = 1, .., 4. \end{cases}$$

De manière plus générale, on peut écrire :

$$\begin{cases} \Sigma_k &= D_k' \Lambda_k D_k \ si \ k \geq 2 \\ \Sigma_k &= D' \Lambda_k D \ si \ k < 2, \end{cases}$$

où les groupes 1 et 2 correspondent aux groupes de même orientation (groupe normal et groupe d'ADN non immunoprécipité) et la matrice Λ_k est une matrice diagonale qui contient les valeurs propres de Σ_k.

L'originalité de ce modèle est de proposer la possibilité d'avoir certains composants avec une orientation fixe et d'autres composants avec une orientation libre. De plus il est possible de fixer seulement l'une des deux valeurs propres dans le choix du volume et de la forme pour un même composant du modèle. Dans le logiciel MIXMOD [4], le choix de fixer ou pas le volume, l'orientation ou la forme est obligatoirement le même pour tous les composants du modèle.

4.2 Estimation des paramètres par l'algorithme EM

Si le label de chaque donnée était observé, l'estimation des paramètres du mélange serait évidente puisque les paramètres de chaque composant $\phi(x_i; \mu_k, \Sigma_k)$ seraient estimés avec les individus de la population k. Mais les labels sont inconnus et l'estimation ne peut être fondée que sur les données observées $x_1, ..., x_n$. Il n'existe pas de formules explicites pour les estimateurs des paramètres d'un mélange, on a besoin de procédures d'estimation itératives. Le vecteur de paramètres $\Psi = (\pi_1, ..., \pi_3, \mu_1, ..., \mu_4, \Sigma_1, ..., \Sigma_4)$ est estimé à l'aide de l'algorithme EM.
Pour trouver l'estimateur des matrices de variance Σ_k, il faut maximiser l'espérance de la log-vraisemblance des données complétées en Σ_k, ce qui revient à minimiser F en D, D_k et Λ_k, où F est définie par :

$$F = \sum_{k=1}^{2} tr(D' W_k D \Lambda_k^{-1}) + \sum_{k=3}^{4} tr(D_k' W_k D_k \Lambda_k^{-1}) + \sum_{k=1}^{4} n_k \log \{\det(\Lambda_k)\},$$

où $W_k = \sum_{i=1}^n \tau_{ik}(x_i - \bar{x}_k)(x_i - \bar{x}_k)'$.
On remarque que seul Λ_k est présent dans les 3 termes de F. Pour D et D_k, minimiser F revient simplement à minimiser le terme où ils apparaissent. L'estimateur de D_k pour $k = 3, 4$ est le même que celui proposé par Celeux et Govaert [10] pour des composants d'orientations différentes, c'est-à-dire $\hat{D}_k$ est la matrice des vecteurs propres de W_k.

Proposition 1 *Soit* $W_k = \sum_{i=1}^n \tau_{ik}(x_i - \bar{x}_k)(x_i - \bar{x}_k)'$ *est une matrice de la forme* $\begin{pmatrix} w_{1k} & w_{2k} \\ w_{2k} & w_{4k} \end{pmatrix}$. *L'estimateur du maximum de vraisemblance de la matrice d'orientation D identique pour les deux premiers composants est de la forme* $\begin{pmatrix} \sqrt{\hat{d}} & -\sqrt{1-\hat{d}} \\ \sqrt{1-\hat{d}} & \sqrt{\hat{d}} \end{pmatrix}$, *où* $\hat{d}$ *est un réel positif défini par :*

$$\hat{d} = \begin{cases} \frac{1}{2} + \frac{\sum_{k=1}^2 (w_{1k} - w_{4k})}{2\left\{\sqrt{(\sum_{k=1}^2 (w_{1k} - w_{4k}))^2 + 4(\sum_{k=1}^2 (w_{2k}))^2}\right\}} & si \ \sum_{k=1}^2 (w_{1k} - w_{4k}) > 0 \\ \frac{1}{2} - \frac{\sum_{k=1}^2 (w_{1k} - w_{4k})}{2\left\{\sqrt{(\sum_{k=1}^2 (w_{1k} - w_{4k}))^2 + 4(\sum_{k=1}^2 (w_{2k}))^2}\right\}} & sinon. \end{cases} \tag{2}$$

Idée de la preuve 1 *Minimiser F en D revient à minimiser* $f(D) = \sum_{k=1}^2 tr(D\Lambda_k^{-1} D' W_k)$. *On peut réécrire* $f(D)$ *sous la forme suivante :*

$$f(D) = \sum_{k=1}^2 \left(\frac{d_1' W_k d_1}{u_{1k}} + \frac{d_2' W_k d_2}{u_2} \right),$$

où d_1' *est le premier vecteur de la matrice D et* d_2' *le second.*
Puisque D est une matrice orthogonale et normée, elle est de la forme $\begin{pmatrix} \sqrt{d} & -\sqrt{1-d} \\ \sqrt{1-d} & \sqrt{d} \end{pmatrix}$.
En développant $f(D)$ *et en dérivant par rapport à d, on obtient un pôlynome de degré 4 en d qui se résout facilement. On remarque alors que D ne dépend plus de* Λ. *Ce résultat analytique n'est valable qu'en dimension 2.* ■

Proposition 2 *Soit* B_k *la matrice définie par* $B_k = D_k' W_k D_k$ *de la forme* $\begin{pmatrix} b_{1k} & b_{3k} \\ b_{4k} & b_{2k} \end{pmatrix}$. *L'estimateur du maximum de vraisemblance de* Λ_k *est de la forme* $\begin{pmatrix} \hat{u}_{1k} & 0 \\ 0 & \hat{u}_2 \end{pmatrix}$, *où*

$$\begin{cases} \hat{u}_{1k} & = b_{1k}/n_k \\ \hat{u}_2 & = \sum_{k=1}^4 b_{2k}/n \end{cases} \tag{3}$$

Idée de la preuve 2 *En développant la trace et le déterminant, on peut réécrire F sous la forme :*

$$F = \sum_{k=1}^4 (b_{1k} u_{1k}^{-1} + b_{2k} u_2^{-1}) + \sum_{k=1}^4 n_k \left\{ \log(u_{1k}) + \log(u_2) \right\},$$

et minimiser F en Λ_k *revient à minimiser F en* u_{1k} *et* u_2. ■

L'estimateur de Σ_k est donc :

$$\hat{\Sigma}_k = \begin{cases} \hat{D}'_k \hat{\Lambda}_k \hat{D}_k \text{ si } k \geq 2 \\ \hat{D}' \hat{\Lambda}_k \hat{D} \text{ si } k < 2, \end{cases}$$

avec $\hat{D}$ défini par (2), $\hat{D}_k$ est la matrice des vecteurs propres de W_k et $\hat{\Lambda}_k$ défini par (3).

5 Application sur un jeu de données réel

Nous appliquons cette méthode sur les données de méthylation d'histone présentées section 2. Les données analysées concernent le chromosome 4 d'*Arabidopsis thaliana* qui est couvert par 111 699 sondes.

5.1 Initialisation de l'algorithme EM

Les résultats fournis par l'algorithme EM sont dépendants de l'initialisation. Il est important de choisir une bonne initialisation afin de ne pas tomber sur un maximum local. En pratique, on peut initialiser l'algorithme avec les résultats fournis par les différents modèles de MIXMOD [4] ou bien définir une classification initiale bien choisie. Il est souvent plus facile de définir des probabilités conditionnelles pour chaque sonde (on peut par exemple s'appuyer sur la Figure 1) que de proposer une matrice initiale Σ_k pertinente. Le critère d'arrêt choisi pour l'algorithme EM est un critère de convergence sur les paramètres avec $\varepsilon = 10^{-6}$.
Nous avons testé 11 initialisations différentes et les résultats obtenus diffèrent selon l'initialisation. Huit des 11 initialisations nous donnent le modèle auquel on s'attend biologiquement représenté schématiquement Figure 1. Mais il reste des différences : les sondes difficiles à classer qui sont au centre des 4 composants sont, selon les modèles, classées soit normales, soit appauvries, soit la moitié est classée appauvrie et l'autre moitié enrichie. Les paramètres estimés des composants ne sont alors pas les mêmes.

5.2 Critères BIC et ICL

La sélection de modèles permet de choisir le modèle minimisant le critère BIC ou le critère ICL donnés section 3.3. Le modèle minimisant à la fois le critère BIC et le critère ICL est le modèle $\lambda_k D_k A_k D'_k$ présenté Figure 2 (cf Table 1). Ce n'est pas le modèle que l'on voudrait sélectionner biologiquement. Ceci est surement dû au fait que les classes ne sont pas des gaussiennes en réalité.

5.3 Estimation des paramètres

Nous présentons les résultats du modèle initialisé avec des probabilités conditionnelles. Les paramètres du mélange estimés par l'algorithme EM sont donnés dans la Table 2. Les proportions de chacun des groupes correspondent à celles attendues par les biologistes. En effet, on sait que la méthylation de cette histone n'est présente qu'en faible proportion dans le génome. Or nous trouvons environ 39% des sondes dans le groupe non immunoprécipité. Nous savons aussi que la différence de méthylation est majoritairement appauvrie chez

	Modèle $\lambda_k D_k A_k D'_k$	Modèle $\lambda D_k A_k D'_k$	Modèle 1	Modèle 2
nb de paramètres	23	20	18	18
BIC	578 171	637 770	606 643	613 470
ICL	607 488	690 126	639 101	640 582

Tab. 1 – Critères BIC et ICL selon les modèles. Modèle 1 correspond à notre modèle initialisé avec des probabilités conditionnelles, Modèle 2 correspond à notre modèle initialisé avec des paramètres bien choisis

le mutant et très rarement enrichie. Le groupe appauvri regroupe 22% des sondes et le groupe enrichi en regroupe seulement 13%.
D'autre part, la matrice d'orientation D estimée pour les groupes 1 et 2 est très proche de la matrice d'orientation attendue pour une direction sur la première bissectrice.

	Groupe 1	Groupe 2	Groupe 3	Groupe 4
$\hat{\pi}$	0.39	0.26	0.13	0.22
$\hat{\mu}$	[7.56 ;7.54]	[12.19 ;12.04]	[8.07 ;9.17]	[9.10 ;7.95]
$\hat{D}$	$\begin{pmatrix} 0.71 & -0.70 \\ 0.70 & 0.71 \end{pmatrix}$	$\begin{pmatrix} 0.71 & -0.70 \\ 0.70 & 0.71 \end{pmatrix}$	$\begin{pmatrix} 0.32 & -0.94 \\ 0.94 & 0.32 \end{pmatrix}$	$\begin{pmatrix} -0.96 & 0.26 \\ -0.26 & -0.96 \end{pmatrix}$
$\hat{\Lambda}$	$\begin{pmatrix} 0.11 & 0 \\ 0 & 0.14 \end{pmatrix}$	$\begin{pmatrix} 9.42 & 0 \\ 0 & 0.14 \end{pmatrix}$	$\begin{pmatrix} 1.41 & 0 \\ 0 & 0.14 \end{pmatrix}$	$\begin{pmatrix} 1.29 & 0 \\ 0 & 0.14 \end{pmatrix}$
$\hat{\Sigma}$	$\begin{pmatrix} 0.12 & -0.01 \\ -0.01 & 0.12 \end{pmatrix}$	$\begin{pmatrix} 4.8 & 4.64 \\ 4.64 & 4.75 \end{pmatrix}$	$\begin{pmatrix} 0.27 & 0.39 \\ 0.39 & 1.28 \end{pmatrix}$	$\begin{pmatrix} 1.21 & 0.29 \\ 0.29 & 0.22 \end{pmatrix}$

Tab. 2 – Estimation des paramètres. Les groupes 1 et 2 correspondent aux groupes normaux (le groupe 1 est le groupe non-immunoprécipité), le groupe 3 correspond au groupe enrichi et le groupe 4 correspond au groupe appauvri.

On obtient quatre groupes en classant chaque sonde dans le groupe pour laquelle la probabilité conditionnelle est la plus grande (cf Figure 4).
D'un point de vue biologique, le plus important est dans un premier temps de distinguer les sondes enrichies ou appauvries (c'est-à-dire là où le taux de méthylation est différent entre le sauvage et le mutant). On peut donc considérer les groupes de même orientation (groupes 1 et 2) comme un seul groupe qui correspond à un taux de méthylation identique dans les deux échantillons (groupe normal). On veut donc classer les sondes en trois groupes : normal, appauvri ou enrichi. Pour cela, on somme les probabilités conditionnelles des groupes 1 et 2. Une autre possibilité est de classer en deux groupes seulement, un groupe qui correspond à une méthylation identique dans les deux échantillons, et l'autre qui correspond à un taux de méthylation différent entre les deux échantillons. Pour cela,

on somme les probabilités conditionnelles des groupes 1 et 2 et celles des groupes 3 et 4. Lorsque l'on classe en 4 groupes, on trouve bien les 4 groupes comme attendus sur la figure 1, mais il est probable que les sondes aux frontières de deux classes aient des probabilités conditionnelles très proches pour les deux classes et soient donc mal classées. Comme nous préférons ne pas avoir d'information sur une sonde plutôt que d'avoir une information fausse, nous fixons un seuil de classification à 0.7, ce qui délimite une marge de non classement autour de chacun des groupes (cf Figure 5). Avec un seuil à 0.7, seulement 12.5% des sondes ne sont pas classées. On peut bien sûr faire de même avec les classements en 2 ou 3 groupes, le nombre de sondes non classées est alors plus faible (11.9% pour un classement en 3 groupes et 9.3% pour un classement en 2 groupes).

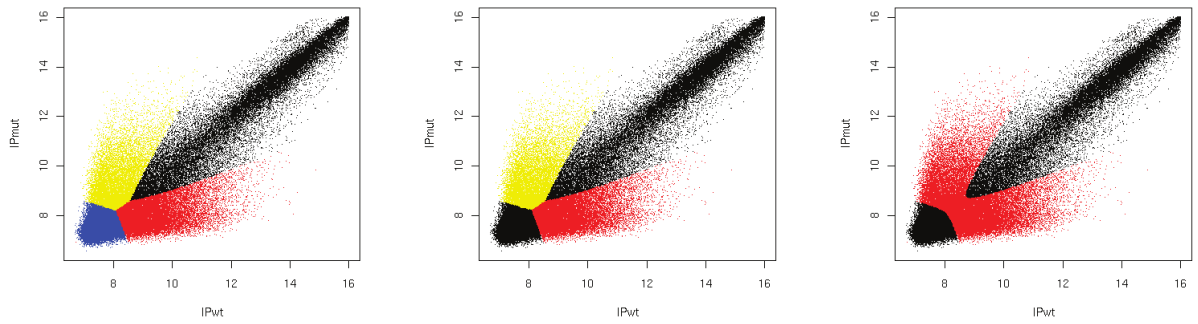

FIG. 4 – Classement des sondes en 4 groupes (gauche), 3 groupes (centre), 2 groupes (droite).

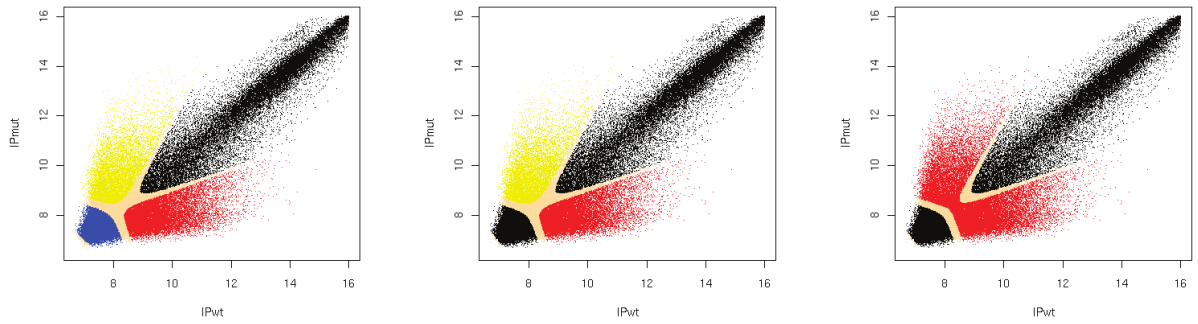

FIG. 5 – Classement des sondes en 4 groupes (gauche) avec un seuil de classification à 0.7 (zone en gris), 3 groupes (centre), 2 groupes (droite).

5.4 Interprétations biologiques

Nous avons ensuite comparé les résultats à l'annotation connue d'*Arabidopsis thaliana* à l'aide du logiciel $SignalMap^{TM}$ fourni par NimbleGen (cf Figure 6). Bien que notre modèle ne prenne pas en compte la structure spatiale des sondes le long du chromosome, les sondes déclarées normales, enrichies ou appauvries chez le mutant sont regroupées sous forme de plage. On s'attend évidemment à ce que des sondes contigües aient le même comportement. D'autre part, la marque H3K9me2 étudiée est une marque hétérochromatinienne présente sur environ 15% du génome. La plupart des régions couvertes par H3K9me2 sont contigües et couvrent plusieurs mégabases dans les régions péricentromériques ou dans l'hétérochromatine interstitielle comme le knob du chromosome 4, mais il existe aussi des régions plus petites (ilôts d'hétérochromatine) situées dans l'euchromatine et qui couvrent majoritairement des éléments transposables (Bernatavichute *et al.* [3]). Nous savons aussi qu'il y a peu de différences entre le sauvage et le mutant pour le taux de méthylation d'H3K9me2. Nos résultats corroborent parfaitement ces connaissances. En effet, on observe une majorité de sondes déclarées non méthylées le long du chromosome 4, mais dans la région péricentromérique (entre les positions 2 800 000 et 5 000 000) et autour du knob (entre les positions 1 600 000 et 2 300 000), on remarque une majorité de sondes du groupe normal et des larges plages de sondes du groupe appauvri ou enrichi. On détecte aussi des plages de sondes, plus petites, appartenant au groupe enrichi ou appauvri situées dans l'euchromatine et qui couvrent des éléments transposables (cf Figure 6). Environ 10% des sondes du génome couvrent des éléments transposables, et on trouve 26% des sondes du groupe normal couvrant un élément transposable. Un test du χ^2 montre une différence significative. Il y a clairement un biais et on peut donc dire que la marque H3K9me2 est majoritairement présente sur les éléments transposables.

FIG. 6 – Comparaison à l'annotation. En bleu les gènes (1ère ligne), en violet les éléments transposables (2ème ligne). En rouge, les sondes où la méthylation est identique entre le sauvage et le mutant. En noir les sondes déclarées enrichies, en bleu les appauvries, en jaunes les non-méthylées (3ème ligne).

6 Conclusion

Nous proposons une méthode fondée sur un mélange de gaussiennes bidimensionnelles contraintes pour l'analyse de données de ChIP-chip IP/IP. La connaissance biologique des données est prise en compte. Les paramètres sont estimés par l'algorithme EM. Cette méthode donne des résultats convaincants pour l'analyse d'un jeu de données réel concernant la méthylation d'une histone. Nous souhaitons aussi analyser d'autres types de données où il n'y aurait que 3 groupes à définir (pas d'appauvri, pas d'enrichi ou pas de non-immunoprécipité). Bien que notre modèle ne prenne pas en compte la structure spatiale des sondes le long du chromosome, les sondes déclarées normales, enrichies ou appauvries sont regroupées sous forme de plage. Une amélioration naturelle consiste à prendre en compte la structure spatiale des sondes en utilisant un modèle de type HMM. D'autre part, on peut aussi rajouter des contraintes de symétrie entre les groupes appauvri et enrichi.

Références

[1] Amaratunga, D. and Cabrera, J. : Exploration and Analysis of DNA Microarray and Protein Array Data. *Wiley Series in Probability and Statistics* (2004).

[2] Banfield, J.D. and Raftery, A.E. : Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** (1993) 803-821.

[3] Bernatavichute, Y.V., Zhang, X., Cokus, S., Pellegrini, M. and Jacobsen, S.E. : Genome-Wide Association of Histone H3 Lysine Nine Methylation with CHG DNA Methylation in Arabidopsis thaliana. *PLoS ONE* **3(9) :e3156** (2008).

[4] Biernacki, C., Celeux, G., Echenim, A., Govaert, G. and Langrognet, F. : Le logiciel MIXMOD d'analyse de mélange pour la classification et l'analyse discriminante. *La Revue de Modulad* **35** (2007) 25-44.

[5] Biernacki, C., Celeux, G. and Govaert, G. : Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence* **22(7)** (2000) 719-725.

[6] Boulicaut, J.F. and Gandrillon O. : Informatique pour l'analyse du transcriptome. *Lavoisier* (2004).

[7] Buck, M.J. and Lieb, J.D. : Chip-chip : considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83** (2004) 349-360.

[8] Cappé, O., Moulines, E. and Rydén, T. : Inference in hidden Markov models. *Springer Series in Statistics, NY : Springer* (2005).

[9] Cawley, S. *et al.* : Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116** (2004) 499-509.

[10] Celeux, G. and Govaert, G. : Gaussian Parsimonious Clustering Models. *Pattern Recognition* **28** (1995) 781-793.

[11] Dempster, A.P., Laird, N.M. and Rubin, D.B. : Maximum likelihood from incomplete data via the EM algorihtm (with discussion). *J. R. Statis. Soc. B* **39** (1977) 1-38.

[12] Ephraim, Y. and Merhav, N. : Hidden Markov processes. *IEEE Transactions on Information Theory* **48(6)** (2002) 1518-1569.

[13] Hupé, P., Stransky, N., Thiery, JP., Radvanyi, F. and Barillot, E. : Analysis of array CGH data : from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20(18)** (2004) 3413-22.

[14] Keles, S. : Mixture modeling for genome-wide localization of transcription factors. *Biometrics* **63** (2007) 10-21.

[15] Martin-Magniette, M-L, Mary-Huard, T., Berard, C. and Robin S. : ChIPmix : mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics* **24 :i181-i186** (2008).

[16] Olshen, AB., Venkatraman, ES., Lucito, R. and Wigler, M. : Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5(4)** (2004) 557-72.

[17] Picard, F., Robin, S., Lavielle, M., Vaisse, C. and Daudin, JJ. : A statistical approach for array CGH data analysis. *BMC Bioinformatics* **6 :27** (2005).

[18] Turck, F., Roudier, F., Farrona, S., Martin-Magniette, M-L. *et al.* : Arabidopsis TFL2/LHP1 Specifically Associates with Genes Marked by Trimethylation of Histone H3 Lysine 27. *PLoS Genet.v* **3 :6** (2007).

Annexe C

Article publié dans *The Plant Journal*

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2011) doi: 10.1111/j.1365-313X.2011.04628.x

Additive inheritance of histone modifications in *Arabidopsis thaliana* intra-specific hybrids

Ali M. Banaei Moghaddam[1], Francois Roudier[2,†], Michael Seifert[1,†], Caroline Bérard[3], Marie-Laure M. Magniette[3], Raheleh Karimi Ashtiyani[1], Andreas Houben[1], Vincent Colot[2] and Michael F. Mette[1,*]

[1]*Leibniz Institute of Plant Genetics and Crop Plant Research, Corrensstraße 3, 06466 Gatersleben, Germany,*
[2]*Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Centre National de la Recherche Scientifique (CNRS) UMR8197, Institut National de la Santé et de la Recherche Médicale (INSERM) U1024, Paris, France, and*
[3]*AgroParisTech, Institut National de la Recherche Agronomique (INRA) UMR518, Paris, France*

Received 14 March 2011; revised 29 April 2011; accepted 4 May 2011.
*For correspondence (fax +49 39482 5137; e-mail mette@ipk-gatersleben.de).
†These authors contributed equally to this work.

SUMMARY

Plant genomes are earmarked with defined patterns of chromatin marks. Little is known about the stability of these epigenomes when related, but distinct genomes are brought together by intra-species hybridization. *Arabidopsis thaliana* accessions and their reciprocal hybrids were used as a model system to investigate the dynamics of histone modification patterns. The genome-wide distribution of histone modifications H3K4me2 and H3K27me3 in the inbred parental accessions Col-0, C24 and Cvi and their hybrid offspring was compared by chromatin immunoprecipitation in combination with genome tiling array hybridization. The analysis revealed that, in addition to DNA sequence polymorphisms, chromatin modification variations exist among accessions of *A. thaliana*. The range of these variations was higher for H3K27me3 (typically a repressive mark) than for H3K4me2 (typically an active mark). H3K4me2 and H3K27me3 were rather stable in response to intra-species hybridization, with mainly additive inheritance in hybrid offspring. In conclusion, intra-species hybridization does not result in gross changes to chromatin modifications.

Keywords: *Arabidopsis thaliana*, epigenome, heterosis, histone methylation, intra-specific hybrids, ChIP on chip.

INTRODUCTION

Extensive studies of DNA methylation and histone modifications in *Arabidopsis thaliana* (Turck *et al.*, 2007; Zhang *et al.*, 2007, 2009; Cokus *et al.*, 2008) and rice (*Oryza sativa*) (He *et al.*, 2010; Zemach *et al.*, 2010) have revealed that plant genomes are earmarked by well-defined patterns of chromatin marks. In the context of their transcriptional activity or inactivity, particular sequence classes such as genes or repeat elements are preferentially associated with distinct patterns of DNA methylation and histone modifications (Roudier *et al.*, 2009; Teixeira and Colot, 2010).

The combining of related but distinct genomes with their respective patterns of chromatin marks in inter-species hybridization and allopolyploid formation often results in changes to chromatin marks. In hybrids of various Arabidopsis species, one parental set of ribosomal RNA genes was shown to be silenced within a few generations, but could be re-activated by interfering with either DNA methylation or histone deacetylation, suggesting a pivotal role for chromatin modification in the regulation of expression of orthologous genes (Lee and Chen, 2001; Lawrence *et al.*, 2004). Gene expression studies in synthetic allopolyploid Arabidopsis (Comai, 2000) and cotton (*Gossypium hirsutum*) (Brubaker *et al.*, 1999) revealed that gene silencing occurs during the first or second generation after hybridization. However, other studies in allopolyploid *Spartina anglica*, *Brassica juncea* and cotton showed that the activity of parental genomes remained unchanged (Axelsson *et al.*, 2000; Liu *et al.*, 2001; Baumel *et al.*, 2002). Induction of DNA methylation changes after hybridization was reported for synthetic *Cucumis* allopolyploids (Chen and Chen, 2008). Similarly, in experimentally synthesized *Brassica napus* (Xu *et al.*, 2009) and Arabidopsis allopolyploids (Madlung *et al.*, 2002), 7 and 8%, respectively, of the tested DNA sites showed changes in cytosine methylation status in comparison with their respective diploid progenitors.

© 2011 Leibniz Institute of Plant Genetics and Crop Plant Research
The Plant Journal © 2011 Blackwell Publishing Ltd

Less is known about the stability or dynamics of chromatin modifications in response to intra-species hybridization. Limited differential DNA methylation (approximately 1% gain or loss) in comparison with the respective progenitors was found for intra-species hybrids of rice (Xiong *et al.*, 1999) and cotton (Zhao *et al.*, 2008). In two rice cultivars from different sub-species, *Oryza sativa japonica* and *O. sativa indica*, variation of DNA methylation, and, to a lower extent, variation of histone modifications H3K4me3 and H3K27me3, was observed between parental lines. In reciprocal hybrids of the two rice sub-species, distinct non-additive patterns of chromatin marks were observed. The level of changes after hybridization was higher for DNA methylation, and both histone modifications were mainly inherited additively in the hybrids (He *et al.*, 2010). Inbred accessions of *A. thaliana* also display substantial DNA methylation variation between each other (Vaughn *et al.*, 2007). Investigation of the DNA methylation pattern in two different accessions of *A. thaliana* and their reciprocal F_1 hybrid progeny showed that DNA methylation polymorphisms are mostly inherited additively, with only limited changes after hybridization (Zhang *et al.*, 2008; Banaei Moghaddam *et al.*, 2010; Groszmann *et al.*, 2011).

Here, we determined whether intra-species crosses between inbred lines lead to changes in chromatin marks other than DNA methylation using accessions of *A. thaliana* as a model. We selected histone H3 dimethylated at lysine 4 (H3K4me2) and histone H3 trimethylated at lysine 27 (H3K27me3) as contrasting histone H3 modifications marks (Fuchs *et al.*, 2006; Kouzarides, 2007; Roudier *et al.*, 2009). Histone H3K4me2 was chosen as a general euchromatic mark that is absent from silent repeat elements, and H3K27me3 was chosen as a euchromatic mark that is mostly associated with genes repressed by polycomb repressive complex 2 (Schubert *et al.*, 2006; Turck *et al.*, 2007; Zhang *et al.*, 2007). A genome-wide 'ChIP on chip' analysis of H3K4me2 distribution in *A. thaliana* indicated that 6% of sequences are targeted by this mark (Zhang *et al.*, 2009). Of these target regions, 93% were genes, with particular enrichment of H3K4me2 in the promoter and 5′ end of transcribed regions, and only 1.3% of the target regions were transposable elements (TEs). In contrast, histone H3K27me3 is associated with silent genes distributed in euchromatic regions that are subject to tissue-specific or developmentally regulated expression (Turck *et al.*, 2007; Zhang *et al.*, 2007). Genome-wide analyses revealed that at least 15–20% of *A. thaliana* genes are targeted by this histone mark and show tissue-specific expression patterns. H3K27me3-marked domains are largely coincident with the entire transcribed region of genes (Turck *et al.*, 2007; Zhang *et al.*, 2007).

To study the stability of histone modification patterns in response to intra-species hybridization of various inbred accessions of *A. thaliana* (Col-0, C24 and Cvi), we performed chromatin immunoprecipitation in combination with genome tiling array hybridization (ChIP on chip) analyses. Changes in the H3K27me3 and H3K4me2 distribution between Col-0, Cvi and C24 were identified, with a greater range of variations for H3K27me3 than H3K4me2. H3K4me2 and H3K27me3 were rather stable after intra-species hybridization, with additive inheritance in Col-0 × Cvi and Col-0 × C24 F_1 hybrid offspring. Changes in the distribution of histone modifications after hybridization were detected in 346 genes for H3K4me2 in Col-0 × Cvi progeny, and in 1233 and 876 genes for H3K27me3 in Col-0 × Cvi and Col-0 × C24 progeny, respectively. However, these changes were rather random and were not associated with particular sequence categories.

RESULTS

Genomic sequence variation among *A. thaliana* accessions resides mainly in transposable elements

Comparative genomic hybridization (CGH) experiments were performed using Arabidopsis whole-genome tiling NimbleGen arrays to identify differences in the genomic sequences of *A. thaliana* accessions Cvi and C24 compared to the reference accession Col-0. The CGH analysis, which detects copy number variation and sequence polymorphisms, was a necessary prerequisite for the comparison of histone modification patterns between Cvi, C24 and Col-0. Based on this information, it was possible to distinguish whether differential hybridization signals of labelled DNA derived from immunoprecipitated chromatin from the various accessions were due to differences in histone modifications rather than variation in the DNA sequence. In total, 6.0 and 5.5% of tiles showed significant CGH polymorphisms

Table 1 CGH analysis for Col-0 versus C24 and Col-0 versus Cvi

	Col-0 versus Cvi	Col-0 versus C24
Percentage of tiles showing CGH polymorphism	6.0	5.5
Percentage of tiles with lower copy number in C24 or Cvi (total size in kb)	5.3 (5395)	5.2 (5221)
Number of tiles per domain	2–369	2–372
Mean size of domain (kb)	3.3	3.9
90% of domains had a size of less than (kb)	6	8
Size of largest domain (kb)	40	57
Percentage of tiles with higher copy number in C24 or Cvi (total size in kb)	0.67 (631)	0.35 (308)
Number of tiles per domain	2–169	2–54
Mean size of domain (kb)	3.0	1.9
90% of domains had a size of less than (kb)	6.5	4
Size of largest domain (kb)	26	9

© 2011 Leibniz Institute of Plant Genetics and Crop Plant Research

for Col-0 versus Cvi and Col-0 versus C24, respectively (Table 1). Most of the CGH polymorphic tiles indicated a decrease in copy number of the corresponding sequence in C24 and Cvi compared to Col-0.

Further analysis focused on CGH polymorphic domains containing two or more consecutive CGH polymorphic tiles (Table 1). Of the tiles present in CGH polymorphic domains identified between Col-0 and Cvi, as well between Col-0 and C24, 93% were identical. However, copy number varied in some cases in opposite directions between accessions, with an increase in Col-0 versus Cvi and a decrease in Col-0 versus C24, and vice versa. Annotation of tiles within CGH polymorphic domains indicated that copy number variation mainly affects TEs, while genic regions and 5′ and 3′ untranslated regions (UTRs) are more conserved between the analysed accessions (Figure 1). Ontology categorization of the genes (excluding TEs) associated with CGH polymorphic domains according to the Munich Information Center for Protein Sequences (MIPS) Functional Catalogue (Ruepp *et al.*, 2004) indicated an excess of genes with functions in signal transduction (Figure 2, category 30) or cell defence (Figure 2, category 32). As annotated TEs were excluded prior to ontology analysis, they are absent from the gene ontology data (Figure 2, category 38). In addition, unclassified proteins (Figure 2, category 99) were common.

In summary, the CGH data revealed sequence polymorphisms among the analysed accessions of *A. thaliana.* These polymorphisms occurred in all types of sequences, but TEs and genes involved in signal transduction and cell defence were over-represented.

Genome-wide patterns of histone modifications H3K4me2 and H3K27me3 show variation among *A. thaliana* accessions

Next, we performed ChIP on chip analysis using Arabidopsis whole-genome tiling NimbleGen arrays to determine the genome-wide distribution of H3K4me2 in accessions Col-0 and Cvi and in Col-0 × Cvi F_1 hybrids, and of H3K27me3 in Col-0, Cvi and C24, and in Col-0 × Cvi and Col-0 × C24 F_1 hybrids. The quality of ChIP assays was confirmed by quantitative PCR using primers specific for sequences known to be associated with H3K4me2 or H3K27me3 (Figure S1) from previous studies (Turck *et al.*, 2007; Zhang *et al.*, 2007, 2009).

Tiles showing polymorphic hybridization signals in CGH were excluded from the ChIP on chip analysis to avoid the influence of sequence differences among accessions. Tiles that were found to be significantly associated with a given modification in any of the analysed genotypes were assigned to histone modification domains of at least three successive tiles that correspond to a region of at least 0.3 kb (Table S1). Given the mean size of chromatin fragments (0.8 kb) and the resolution provided by the NimbleGen microarrays (165 bp) used in the ChIP on chip experiments,

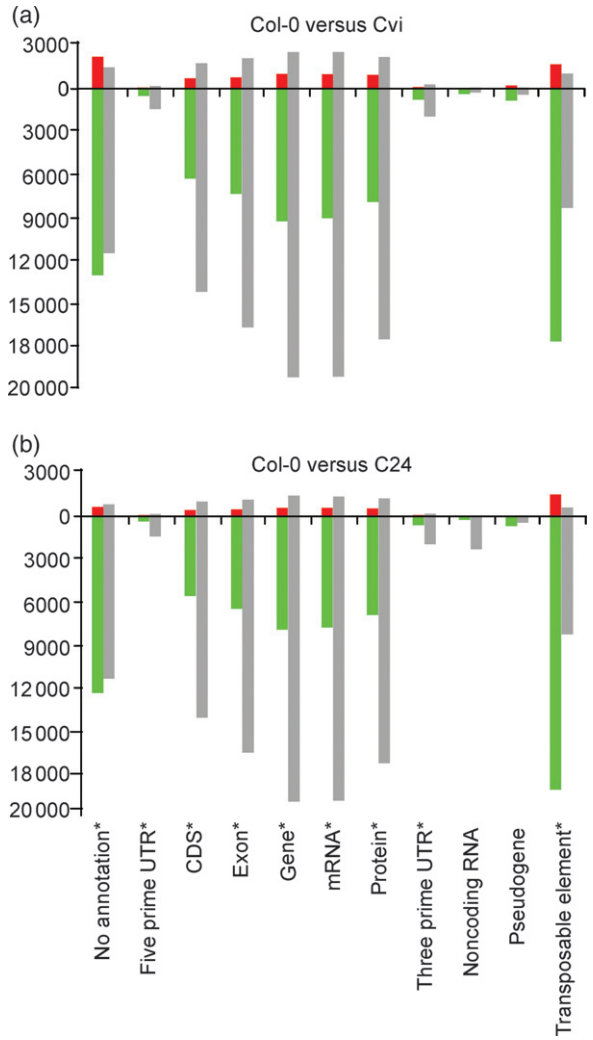

Figure 1. Classification of CGH polymorphic tiles based on their sequence types.
CGH polymorphic tiles between *A. thaliana* accessions (a) Col-0 versus Cvi and (b) Col-0 versus C24, were classified based on annotation (TAIR8) of their underlying sequences. Red and green bars represent tiles with higher and lower copy numbers in Cvi (a) and C24 (b) relative to Col-0, respectively. Grey bars indicate controls by random counts. Cases of significant deviation between CGH data (green or red bars) and random counts (grey bars) as indicated by a *P* value <0.01 are indicated by asterisks.

tiles that could not be assigned to such domains were not considered for further analysis. Subsequently, histone modification domains were categorized according to the TAIR8 annotation into genes or TEs. In the case of large domains, it may be that one domain simultaneously harbours genic sequences and TEs.

First the distribution of histone modifications was compared between parental accessions Col-0 and Cvi, as well as Col-0 and C24 (Figure S2). H3K4me2, a classical euchromatic histone mark, was associated with domains ranging in length from 314 bp to 31.2 and 21.9 kb in Col-0 and Cvi, respectively (Table 2 and Figure S3). Overall, 94% of these

© 2011 Leibniz Institute of Plant Genetics and Crop Plant Research
The Plant Journal © 2011 Blackwell Publishing Ltd,

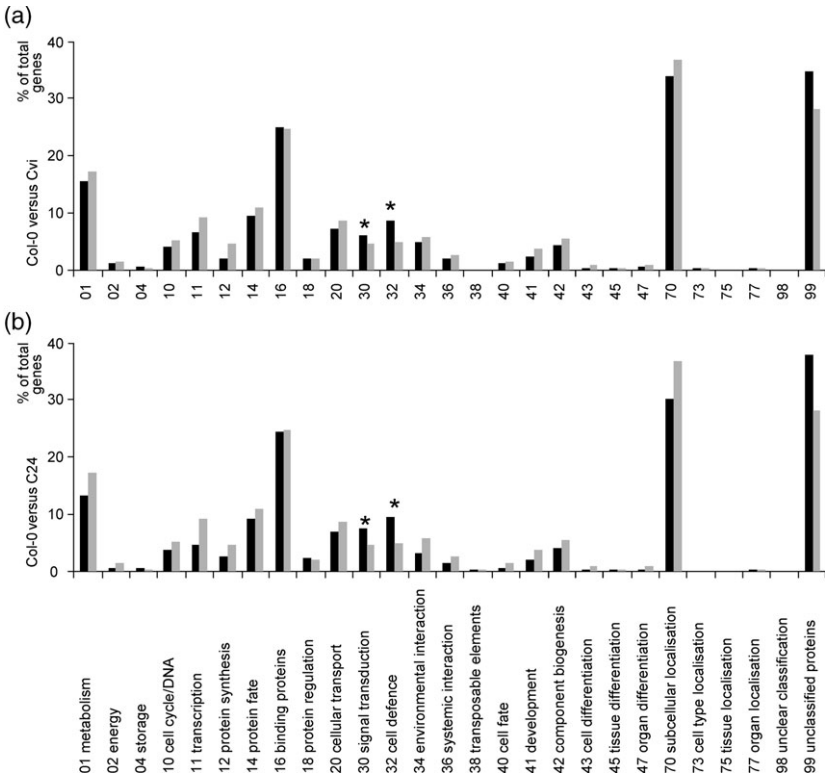

Figure 2. Ontology classification of genes localized in CGH polymorphic regions.
Genic sequences (according to TAIR8) that were found to be localized in CGH polymorphic regions indentified between (a) Col-0 and Cvi and (b) Col-0 and C24 were classified into gene ontology groups according to the MIPS Functional Catalogue (Ruepp *et al.*, 2004) to identify potentially over-represented categories. Black bars indicate the frequencies of CGH polymorphic genic sequences; as a control, grey bars indicate the frequency of randomly selected genic sequences in the various gene ontology classes as a percentage of the total number of genes in each class. Asterisks indicate cases of significant deviation between the frequencies of polymorphic and randomly selected sequences in a given category with a *P* value <0.001.

domains coincided with genes (Table S2) and 9% with TE sequences (Table S3). Intersection analysis of H3K4me2-marked domains in Col-0 and Cvi revealed that 87% of them were common between both accessions. These common domains coincided with 93% (21 908) of the genes and 60% (1526) of the TE sequences associated with H3K4me2 (Figure 3). The length of the 13% of H3K4me2-marked regions that differed between the two accessions (Table 3) ranged from 314 bp to 6.7 kb, and corresponded to 4.6% (1053) and 2.6% (581) of the genes and 26% (537) and 23.9% (479) of the TE sequences in Col-0 and Cvi, respectively (Table 3). Overall, a larger proportion of TEs than genes showed H3K4me2 polymorphisms. However, the absolute numbers of genes and TEs showing differential association with H3K4me2 between Col-0 and Cvi were similar, as fewer TEs than genes are associated with this histone modification.

Genes that were differentially associated with H3K4me2 in either of the parental lines were randomly distributed in various gene ontology groups according to the MIPS Functional Catalogue (Figure S4) (Ruepp *et al.*, 2004) and the DAVID tool (Table S4) (Huang *et al.*, 2009). Similar results were obtained for the distribution of H3K4me2 in and between Col-0 and C24 accessions analysed using a chromosome 4 tiling array (data not shown) (Turck *et al.*, 2007).

H3K27me3, a histone mark associated with the repression of genes, was found over domains ranging from 315 bp to 26.7 kb in length (Table 2 and Figure S3) in Col-0, Cvi and C24. In Col-0, 67% of these domains coincided with genes and 36% with TEs. Similar results were obtained for Cvi and

Table 2 H3K4me2- and H3K27me3-associated domains in Col-0, Cvi and C24

Experiment[a]	Genotype	Modification	Associated tiles (from total of 717 235)	Number of domains	Domain size (kb)			Annotation	
					Maximum	Minimum	Mean	Genes	TEs
A	Col-0	H3K4me2	287 285	21 539	31.2	0.3	2.2	22 961	2063
	Cvi	H3K4me2	281 000	20 972	21.9	0.3	2.2	22 489	2005
B	Col-0	H3K27me3	140 035	11 182	22.7	0.3	2.1	9125	4950
	Cvi	H3K27me3	138 791	12 585	20	0.3	1.9	9326	5357
C	Col-0	H3K27me3	147 318	9834	26.7	0.3	2.5	9427	4704
	C24	H3K27me3	125 947	10 068	22.5	0.3	2.1	8118	4471

[a]Chromatin preparations for experiments A and B were performed in parallel but the one for experiment C was performed independently.

© 2011 Leibniz Institute of Plant Genetics and Crop Plant Research

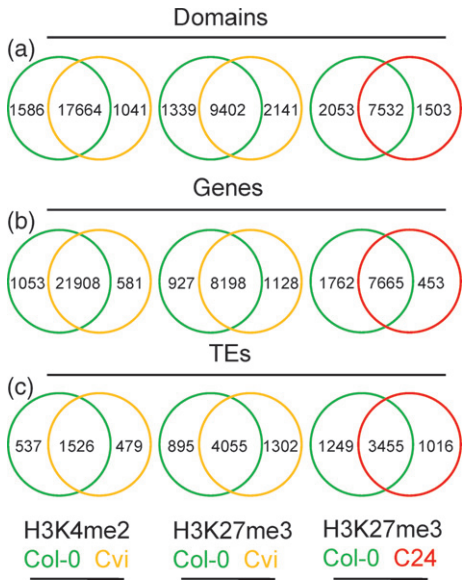

Figure 3. Intersection analysis of domains, genes and TEs associated with H3K4me2 or H3K27me3 in three *A. thaliana* accessions.
The sets of (a) domains, (b) genes and (c) TEs that were found to be associated with H3K27me3 or H3K4me2 in ChIP on chip experiments using whole-genome NimbleGen tiling arrays in *A. thaliana* accessions Col-0 (green circles), Cvi (yellow circles) and C24 (red circles) were subjected to intersection analysis to determine common elements (intersection area of two circles) and unique elements (areas outside intersections). Due to the analysis method, the sums of domain numbers do not necessarily correspond to the total numbers of domains in Table 2.

C24 (Figure 3, Table 2, and Tables S2 and S3). Intersection analysis of H3K27me3-marked domains revealed that 87% of them were common between Col-0 and Cvi. These common domains coincided with 80% (8198) of the genes and 58% (4055) of the TEs marked by H3K27me3 (Figure 3). Compared to H3K4me2, fewer genes and more TEs were associated with H3K27me3 in Col-0 and Cvi (Figure 3). Some H3K27me3 domains were unique to one of the analysed accessions (Table 3). The length of these H3K27me3 polymorphic domains ranged from 314 bp to 6.6 kb in Col-0 and 5.7 kb in Cvi. These domains coincided with 9% (927) and 11% (1128) of the genes, and 17% (895) and 25% (1302) of the TEs associated with H3K27me3, respectively. Consistent results were obtained for comparison of H3K27me3 polymorphic domains between Col-0 and C24 (Figure 3 and Table 3). Thus, in general, more TEs coincided with H3K27me3 than with H3K4me2 polymorphic domains. Similar numbers of genes and TEs were differentially associated with H3K27me3 in Col-0 compared to Cvi and Col-0 compared to C24. These genes were randomly distributed in various gene ontology groups (Figure S4 and Table S4).

Comparison of genes associated with H3K27me3 in the reference accession Col-0 with those found in previous studies (Turck *et al.*, 2007; Zhang *et al.*, 2007) revealed approximately 80% overlap, indicating good reproducibility despite the differences in Plant materials and Experimental Procedures (Figure S5). In addition, our study identified H3K27me3-associated genes that were not found in previous studies.

Taken together, our findings indicate that H3K4me2 and H3K27me3 distribution patterns are highly conserved in different A. *thaliana* accessions, with few local variations in a random manner regardless of TEs and gene ontology.

Intra-species hybridization mediates limited alterations in H3K4me2 and H3K27me3 distribution patterns

We next investigated whether intra-species hybridization could generate alternative distribution patterns of H3K4me2 in hybrid offspring between Col-0 and Cvi (Figure 4a–c, Figure S2, and Tables S1–S3). In order to focus on the most relevant cases, the analysis was performed on domains that were associated with H3K4me2 exclusively in both parental accessions, but not in F_1 hybrid progeny (Figure 4d, category IV), or in the F_1 hybrid progeny, but not in the parental accessions (Figure 4d, category III). The few hybridization-responsive domains identified (3% of all H3K4me2-associated domains) corresponded to 346 genes and 226 TEs (1.5 and 8% of H3K4me2-associated genes and TEs, respectively) and their length ranged from 0.3 to 1.6 kb (Table 4). In comparison to genes, TEs were over-represented among hybridization-responsive H3K4me2 domains. The hybridization-responsive H3K4me2-associated genes were randomly distributed between various ontology groups and gene families (Figure S6 and Table S4). Similar results were

Table 3 H3K4me2 and H3K27me3 polymorphic domains in Col-0 versus Cvi and Col-0 versus C24

Experiment[a]	Specific for accession	Modific ation	Number of domain	Domain size (kb)			Annotation	
				Maximum	Minimum	Mean	Genes	TEs
A	Col-0	H3K4me2	1585	6.6	0.3	0.67	1053	537
	Cvi	H3K4me2	1041	4.8	0.3	0.72	581	479
B	Col-0	H3K27me3	1339	4.6	0.3	0.72	927	895
	Cvi	H3K27me3	2141	5.7	0.3	0.81	1128	1302
C	Col-0	H3K27me3	2053	7.4	0.3	0.9	1762	1249
	C24	H3K27me3	1503	6.7	0.3	0.77	453	1016

[a]Chromatin preparations for experiments A and B were performed in parallel but the one for experiment C was performed independently.

© 2011 Leibniz Institute of Plant Genetics and Crop Plant Research
The Plant Journal © 2011 Blackwell Publishing Ltd,

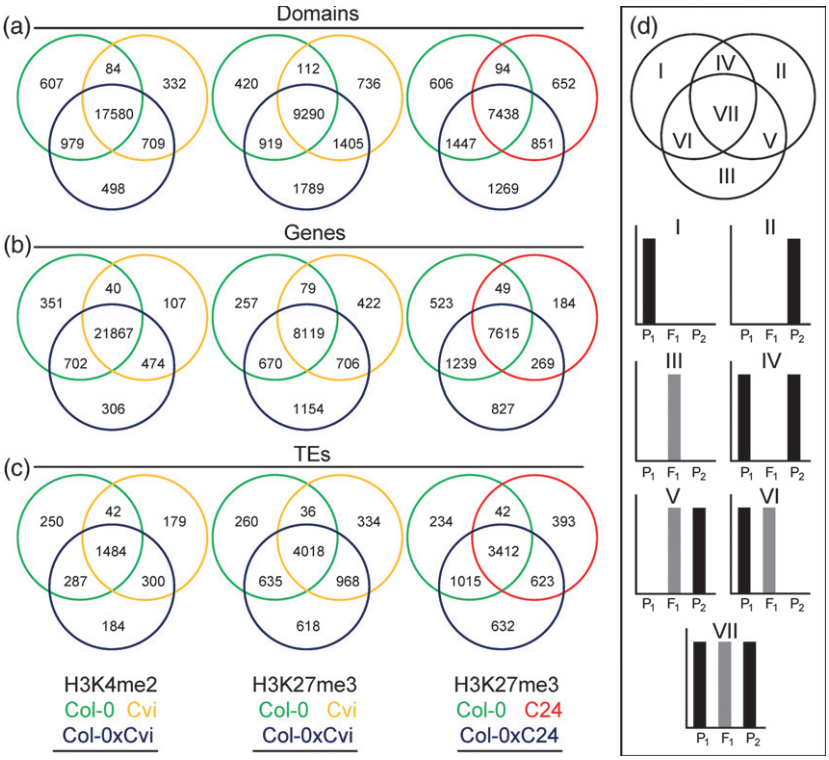

Figure 4. Intersection analysis of domains, genes and TEs associated with H3K4me2 or H3K27me3 in *A. thaliana* accessions and their F_1 hybrids.
(a–c) The sets of (a) domains, (b) genes and (c) TEs that were found to be associated with H3K27me3 or H3K4me2 in ChIP on chip experiments using whole-genome NimbleGen tiling arrays in *A. thaliana* parental accessions Col-0 (green circles), Cvi (yellow circles) and C24 (red circles), and their F_1 hybrids Col-0 × Cvi, and Col-0 × C24, respectively (blue circles), were subjected to intersection analysis.
(d) The seven areas of each three-circle intersection diagram define seven categories of elements. Categories I and II comprise cases in which a histone mark is detected in one of the parental accessions (P_1 or P_2, respectively), but not in the other parental accession or the F_1 hybrid. Category III comprises cases in which a histone mark is detected in neither parental accession, but is present in the F_1 hybrid. Category IV comprises cases in which a histone mark is detected in both parental accessions, but not in the F_1 hybrid. Categories V and VI comprise cases in which a histone mark is detected in one of the parental accessions (P_1 or P_2, respectively) and in the F_1 hybrid, but not in the other parental accession. Category VII comprises cases in which a histone mark is detected in both parental accessions and the F_1 hybrid. Categories III and IV indicate changes in histone marks in F_1 hybrids in comparison with the parental accessions, while the other categories either indicate no change (category VII) or are not conclusive with regard to changes (categories I, II, V and VI).

Table 4 H3K4me2 and H3K27me3 hybridization-responsive domains in Col-0 × Cvi and Col-0 × C24

			Hybridization-responsive domains (Figure 4, categories III and IV)					
				Domain size (kb)			Annotation	
Experiment[a]	Crosses	Modifications	Number of domains	Maximum	Minimum	Mean	Genes	TEs
A	Col-0 × Cvi	H3K4me2	582	1.6	0.3	0.6	346	226
B	Col-0 × Cvi	H3K27me3	1901	2.2	0.3	0.6	1233	654
C	Col-0 × C24	H3K27me3	1363	2.4	0.3	0.6	876	674

[a]Chromatin preparations for experiment A and B were performed in parallel but the one for experiment C was performed independently.

found for Col-0 × C24 F_1 hybrids in ChIP on chip analysis using the chromosome 4 tiling array (data not shown).

Whole-genome comparison (Figure S2) and intersection analysis of H3K27me3-marked domains, genes and TE sequences involved were also performed for Col-0, Cvi and Col-0 × Cvi (Figure 4). Hybridization-responsive domains (13% of all H3K27me3-associated domains; Figure 4d, categories III and IV) corresponded to 1233 genes and 654 TEs (11 and 10% of total H3K27me3-associated genes and TEs, respectively), and ranged in size from 0.3 to 2.2 kb (Table 4). No over-representation of either TEs or genes was detected. The hybridization-responsive H3K27me3-marked genes were randomly distributed between ontology groups and gene families (Figure S6 and Table S4). Similar

© 2011 Leibniz Institute of Plant Genetics and Crop Plant Research

H3K27me3 dynamics were found in Col-0 × C24 hybrids (Figure 4a–c and Tables S1–S3).

In conclusion, the genome-wide distribution of H3K4me2 is rather stable and that of H3K27me3 is slightly more dynamic in response to intra-species hybridization. TEs as well as genes showed changes in both histone modifications. No obvious over-representation of particular gene ontology groups or families was found. The changes in the two marks analysed happened largely independently. For very few genes, histone modifications changed simultaneously in response to hybridization (Figure S7).

DISCUSSION

By combining comparative genomic hybridization and epigenomic profiling, we have shown that H3K4me2 and H3K27me3 distribution patterns are overall very similar in various *Arabidopsis thaliana* accessions, and remain largely unchanged in their F_1 progeny.

Sequence polymorphisms between *A. thaliana* accessions are mainly found in transposable elements and genes undergoing rapid evolution

Our CGH analysis of the genomes of Cvi and C24 relative to Col-0 identified regions showing DNA sequence polymorphism between these accessions of *A. thaliana*. The number of polymorphic regions identified by CGH was slightly higher for Cvi than for C24 in comparison with Col-0. This is consistent with a previous analysis based on single nucleotide polymorphisms, which indicated that Cvi is more divergent from Col-0 than C24 is (Schmid *et al.*, 2003; Clark *et al.*, 2007).

As TE sequences change more rapidly than genes (Kazazian, 2004), sequence differences between *A. thaliana* accessions are not randomly distributed, and TEs differ more than other parts of the genome (Figure 1). A similar non-random distribution of sequence polymorphisms was previously found by CGH analysis between *A. thaliana* accessions (Clark *et al.*, 2007) and between maize (*Zea mays*) inbred lines B73 and Mo17 (Springer *et al.*, 2009). In maize, approximately 70% of polymorphic regions were located in intergenic regions.

Among the CGH polymorphic regions, gene ontology analysis indicated over-representation of functions associated with signal transduction and cell defence (Figure 2). Copy number variation between accessions of *A. thaliana* has previously been reported for loci involved in plant disease resistance (Noel *et al.*, 1999). Furthermore, consistent with our findings, a comparison of 20 accessions of *A. thaliana* by microarray-based whole-genome re-sequencing revealed that members of defence-related families such as nucleotide-binding leucine-rich repeat genes and receptor-like kinase genes are over-represented among genes that are affected by sequence polymorphisms (Clark *et al.*, 2007). The enhanced rate of sequence variation indicates rapid evolution of these gene families.

Fewer CGH polymorphic tiles indicated an increase in copy number in C24 and Cvi compared with those that indicated a decrease in copy number in both accessions (Figure 1). This could be due to the fact that the tiling arrays used in this study are based on the Col-0 reference sequence. Thus, sequences that solely exist in Col-0 but not in C24 or Cvi are classified as absent from C24 or Cvi, whereas sequences that solely exist in C24 or Cvi but not in Col-0 are not detected. Therefore, our CGH results probably under-estimate the level of sequence polymorphism between *A. thaliana* accessions. CGH-identified polymorphic sequences were excluded from the ChIP on chip analysis, as it was impossible to distinguish whether differences in signal intensities between the accessions were due to differences in the DNA sequence or the histone modification status. Thus, the possibility cannot be excluded that some fast-evolving genomic regions that were removed from analysis showed differential chromatin marking between accessions.

Histone modification patterns are conserved between accessions of *A. thaliana*

Several previous studies have detected differential DNA methylation in various *A. thaliana* accessions (Vaughn *et al.*, 2007; Zhang *et al.*, 2008; Banaei Moghaddam *et al.*, 2010). Not only sequence conservation, but also DNA methylation patterns, were found to be more similar between Col-0 and C24 than between Col-0 and Cvi (Vaughn *et al.*, 2007). Our ChIP on chip data extend these observations to histone H3K4me2 and H3K27me3 distribution patterns. These differ between accessions of *A. thaliana*, consistent with the situation observed in cultivars of rice (He *et al.*, 2010).

Among domains with different histone modification status between accessions, the mean length was <1 kb for both histone marks, while the mean length of the conserved H3K4me2- and H3K27me3-marked domains exceeded 2 kb. Thus, polymorphic histone modification domains are restricted to rather small regions. Despite their shorter length, these differentially marked domains may be associated with locus-specific differential regulation in the various accessions. For instance, TE sequences next to genes can affect their transcriptional regulation through deposition of repressive chromatin modifications. A TE in an intron of *FLOWERING LOCUS C* (*FLC*) in *A. thaliana* accession Landsberg *erecta* causes transcriptional inactivation of this locus, and consequently earlier flowering of L*er* in comparison to Col-0 (Liu *et al.*, 2004). TEs also alter the expression of adjacent genes in wheat (*Triticum aestivum*) (Kashkush *et al.*, 2003). It remains to be determined whether the regions differentially marked by the repressive H3K27me3 or active H3K4me2 modifications in Cvi or C24 in comparison to Col-0 also show differential expression between accessions.

© 2011 Leibniz Institute of Plant Genetics and Crop Plant Research
The Plant Journal © 2011 Blackwell Publishing Ltd,

Consistently, we found that polymorphisms in H3K27me3 (typically a repressive mark) were associated with both TEs and genes. In contrast, polymorphisms in H3K4me2 (typically an active mark) were mainly restricted to genes. This agrees well with genome-wide high-resolution analyses of histone modifications in *Saccharomyces cerevisiae* and mammals, which detected H3K4me2 in regions undergoing transcription across the body (Pokholok *et al.*, 2005) and in the vicinity of active genes (Bernstein *et al.*, 2005). Histone modification polymorphisms in genes were not associated with particular gene ontology classes.

The mechanisms by which histone modification polymorphisms are maintained over generations are not clear (Saze, 2008). On the one hand, heritable maintenance of particular chromatin states cannot be excluded. On the other hand, chromatin modifications may alter as consequences of changes in transcriptional activity. In *A. thaliana*, H3K27me3 is deposited by polycomb repressive complex 2 (Schubert *et al.*, 2006), and has been found to be associated with approximately 4400 genes, many of which are differentially expressed during development (Zhang *et al.*, 2007). As accessions may have evolved specific developmental programs (Chen, 2010), some of the H3K27me3 polymorphisms observed in our study could correspond to differences in gene expression patterns. Indeed, gene activity and H3K4 and H3K27 methylation levels were correlated in different rice cultivars (He *et al.*, 2010).

The observed similarities and differences suggest a role for chromatin modifications in addition to that of DNA sequence polymorphisms in the diversity of various accessions of *A. thaliana*.

Inheritance of H3K27me3 and H3K4me2 distribution patterns in hybrid offspring is additive and only to a small extent responsive to intra-specific hybridization

Comparison of H3K4me2 and H3K27me3 distribution patterns between hybrid offspring and parental inbred lines revealed limited changes in intra-specific hybrids. These results are in agreement with a previous study on *A. thaliana* accessions and their intra-specific hybrids regarding the distribution of histone methylation marks at the microscopic level (Banaei Moghaddam *et al.*, 2010), as well as a study that compared histone methylation patterns in hybrids between rice cultivars (He *et al.*, 2010). Similarly, inheritance of DNA methylation polymorphisms has been shown to be additive (Zhang *et al.*, 2008; Banaei Moghaddam *et al.*, 2010).

We conclude that intra-specific hybridization in *A. thaliana* does not result in global epigenomic rearrangements. More investigations are required to analyse whether this conclusion is also valid for other chromatin modifications. More generally, it would be interesting to determine whether heritable variations in epigenomic patterns between inbred lines contribute to hybrid performance in addition to sequence polymorphisms.

EXPERIMENTAL PROCEDURES

Plant materials

A. thaliana accessions Col-0, C24 and Cvi and their reciprocal F_1 hybrid offspring Col-0 × C24, C24 × Col-0, Col-0 × Cvi and Cvi × Col-0 were used (Banaei Moghaddam *et al.*, 2010). Approximately 400 seeds of each sample were surface-sterilized and cultured in liquid medium, and grown for 10 days under controlled conditions with 16 h light per day (light intensity of approximately 100 μE), 22°C day temperature and 18°C night temperature (Lippman *et al.*, 2004).

DNA extraction, chromatin immunoprecipitation and array hybridization

Plant genomic DNA used for CGH was extracted using the Qiagen DNeasy plant DNA extraction system (http://www.qiagen.com/) according to manufacturer's instructions. ChIP assays were performed essentially as described previously (Gendrel *et al.*, 2005) using anti-H3K4me2 (07-030) and H3K27me3 (07-449) antibodies from Upstate/Millipore (http://www.millipore.com). Each experiment was performed in two biological replicates.

DNA recovered after immunoprecipitation (IP) and directly from input Col-0 chromatin (INPUT), or genomic DNA extracted from the various genotypes, was amplified, differentially labelled and co-hybridized in dye-swap experiments to correct for dye biases, as previously described (Lippman *et al.*, 2004; Turck *et al.*, 2007) for the chromosome 4 tiling microarray, or according to the manufacturer's instructions for the whole-genome tiling arrays (Roche NimbleGen, http://www.nimblegen.com). The Arabidopsis chromosome 4 tiling microarray comprised 21 800 printed features, with a mean size of 1 kb, covering the main part of chromosome 4. The heterochromatic knob on the short arm and several megabases of pericentromeric heterochromatin are included, and account for 16% of the 18.6 Mb covered by the array (Martienssen et al., 2005). Details of array design and production are described by Vaughn *et al.* (2007). This platform has been deposited to the Gene Expression Omnibus (GEO) under accession number GPL10172. The whole-genome tiling microarray consists of 50–75 nt tiles, with a mean spacing of 165 nt, that are distributed across the entire genome sequence (TAIR7) without repeat masking. These tiles have a mean melting temperature of 74°C, and 88% of them match a unique position in the genome. This custom design was split into two arrays of 360 718 tiles each, using every other tile in each array (GEO accessions GPL10919 and GPL10920).

Comparative genomic hybridization (CGH) analysis

CGH experiments were performed for Col-0 versus Cvi and Col-0 versus C24. Data obtained using Arabidopsis whole-genome tiling NimbleGen arrays were analysed using hidden Markov models (Seifert *et al.*, 2009). A fully connected three-state hidden Markov model with state-specific Gaussian emission densities was adapted to each CGH experiment using a Bayesian Baum–Welch algorithm. Decoding of the status of each tile (deleted, unchanged or amplified) was performed using the Viterbi algorithm. Groups of contiguous tiles with log ratios that were significant different from zero for each set of compared accessions were interpreted as representing regions of copy number variation. For interpretation of CGH data, only contiguous CGH polymorphic tiles that were consistently found in both directions of dye-swap experiments were included.

ChIP on chip analysis

Raw hybridization data obtained with the chromosome 4 tiling array were normalized as described previously (Turck *et al.*, 2007),

© 2011 Leibniz Institute of Plant Genetics and Crop Plant Research

and an ANOVA model was applied to whole-genome data to remove technical biases. Normalized data were analysed using the ChIPmix method (Martin-Magniette *et al.*, 2008), which was adapted to handle multiple biological replicates simultaneously. This method is based on a mixture model of regressions, the parameters of which are estimated using an expectation-maximization (EM) algorithm. For each tile, a posterior probability, defined as the probability of enrichment given the log(INPUT) and log(IP) intensities, is used to classify the tile into the normal or enriched class. A false-positive risk is determined by defining the probability of obtaining a posterior probability at least as extreme as the one that is actually observed when the tile is normal. False-positive risks are then adjusted by the Benjamini–Hochberg procedure, and tiles for which the adjusted false-positive risk is lower than 0.01 are considered enriched. Previously published data (Turck *et al.*, 2007; Vaughn *et al.*, 2007) were re-analysed using the same procedure. Neighbouring enriched tiles are combined into domains, requiring minimal runs of 1.6 kb or 300 bp and allowing maximal gaps of 800 or 200 bp for chromosome 4 or whole-genome data, respectively. Enriched but isolated tiles were not considered for further analyses.

Scatter plots and Pearson correlation analyses revealed higher correlation between data sets for histone modification polymorphic tiles among biological replicates for one accession than between datasets of different accessions (Figure S8). ChIP assays were validated by quantitative PCR for sequences known to be associated or not associated with the histone modification of interest (Figure S1 and Table S5) (Turck *et al.*, 2007; Zhang *et al.*, 2007, 2009). For subsequent analysis of ChIP on chip data for Col-0, Cvi and C24 and their intra-specific hybrids, all CGH polymorphic tiles (including singletons and those tiles that were detected only in one of two CGH technical replicates) were excluded.

Data availability and computational analyses

Raw and processed data have been deposited to the National Center for Biotechnology Information Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession GSE24836, and to CATdb (http://urgv.evry.inra.fr/CATdb) (Samson *et al.*, 2004; Gagnot *et al.*, 2008). In addition, array data and genome annotations are available for visualization at http://epigara.biologie.ens.fr/cgi-bin/gbrowse/a2e/. Gene ontology and gene family analysis were performed using the MIPS Functional Catalogue (http://mips.helmholtz-muenchen.de/proj/funcatDB/) (Ruepp *et al.*, 2004) and the DAVID tool (http://david.abcc.ncifcrf.gov/) (Huang *et al.*, 2009).

ACKNOWLEDGEMENTS

This work was supported by Deutsche Forschungsgemeinschaft grant HO 1779/7-1/2 to M.F.M. and A.H. within the framework program SPP 1149 'Heterosis in Plants'. Bioinformatics analyses by M.S. were supported by the Ministry of Culture of the State of Saxony Anhalt, Germany (XP3624HP/0606T) and the Deutscher Akademischer Austauschdienst (PROCOPE 50748812). We would like to thank O. Weiß for excellent technical assistance, and I. Schubert, M. Strickert and S. Scholten (Department of Biology, University of Hamburg) for discussions and helpful comments on the manuscript. We also thank E. Duvernois-Berthet for data visualization on Genome Browser (GB).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Confirmation of ChIP preparations by quantitative PCR of reference sequences.
Figure S2. Sample comparison of H3K4me2 and H3K27me3 patterns.
Figure S3. Size distribution of histone modification domains.
Figure S4. Ontology classification of histone modification polymorphic genes.
Figure S5. H3K27me3-associated genes in this study in comparison with previous analyses.
Figure S6. Ontology classification of hybridization-responsive genes.
Figure S7. Intersection analysis of genes for which H3K4me2 or H3K27me3 changed after hybridization.
Figure S8. Reproducibility of histone modification data for H3K4me2 and H3K27me3.
Table S1. H3K4me2 and H3K27me3 domains in parental accessions and F_1 hybrid offspring.
Table S2. Genes associated with H3K4me2 and H3K27me3 in parental accessions and F_1 hybrid offspring.
Table S3. TEs associated with H3K4me2 and H3K27me3 in parental accessions and F_1 hybrid offspring.
Table S4. Gene ontology and family analysis using the DAVID tool.
Table S5. Protocol for quantitative PCR of ChIP reference sequences.
Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

REFERENCES

Axelsson, T., Bowman, C.M., Sharpe, A.G., Lydiate, D.J. and Lagercrantz, U. (2000) Amphidiploid *Brassica juncea* contains conserved progenitor genomes. *Genome*, **43**, 679–688.

Banaei Moghaddam, A.M., Fuchs, J., Czauderna, T., Houben, A. and Mette, M.F. (2010) Intraspecific hybrids of *Arabidopsis thaliana* revealed no gross alterations in endopolyploidy, DNA methylation, histone modifications and transcript levels. *Theor. Appl. Genet.* **120**, 215–226.

Baumel, A., Ainouche, M., Kalendar, R. and Schulman, A.H. (2002) Retrotransposons and genomic stability in populations of the young allopolyploid species *Spartina anglica* C.E. Hubbard (Poaceae). *Mol. Biol. Evol.* **19**, 1218–1227.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K. *et al.* (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, **120**, 169–181.

Brubaker, C., Brown, A., Stewart, J., Kilby, M. and Grace, J. (1999) Production of fertile hybrid germplasm with diploid Australian *Gossypium* species for cotton improvement. *Euphytica*, **108**, 199–213.

Chen, Z.J. (2010) Molecular mechanisms of polyploidy and hybrid vigor. *Trends Plant Sci.* **15**, 57–71.

Chen, L. and Chen, J. (2008) Changes of cytosine methylation induced by wide hybridization and allopolyploidy in *Cucumis*. *Genome*, **51**, 789–799.

Clark, R.M., Schweikert, G., Toomajian, C. *et al.* (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science*, **317**, 338–342.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M. and Jacobsen, S.E. (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature*, **452**, 215–219.

Comai, L. (2000) Genetic and epigenetic interactions in allopolyploid plants. *Plant Mol. Biol.* **43**, 387–399.

Fuchs, J., Demidov, D., Houben, A. and Schubert, I. (2006) Chromosomal histone modification patterns – from conservation to diversity. *Trends Plant Sci.* **11**, 199–208.

Gagnot, S., Tamby, J.P., Martin-Magniette, M.L., Bitton, F., Taconnat, L., Balzergue, S., Aubourg, S., Renou, J.P., Lecharny, A. and Brunaud, V. (2008) CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Res.* **36**, D986–D990.

Gendrel, A.V., Lippman, Z., Martienssen, R. and Colot, V. (2005) Profiling histone modification patterns in plants using genomic tiling microarrays. *Nat. Methods*, **2**, 213–218.

© 2011 Leibniz Institute of Plant Genetics and Crop Plant Research

Groszmann, M., Greaves, I.K., Albertyn, Z.I., Scofield, G.N., Peacock, W.J. and Dennis, E.S. (2011) Changes in 24_nt siRNA levels in Arabidopsis hybrids suggest an epigenetic contribution to hybrid vigor. *Proc. Natl Acad. Sci. USA*, **108**, 2617–2622.

He, G., Zhu, X., Elling, A.A. *et al.* (2010) Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. *Plant Cell*, **22**, 17–33.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57.

Kashkush, K., Feldman, M. and Levy, A.A. (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat. Genet.* **33**, 102–106.

Kazazian, H.H. Jr (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626–1632.

Kouzarides, T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.

Lawrence, R.J., Earley, K., Pontes, O., Silva, M., Chen, Z.J., Neves, N., Viegas, W. and Pikaard, C.S. (2004) A concerted DNA methylation/histone methylation switch regulates rRNA gene dosage control and nucleolar dominance. *Mol. Cell*, **13**, 599–609.

Lee, H.S. and Chen, Z.J. (2001) Protein-coding genes are epigenetically regulated in Arabidopsis polyploids. *Proc. Natl Acad. Sci. USA*, **98**, 6753–6758.

Lippman, Z., Gendrel, A.V., Black, M. *et al.* (2004) Role of transposable elements in heterochromatin and epigenetic control. *Nature*, **430**, 471–476.

Liu, B., Brubaker, C.L., Mergeai, G., Cronn, R.C. and Wendel, J.F. (2001) Polyploid formation in cotton is not accompanied by rapid genomic changes. *Genome*, **44**, 321–330.

Liu, J., He, Y., Amasino, R. and Chen, X. (2004) siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis. *Genes Dev.* **18**, 2873–2878.

Madlung, A., Masuelli, R.W., Watson, B., Reynolds, S.H., Davison, J. and Comai, L. (2002) Remodeling of DNA methylation and phenotypic and transcriptional changes in synthetic Arabidopsis allotetraploids. *Plant Physiol.* **129**, 733–746.

Martienssen, R.A., Doerge, R.W. and Colot, V. (2005) Epigenomic mapping in Arabidopsis using tiling microarrays. *Chromosome Res.* **13**, 299–308.

Martin-Magniette, M.L., Mary-Huard, T., Bérard, C. and Robin, S. (2008) ChIPmix: mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics*, **24**, i81–i86.

Noel, L., Moores, T.L., van Der Biezen, E.A., Parniske, M., Daniels, M.J., Parker, J.E. and Jones, J.D. (1999) Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of Arabidopsis. *Plant Cell*, **11**, 2099–2112.

Pokholok, D.K., Harbison, C.T., Levine, S. *et al.* (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, **122**, 517–527.

Roudier, F., Teixeira, F.K. and Colot, V. (2009) Chromatin indexing in Arabidopsis: an epigenomic tale of tails and more. *Trends Genet.* **25**, 511–517.

Ruepp, A., Zollner, A., Maier, D. *et al.* (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.* **32**, 5539–5545.

Samson, F., Brunaud, V., Duchêne, S., De Oliveira, Y., Caboche, M., Lecharny, A. and Aubourg, S. (2004) FLAGdb++ : a database for the functional analysis of the Arabidopsis genome. *Nucleic Acids Res.* **32**, D347–D350.

Saze, H. (2008) Epigenetic memory transmission through mitosis and meiosis in plants. *Semin. Cell Dev. Biol.* **19**, 527–536.

Schmid, K.J., Sorensen, T.R., Stracke, R., Torjek, O., Altmann, T., Mitchell-Olds, T. and Weisshaar, B. (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res.* **13**, 1250–1257.

Schubert, D., Primavesi, L., Bishopp, A., Roberts, G., Doonan, J., Jenuwein, T. and Goodrich, J. (2006) Silencing by plant Polycomb-group genes requires dispersed trimethylation of histone H3 at lysine 27. *EMBO J.* **25**, 4638–4649.

Seifert, M., Banaei, A., Keilwagen, J., Mette, M.F., Houben, A., Roudier, F., Colot, V., Grosse, I. and Strickert, M. (2009) Array-based genome comparison of Arabidopsis ecotypes using hidden Markov models. In *Proceedings of Biosignals: Second International Conference on Bio-inspired Systems and Signal Processing* (Encarnacao, P. and Veloso, A., eds). Setúbal: INSTICC press, pp. 3–11. http://dig.ipk-gatersleben.de/HMMs/ACGH/ACGH.html [accessed 23 April 2009].

Springer, N.M., Ying, K., Fu, Y. *et al.* (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* **5**, e1000734.

Teixeira, F.K. and Colot, V. (2010) Repeat elements and the Arabidopsis DNA methylation landscape. *Heredity*, **105**, 14–23.

Turck, F., Roudier, F., Farrona, S., Martin-Magniette, M.L., Guillaume, E., Buisine, N., Gagnot, S., Martienssen, R.A., Coupland, G. and Colot, V. (2007) Arabidopsis TFL2/LHP1 specifically associates with genes marked by trimethylation of histone H3 lysine 27. *PLoS Genet.* **3**, e86.

Vaughn, M.W., Tanurdzic, M., Lippman, Z. *et al.* (2007) Epigenetic natural variation in *Arabidopsis thaliana*. *PLoS Biol.* **5**, e174.

Xiong, L.Z., Xu, C.G., Saghai Maroof, M.A. and Zhang, Q. (1999) Patterns of cytosine methylation in an elite rice hybrid and its parental lines, detected by a methylation-sensitive amplification polymorphism technique. *Mol. Gen. Genet.* **261**, 439–446.

Xu, Y., Zhong, L., Wu, X., Fang, X. and Wang, J. (2009) Rapid alterations of gene expression and cytosine methylation in newly synthesized *Brassica napus* allopolyploids. *Planta*, **229**, 471–483.

Zemach, A., McDaniel, I.E., Silva, P. and Zilberman, D. (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, **328**, 916–919.

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y.V., Pellegrini, M., Goodrich, J. and Jacobsen, S.E. (2007) Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS Biol.* **5**, e129.

Zhang, X., Shiu, S., Cal, A. and Borevitz, J.O. (2008) Global analysis of genetic, epigenetic and transcriptional polymorphisms in *Arabidopsis thaliana* using whole genome tiling arrays. *PLoS Genet.* **4**, e1000032.

Zhang, X., Bernatavichute, Y.V., Cokus, S., Pellegrini, M. and Jacobsen, S.E. (2009) Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis thaliana*. *Genome Biol.* **10**, R62.

Zhao, Y., Yu, S., Xing, C., Fan, S. and Song, M. (2008) Analysis of DNA methylation in cotton hybrids and their parents. *Mol. Biol.* **42**, 169–178.

© 2011 Leibniz Institute of Plant Genetics and Crop Plant Research

Annexe D

Article publié dans *EMBO*

The EMBO Journal (2011) 30, 1928–1938 | © 2011 European Molecular Biology Organization | Some Rights Reserved 0261-4189/11
www.embojournal.org

THE EMBO JOURNAL

EMBO open

Integrative epigenomic mapping defines four main chromatin states in Arabidopsis

This is an open-access article distributed under the terms of the Creative Commons Attribution Noncommercial Share Alike 3.0 Unported License, which allows readers to alter, transform, or build upon the article and then distribute the resulting work under the same or similar license to this one. The work must be attributed back to the original author and commercial use is not permitted without specific permission.

François Roudier[1,*], Ikhlak Ahmed[1], Caroline Bérard[2], Alexis Sarazin[1], Tristan Mary-Huard[2], Sandra Cortijo[1], Daniel Bouyer[3], Erwann Caillieux[1], Evelyne Duvernois-Berthet[1], Liza Al-Shikhley[1], Laurène Giraut[4], Barbara Després[1], Stéphanie Drevensek[1], Frédy Barneche[1], Sandra Dèrozier[4], Véronique Brunaud[4], Sébastien Aubourg[4], Arp Schnittger[3], Chris Bowler[1], Marie-Laure Martin-Magniette[2,4], Stéphane Robin[2], Michel Caboche[4] and Vincent Colot[1,*]

[1]Institut de Biologie de l'Ecole Normale Supérieure, Centre National de la Recherche Scientifique (CNRS) UMR8197, Institut National de la Santé et de la Recherche Médicale (INSERM) U1024, Paris, France, [2]AgroParisTech/Institut National de la Recherhe Agronomique (INRA), UMR518 Mathématiques et Informatiques Appliquées, Paris, France, [3]Institut de Biologie Moléculaire des Plantes, CNRS UPR2357, Université de Strasbourg, Strasbourg, France and [4]Unité de Recherche en Génomique Végétale, UMR8114 INRA/CNRS/Université d'Evry Val d'Essonne (UEVE), Evry, France

Post-translational modification of histones and DNA methylation are important components of chromatin-level control of genome activity in eukaryotes. However, principles governing the combinatorial association of chromatin marks along the genome remain poorly understood. Here, we have generated epigenomic maps for eight histone modifications (H3K4me2 and 3, H3K27me1 and 2, H3K36me3, H3K56ac, H4K20me1 and H2Bub) in the model plant Arabidopsis and we have combined these maps with others, produced under identical conditions, for H3K9me2, H3K9me3, H3K27me3 and DNA methylation. Integrative analysis indicates that these 12 chromatin marks, which collectively cover ~90% of the genome, are present at any given position in a very limited number of combinations. Moreover, we show that the distribution of the 12 marks along the genomic sequence defines four main chromatin states, which preferentially index active genes, repressed genes, silent repeat elements and intergenic regions. Given the compact nature of the Arabidopsis genome, these four indexing states typically translate into short chromatin domains interspersed with each other. This first combinatorial view of the Arabidopsis epigenome points to simple principles of organization as in metazoans and provides a framework for further studies of chromatin-based regulatory mechanisms in plants.

The EMBO Journal (2011) **30,** 1928–1938. doi:10.1038/emboj.2011.103; Published online 12 April 2011
Subject Categories: chromatin and transcription; plant biology
Keywords: Arabidopsis; chromatin; DNA methylation; epigenome; histone modifications

*Corresponding authors. F Roudier or V Colot, Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, INSERM U1024, 46 rue d'Ulm, 75230 Paris Cedex 05, France. Tel.: +33 014 432 3538; Fax: +33 014 432 3935; E-mails: roudier@biologie.ens.fr or colot@biologie.ens.fr

Received: 23 November 2010; accepted: 10 March 2011; published online: 12 April 2011

Introduction

Packaging of DNA into chromatin is pivotal for the regulation of genome activity in eukaryotes. The basic unit of chromatin is the nucleosome, which is composed of 147 bp of DNA wrapped around a protein octamer composed of two molecules each of the core histones H2A, H2B, H3 and H4. Covalent modifications of histones, DNA methylation, incorporation of histone variants, and other factors, such as chromatin-remodelling enzymes or small RNAs, all contribute to defining distinct chromatin states that modulate access to DNA (Berger, 2007; Kouzarides, 2007). In particular, different histone modifications are thought to act sequentially or in combination in order to confer distinct transcriptional outcomes (Strahl and Allis, 2000; Jenuwein and Allis, 2001; Berger, 2007; Lee *et al*, 2010a). More generally, it is now well established that the precise composition of chromatin along the genome, which defines the epigenome, participates in the selective readout of the genomic sequence.

Thanks in part to a compact, almost fully sequenced and well-annotated genome, the flowering plant *Arabidopsis thaliana* has become a model of choice for exploring the epigenomes of multicellular organisms and the contribution of chromatin to the regulation of genome activity during development or in response to the environment. Indeed, epigenomic profiling in Arabidopsis has begun to provide insights into the relationship between transcriptional activity and localization of chromatin marks or histone variants (Roudier *et al*, 2009; Feng and Jacobsen, 2011). For instance, H3K4me3 and H3K36me2 are detected at the 5′- and 3′-ends of actively transcribed genes, respectively (Oh *et al*, 2008; Zhang *et al*, 2009), while H3K27me3 broadly marks repressed genes (Turck *et al*, 2007; Zhang *et al*, 2007; Oh *et al*, 2008). In contrast, cytosine methylation (5mC) has a dual localization. It is present predominantly over silent transposable elements (TEs) and other repeats, where it is associated with H3K9me2 and

 ©2011 European Molecular Biology Organization

H3K27me1, but also in the body of ~30% of genes, many of which are characterized by moderate expression levels (Lippman *et al*, 2004; Zhang *et al*, 2006; Zilberman *et al*, 2006; Turck *et al*, 2007; Vaughn *et al*, 2007; Bernatavichute *et al*, 2008; Cokus *et al*, 2008; Lister *et al*, 2008; Jacob *et al*, 2010). Furthermore, the variant histone H2A.Z, which is preferentially deposited near the 5′-end of genes and promotes transcriptional competence, antagonizes DNA methylation and vice versa (Zilberman *et al*, 2008). However, extensive combinatorial analyses of these and other chromatin marks have not been performed so far in Arabidopsis and meta-analysis of published data is complicated by the fact that biological materials and methodologies often differ between studies.

Here, we report the epigenomic profiles of eight histone modifications (H3K4me2, H3K4me3, H3K27me1, H3K27me2, H3K36me3, H3K56ac, H4K20me1 and H2Bub). Integrative analyses of these and other profiles, previously obtained under identical conditions for DNA methylation, H3K9me2, H3K9me3 and H3K27me3 (Turck *et al*, 2007; Vaughn *et al*, 2007), indicate a low combinatorial complexity of chromatin marks in Arabidopsis, as recently reported for metazoans (Wang *et al*, 2008; Hon *et al*, 2009; Ernst and Kellis, 2010; Gerstein *et al*, 2010; Roy *et al*, 2010; Kharchenko *et al*, 2011; Liu *et al*, 2011; Riddle *et al*, 2011; Zhou *et al*, 2011). Furthermore, our study identifies four main chromatin states in Arabidopsis, which have distinct indexing functions and which typically form short domains interspersed with each other. This first comprehensive view of the Arabidopsis epigenome suggests simple principles of organization, as recently proposed for Drosophila (Filion *et al*, 2010), and provides a resource to refine our understanding of the control of genome activity at the level of chromatin.

Results

Epigenomic profiling of 12 chromatin marks

Epigenomic maps were generated for eight histone modifications (H3K4me2, H3K4me3, H3K27me1, H3K27me2, H3K36me3, H3K56ac, H2Bub and H4K20me1) using chromatin extracted from young seedlings and immunoprecipitation followed by hybridization to a tiling microarray that covers the entire chromosome 4 of Arabidopsis at ~900 bp resolution (Turck *et al*, 2007). Data previously obtained for 5mC (Vaughn *et al*, 2007), H3K9me2, H3K9me3 and H3K27me3 (Turck *et al*, 2007) using similar materials and methodologies were also considered. Epigenomic profiling was additionally performed for seven of these marks (H3K4me2, H3K4me3, H3K27me1, H3K27me3, H3K36me3, H2Bub and 5mC) using a tiling microarray covering the whole-genome sequence at 165 bp resolution. Chromosome 4 and whole-genome maps were also obtained for histone H3 to control for nucleosome occupancy. The 12 marks were chosen because they were shown in previous studies to be associated with distinct transcriptional activities or subnuclear localization in Arabidopsis. In addition, our selection was focussed to a large extent on histone lysine methylation, which exists in three forms (mono-, di- and tri-methylation) and therefore has a versatile indexing potential (Sims and Reinberg, 2008).

Collectively, the 12 chromatin marks cover almost all of the regions that are detectably associated with histone H3, which amount to ~90% of the total genome sequence (data not shown; Chodavarapu *et al*, 2010). The distribution of each chromatin modification was characterized in detail along chromosome 4. In agreement with previous reports (Lippman *et al*, 2004; Turck *et al*, 2007; Zhang *et al*, 2007, 2009; Oh *et al*, 2008; Tanurdzic *et al*, 2008), H3K4me2, H3K4me3, H3K9me3, H3K27me3 and H3K56ac are mostly found in euchromatin (Figure 1A; Supplementary Figure S1; Supplementary Table I), which reflects the fact that these different modifications are associated almost exclusively with genes (Figure 1B). H2Bub and H3K36me3, for which no epigenomic maps have been reported to date in plants, are also characterized by a predominant distribution over genes. In contrast, H4K20me1 is found in heterochromatin mainly and associates with TE and other repeat element sequences (Figure 1B), like H3K9me2 (Lippman *et al*, 2004; Bernatavichute *et al*, 2008). The present analysis reveals in addition that, like 5mC (Zhang *et al*, 2006; Zilberman *et al*, 2006), H3K27me1 and H3K27me2 are dual marks associated not only with TEs but also with a fraction of genes (Supplementary Tables II–IV).

Each chromatin mark defines domains of contiguous tiles and the number of these domains ranges from 306 for H3K9me2 to 1163 for H3K4me3. For H3K4me3, H3K36me3, H3K56ac, H3K9me3, H2Bub or H3K27me3, domains have similar median length between euchromatin and heterochromatin and mostly coincide with single transcription units (Supplementary Table II; Supplementary Figure S2). By contrast, H3K9me2, H4K20me1, H3K27me1, H3K27me2 and 5mC form small domains in euchromatin but large domains in heterochromatin, as a result of the dense clustering of TE and other repeat sequences in the latter (Supplementary Figure S2; Supplementary Table II).

Combinatorial analysis of chromatin marks

As a first step in exploring the combinatorial deposition patterns of chromatin marks, unbiased pairwise association analyses were carried out. A heat map generated from the calculated association values (Supplementary Table V) and organized by hierarchical clustering reveals two clear groups of correlated pairs that distinguish genes from TE sequences (Figure 1C). Next, co-occurrence of marks was registered over each of the ~20 000 tiles of the chromosome 4 array. Of the $2^{12} = 4096$ combinations theoretically possible, only 665 were observed and among these, only 38 concerned at least 100 tiles (Supplementary Figure S3A). This indicates therefore a limited repertoire of chromatin signatures in Arabidopsis, as in other eukaryotes (Ernst and Kellis, 2010; Kharchenko *et al*, 2011; Liu *et al*, 2011). The four prevalent combinations of marks are H3K27me1 + 5mC + H3K9me2 + H4K20me1 + H3K27me2, H3K56Ac + H2Bub + H3K4me3 + H3K4me2 + H3K9me3 + H3K36me3, H3K27me3 + H3K27me2 + H3K4me2 and H3K27me3 + H3K27me2, which cover 10.9, 6.8, 4.7 and 4.6% of the tiling array, respectively. Whereas the first combination is almost exclusively associated with TE sequences, the other three are mainly present over genes (Supplementary Figure S3B). Furthermore, like H3K27me3 + H3K27me2, most of the remaining combinations represented by at least 100 tiles are subcombinations of the three prevalent ones (Supplementary Figure S3B and data not shown).

©2011 European Molecular Biology Organization

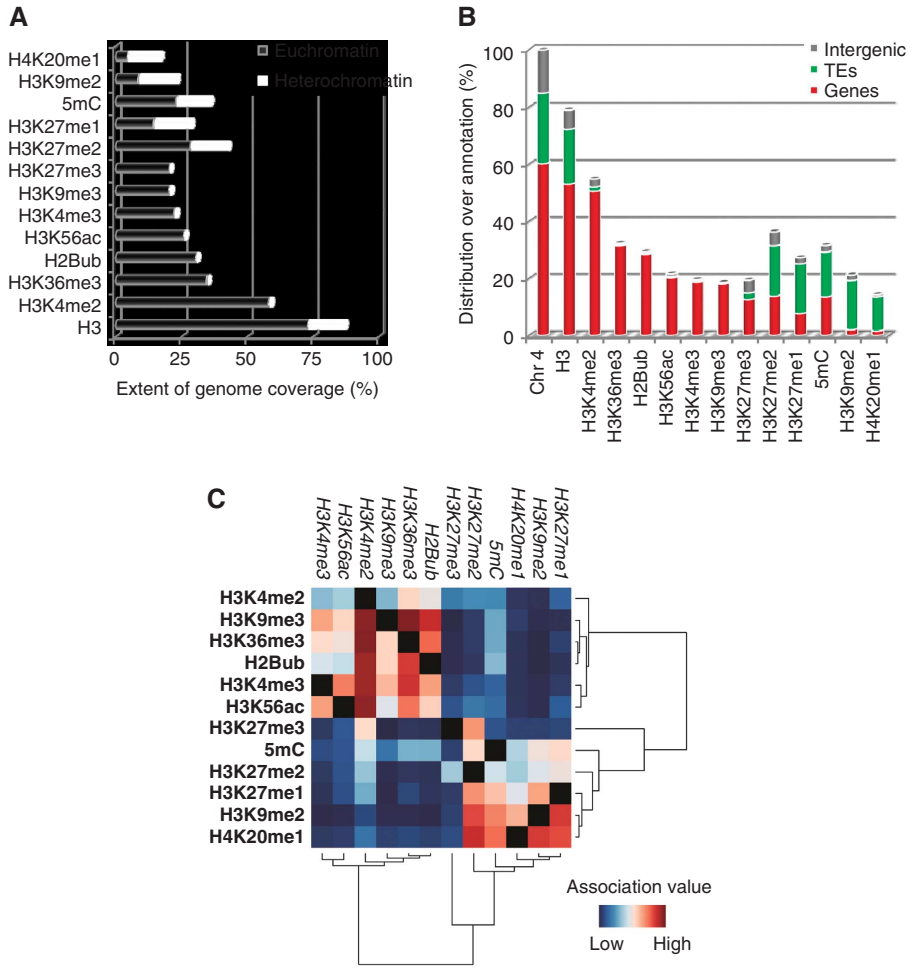

Figure 1 Genomic distribution of chromatin marks. (**A**) Relative coverage of chromatin marks in the euchromatin and heterochromatin of chromosome 4. Coordinates for heterochromatin are 1.61–2.36 Mb (knob) and 2.78–5.15 Mb (pericentromeric regions). (**B**) Chromosome-wide distribution of chromatin marks over annotated features. Tiles that overlap annotated genes or transposable elements (TAIR8) by at least 50 bp were assigned to the corresponding annotation and otherwise called 'intergenic'. (**C**) Pairwise association analysis of the 12 chromatin marks along chromosome 4. Mean association values were calculated for each pair of modifications over all marked tiles and are shown as a directional heat map organized by hierarchical clustering using Pearson's correlation distances.

To complement this tile-centric analysis and to identify the prevalent combinatorial patterns of the 12 chromatin marks, unsupervised *c*-means clustering was performed. The number of clusters (k) was varied from 2 to 11 and $k=4$ was determined to be optimal in maximizing homogeneity within clusters and heterogeneity between them. The four chromatin states (CS1–CS4) defined by these four clusters are also identified by PCA analysis (data not shown), thus reinforcing their significance. Whereas CS1 regroups ~90% of the tiles associated with H3K4me3, H3K36me3, H3K9me3 and H2Bub as well as the majority of H3K4me2- and H3K56ac-marked sequences, H3K27me3 and H3K27me2 are the most prevalent modifications in CS2 (Figure 2A). As expected from their composition, CS1 and CS2 are mainly associated with genes (Figure 2B) and have antagonistic indexing functions, being prevalent among active and repressed/lowly expressed genes, respectively (Figure 2C). CS3, which is associated predominantly with TE sequences (Figure 2B), regroups most of the tiles marked by H3K9me2, H4K20me1 and H3K27me1 as well as ~50% of those marked by H3K27me2 and 5mC (Figure 2A). In contrast to the other three chromatin states, CS4 is not particularly enriched in any chromatin mark (Figure 2A) and is found mainly outside of genes and TE sequences (Figure 2B). Nonetheless, CS4 also marks ~10% of genes, most of which display low expression (Figure 2C). In keeping with the domain layout of individual marks, CS1–CS4 typically form small domains interspersed with each other, except in cytologically defined heterochromatin, where CS3 forms larger domains as a result of the clustering of TE sequences (Figure 2D; Supplementary Figure S4).

Chromatin signatures of genes

To investigate further the chromatin indexing of genes, pairwise analysis of chromatin modifications was carried out specifically over genic tiles, which revealed a tight association between H3K4me3 and H3K56ac, between H3K36me3, H3K9me3 and H2Bub and between H3K27me2 and H3K27me3 (Figure 3A). Next, average enrichment levels were calculated within and around genes for all marks except H3K9me2 and H4K20me1, which are almost exclusively associated with TE and other repeat sequences. As shown

 ©2011 European Molecular Biology Organization

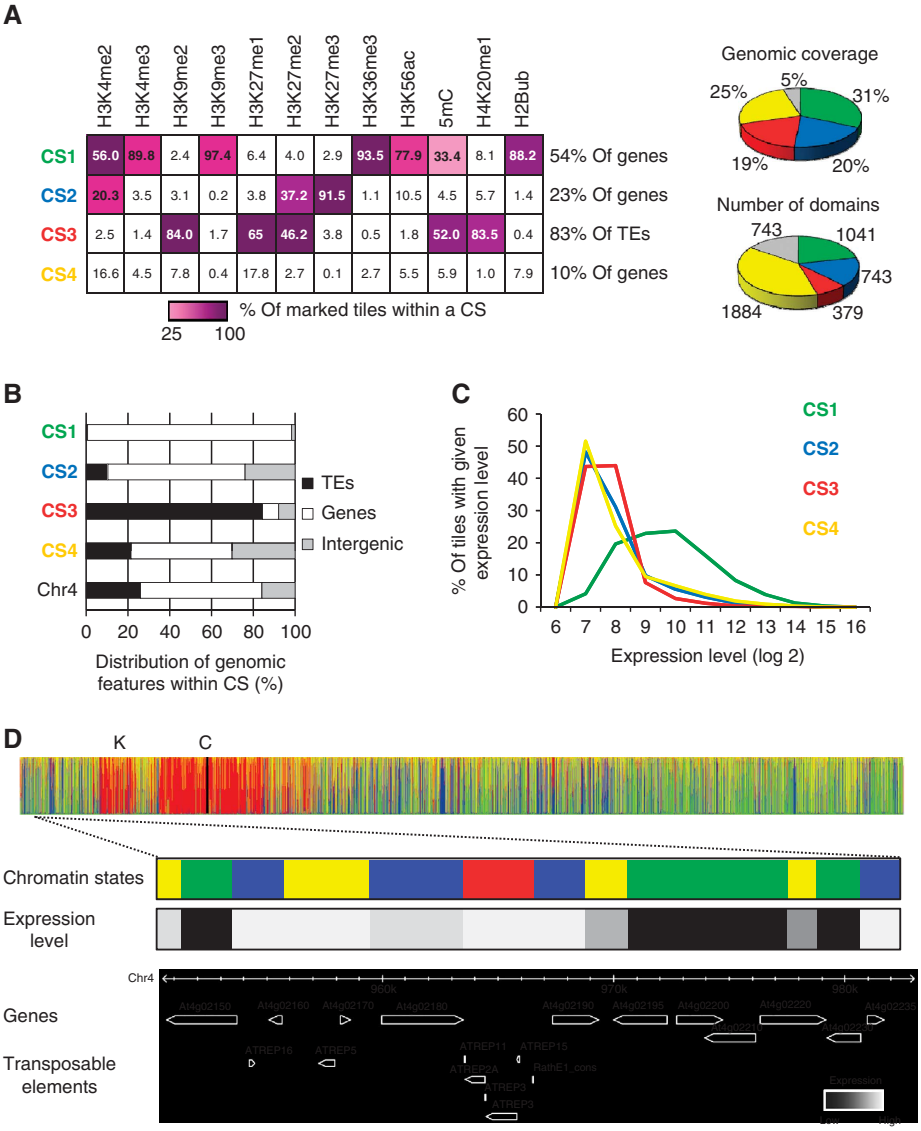

Figure 2 The Arabidopsis epigenome contains four predominant chromatin states. (**A**) The table on the left indicates the composition of the four predominant chromatin states (CS) identified by *c*-means clustering. The distribution of the 12 chromatin marks over the four CS is indicated as a heat map for values ranging from 25% (light purple) to 100% (dark purple). The degree of homogeneity of each CS is indicated by the percentage of tiles assigned to it that are associated with each of the 12 chromatin marks (numbers inside cells). Note that no single mark is present over >20% of the tiles assigned to CS4, in contrast to what is observed for CS1–CS3. The percentage of genes indexed by CS1, CS2 and CS4 and the percentage of TE annotations indexed by CS3 are also shown. Pie charts indicate the relative genomic coverage of the four CS and the number of domains that they each form. Grey colour corresponds to tiles that cannot be unambiguously assigned to any of the four CS (see Materials and methods). (**B**) Relative proportion of genomic features within each CS. Tiles that overlap annotated genes or transposable elements (TAIR8) by at least 50 bp were assigned the corresponding annotation. All other tiles were considered as 'intergenic'. (**C**) Relationship between chromatin states and gene expression level. The percentage of tiles associated with a given CS is represented according to expression level. The dashed line represents the distribution of all annotated genes of chromosome 4. Expression data (Schmid *et al*, 2005) were obtained by averaging appropriate developmental stages. (**D**) Distribution of the four CS along chromosome 4. For each tile, membership to a given CS is colour coded. K: heterochromatic knob. The non-sequenced part of the centromere (C) is represented by the vertical black line. The high interspersion of chromatin states seen outside of heterochromatin is highlighted in a genome browser view of a 30-kb euchromatic region (positions 0.95–0.98 Mb).

in Figure 3B, values are highest within the transcribed region for the 10 chromatin modifications considered and are typically lowest upstream or downstream of it. However, distribution patterns vary substantially between marks, as previously established in several instances (Turck *et al*, 2007; Zhang *et al*, 2007; Jacob *et al*, 2010). H3K4me3, H3K56ac, H3K4me2, H3K36me3 and H3K9me3 all peak at the 5′-end of the transcribed region, but the first two marks more sharply than the other three (Figure 3B). In contrast, H2Bub as well as H3K27me1 are highest more centrally, 5mC is most enriched in the 3′-half of the transcribed region and both H3K27me2 and H2K27me3 show an even distribution across transcribed regions. Finally, H3K27me2 differs from all other marks including H3K27me3 in that it remains high in flanking regions, a difference which does not result from the presence of H3K27me2-marked TE sequences

©2011 European Molecular Biology Organization

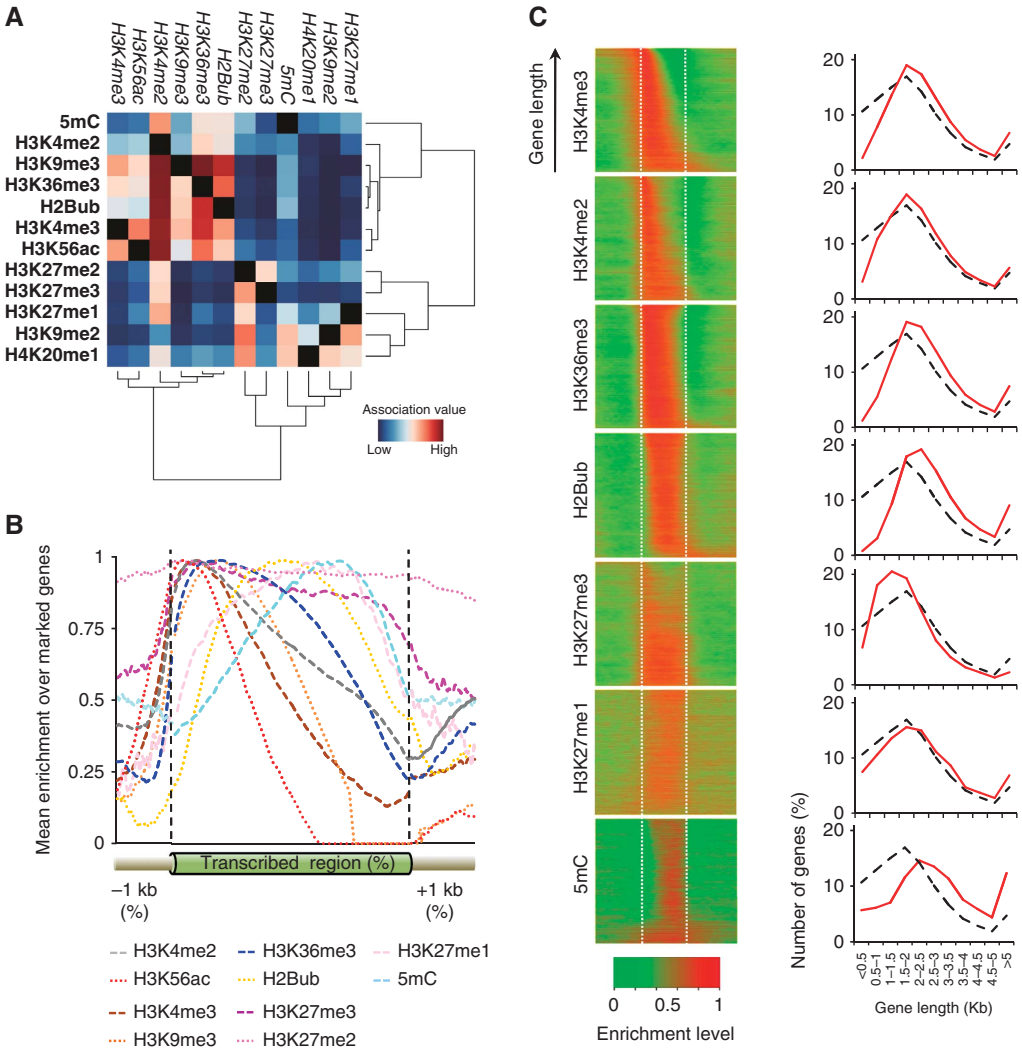

Figure 3 Distribution of chromatin marks over genes. (**A**) Pairwise association analysis of the 12 chromatin marks along chromosome 4. Mean association values were calculated for each pair of modifications over all marked genic tiles and are shown as a directional heat map organized by hierarchical clustering using Pearson's correlation distances metric. (**B**) Mean enrichment levels relative to histone H3 are plotted along marked genes (transcribed region, scaled to accommodate for different gene lengths, bin size of 1%) as well as up to 1 kb of upstream and downstream sequences (bin size of 10 bp). Maximum value for any given mark is arbitrarily set to 1. Data were obtained using the chromosome 4 tiling array. Note that values for H3K27me2 in upstream and downstream regions are significantly higher than for unmarked genes (>0.9 versus ~0.6, not shown). (**C**) Left panels: Enrichment levels relative to histone H3 for marked genes sorted by length. Each line represents a single gene as well as 1 kb of upstream and downstream sequences. Enrichment is indicated as a heat map, with maximal (red) and minimal (green) values set to 1 and 0, respectively. Right panels: Frequency distribution of marked (red line) and all genes (black dashed line) according to their length. Data were obtained using the whole-genome tiling array.

adjacent to genes nor from the lower signal to noise ratio measured for this mark (see legend of Figure 3A, data not shown). Using the genome-wide profiles obtained for seven chromatin modifications, we could show in addition that contrary to H3K27me3, which preferentially marks small genes as noted before (Luo and Lam, 2010), H2Bub, H3K36me3, 5mC and, to a lesser extent, H3K4me2 as well as H3K4me3 tend to be associated with longer genes (Figure 3C). Unlike these chromatin modifications, H3K27me1 does not exhibit preferential association in relation to gene length (Figure 3C).

It has been established that H3K4me3 and H3K56ac mark genes that are highly and broadly expressed (Oh *et al*, 2008; Tanurdzic *et al*, 2008; Zhang *et al*, 2009). Conversely, H3K27me3 is preferentially associated with genes that are expressed at low levels or in a tissue-specific manner (Turck *et al*, 2007; Zhang *et al*, 2007; Oh *et al*, 2008; Jacob *et al*, 2010) and 5mC tends to mark moderately expressed genes (Zilberman *et al*, 2006; Vaughn *et al*, 2007). Our analysis confirms these results and indicates in addition that H2Bub, H3K36me3 and H3K9me3 tend to mark highly expressed genes, like H3K4me3 and H3K56ac (Figure 4A). On the other hand, H3K4me2 does not appear to index genes in relation to their expression level and H3K27me1 as well as H3K27me2 tend to be associated with genes that are expressed at low level or in a tissue-specific manner, like H3K27me3 (Figure 4A and B). However, H3K27me1 and H3K27me2/3 mark largely non-overlapping sets of genes with different ontologies (Figure 3A; Supplementary Tables III, IV, VI and VII), which suggests the existence of two

 ©2011 European Molecular Biology Organization

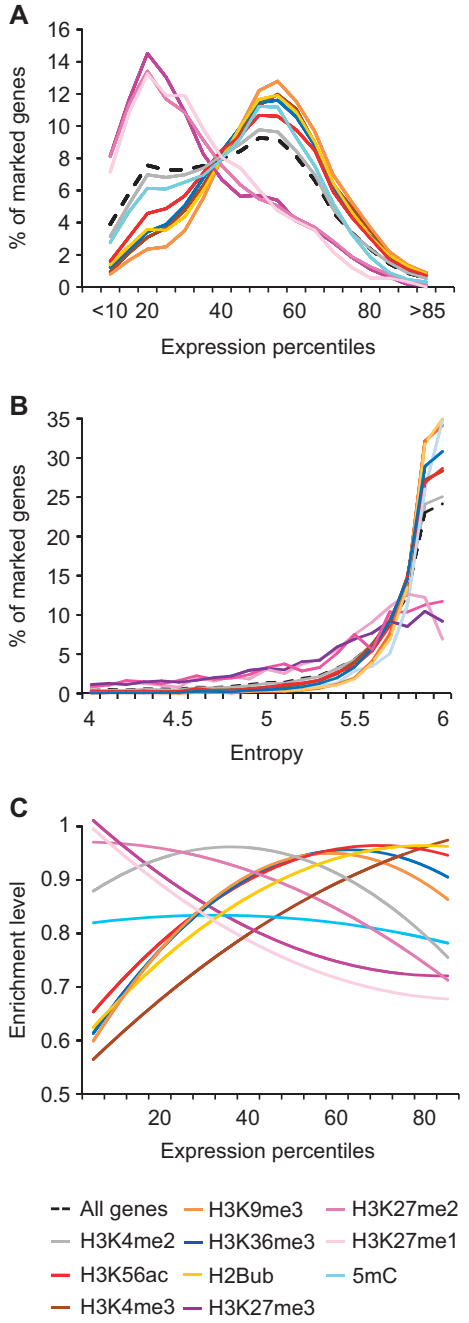

Figure 4 Chromatin indexing in relation to gene expression. (**A**) Distribution density of marked genes according to expression percentiles. Genes were binned according to their absolute expression values in whole seedlings. The dashed line indicates the distribution of all annotated genes on chromosome 4 across all expression percentiles. Expression data (Schmid *et al*, 2005) were obtained by averaging appropriate developmental stages. (**B**) Tissue specificity of marked genes as estimated by Shannon entropy calculation. Low entropy values indicate high tissue specificity. The fraction of marked genes associated with a given entropy value is plotted for each chromatin modification. (**C**) Relationship between gene expression and enrichment level for each chromatin modification. Maximum enrichment level is set to 1 in each case.

distinct gene repression systems associated with methylation of H3K27. For most chromatin marks, average enrichment levels correlate either positively or negatively with expression levels (Figure 4C). Thus, values for H3K4me3, H3K56ac, H3K36me3, H2Bub and H3K9me3 increase gradually with gene expression, at least up to mid expression levels, whereas values for H3K27me1, H3K27me2 and H3K27me3 show an opposite trend. Whether these correlations reflect expression of genes in a variable number of cells, or true differential enrichment in relation to expression level, remains to be determined.

Collectively, our findings indicate that H3K4me3 and H3K27me3 are diagnostic of two antagonist chromatin states that are associated with most active and repressed genes, respectively. However, ~13% (3433 out of 27 294) of genes marked by H3K4me3 or H3K27me3 in whole seedlings present both marks, in agreement with previous observations (Oh *et al*, 2008; Zhang *et al*, 2009). To explore this further, H3K4me3 and H3K27me3 were mapped genome-wide using chromatin extracted from roots and profiles were compared with those obtained for whole seedlings (this study) or aerial parts only (Oh *et al*, 2008). Out of the 3433 genes with both marks in whole seedlings, 284 genes (8.3%) are only marked by H3K4me3 in roots and by H3K27me3 in aerial parts or vice versa (Figure 5A; Supplementary Table VIII). Correspondingly, a majority of these genes show differential expression between roots and aerial parts (Figure 5B), which is in contrast to genes with persistent co-marking in both plant parts (Figure 5A and C). Thus, it can be concluded that co-marking in whole seedlings results for a number of genes from the mixing of cells with opposite chromatin indexing in the two plant parts. By extension, it is likely that persistent co-marking in one or the other plant parts (Figure 5A) reflects similar mixing of cells with distinct epigenomes, but this time within organs. Co-marking could nevertheless correspond to *bona fide* bivalent marking in some cases, as originally reported in mammals for key regulatory genes poised for activation (Wang *et al*, 2009) and as also described in Arabidopsis for a small number of genes encoding transcription factors (Jiang *et al*, 2008; Berr *et al*, 2010). In this respect, it is noteworthy that ontology analysis of the 224 genes with persistent co-marking in both roots and aerial parts (Figure 5A) indicates significant enrichment for terms associated with regulation of transcription (data not shown).

Discussion

A small number of prevalent chromatin states index the Arabidopsis genome

Using an integrative analysis of the distribution of 12 chromatin marks, we show that the Arabidopsis epigenome is organized around four predominant chromatin states with distinct biochemical, transcriptional and sequence properties. This representation refines the classical segmentation between cytologically defined heterochromatin and euchromatin. A first chromatin state (CS1) corresponds to transcriptionally active genes and is typically enriched in the trimethylated forms of H3K4 and H3K36. Two further states correspond to two distinct types of repressive chromatin. H3K27me3-marked repressive chromatin (CS2) is mainly associated with genes under PRC2-mediated repression (Turck *et al*, 2007; Zhang *et al*, 2007), while H3K9me2- and H4K20me1-marked repressive chromatin (CS3) corresponds to classical heterochromatin and is almost exclusively located

©2011 European Molecular Biology Organization

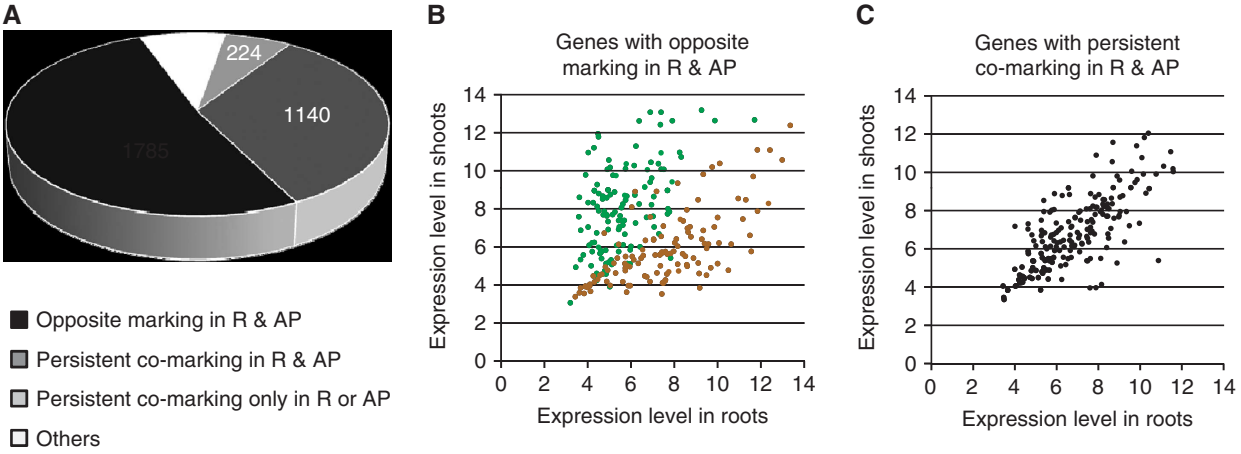

Figure 5 Analysis of genes co-marked with H3K27me3 and H3K4me3 in whole seedlings. (**A**) The 3433 genes co-marked in whole seedlings were split into different classes according to their marking in roots (R; this study) and aerial parts (AP; Oh *et al*, 2008). 'Others' indicate genes with other marking patterns in the two plant parts. This class, which is not expected based on the co-marking observed in whole seedlings, can be explained in part by the fact that the different data sets were not all generated using the same conditions and methodologies. (**B**) Expression analysis in roots and shoots (Schmid *et al*, 2005) for the 284 genes showing opposite marking in roots and aerial parts. Brown dots indicate genes with H3K4me3 in roots and H3K27me3 in aerial parts, green dots indicate genes with the opposite marking pattern. (**C**) Expression analysis in roots and shoots (Schmid *et al*, 2005) for the 224 genes showing persistent co-marking in roots and aerial parts.

over silent TEs (Lippman *et al*, 2004; Bernatavichute *et al*, 2008). A fourth chromatin state (CS4) is characterized by the absence of any prevalent mark and is associated with weakly expressed genes and intergenic regions.

This rather simple organization of Arabidopsis chromatin into four main states shows similarities with that recently reported for Drosophila cells. Indeed, based on the integration of epigenomic maps obtained for 53 chromatin proteins, it was concluded that the Drosophila epigenome is organized into a mosaic of five principal chromatin types that display distinct functional properties (Filion *et al*, 2010). Specifically, Arabidopsis CS2 and CS3 are similar to Drosophila 'BLUE' and 'GREEN' chromatin types, which correspond to repressive chromatin associated with the Polycomb pathway and classical heterochromatin, respectively. Furthermore, CS4, which has no prevalent chromatin mark and indexes some weakly expressed genes as well as intergenic regions is reminiscent of Drosophila 'BLACK' chromatin, which is relatively gene poor and constitutes a repressive environment distinct from heterochromatin. In contrast, transcriptionally active chromatin is represented by a single chromatin state in Arabidopsis (CS1) but by two distinct types in Drosophila that differ in several ways, including the enrichment of H3K36me3 in 'YELLOW' but not in 'RED' chromatin.

Other large-scale epigenomic studies have been performed in yeast (Liu *et al*, 2005), *C. elegans* (Gerstein *et al*, 2010; Liu *et al*, 2011), Drosophila (Kharchenko *et al*, 2011; Roy *et al*, 2010; Riddle *et al*, 2011) and human cells (Wang *et al*, 2008; Hon *et al*, 2009; Ernst and Kellis, 2010; Zhou *et al*, 2011), which all indicate a relatively low combinatorial complexity of chromatin marks. Furthermore, the two main repressive chromatin states defined in Arabidopsis (CS2 and CS3) have similar counterparts in metazoans, indicating that they are highly conserved between plants and animals. On the other hand, the single predominant chromatin state (CS1) that we have identified for transcriptionally active genes in Arabidopsis has no obvious equivalent in these other organisms. Instead, several chromatin states have been associated with expressed genes in other organisms. This discrepancy likely results from the smaller size of genes and intergenic regions in Arabidopsis (~2 kb each on average), as well as the relatively lower resolution of our data. Indeed, our analysis shows that distribution patterns vary substantially between chromatin marks associated with active genes (Figure 3B), which suggests that CS1 could be further refined into at least two additional chromatin signatures, specific to the promoter and transcribed region of these genes.

Although the number of chromatin states identified via this type of integrative approach may appear surprisingly low, such analyses aim to identify prevalent combinations of chromatin marks or chromatin proteins. Furthermore, the heterogeneity of the biological material used in many of these studies, including ours, likely hampered the detection of certain chromatin states such as those that are specific to rare cell types. Ultimately, only a knowledge of the epigenomes of individual cell types will enable a full understanding of the functional impact of chromatin-level regulation on genome activity.

Chromatin indexing of genes in Arabidopsis

Our work indicates that the Arabidopsis epigenome is mainly organized at the level of single transcription units and that the distribution of chromatin marks along genes is linked to the transcription process (Figures 2 and 3). For example, H3K4me3 peaks around the transcription start site of actively expressed genes, as observed in all other eukaryotes examined to date (Rando and Chang, 2009). Similarly, H3K56ac is specifically located at gene promoters and shows preferential marking of active genes, suggesting that, like in yeast, it could facilitate rapid transcriptional activation (Williams *et al*, 2008). In contrast to H3K4me3, H3K4me2 shows no particular association with highly expressed genes

 ©2011 European Molecular Biology Organization

or with specific parts of genes. Rather than being a constitutive mark of transcription, H3K4me2 may be implicated in fine tuning of tissue-specific expression, as recently reported in mammals (Pekowska *et al*, 2010).

The distribution of H3K36me3, H3K9me3 and H2Bub over the transcribed regions of expressed genes suggests that these modifications are linked with transcriptional elongation. In the case of H2Bub, this is in agreement with the distribution reported in mammals and yeast (Minsky *et al*, 2008; Schulze *et al*, 2009). For H3K9me3, enrichment over the coding region of expressed genes in Arabidopsis (this study; Caro *et al*, 2007; Turck *et al*, 2007; Charron *et al*, 2009) contrasts with the enrichment predominantly over heterochromatin in animals. However, association with the transcribed regions of some active genes has been reported in mammals (Vakoc *et al*, 2005, 2006; Squazzo *et al*, 2006). Whether H3K9me3 could serve different outcomes depending on genomic and or chromatin context and whether it has any role in transcription regulation in plants remains to be determined. Given the discrepancy between the low amounts of H3K9me3 reported in bulk histones (Jackson *et al*, 2004; Johnson *et al*, 2004) and its apparent abundance reported by ChIP-chip, it is also possible that the H3K9me3 antibody we used recognizes another modification in Arabidopsis, which would be H3K36me3 based on our epigenomic analysis. However, *in vitro* competition assays using an H3K36me3 peptide suggest that this is unlikely (Supplementary Figure S5).

H3K36me3 preferentially marks exons of transcribed genes in yeast, *C. elegans* and mammals (Kolasinska-Zwierz *et al*, 2009) and it was shown to be involved in the control of alternative splicing in mammals (Luco *et al*, 2010). In Arabidopsis, however, H3K36me3 peaks in the first half of the coding region, which is in contrast to the 3′-end enrichment reported in other organisms (Wang *et al*, 2009). This preferential enrichment at the 5′-end, which is not dependent on gene length, could indicate that the principles governing H3K36me3 deposition differ between plants and other eukaryotes. In fact, H3K36me3 distribution in Arabidopsis resembles that of H3K79me3 in mammals (Wang *et al*, 2009). As Arabidopsis lacks a clear homologue of the H3K79 methyltransferase Dot1 and has no H3K79me3 (Zhang *et al*, 2007), it is possible that H3K36me3 in plants serves a function equivalent to H3K79me3 in other eukaryotes. Furthermore, H3K36me2 could have a role similar to that attributed to H3K36me3 in other eukaryotes, as it peaks at the 3′-end of expressed genes in Arabidopsis (Oh *et al*, 2008).

Chromatin marks associated with transcription have been proposed to cross talk and serve as checkpoints in budding yeast and mammals (Suganuma and Workman, 2008; Weake and Workman, 2008; Lee *et al*, 2010a). A similar scenario could be envisioned in Arabidopsis based on the chromatin marks that predominate in CS1, whereby the RNA polymerase II-associated factor 1 complex would induce monoubiquitylation of H2B via the activity of the Rad6-Bre1 ubiquitin ligase homologues UBC1, 2 and 3 as well as HUB1 and 2, as shown at the *FLC* gene (Cao *et al*, 2008; Gu *et al*, 2009; Schmitz *et al*, 2009). H2Bub deposition would in turn help recruit COMPASS (COMplex Proteins ASsociated with Set1), thus mediating deposition of H3K4me3 and potentially H3K36me3 (in place of H3K79me3) as well as H3K36me2. Similarly to other eukaryotes, initiation of another round of transcription would require the activity of the Ubp8 ubiquitin protease homologue, UBP26, which catalyses H2B deubiquitylation (Sridhar *et al*, 2007). Consistent with this, H3K36me3 but not H3K36me2 nor H3K4me3 is almost lost at the 5′-end of the gene *FLC* in *ubp26* mutant plants and this loss is associated with a reduction of *FLC* expression (Schmitz *et al*, 2009). The steady-state distribution pattern of H2Bub observed over expressed genes presumably results from targeted deubiquitylation of H2B at the 5′-end and probably 3′-end of the transcribed region, rather than from an increased ubiquitylation of H2B towards the middle of the gene.

Our epigenomic profiling of the three forms of H3K27 indicates that methylation of this lysine residue is generally associated with repressive chromatin and that its indexing function depends on the degree of modification (mono-, di- and tri-methylation). Thus, in agreement with previous studies, H3K27me3, which is the hallmark of CS2, is almost exclusively present over transcriptionally repressed genes (Turck *et al*, 2007; Zhang *et al*, 2007), while H3K27me1 is prevalent over silent TEs in pericentromeric regions, where it is thought to prevent over-replication (Jacob *et al*, 2009, 2010). Our analysis reveals in addition that H3K27me2 is enriched over H3K27me3-marked genes, as well as of over TE sequences. Although immunolocalization of H3K27me2 at chromocenters (Fuchs *et al*, 2006) was proposed to result from cross-reactivity of antibodies with H3K27me1 in Arabidopsis (Jacob *et al*, 2009), we did not observe extensive cross-reactivity of the H3K27me2 antibodies used in our study with H3K27me1 (Supplementary Figure S5). Moreover, while all forms of methylated H3K27 can be found over genes and are associated with transcriptional repression, little overlap is observed between the small group of genes marked by H3K27me1 and the much larger set of genes marked by H3K27me2/3, suggesting that these modifications define two repressive pathways with distinct gene targets (Supplementary Tables VI and VII). Whereas H3K27me3 deposition is catalysed by the evolutionarily conserved Polycomb Repressive Complexes 2 (Kohler and Hennig, 2010; Bouyer *et al*, 2011), H3K27me1 deposition over TE sequences is partly dependent on the activity of the two SET-domain proteins ATXR5 and ATXR6 (Jacob *et al*, 2009). Whether H3K27me1 deposition over genes requires the same or different histone methyltransferases and whether it is associated with the control of DNA replication remain to be determined. Irrespective of the mechanisms involved, it is noteworthy that whereas H3K27me1-marked TE sequences are also co-marked with H3K9me2 and 5mC, this is not the case for H3K27me1-marked genes.

Acetylation of H3K56 is another chromatin mark that has been linked with the replication process. In Arabidopsis cell cultures, early replicating sequences form broad domains of H3K56ac (Lee *et al*, 2010b). Our epigenomic profiling of H3K56ac reveals mostly short domains located at the 5′-end of expressed genes, which correspond to the replication-independent incorporation of acetylated H3K56. However, a few large domains ($\sim$20 kb) are also detected, which span several genes, intergenic regions and TEs. As our epigenomic maps have been derived from whole seedlings that comprise only a small proportion of mitotic cells, these large H3K56ac domains might correspond to sequences frequently used as endoreplication origins.

©2011 European Molecular Biology Organization

Although most Arabidopsis genes are associated with chromatin states CS1 or CS2, ~10% are instead associated with CS4, which is characterized by the absence of any prevalent chromatin mark among the 12 that were analysed in this work (Figure 3). Analysis of additional chromatin marks and proteins will be required to determine more precisely the nature of CS4 and notably the extent of its similarity to the repressive chromatin type BLACK of Drosophila (Filion *et al*, 2010).

To conclude, the first integrative view of the Arabidopsis epigenome provided here could be compared with a first sketch, which is progressively refined until a complete blueprint is produced. Importantly, key aspects of the Arabidopsis epigenome are already apparent in this first sketch, like the relative simplicity of designing principles, which appears to be shared with metazoans.

Materials and methods

Immunoprecipitation of chromatin and methylated DNA, labelling and microarray hybridization

All experiments were performed using wild-type *Arabidopsis thaliana* accession Columbia seedlings grown for 10 days either in liquid MS (whole seedlings) or on MS agar plates (roots and aerial parts) supplemented with 1% sucrose under long day conditions. ChIP and Me-DIP assays were carried out essentially as described (Lippman *et al*, 2005) using commercially available antibodies (Supplementary Table IX; Supplementary Figure S5). Specificity of the H3K27me2 and H3K9me3 antibodies was tested by peptide competition and western blotting analysis on nuclear extracts (Supplementary Figure S5) as described in Bouyer *et al* (2011) using H3K27me3, H3K27me2, H3K27me1, H3K9me3 and H3K36me3 peptides (Millipore, 12-565, 12-566, 12-567, 12-568 and Diagenode sp-058-050, respectively). Immunoprecipitated DNA (IP) and input DNA (INPUT) were amplified, differentially labelled and co-hybridized in dye-swap experiments as described (Lippman *et al*, 2004; Turck *et al*, 2007) for the chromosome 4 tiling array or according to the manufacturer's instructions for the Roche NimbleGen whole-genome tiling array. Two biological replicates were analySed (two dye-swaps). The chromosome 4 array contains 21 800 printed features, on average ~900 bp in size. The heterochromatic knob on the short arm and several megabases of pericentromeric heterochromatin are included and account for 16% of the 18.6 Mb covered by the array. Details of array design and production are described in Vaughn *et al* (2007). This platform has been deposited to GEO under accession number GPL10172. The whole-genome tiling array consists of 50–75 nt tiles, with 110 nt spacing on average, that are tiled across the entire genome sequence (TAIR7), without repeat masking. Tiles have a melting temperature of 74°C on average and 88% of them match a unique position in the genome. This custom design was either split into two arrays of 360 718 tiles each, with every other tile on each array (GEO accessions GPL10911 and GPL10918) or synthesized in triplicates of 711 320 tiles each on a single array (GEO accession GPL11005).

ChIP- and Me-DIP-chip data analysis

Hybridization data were normalized as described previously for the chromosome 4 array (Turck *et al*, 2007) or using an ANOVA model was applied to remove technical biases from data obtained using the whole-genome array. Data were averaged on the dye-swap to remove tile-specific dye bias. Normalized data were analysed using the ChIPmix method (Martin-Magniette *et al*, 2008), which was adapted to handle multiple biological replicates simultaneously. This method is based on a mixture model of regressions, the parameters of which are estimated using the EM algorithm. For each tile, a posterior probability is defined as the probability to be enriched given the log(Input) and log(IP) intensities, and is used to assign each tile into a normal or enriched class. A false-positive risk is determined by defining the probability of obtaining a posterior probability at least as extreme as the one that is actually observed when the tile is normal. False-positive risks are then adjusted by the Benjamini–Hochberg procedure and tiles for which the adjusted false-positive risk is <0.01 are declared enriched. Previously published data (Turck *et al*, 2007; Vaughn *et al*, 2007) were re-analysed using the same procedure. Neighbouring enriched tiles are joined into domains by requiring a minimal run of 1.6 kb or 400 bp and allowing a maximal gap of 800 or 200 bp for data obtained using the chromosome 4 or whole-genome arrays, respectively. Thus, 'singletons' are not considered for further analyses.

Computational analyses

General bioinformatics methods including positional, quantitative and class-based computations were conducted in Excel and using *ad hoc* scripts written in R, PERL or Python. Genes and transposable elements were annotated based on TAIR8 and other sequences are assumed to be intergenic. Gene Ontology analyses were done using the GOrilla (Eden *et al*, 2009) with an additional correction for multiple testing of the *P*-values. Pairwise association analysis, which is directional unlike correlation analysis, was calculated by scoring the frequency of co-occurrence of pairs of chromatin modifications among the 12 marks analysed on the chromosome 4 tiling array.

Whole seedlings transcriptome data were retrieved from Schmid *et al* (2005) and genes were binned into 20 expression percentiles according to their absolute expression values. Within each expression percentile, the number of genes marked by a given chromatin modification was calculated and represented as a percentage of all the genes marked by this modification. Shannon entropy for each set of marked genes was calculated as described (Zhang *et al*, 2006) using publicly available developmental expression series (Schmid *et al*, 2005), after filtering genes that showed no expression in any conditions.

Fuzzy *c*-means clustering using R MCLUST package was performed to classify tiles into principal chromatin states based on the 12 epigenomic maps. *c*-means clustering computes membership values for each tile towards all the clusters and all the membership values add up to 1. Each tile was assigned to one cluster only, based on a membership value equal or higher to 0.5. To identify the optimal number of clusters (k), cluster validity value, which is an estimate of homogeneity within the clusters and heterogeneity between them, was calculated for clusters from $k=2$–11.

Data availability

Raw and processed data have been deposited to NCBI's Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under the super-series accession GSE24710 and to CATdb (http://urgv.evry.inra.fr/CATdb) (Samson *et al*, 2004; Gagnot *et al*, 2008). In addition, array data and genome annotation are displayed using a Generic Genome Browser, available for visualization at http://epigara.biologie.ens.fr/index.html.

Supplementary data

Supplementary data are available at *The EMBO Journal* Online (http://www.embojournal.org).

Acknowledgements

We thank members of the Colot group and Edith Heard for critical reading of the manuscript. This work was supported by grants from the Agence Nationale de la Recherche (ANR Genoplante TAG and REGENEOME, ANR blanc DDB1, ANR Sysbio) and by the European Union Network of Excellence 'The Epigenome'. IA, AS and SC were supported by PhD studentships from the ANR, the Centre National de la Recherche Scientifique (CNRS) and the Ministère de l'Enseignement Supérieur et de la Recherche (MESR), respectively.

Author contributions: FR and VC conceived and designed the experiments. FR, SC, DB, EC, LG, BD, SDr and FB performed the experiments. FR, IA, AS and VC analysed the data. CBe, TM-H, ED-B, LA-S, SDe, VB, SA, AS, CBo, M-LMM, SR and MC contributed reagents/materials/analysis tools. FR and VC wrote the paper.

Conflict of interest

The authors declare that they have no conflict of interest.

 ©2011 European Molecular Biology Organization

References

Berger SL (2007) The complex language of chromatin regulation during transcription. *Nature* **447:** 407–412

Bernatavichute YV, Zhang X, Cokus S, Pellegrini M, Jacobsen SE (2008) Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana. *PLoS One* **3:** e3156

Berr A, McCallum EJ, Menard R, Meyer D, Fuchs J, Dong A, Shen WH (2010) Arabidopsis SET DOMAIN GROUP2 is required for H3K4 trimethylation and is crucial for both sporophyte and gametophyte development. *Plant Cell* **22:** 3232–3248

Bouyer D, Roudier F, Heese M, Ellen D, Andersen ED, Gey D, Nowack MK, Goodrich J, Renou J-P, Grini PE, Colot V, Schnittger A (2011) Polycomb Repressive Complex 2 controls the embryo to seedling phase transition. *PLoS Genet* **7:** e1002014

Cao Y, Dai Y, Cui S, Ma L (2008) Histone H2B monoubiquitination in the chromatin of FLOWERING LOCUS C regulates flowering time in Arabidopsis. *Plant Cell* **20:** 2586–2602

Caro E, Castellano MM, Gutierrez C (2007) A chromatin link that couples cell division to root epidermis patterning in Arabidopsis. *Nature* **447:** 213–217

Charron JB, He H, Elling AA, Deng XW (2009) Dynamic landscapes of four histone modifications during deetiolation in Arabidopsis. *Plant Cell* **21:** 3732–3748

Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ, Casero D, Bernal M, Huijser P, Clark AT, Kramer U, Merchant SS, Zhang X, Jacobsen SE, Pellegrini M (2010) Relationship between nucleosome positioning and DNA methylation. *Nature* **466:** 388–392

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452:** 215–219

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10:** 48

Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28:** 817–825

Feng S, Jacobsen SE (2011) Epigenetic modifications in plants: an evolutionary perspective. *Curr Opin Plant Biol* (advance online publication; doi:10.1016/j.pbi.2010.12.002)

Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, van Steensel B (2010) Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143:** 212–224

Fuchs J, Demidov D, Houben A, Schubert I (2006) Chromosomal histone modification patterns–from conservation to diversity. *Trends Plant Sci* **11:** 199–208

Gagnot S, Tamby JP, Martin-Magniette ML, Bitton F, Taconnat L, Balzergue S, Aubourg S, Renou JP, Lecharny A, Brunaud V (2008) CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Res* **36:** D986–D990

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K *et al* (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330:** 1775–1787

Gu X, Jiang D, Wang Y, Bachmair A, He Y (2009) Repression of the floral transition via histone H2B monoubiquitination. *Plant J* **57:** 522–533

Hon G, Wang W, Ren B (2009) Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol* **5:** e1000566

Jackson JP, Johnson L, Jasencakova Z, Zhang X, PerezBurgos L, Singh PB, Cheng X, Schubert I, Jenuwein T, Jacobsen SE (2004) Dimethylation of histone H3 lysine 9 is a critical mark for DNA methylation and gene silencing in Arabidopsis thaliana. *Chromosoma* **112:** 308–315

Jacob Y, Feng SH, LeBlanc CA, Bernatavichute YV, Stroud H, Cokus S, Johnson LM, Pellegrini M, Jacobsen SE, Michaels SD (2009) ATXR5 and ATXR6 are H3K27 monomethyltransferases required for chromatin structure and gene silencing. *Nat Struct Mol Biol* **16:** 763–796

Jacob Y, Stroud H, Leblanc C, Feng S, Zhuo L, Caro E, Hassel C, Gutierrez C, Michaels SD, Jacobsen SE (2010) Regulation of heterochromatic DNA replication by histone H3 lysine 27 methyltransferases. *Nature* **466:** 987–991

Jenuwein T, Allis CD (2001) Translating the histone code. *Science* **293:** 1074–1080

Jiang D, Wang Y, He Y (2008) Repression of FLOWERING LOCUS C and FLOWERING LOCUS T by the Arabidopsis Polycomb repressive complex 2 components. *PLoS One* **3:** e3404

Johnson L, Mollah S, Garcia BA, Muratore TL, Shabanowitz J, Hunt DF, Jacobsen SE (2004) Mass spectrometry analysis of Arabidopsis histone H3 reveals distinct combinations of post-translational modifications. *Nucleic Acids Res* **32:** 6511–6518

Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, Linder-Basso D, Plachetka A, Shanower G, Tolstorukov MY, Luquette LJ, Xi R, Jung YL, Park RW, Bishop EP, Canfield TP *et al* (2011) Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471:** 480–485

Kohler C, Hennig L (2010) Regulation of cell identity by plant Polycomb and trithorax group proteins. *Curr Opin Genet Dev* **20:** 541–547

Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet* **41:** 376–381

Kouzarides T (2007) Chromatin modifications and their function. *Cell* **128:** 693–705

Lee JS, Smith E, Shilatifard A (2010a) The language of histone crosstalk. *Cell* **142:** 682–685

Lee TJ, Pascuzzi PE, Settlage SB, Shultz RW, Tanurdzic M, Rabinowicz PD, Menges M, Zheng P, Main D, Murray JA, Sosinski B, Allen GC, Martienssen RA, Hanley-Bowdoin L, Vaughn MW, Thompson WF (2010b) Arabidopsis thaliana chromosome 4 replicates in two phases that correlate with chromatin state. *PLoS Genet* **6:** e1000982

Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R (2004) Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430:** 471–476

Lippman Z, Gendrel AV, Colot V, Martienssen R (2005) Profiling DNA methylation patterns using genomic tiling microarrays. *Nat Methods* **2:** 219–224

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133:** 523–536

Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, Rando OJ (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biol* **3:** e328

Liu T, Rechtsteiner A, Egelhofer TA, Vielle A, Latorre I, Cheung MS, Ercan S, Ikegami K, Jensen M, Kolasinska-Zwierz P, Rosenbaum H, Shin H, Taing S, Takasaki T, Iniguez AL, Desai A, Dernburg AF, Lieb JD, Ahringer J, Strome S *et al* (2011) Broad chromosomal domains of histone modification patterns in C. elegans. *Genome Res* **21:** 227–236

Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T (2010) Regulation of alternative splicing by histone modifications. *Science* **327:** 996–1000

Luo C, Lam E (2010) ANCORP: a high-resolution approach that generates distinct chromatin state models from multiple genome-wide datasets. *Plant J* **63:** 339–351

Martin-Magniette ML, Mary-Huard T, Berard C, Robin S (2008) ChIPmix: mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics* **24:** i181–i186

Minsky N, Shema E, Field Y, Schuster M, Segal E, Oren M (2008) Monoubiquitinated H2B is associated with the transcribed region of highly expressed genes in human cells. *Nat Cell Biol* **10:** 483–488

Oh S, Park S, van Nocker S (2008) Genic and global functions for Paf1C in chromatin modification and gene expression in Arabidopsis. *PLoS Genet* **4:** e1000077

Pekowska A, Benoukraf T, Ferrier P, Spicuglia S (2010) A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Res* **20:** 1493–1502

©2011 European Molecular Biology Organization

Rando OJ, Chang HY (2009) Genome-wide views of chromatin structure. *Annu Rev Biochem* **78:** 245–271

Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, Tolstorukov MY, Gorchakov AA, Jaffe JD, Kennedy C, Linder-Basso D, Peach SE, Shanower G, Zheng H, Kuroda MI, Pirrotta V, Park PJ, Elgin SC, Karpen GH (2011) Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. *Genome Res* **21:** 147–163

Roudier F, Teixeira FK, Colot V (2009) Chromatin indexing in Arabidopsis: an epigenomic tale of tails and more. *Trends Genet* **25:** 511–517

Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S, Arshinoff BI, Ay F, Meyer PE, Robine N, Washington NL, Di Stefano L, Berezikov E, Brown CD, Candeias R *et al* (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330:** 1787–1797

Samson F, Brunaud V, Duchene S, De Oliveira Y, Caboche M, Lecharny A, Aubourg S (2004) FLAGdb++: a database for the functional analysis of the Arabidopsis genome. *Nucleic Acids Res* **32:** D347–D350

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005) A gene expression map of Arabidopsis thaliana development. *Nat Genet* **37:** 501–506

Schmitz RJ, Tamada Y, Doyle MR, Zhang X, Amasino RM (2009) Histone H2B deubiquitination is required for transcriptional activation of FLOWERING LOCUS C and for proper control of flowering in Arabidopsis. *Plant Physiol* **149:** 1196–1204

Schulze JM, Jackson J, Nakanishi S, Gardner JM, Hentrich T, Haug J, Johnston M, Jaspersen SL, Kobor MS, Shilatifard A (2009) Linking cell cycle to histone modifications: SBF and H2B monoubiquitination machinery and cell-cycle regulation of H3K79 dimethylation. *Mol Cell* **35:** 626–641

Sims III RJ, Reinberg D (2008) Is there a code embedded in proteins that is based on post-translational modifications? *Nat Rev Mol Cell Biol* **9:** 815–820

Squazzo SL, O'Geen H, Komashko VM, Krig SR, Jin VX, Jang SW, Margueron R, Reinberg D, Green R, Farnham PJ (2006) Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res* **16:** 890–900

Sridhar VV, Kapoor A, Zhang K, Zhu J, Zhou T, Hasegawa PM, Bressan RA, Zhu JK (2007) Control of DNA methylation and heterochromatic silencing by histone H2B deubiquitination. *Nature* **447:** 735–738

Strahl BD, Allis CD (2000) The language of covalent histone modifications. *Nature* **403:** 41–45

Suganuma T, Workman JL (2008) Crosstalk among histone modifications. *Cell* **135:** 604–607

Tanurdzic M, Vaughn MW, Jiang H, Lee TJ, Slotkin RK, Sosinski B, Thompson WF, Doerge RW, Martienssen RA (2008) Epigenomic consequences of immortalized plant cell suspension culture. *PLoS Biol* **6:** 2880–2895

Turck F, Roudier F, Farrona S, Martin-Magniette ML, Guillaume E, Buisine N, Gagnot S, Martienssen RA, Coupland G, Colot V (2007) Arabidopsis TFL2/LHP1 specifically associates with genes marked by trimethylation of histone H3 lysine 27. *PLoS Genet* **3:** e86

Vakoc CR, Mandat SA, Olenchock BA, Blobel GA (2005) Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin. *Mol Cell* **19:** 381–391

Vakoc CR, Sachdeva MM, Wang H, Blobel GA (2006) Profile of histone lysine methylation across transcribed mammalian chromatin. *Mol Cell Biol* **26:** 9185–9195

Vaughn MW, Tanurd Ic M, Lippman Z, Jiang H, Carrasquillo R, Rabinowicz PD, Dedhia N, McCombie WR, Agier N, Bulski A, Colot V, Doerge RW, Martienssen RA (2007) Epigenetic natural variation in Arabidopsis thaliana. *PLoS Biol* **5:** e174

Wang Z, Schones DE, Zhao K (2009) Characterization of human epigenomes. *Curr Opin Genet Dev* **19:** 127–134

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40:** 897–903

Weake VM, Workman JL (2008) Histone ubiquitination: triggering gene activity. *Mol Cell* **29:** 653–663

Williams SK, Truong D, Tyler JK (2008) Acetylation in the globular core of histone H3 on lysine-56 promotes chromatin disassembly during transcriptional activation. *Proc Natl Acad Sci USA* **105:** 9000–9005

Zhang X, Bernatavichute Y, Cokus S, Pellegrini M, Jacobsen S (2009) Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana. *Genome Biol* **10:** R62

Zhang X, Clarenz O, Cokus S, Bernatavichute YV, Pellegrini M, Goodrich J, Jacobsen SE (2007) Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS Biol* **5:** e129

Zhang XY, Yazaki J, Sundaresan A, Cokus S, Chan SWL, Chen HM, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, Ecker JR (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell* **126:** 1189–1201

Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* **12:** 7–18

Zilberman D, Coleman-Derr D, Ballinger T, Henikoff S (2008) Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature* **456:** 125–129

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2006) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* **39:** 61–69

SOME RIGHTS RESERVED

***The EMBO Journal* is published by *Nature Publishing Group* on behalf of *European Molecular Biology Organization*. This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported License. [http://creativecommons.org/licenses/by-nc-sa/3.0/]**

©2011 European Molecular Biology Organization

Annexe E

Article publié dans *SAGMB*

Statistical Applications in Genetics and Molecular Biology

Volume 10, *Issue* 1 2011 *Article* 50

Unsupervised Classification for Tiling Arrays: ChIP-chip and Transcriptome

Caroline Bérard, *UMR AgroParisTech/INRA MIA 518*
Marie-Laure Martin-Magniette, *UMR AgroParisTech/ INRA MIA 518, URGV UMR INRA/CNRS/UEVE*
Véronique Brunaud, *URGV UMR INRA/CNRS/UEVE*
Sébastien Aubourg, *URGV UMR INRA/CNRS/UEVE*
Stéphane Robin, *UMR AgroParisTech/INRA MIA 518*

Recommended Citation:
Bérard, Caroline; Martin-Magniette, Marie-Laure; Brunaud, Véronique; Aubourg, Sébastien; and Robin, Stéphane (2011) "Unsupervised Classification for Tiling Arrays: ChIP-chip and Transcriptome," *Statistical Applications in Genetics and Molecular Biology*: Vol. 10: Iss. 1, Article 50.
DOI: 10.2202/1544-6115.1692
Available at: http://www.bepress.com/sagmb/vol10/iss1/art50

©2011 De Gruyter. All rights reserved.

Unsupervised Classification for Tiling Arrays: ChIP-chip and Transcriptome

Caroline Bérard, Marie-Laure Martin-Magniette, Véronique Brunaud, Sébastien Aubourg, and Stéphane Robin

Abstract

Tiling arrays make possible a large-scale exploration of the genome thanks to probes which cover the whole genome with very high density, up to 2,000,000 probes. Biological questions usually addressed are either the expression difference between two conditions or the detection of transcribed regions. In this work, we propose to consider both questions simultaneously as an unsupervised classification problem by modeling the joint distribution of the two conditions. In contrast to previous methods, we account for all available information on the probes as well as biological knowledge such as annotation and spatial dependence between probes. Since probes are not biologically relevant units, we propose a classification rule for non-connected regions covered by several probes. Applications to transcriptomic and ChIP-chip data of *Arabidopsis thaliana* obtained with a NimbleGen tiling array highlight the importance of a precise modeling and of the region classification. The "TAHMMAnnot" package is implemented in R and C and is freely available from CRAN.

KEYWORDS: bivariate Gaussian mixture, hidden Markov model, tiling arrays, unsupervised classification

Author Notes: The authors thank F. Roudier and V. Colot from IBENS for providing ChIP-chip data and for helpful discussions of biological interpretation, and S. Derozier and A. Lecharny from URGV for their help in bioinformatic analysis. The authors are grateful to S. Balzergue from URGV for providing transcriptomic tiling array data. This work was funded by MIA, GAP and MICA departments of INRA.

1 Introduction

For 15 years, the study of large-scale genomes has been possible thanks to DNA microarrays which originally had probes designed on genes. The tiling arrays now propose probes which cover the whole genome without *a priori* knowledge of structural annotation. The density is still increasing and companies now offer tiling arrays with 2 million probes. Thanks to technological advances and to the miniaturization of the support, tiling arrays have become a usual tool in biology laboratories. They make possible a large-scale exploration of the genome with a reasonable cost. Recently, Next Generation Sequencing (NGS) technology has revolutionized the domain because it directly produces nucleotide sequences. However, like any new technology, it remains expensive and suffers for now from uncontrolled technical biases (Oshlack, Robinson, and Young, 2010). NGS technology also raises new questions on read mapping or genome assembly. For all these reasons, tiling arrays, with a technology which is well controlled, remain widely used. They are a powerful tool for analyzing all kinds of experiments and are used in a wide range of studies such as DNA methylation, chromatin modification or transcription factor analysis with ChIP-chip experiments (Buck and Lieb, 2004), DNA copy number variation detection with CGH (Pinkel, Segraves, Sudar, Clark, Poole, Kowbel, Collins, Kuo, Chen, and Zhai, 1998, Snijders, Nowak, Segraves, Blackwood, and *et al.*, 2001) and surveys of genomic transcriptional activities or transcript mapping with transcriptional experiments (Mockler, Chan, Sundaresan, and *et al.*, 2005, Yamada, Lim, Dale, and *et al.*, 2003, Hanada, Zhang, Borevitz, Li, and Shiu, 2007).

For comparative genomic hybridization, many different approaches exist for determining DNA copy number variations in CGH data such as segmentation (Hupé, Stransky, Thiery, Radvanyi, and Barillot, 2004, Picard, Robin, Lavielle, Vaisse, and Daudin, 2005) or Hidden Markov Models (Fridlyand, Snijders, Pinkel, Albertson, and Jain, 2004, Seifert, Banaei, Keilwagen, Mette, Houben, Roudier, Colot, Grosse, and Strickert, 2009).

Transcriptomic experiments may have one of two different purposes: the detection of transcribed regions or the study of gene expression across several conditions (also called differential analysis). Most methods previously developed for tiling array transcriptomic data deal with the first purpose. Among them, some methods are based on probe-by-probe statistical tests (e.g. Fisher test developed by Halasz, van Batenburg, Perusse, Hua, Lu, White, and Bussemaker (2006)) and others are based on segmentation methods such as Huber, Toedling, and Steinmetz (2006) or Zeller, Henz, Laubinger, Weigel, and Rätsch (2008) or HMM (Nicolas, Leduc, Robin, Rasmussen, Jarmer, and Bessières, 2009). The incorporation of annotation knowledge has also been proposed in a supervised framework (Du, Rozowsky, Korbel, Zhang, Royce, Schultz, Snyder, and Gerstein, 2006; Munch,

Gardner, Arctander, and Krogh, 2006). Surprisingly few methods are devoted to the study of gene expression profiles across samples based on tiling arrays. Some methods aggregate probes within regions and then apply hypothesis testing. The method gSAM (Ghosh, Hirsch, Sekinger, Kapranov, Struhl, and Gingeras, 2007) is an extension of SAM, which models the differential expression of a given region by a constant piece-wise function. In the TileMap method (Ji and Wong, 2005) each probe is used separately and a test statistic is proposed, based on a hierarchical empirical Bayes model.

For ChIP-chip experiments where the chromatin immunoprecipitation sample (ChIP) and the reference sample of genomic DNA are compared, the main goal is to detect regions enriched by ChIP. Johnson, Li, Meyer, Gottardo, Carroll, Brown, and Liu (2006) proposed a Model-based Analysis of Tiling arrays (MAT) algorithm dedicated to Affymetrix arrays. MAT models the baseline probe behavior based on probe sequence characteristics and genome copy number. Li, Meyer, and Liu (2005) proposed to model the behavior of each probe and a 2-state HMM is then used to estimate the enrichment probability at each probe location. In these two methods it is assumed that only a small proportion of probes is enriched by ChIP. This assumption is reasonable for ChIP-chip experiments dealing with transcription factors but not for histone modification or DNA methylation where a large enrichment is expected. Humburg, Bulger, and Stone (2008) have suggested a parameter estimation procedure for robust HMM analysis of chromatin structure where several long regions of interest are expected. ChIP-chip data can also be seen as one signal along the genome when using the log-ratio between the intensities of the ChIP and the reference samples. Analyses are then usually done using a sliding window (Cawley, Bekiranov, Ng, and *et al.*, 2004) and statistic tests. Keles, van de Laan, Dudoit, and Cawley (2004) and He, Li, Zhou, Deng, Zhao, and Luo (2009) have respectively proposed the Welch t-statistic and the non parametric Wilcoxon rank-sum method.

When two samples are compared, most methods rely on the log-ratios. But this can mask the multimodality of the data due to the dimension reduction. To overcome this problem, Martin-Magniette, Mary-Huard, Bérard, and Robin (2008) have argued that is worth working directly with the two measurements of each probe. For ChIP-chip data, they have proposed to model the distribution of the ChIP sample conditionally to the reference sample by a mixture of two linear regressions. The same idea was used by Johannes, Wardenaar, Colomé-Tatché, and *et al.* (2010) to directly compare two ChIP samples. Assuming that the samples play a symmetric role, they introduced a bidimensional mixture model of four components with constraints on the mean parameters to study the differential enrichment.

In this article, we focus on the modeling of the joint distribution with an unsupervised classification point of view to study the difference between two ChIP

DOI: 10.2202/1544-6115.1692

or transcriptomic samples. Comparing the two samples requires distinguishing four different biologically interpretable groups of probes: a group with similar behavior in both samples, a group with higher intensity in the first sample than in the second sample, a symmetric group with higher intensity in the second sample and a last group with low intensity in both samples which can be viewed as noise, corresponding to the non-transcribed regions (cf Figure 1).

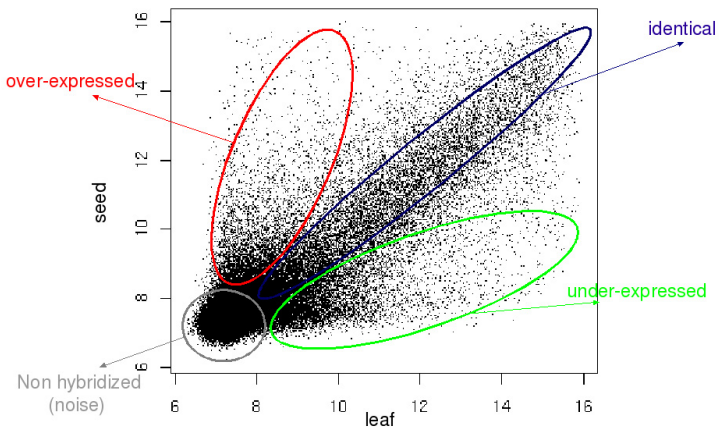

Figure 1: Schematic explanation of the 4 groups to consider when comparing two samples. Example of transcriptomic data.

A parametric classification method based on multivariate mixture models permits a direct comparison of two (or more) samples. This differs from a log-ratio based study which does not distinguish the group of identical behavior from the noise group. Our method simultaneously analyses the expression difference between two conditions and detects hybridized regions. In contrast to previous methods, we consider all the available information: the intensity of the two signals, the position of the probe along the genome and its structural annotation. The position of the probe is important because there is a signal dependence between adjacent probes due to the high resolution of tiling arrays. Structural annotation informs us about the location of the probes in intergenic, exonic or intronic regions (see Figure 2, screen capture of FLAGdb++ (Samson, Brunaud, Duchêne, De Oliveira, Caboche, Lecharny, and Aubourg, 2004)). This must be accounted for as, in a transcriptomic experiment, probes annotated as exonic are more likely to be hybridized whereas intergenic or intronic (non-coding) probes should be mainly in the noise group.

We use a 4-state heterogenous hidden Markov model with bidimensional Gaussian emission densities to gather all this information. Finally since genome annotation is an on-going process with possible errors, we will discuss the relevance of its use for each specific application.

Figure 2: Example of genome annotation. Yellow squares: probes; blue arrows: known genes. The arrows correspond to exons and the fine lines between arrows correspond to introns.

Most methods provide probe-by-probe results. As for the classification purpose, the HMM provides an answer for each probe, via the posterior probabilities. However probes are not relevant units from a biological point of view. Although HMM are widely used for classification problems in genomic data, the classification of biologically interesting regions is not a common practice. To get a result by region, the most commonly used method is a sliding window approach where the probe signals are merged *a priori*. Another method proposed by Li et al. (2005) is to define a region as at least two probes with positive log-odds enrichment values in the ChIP sample and at least one probe with a log-odds enrichment value lower than -15 in the control sample. But these methods do not deal with regions covered by several non-adjacent probes, such as genes with exons and introns. We propose a new solution deriving a posterior probability for a region given *a priori* with arbitrary structure (such as a non-connected region) and also a procedure of gene classification which allows us to quickly get a list of differentially expressed genes. This calculation clearly improves the classification of regions compared to the results derived from the classification of the probes.

The article is organized as follows. The statistical model is described in Section 2.1. The inference is given in Section 2.2. Section 2.3 describes the classification method for a probe and a gene. In Section 3, we discuss the different sub-models and the method is illustrated on NimbleGen tiling arrays for transcriptomic and ChIP-chip data of the plant *Arabidopsis thaliana*. We also perform a simulation study in Section 3.3 to compare our approach with three existing methods. The main conclusions and some possible extensions are discussed in Section 4. The method is implemented in R and C and is freely available from CRAN, with the "TAHMMAnnot" package.

DOI: 10.2202/1544-6115.1692

2 Methods

We propose a non-supervised classification model to compare the intensities of the two samples hybridized on the array for each probe. It accounts for all available information for each probe: the two intensities to be compared, the position of the probe along the chromosome and the current annotation of the probe (for example exonic, intronic, intergenic, transposable element, etc). As in Johannes et al. (2010), our method does not deal with the usual log-ratio but rather considers the two intensities separately and the joint signal of the two samples is modeled.

2.1 Model

For probe t, we denote

- $X_t = (X_{t1}, X_{t2})$ the log-intensities for both samples,
- $C_t \in \{1, ...P\}$ the annotation category,
- $Z_t \in \{1, ...K\}$ the unknown status.

In our case, $K = 4$, Groups 1 and 2 will refer respectively to 'noise' and 'identical' probes, whereas Groups 3 and 4 will refer to differentially hybridized probes. To account for the dependence between adjacent probes, we assume that the process $\{Z_t\}$ is a first order Markov chain with heterogenous transition π^p depending on the annotation category:

$$P(Z_t = l | Z_{t-1} = k, C_t = p) \quad = \quad \pi^p_{kl}$$

We then assume that the $\{X_t\}$ are independent conditionally to the $\{Z_t\}$ with distribution

$$(X_t | Z_t = k) \quad \sim \quad \mathcal{N}(\mu_k, \Sigma_k).$$

The parameters μ_k and Σ_k of the Gaussian distribution do not depend on the annotation category.

If there is no spatial dependence, the $\{Z_t\}$ are independent and distributed according to a multinomial of parameter π^p_k which corresponds to the proportion of probes from group k in annotation category p. If there is no annotation and no spatial dependence, the model comes down to a mixture model with four components. All these sub-models are discussed in Section 3. We focus on the model with $K = 4$ groups, but the models with $K = 2$ and $K = 3$ are also possible if a differential expression does not exist in one (or two) directions. The constraint of the orientation of the first two components remains unchanged. If there is no biological information about the number of groups, the choice between a model with 2, 3 or 4 groups can be made with a selection criterion such as BIC.

2.2 Inference

We use the parametrization proposed by Banfield and Raftery (1993) which enables us to characterize geometric properties of the Gaussian density (volume, shape, orientation). This parametrization considers the eigenvalue decomposition of the variance matrix of the group k:

$$\Sigma_k = D_k \Lambda_k D_k'.$$

The matrix Λ_k describes both the volume and the shape of the ellipse associated with the Gaussian distribution. The matrix D_k describes the orientation of this ellipse. A similar decomposition of variance matrix is studied by Celeux and Govaert (1995) in the Gaussian mixture context and is implemented in the Mixmod software (Biernacki, Celeux, Echenim, Govaert, and Langrognet, 2007) and in the Mclust R package (Fraley and Raftery, 2006, revised 2010). In their approach, each term of the decomposition is either equal in all groups or specific to each group.

In our case we need an intermediate modeling. By definition groups 1 and 2 should have the same orientation (see Figure 1), which implies that $D_1 = D_2$. Furthermore the dispersion around the main axis is expected to be similar in all groups, which amounts to fixing the second eigenvalue of Σ_k for all groups. This can be summarized as

$$\begin{aligned} \Sigma_k &= D_k \Lambda_k D_k', \qquad \text{for } k = 1, \ldots, 4; \\ D_1 &= D_2; \\ \Lambda_k &= \begin{pmatrix} u_{1k} & 0 \\ 0 & u_2 \end{pmatrix}, \qquad \text{with } u_{1k} > u_2, \text{ for } k = 1, \ldots, 4. \end{aligned}$$

The parameters $\{\pi^p\}$, $\{\mu_k\}$, $\{D_k\}$, $\{u_{1k}\}$ and u_2 are estimated using the EM algorithm. The E-step is achieved with the Forward-Backward algorithm (Baum (1972), Rabiner (1989)). This model requires a specific M-step to satisfy the prescribed constraints on the variance matrices (see Appendix for formulas). These constraints cannot be satisfied with Mixmod or Mclust. In Johannes et al. (2010) the constraints are related to the means which assumes a strong symmetry in the distribution of the data.

2.3 Posterior probabilities for a region

The posterior probability for each probe to belong to each group

$$\tau_{tk,X} = P(Z_t = k | X, C), \qquad \text{where } X = \{X_t\},$$

is obtained as a by-product of the Forward-Backward algorithm and can be used for probe classification. Nevertheless, the probe may not be the relevant biological entity and we would rather look at the status of a region such as a gene or a

DOI: 10.2202/1544-6115.1692

transposable element. We propose the analogous calculation of the posterior probabilities for a region. Many quantities could be calculated but our criterion is the natural classification criterion for a region in the context of HMM. It is a generalization of the posterior probabilities, which allows us to locally restore the hidden path. A region is declared in the group k if the hidden path remains in the k-th state throughout this region.

We define a region as a set of probes that can be decomposed into sub-regions of adjacent probes. As a reference to the gene structure and without loss of generality, we will refer to these sub-regions as 'exons' and to the spaces between them as 'introns'. In eukaryotic genes, exons correspond to coding regions that are spliced together in the transcript to become the mRNA, after removal of introns, which are not expressed. We define the posterior probability for such a region g to belong to group k as the probability for all its probes to belong to group k:

$$Q_{gk,X} = P(\forall t \in g, Z_t = k | X, C) \tag{1}$$

A region is covered by several probes and our definition considers the case of a homogeneous region, which is when all probes have the same status. We compute this probability for a gene g with Q exons (and $Q-1$ introns). We denote e_q the position of the first probe of exon q and i_q the position of the first probe of intron q; thus $i_q - 1$ refers to the last probe of exon $q-1$. As convention, we denote i_Q the position of the first probe after the end of the gene. We also denote $X_u^v = \{X_t\}_{u \le t \le v}$. We get

$$\begin{aligned}
Q_{gk,X} &= P(\forall t \in g, Z_t = k | X, C) \\
&= P(Z_{e_1} = k | X_1^{e_1}) \times \left(\prod_{t=e_1+1}^{i_1-1} A_{k,t} \right) \times \prod_{q=2}^{Q-1} \left(B_{k,q} \times \prod_{t=e_q+1}^{i_q-1} A_{k,t} \right) \\
&\quad \times B_{k,Q} \times \left(\prod_{t=e_Q+1}^{i_Q-2} A_{k,t} \right) \times P(Z_{i_Q-1} = k | Z_{i_Q-2} = k, X_{i_Q-1}^n), \\
\text{where} \quad A_{k,s} &= P(Z_s = k | Z_{s-1} = k, X_s, C), \\
B_{k,q} &= P(Z_{e_q} = k | Z_{i_{q-1}-1} = k, X_{i_{q-1}}^{e_q}, C), \text{ with } C = \{C_t\}.
\end{aligned}$$

All these terms can be calculated with the Forward recursion of the Forward / Backward algorithm. Note that the sum of the $Q_{gk,X}$ for $k \in \{1, ...4\}$ is not equal to one, as all probes from the same gene may not have the same status. Changing the list of exons associated to a gene allows us to account for alternative splicing or to exclude the last exon for which the expression level could be lower due to the labeling protocol (Nicolas et al., 2009).

3 Applications

We now illustrate the use of the proposed modeling on both ChIP-chip and transcriptomic data. All experiments have been carried out on a two-color NimbleGen array of about 700 000 probes designed to insure a constant hybridization temperature. For each dataset two biological replicates are available, for which hybridizations are performed in dye-swap. The normalization step is done by averaging on the dye-swap the two signals of each technical replicate to remove the gene-specific dye bias (Mary-Huard, Picard, and Robin, 2006). Analyses are performed per chromosome on the normalized data. Then we present a simulation study to compare our approach with 3 existing methods.

3.1 ChIP-chip dataset

We analyse the data from a histone modification (H3K9me2) study in *Arabidopsis thaliana* for a wildtype and a mutant (polIV). We directly compare the ChIP samples of the wildtype and the mutant to study their difference in methylation.

The methylation mainly affects transposable elements but also large adjacent regions (Humburg et al., 2008). Therefore the enriched probes are expected to be found both in the transposable elements and in wide neighboring regions. As the methylation does not affect a specific annotation category, the standard annotation information is not useful to detect enriched probes. This suggests using an HMM model without the annotation knowledge.

The histone methylation under study is known to be weakly present in the genome and the mutant is known to have a loss of methylation compared to the wildtype (Bernatavichute, Zhang, Cokus, Pellegrini, and Jacobsen, 2008). We find consistent results as shown by the estimated proportions in each group given by our model: 43% noise, 21% identical, 22% loss in mutant, 14% gain in mutant. The studied histone modification is also known as a heterochromatin mark. Most regions covered by H3K9me2 are adjacent and cover several megabases in pericentromeric regions or in interstitial heterochromatin regions (a tightly packed form of chromatin) as the knob of chromosome 4, but there are also smaller regions (islands of heterochromatin) located in euchromatin (a lightly packed form of chromatin) and covering mainly transposable elements (Bernatavichute et al., 2008). The results obtained using our method corroborate this information: 91.3% of probes in heterochromatin are methylated whereas only 49.5% of probes in euchromatin are methylated. In heterochromatin, 82% of probes have identical behaviour between wildtype and mutant whereas only 9.5% of probes are identical in euchromatin. Moreover 56% of methylated probes cover transposable elements or a 500 base-pair (*bp*) surrounding region.

DOI: 10.2202/1544-6115.1692

The transition probabilities provide insights about the length of regions from each group through mean sojourn time. The average size of the binding sites is 14.3 probes (corresponding to 3289*bps*) for the identical group, 4.5 probes (1035*bps*) for the group with lost in mutant, 3.7 probes (851*bps*) for the group enriched in mutant and 7.7 probes (1771*bps*) for the noise group. These calculations show that impoverished or enriched regions are three times smaller than regions with identical behaviour between wildtype and mutant. Moreover, the transposable elements are 2 to 3 times smaller in the euchromatin compared to the heterochromatin. This suggests that most of the methylation losses of the mutant occur in transposable elements from the euchromatin. The transposable element META1 (located between positions 5326458 and 5331580 on chromosome 4) is known to have a loss of methylation in the mutant. The regulatory region of META1 is located at the beginning of the transposable element with small RNAs which are involved in the methylation process. Our method declared the first half of the probes covering META1 (near the start position) in the group where methylation is lost. The other probes are declared identically methylated between the two samples. This example shows the advantage of the high resolution of the tiling array.

Comparison with the models of Johannes et al. (2010) As in Section 2.1, Groups 1 and 2 refer respectively to 'noise' and 'identical' probes, whereas Groups 3 and 4 correspond to differentially enriched probes. Johannes et al. (2010) proposed two mixture models of 4 bidimensional Gaussian distributions with constraints on the mean parameters. The first model is a full-switching model (Model 2) where the component means are constrained as follows: $\vec{\mu}_1 = (\mu_1, \mu_1)$, $\vec{\mu}_2 = (\mu_2, \mu_2)$, $\vec{\mu}_3 = (\mu_2, \mu_1)$, $\vec{\mu}_4 = (\mu_1, \mu_2)$ and the covariance matrices of Groups 3 and 4 are equal ($\Sigma_3 = \Sigma_4$). The flexible-switching model (so-called Model 3) is the full-switching model (Model 2) with less restrictive constraints on $\vec{\mu}_3 = (\mu_4, \mu_3)$ and $\vec{\mu}_4 = (\mu_3, \mu_4)$. We compared the HMM model with their models on the H3K9me2 dataset. Model 2 leads to a smaller proportion of differentially enriched regions (7.8% lost in mutant and 1.2% gain in mutant, see Figure 3) than the HMM (22% and 14% respectively). The transposable element META1 that is declared differentially enriched with our model (see above) is found in the identical group according to their Model 2. The classification of Model 3 seems to be unsuitable for probes with similar intensities between 8 and 10 where more probes are expected to be declared in the identical group (see Figure 3). By comparing the results of the HMM with the results given by our simplest model (mixture model, without dependence), we note that the good definition of the differentially enriched probes is a consequence of the constraints on the variance matrices we put in the model. In fact, the added advantage of HMM is only to provide blurred boundaries between groups

(due to the dependence assumption). In conclusion, it seems that the independence assumption, the symmetrical constraints on the means, and the equal variances for the differentially enriched probes lead to a model which is too simple to analyze such data. These two models also do not fit well the transcriptional dataset defined in Section 3.2 (results not shown).

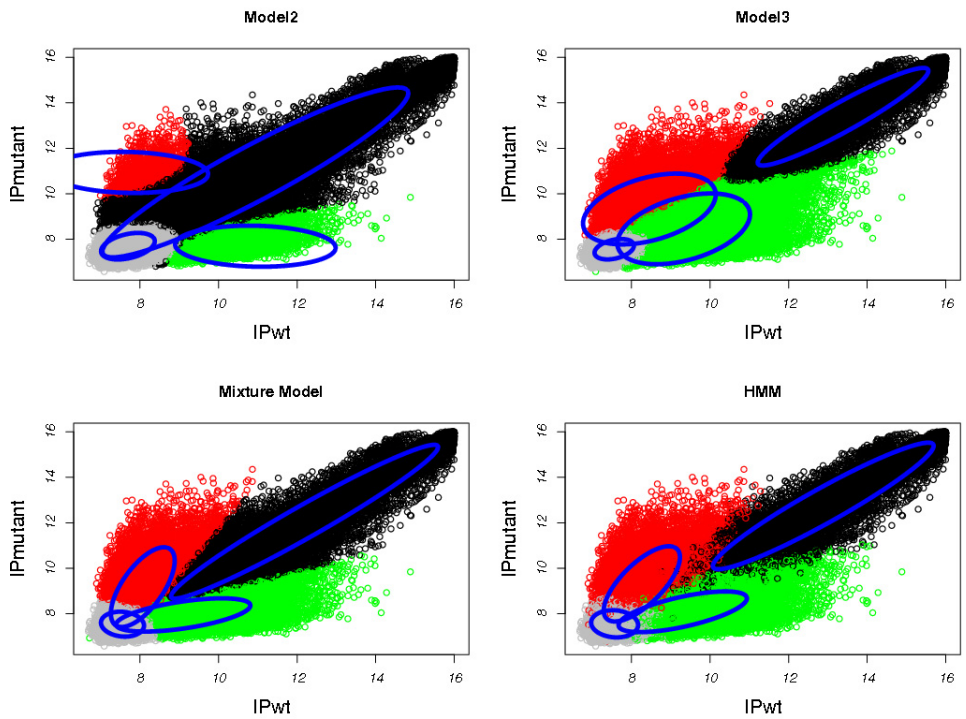

Figure 3: Classification comparison between the two models of Johannes et al. (2010) and the mixture model and HMM on H3K9me2 dataset. top left: full-switching model (Model 2), top right: flexible-switching model (Model 3), bottom left: mixture model, bottom right: HMM.

3.2 Transcriptional dataset

We now study the gene differential expression between the leaf and the seed 10 days after pollination of the plant *Arabidopsis thaliana*. First we compare the 4 sub-models, second we present the results by gene, then we consider the detection of new transcribed regions. In the results, over-expressed (under-expressed) refers to probes with a higher (smaller) signal in the seed.

DOI: 10.2202/1544-6115.1692

Table 1: Fit of the 4 models. $\mathscr{M}_1$ = mixture, $\mathscr{M}_2$ = HMM, $\mathscr{M}_3$= mixture + annotation, $\mathscr{M}_4$ = HMM + annotation.

	$\mathscr{M}_1$	$\mathscr{M}_2$	$\mathscr{M}_3$	$\mathscr{M}_4$
number of parameters	19	31	25	61
-2 log-likelihood	406249	371309	373283	356617
BIC	406469	371668	373573	357323
ICL	436197	412706	399986	398272

3.2.1 Comparison of the 4 models

The model presented in Section 2.1 uses all available information and is referred to as model $\mathscr{M}_4$. For the annotation, $P = 3$ categories are considered: intergenic, intron and exon, and only exonic RNA is expected to be found in the sample. Model $\mathscr{M}_4$ can be simplified if either the structural annotation (Model $\mathscr{M}_2$) or spatial dependence between probes (Model $\mathscr{M}_3$) is not taken into account. Model $\mathscr{M}_1$ is the simplest model with neither annotation nor spatial dependence. It comes down to an independent mixture model. The constraints on the variance matrices detailed in Section 2.2 are kept in all models. Table 1 presents the fit of the four models for chromosome 4. First we note that combining the HMM with the annotation information leads to a real improvement in terms of likelihood. This can be seen when comparing models $\mathscr{M}_1$ and $\mathscr{M}_3$ or models $\mathscr{M}_2$ and $\mathscr{M}_4$. In both cases, the best BIC is obtained when adding the annotation in the model. The full model $\mathscr{M}_4$ achieves the best BIC criterion, suggesting that all available information should be taken into account. Biernacki, Celeux, and Govaert (2000) proposed an alternative selection criterion named ICL dedicated to classification purposes. In contrast to BIC which aims at finding the best fitting of the data distribution, the purpose of the ICL criterion is to assess the number of mixture components that leads to the most reliable clustering (with small entropy of the posterior distribution). It is a penalised criterion based on the integrated likelihood which corresponds to the BIC penalised by the entropy. ICL has been established in the independent mixture context but Celeux and Durand (2008) have shown in a simulation study that it seems to have the same behaviour in the HMM context. According to ICL, model $\mathscr{M}_4$ is also chosen, so we use model $\mathscr{M}_4$ to compare the two transcriptomic samples.

As expected, intergenic probes mostly belong to the noise group (84%) and few belong to expressed groups: 9% in the under-expressed group and 6% in the over-expressed. Intronic probes display a similar, although different, repartition: 60% noise, 7% identical, 24% under-expressed and 9% over-expressed (cf Section 3.2.3 for discussion about expressed probes in intergenic and intronic regions). As expected, most exonic probes (78%) belong to the expressed groups: 41% identi-

cal, 23% under-expressed and 14% over-expressed. The transition matrices for the intronic and intergenic categories are very similar (not shown): whatever the status of probe t, probe $t+1$ has a 70% to 95% chance of being noise. This is different for the exonic probes where the transition matrix has high probabilities on the diagonal meaning that probe $t+1$ has high probability (80% to 90%) of having the same status as probe t. All these results seem to be coherent with what is expected for transcriptomic data.

3.2.2 Gene classification

We now consider the classification of each gene. To this end, we compute the posterior probability $Q_{gk,X}$ defined in Equation (1). The advantage of our classification criterion is to offer the possibility not to classify the gene if it is too heterogeneous in terms of probe status covering the gene. In fact we propose to classify the genes via a two-step procedure. The probability for a gene to be homogeneous whatever the status is $\sum_k Q_{gk,X}$. We first verify whether the gene has homogeneous status by considering a ratio similar to a Bayes factor: $\sum_k Q_{gk,X} / \sum_k Q_{gk}$, where $Q_{gk} = P(\forall t \in g, Z_t = k|C)$ is the non-conditional version of $Q_{gk,X}$, which is for a gene

$$Q_{gk} = m_k^E \times (\pi_{kk}^E)^{\sum_{q=1}^{Q}(i_q - e_q) - 1} \times \prod_{q=1}^{Q-1} [(\pi^I)^{e_{q+1} - i_q}]_{kk} \ ,$$

where superscripts E and I refer to the exonic and intronic categories, respectively.

As its computation involves a product of probabilities with as many terms as the number of exonic probes in the gene, $Q_{gk,X}$ goes to zero for long genes. The ratio with Q_{gk} does not correct this effect; therefore we apply an additional linear correction on the log-ratio with respect to the length of exons and the number of exons in the gene. We define this corrected log-ratio as a *unistatus* value which is a tool for decision support. If the homogeneous assumption seems verified, the second step is to calculate the conditional posterior probability $Q_{gk,X} / \sum_l Q_{gl,X}$ to assign the gene to the group k for which this posterior probability is the highest.

We found 80% of genes which have a unistatus value higer than 0 (corresponding to 22528 genes). Among these 22528 genes, 11900 are declared identically expressed in the seed and in the leaf, 3632 are declared under-expressed in the seed and 2667 are declared over-expressed in the seed. It is difficult to interpret biologically these results given the fact that the functional annotation is still unclear. Available tools are databases such as Genevestigator (Zimmermann, Hirsch-Hoffmann, Hennig, and Gruissem, 2004) which makes possible the visualization of gene expression across thousands of experimental conditions through data from Affymetrix microarrays. To illustrate how our results confirm the actual knowledge

DOI: 10.2202/1544-6115.1692

in biology, we identified 96 genes linked to seed using Flagdb++ which summarizes all available information on annotation. Among them, 70 have a probe on Affymetrix microarray and are so represented in Genevestigator. Only 8 genes are known for sure to be specifically expressed in the mature silique which is the experimental condition under study hybridized on the tiling array. For the other 62 genes, their expression is not located specifically in the mature silique but also in other development stage of seed. Among these 8 genes, 7 have a unistatus value higer than 0 and are declared over-expressed in the seed with our calculation. For the others, the expression is not clearly located in seed, which makes it difficult the comparison.

3.2.3 Detection of new transcripts

Although our model is built for the comparison of two samples, it also allows the detection of previously unknown transcription sites thanks to the high resolution of the tiling array. To this aim, the model without annotation seems more suitable, since we are bringing it into question. A lot of regions with expressed probes are found in intergenic regions: 1328 small regions with 2 or 3 consecutive expressed probes, 185 regions with 4 or 5 consecutive expressed probes and 90 regions with more than 5 consecutive probes (including 25 regions with more than 10 consecutive probes). For the 90 regions with more than 5 consecutive probes, we check with other annotation information such as Expressed Sequence TAG (EST) or genes predicted by the Eugene software (Schiex, Moisan, and Rouzé, 2001) which are not yet in the official TAIR annotation. We found 39 regions matching with annotation like small RNA, rRNA, tRNA, including 12 regions corresponding to a coding sequence defined in Eugene and 10 corresponding to transcriptional units recently annotated due to the presence of EST. Figure 4 (from FLAGdb++ (Samson et al., 2004)) shows examples of results for two annotated genes and also for two expressed regions which correspond to EST and Eugene genes. Moreover the obtained results show many other interesting things, such as surprisingly many transcriptions in the introns in 5'UTR (40% of intronic probes declared expressed in Section 3.2.1). This seems to be consistent with a recent article of Cenik, Derti, Mellor, Berriz, and Roth (2010) assuming a functional role of 5'UTR short introns.

3.3 Simulation study

We performed a simulation study to compare our approach with 3 existing methods: ChIPOTle (Buck, Nobel, and Lieb, 2005), the method of Johannes et al. (2010) and the one of Seifert et al. (2009). As there is no annotation in the simulation datasets,

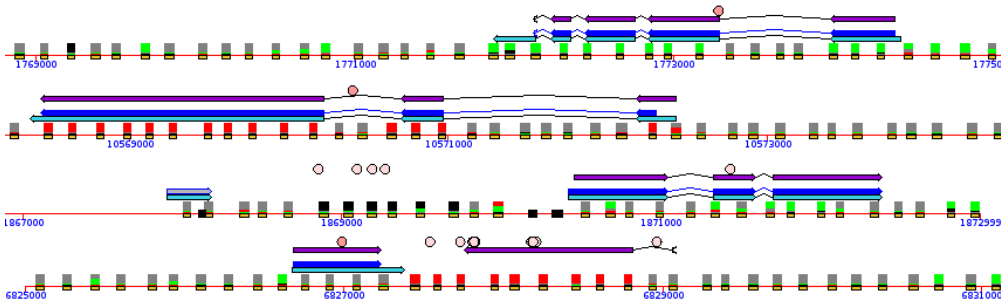

Figure 4: Presentation of results in Flagdb++. Circles represent EST, blue arrows are official TAIR annotation genes and violet arrows represent Eugene genes. The small rectangles are probes colored according to their status: grey if not expressed, black if identically expressed and red or green if differentially expressed. On the first 2 lines, there are 2 expressed genes covered by a majority of probes with the same status except the intronic probes which are reasonably declared as not expressed. The last 2 lines present expressed probes where there are no official genes but the signal coincides with EST and/or Eugene genes.

we applied our model $\mathscr{M}_2$ corresponding to the simple HMM with constraints on the variance matrices.

3.3.1 Design

We generated two datasets of size $n = 90\,000$ for which we are interested in retrieving the differentially expressed probes. The first two datasets are simulated with a latent variable Z being a first order Markov chain taking its values in $\{1,...,4\}$. The transition matrix π and the stationary distribution m have been adjusted on real datasets described below. To overcome the Gaussian assumption, the observations X were sampled according to an empirical distribution in each of the four groups of a real dataset. More precisely, the observations are sampled from real datasets in order to be similar to realistic tiling array data. The resampling is done using the posterior probabilities as weight for each probe. Two real datasets have been chosen with different expected proportions of differentially expressed probes. The first dataset, presented in Section 3.1, concerns the study of a histone mark (H3K9me2) in *Arabidopsis thaliana* for a wildtype and a mutant (polIV). In this dataset, about 30% of probes are expected to be differentially expressed. The second dataset is a publicly available ChIP-chip dataset coming from Penterman, Zilberman, Huh, Ballinger, Henikoff, and R.L. (2007) which compares the methylation profile of a

DOI: 10.2202/1544-6115.1692

wild-type *Arabidopsis* plant to that of a triple mutant. It leads to low proportions of differentially expressed probes.

We analysed the synthetic datatsets with Model $\mathscr{M}_2$ and with three other methods.

- ChIPOTle is a method dedicated to peak-finding in classical ChIP-chip experiments. Therefore it only provides two populations. It uses a sliding window approach based on the log-ratio. The window size and step parameters have to be tuned. With default parameters, ChIPOTle detects only one peak. We put window=200 and step=50, which seems to be a good combination giving a reasonable number of peaks for each simulated dataset. We used the absolute value of the log-ratio to mimic situations usually analysed.
- Seifert et al. (2009) proposed a three-state HMM that models the log-ratios of the two intensities. It requires the incorporation of *a priori* knowledge using prior distributions. The choice of the priors is not easy and those given by default do not provide three populations. Hence we modified them. We put startDistribution = (0.1,0.7,0.2), means = (-1,0.0,1), stds = (0.3,1,0.5), scaleOfAprioriMeans = (0.1,1,75) and shapeOfStandardDeviations = (20;1;100).
- Johannes et al. (2010) proposed two mixture models of four bidimensional Gaussian distributions with constraints on the mean parameters for the simultaneous analysis of two samples. These two models are described more precisely in Section 3.1.

In the simulated datasets, we are looking for the four groups defined in Section 1. The classification is done with the MAP rule for our model $\mathscr{M}_2$ and the model of Johannes et al. (2010). However, ChIPOTle and the method of Seifert et al. (2009) do not provide four groups: they merge the noise and the identical groups. The method of Seifert et al. (2009) applies the Viterbi algorithm to determine the most probable state sequence and give a classification of probes into three groups. For ChIPOTle, the differentially expressed probes are deduced from the detected peaks and there are only two groups. To compare our method with the ones of Seifert et al. (2009) and ChIPOTle, we summed the posterior probabilities of the noise and identical groups to obtain a classification into three groups, and also those of the over-expressed and under-expressed groups to obtain a classification into two groups.

The methods are compared using the classification results in terms of sensitivity, specificity and False Discovery Rate (FDR) for a given group k. The sensitivity is defined as: $\frac{TP_k}{TP_k+FN_k}$, the specificity is defined as $\frac{TN_k}{TN_k+FP_k}$, and the FDR is defined as $\frac{FP_k}{TP_k+FP_k}$, where:

$$TP_k = \sum_t \mathbf{1}_{(\hat{Z}_t=k)}\mathbf{1}_{(Z_t=k)} \qquad FN_k = \sum_t \mathbf{1}_{(\hat{Z}_t=k)}\mathbf{1}_{(Z_t\neq k)}$$
$$FP_k = \sum_t \mathbf{1}_{(\hat{Z}_t\neq k)}\mathbf{1}_{(Z_t=k)} \qquad TN_k = \sum_t \mathbf{1}_{(\hat{Z}_t\neq k)}\mathbf{1}_{(Z_t\neq k)}$$

We hence focus on the probes assigned as differentially expressed. Both sensitivity and specificity are expected to be large whereas FDR is hoped to be small.

3.3.2 Results

The results are presented in Tables 2 and 3. The peaks detected with the ChIPOTle method only represent between 31% and 35% of differentially expressed probes for the two datasets. The flexible-switching model (Model 3) of Johannes et al. (2010) provides better results than the full-switching model (Model 2); therefore we focus on Model 3. In the first dataset, Model 3 of Johannes et al. (2010) and the method of Seifert et al. (2009) have similar behavior. They respectively find 85% or 82% of under-expressed probes and 63% or 72% of over-expressed probes. In the second dataset where few probes are differentially expressed, the two methods behave differently. The method of Seifert et al. (2009) has difficulty finding the over-expressed group (only 55% of detected probes). On the contrary, Model 3 of Johannes et al. (2010) finds 100% of differentially expressed probes but detects a lot of false positives in return (FDR=99%). About 40 000 probes with similar intensities between 8 and 10 are declared differentially expressed whereas they are expected to be declared in the identical or in the noise group. In the two simulated datasets, our model $\mathscr{M}_2$ identifies more than 83% of differentially expressed probes, with fewer false positives (FDR between 2 and 13%). Among the four methods, Model $\mathscr{M}_2$ provided the best triplets of sensitivity, specificity and FDR whatever the proportion of differentially expressed probes in the dataset.

In order to overcome the Markovian dependency assumption, we also simulated two other datasets where the hidden path Z was sampled in a 4-state jump process with Markovian between-state transitions and Negative Binomial sojourn times (instead of Geometric). Similar results are obtained. The triplets of sensitivity, specificity and FDR obtained with Model $\mathscr{M}_2$ are (92,99,4) and (91,99,5) for the two differentially expressed groups in the first dataset respectively, and (99,100,3) and (85,100,16) in the second dataset. This ensures that our model is not too dependent on the Markovian assumption.

DOI: 10.2202/1544-6115.1692

Table 2: Dataset derived from H3K9me2 with a large proportion of differentially expressed probes. The notation corresponds to the order sensitivity, specificity, FDR in %.

	noise	identical	under-expressed	over-expressed
Z	38135	19527	19413	12925
ChIPOTle	99, 35, 27		35, 99, 6	
$\mathscr{M}_2$ with 2 groups	97, 94, 4		94, 94, 5	
Seifert *et al.*	92, 79, 11		82, 95, 18	72, 98, 13
$\mathscr{M}_2$ with 3 groups	98, 93, 4		92, 99, 6	90, 99, 6
Johannes *et al.* M3	94, 90, 12	82, 100, 0.1	85, 89, 31	63, 99, 11
TAHMMAnnot $\mathscr{M}_2$	96, 96, 5	99, 100, 1	92, 98, 6	90, 99, 6

Table 3: Dataset derived from Penterman et al. (2007) with a small proportion of differentially expressed probes. The notation corresponds to the order sensitivity, specificity, FDR, in %.

	noise	identical	under-expressed	over-expressed
Z	45300	43401	782	517
ChIPOTle	100, 31, 1		31, 100, 0	
$\mathscr{M}_2$ with 2 groups	100, 91, 0.1		91, 100, 6	
Seifert *et al.*	98, 99, 0		100, 99, 48	55, 98, 84
$\mathscr{M}_2$ with 3 groups	100, 91, 0.1		97, 100, 2	82, 100, 12
Johannes *et al.* M3	23, 96, 13	74, 79, 12	100, 100, 31	100, 56, 99
TAHMMAnnot $\mathscr{M}_2$	85, 84, 16	83, 85, 16	97, 100, 2	83, 100, 13

4 Discussion

Tiling array is a powerful technology which requires adapted statistical methods to deal with the large quantity and the variability of data. We focus on the comparison of two samples from transcriptomic or ChIP-chip experiments and also on the detection of transcribed regions by directly modeling the joint distribution of the two sample intensities. We consider all the available information from the probes: the intensity of the two signals, the dependence between neighboring probes and the structural annotation. The annotation knowledge is very useful information with the aim of classification because of the intrinsic difference between exonic or intergenic probes. This method can be used for ChIP-chip or transcriptomic data whenever there are two conditions to compare. Both one-color and two-color tiling arrays can be analysed. This method could be adapted for the comparison of $d > 2$ samples. The number of parameters would be linear in d and quadratic in K but

still very small compared to the number of observations. However, the definition of generic geometrical constraints in dimension d is not straightforward and would need to be adapted to the experimental design. A simulation study highlighted the performance of our approach and applications on *Arabidopsis thaliana* tiling array show the ability of the model to interpret the data and provide a new insight on gene expression or gene expression control as well as new biological hypotheses. *Arabidopsis thaliana* is a model plant with a very well-known genome annotation but for many organisms the annotation is not available or unreliable. That is why the sub-models are also useful. The model without annotation allows us not to be limited by the quality of the available annotation and this model is also useful to detect new genes to improve the current official annotation.

This work also raises the question of classification. The results are given by probe and by region. We compute a posterior probability by region and we propose a procedure for region classification. The most common regions are the genes which are non-connected regions, but any other region can be defined. It would be interesting to control the False Discovery Rate, *i.e.* the expected proportion of misclassifications, in the case of having 4 groups and under the dependence hypothesis, and also for the results given by region. Moreover it is clear that the assumption of a normal distribution for the emission distribution may not be realistic. We are now working on a model where the emission distributions are themselves mixtures, in order to get more flexible distributions and therefore a better fit to real data.

A Appendix: Computation of the estimates of D and Λ.

Recall that $X_t = (X_{t1}, X_{t2})$ are the log-intensities for both samples, t varies from 1 to n, where n is the total number of observations.

Let $\bar{X}_k = \frac{\sum_{t=1}^{n} \tau_{tk,X} X_t}{n_k}$, where $n_k = \sum_{t=1}^{n} \tau_{tk,X}$.

Let $W_k = \sum_{t=1}^{n} \tau_{tk,X} (X_t - \bar{X}_k)(X_t - \bar{X}_k)'$ be a matrix like $\begin{pmatrix} w_{1k} & w_{2k} \\ w_{2k} & w_{4k} \end{pmatrix}$ and $\Lambda_k = \begin{pmatrix} u_{1k} & 0 \\ 0 & u_2 \end{pmatrix}$, with $u_{1k} > u_2$, for $k = 1, ..., 4$. The maximum likelihood estimator of the orientation matrix D identical for the first two components with the same orientation is in the form of $\begin{pmatrix} \hat{d} & -\sqrt{1-\hat{d}^2} \\ \sqrt{1-\hat{d}^2} & \hat{d} \end{pmatrix}$, where $\hat{d}$ is the minimum of the function:

$$f(d) = \sum_{k=1}^{2} \left\{ \frac{d^2 w_{1k} + 2 w_{2k} d \sqrt{1-d^2} + w_{4k}(1-d^2)}{u_{1k}} + \frac{d^2 w_{4k} + 2 w_{2k} d \sqrt{1-d^2} + w_{1k}(1-d^2)}{u_2} \right\}$$

DOI: 10.2202/1544-6115.1692

The estimator of $\hat{d}$ is defined by:

$$\widehat{d^2} - \frac{1}{2} \quad = \quad \pm \frac{N_{1,4}}{2\left[\left\{N_{1,4}\right\}^2 + 4\left\{N_2\right\}^2\right]^{1/2}}, \quad \text{with } \hat{d} > 0,$$

where $N_{1,4} = \sum_{k=1}^{2} (w_{1k} - w_{4k})(u_2 - u_{1k})/u_{1k}u_2$ and $N_2 = \sum_{k=1}^{2} (w_{2k})(u_2 - u_{1k})/u_{1k}u_2$.

Let B_k be a matrix defined by $B_k = D_k' W_k D_k$ like $\begin{pmatrix} b_{1k} & b_{3k} \\ b_{4k} & b_{2k} \end{pmatrix}$.

The maximum likelihood estimator of Λ_k is in the form of $\begin{pmatrix} \hat{u}_{1k} & 0 \\ 0 & \hat{u}_2 \end{pmatrix}$, where

$$\begin{cases} \hat{u}_{1k} & = b_{1k}/n_k \\ \hat{u}_2 & = \sum_{k=1}^{4} b_{2k}/n. \end{cases}$$

References

Banfield, J. and A. Raftery (1993): "Model-based gaussian and non-gaussian clustering." *Biometrics*, 49, 803–821.

Baum, L. (1972): "An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes." *Inequalities*, 3, 1–8.

Bernatavichute, Y., X. Zhang, S. Cokus, M. Pellegrini, and S. Jacobsen (2008): "Genome-wide association of histone h3 lysine nine methylation with chg dna methylation in arabidopsis thaliana." *PLoS ONE*, 3(9):e3156.

Biernacki, C., G. Celeux, A. Echenim, G. Govaert, and F. Langrognet (2007): "Le logiciel mixmod d'analyse de mélange pour la classification et l'analyse discriminante." *La Revue de Modulad*, 35, 25–44.

Biernacki, C., G. Celeux, and G. Govaert (2000): "Assessing a mixture model for clustering with the integrated completed likelihood." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719–725.

Buck, M. and J. Lieb (2004): "Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments," *Genomics*, 83, 349–360.

Buck, M., A. Nobel, and J. Lieb (2005): "Chipotle: a user-friendly tool for the analysis of chip-chip data." *Genome Biol.*, 6(11).

Cawley, S., S. Bekiranov, H. Ng, and *et al.* (2004): "Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas," *Cell*, 116, 499–509.

Celeux, G. and J. Durand (2008): "Selecting hidden markov model state number with cross-validated likelihood," *Computational Statistics*, 23(4), 541–564.

Celeux, G. and G. Govaert (1995): "Gaussian parsimonious clustering models," *Pattern Recognition*, 28, 781–793.

Cenik, C., A. Derti, J. Mellor, G. Berriz, and F. Roth (2010): "Genome-wide functional analysis of human 5'untranslated region introns." *Genome Biology*, 11:R29.

Du, J., J. Rozowsky, J. Korbel, Z. Zhang, T. Royce, M. Schultz, M. Snyder, and M. Gerstein (2006): "A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chip-chip experiments: Systematically incorporating validated biological knowlegdge." *Bioinformatics*, 22(24), 3016–3024.

Fraley, C. and A. Raftery (2006, revised 2010): "Mclust version 3 for r: Normal mixture modeling and model-based clustering." *Technical Report no. 504, Department of Statistics, University of Washington*.

Fridlyand, J., A. Snijders, D. Pinkel, D. Albertson, and A. Jain (2004): "Hidden markov models approach to the analysis of array cgh data." *J Multivariate Analysis*, 90, 132–153.

Ghosh, S., H. Hirsch, E. Sekinger, P. Kapranov, K. Struhl, and T. Gingeras (2007): "Differential analysis for high density tiling microarray data." *BMC Bioinformatics*, 8:359.

Halasz, G., M. van Batenburg, J. Perusse, S. Hua, X. Lu, K. White, and H. Bussemaker (2006): "Detecting transcriptionally active regions using genomic tiling arrays." *Genome Biology*, 7.

Hanada, K., X. Zhang, J. Borevitz, W. Li, and S. Shiu (2007): "A large number of novel coding small open reading frames in the intergenic regions of the arabidopsis thaliana genome are transcribed and/or under purifying selection." *Genome Research*, 17, 632–640.

He, K., X. Li, J. Zhou, X. Deng, H. Zhao, and J. Luo (2009): "Ntap: for nimblegen tiling array chip-chip data analysis." *Bioinformatics*, 25, 1838–1840.

Huber, W., J. Toedling, and L. Steinmetz (2006): "Transcript mapping with high-density oligonucleotide tiling arrays." *Bioinformatics*, 22(6), 1963–1970.

Humburg, P., D. Bulger, and G. Stone (2008): "Parameter estimation for robust hmm analysis of chip-chip data." *BMC Bioinformatics*, 9:343.

Hupé, P., N. Stransky, J. Thiery, F. Radvanyi, and E. Barillot (2004): "Analysis of array cgh data: from signal ratio to gain and loss of dna regions," *Bioinformatics*, 20(18), 3413–3422.

DOI: 10.2202/1544-6115.1692

Ji, H. and W. Wong (2005): "Tilemap: create chromosomal map of tiling array hybridizations." *Bioinformatics*, 21, 3629–3636.

Johannes, F., R. Wardenaar, M. Colomé-Tatché, and *et al.* (2010): "Comparing genome-wide chromatin profiles using chip-chip or chip-seq." *Bioinformatics*, 26, 1000–1006.

Johnson, W., W. Li, C. Meyer, R. Gottardo, J. Carroll, M. Brown, and X. Liu (2006): "Model-based analysis of tiling-arrays for chip-chip," *PNAS*, 103, 12457–12462.

Keles, S., M. van de Laan, S. Dudoit, and S. Cawley (2004): "Multiple testing methods for chip-chip high density oligonucleotide array data." *University of California Berkeley Division of Biostatistics Working Paper Series*, 147.

Li, W., A. Meyer, and X. Liu (2005): "A hidden markov model for analyzing chip-chip experiments on genome tiling arrays and its application to p53 binding sequences." *Bioinformatics*, 211, 274–282.

Martin-Magniette, M., T. Mary-Huard, C. Bérard, and S. Robin (2008): "Chipmix: mixture model of regressions for two-color chip-chip analysis," *Bioinformatics*, 24:i181-i186.

Mary-Huard, T., F. Picard, and S. Robin (2006): *Introduction to Statistical Methods for Microarray Data Analysis, Mathematical and Computational Methods in Biology.*

Mockler, T., S. Chan, A. Sundaresan, and *et al.* (2005): "Applications of dna tiling arrays for whole genome analysis." *Genomics*, 85, 1–15.

Munch, K., P. Gardner, P. Arctander, and A. Krogh (2006): "A hidden markov model approach for determining expression from genomic tiling micro arrays." *BMC Bioinformatics*, 7:239.

Nicolas, P., A. Leduc, S. Robin, S. Rasmussen, H. Jarmer, and P. Bessières (2009): "Transcriptional landscape estimation from tiling array data using a model of signal shift and drift." *Bioinformatics*, 25(18), 2341–2347.

Oshlack, A., M. Robinson, and M. Young (2010): "From rna-seq reads to differential expression results." *Genome Biology*, 11:220.

Penterman, J., D. Zilberman, J. Huh, T. Ballinger, S. Henikoff, and F. R.L. (2007): "Dna demethylation in the arabidopsis genome." *PNAS*, 104, 6752–6757.

Picard, F., S. Robin, M. Lavielle, C. Vaisse, and J. Daudin (2005): "A statistical approach for array cgh data analysis," *BMC Bioinformatics*, 6:27.

Pinkel, D., R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. Kuo, C. Chen, and Y. Zhai (1998): "High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays." *Nat.Genet.*, 20, 207–211.

Rabiner, L. (1989): "A tutorial on hidden markov models and selected applications in speech recognition." *Proc. IEEE*, 77, 257–286.

Samson, F., V. Brunaud, S. Duchêne, Y. De Oliveira, M. Caboche, A. Lecharny, and S. Aubourg (2004): "Flagdb++: a database for the functional analysis of the arabidopsis genome." *Nucleic Acids Res.*, Jan 1;32 Database issue: D347-50.

Schiex, T., A. Moisan, and P. Rouzé (2001): "Eugene: An eucaryotic gene finder that combines several sources of evidence." *Computational Biology*, Eds. O. Gascuel and M-F. Sagot, LNCS 2066, 111–125.

Seifert, M., A. Banaei, J. Keilwagen, M. Mette, A. Houben, F. Roudier, V. Colot, I. Grosse, and M. Strickert (2009): "Array-based genome comparison of arabidopsis ecotypes using hidden markov models." in *Biosignals*.

Snijders, A., N. Nowak, R. Segraves, S. Blackwood, and *et al.* (2001): "Assembly of microarrays for genome-wide measurement of dna copy number." *Nat.Genet.*, 29, 263–264.

Yamada, K., J. Lim, J. Dale, and *et al.* (2003): "Empirical analysis of transcriptional activity in the arabidopsis genome." *Science*, 302, 842–846.

Zeller, G., S. Henz, S. Laubinger, D. Weigel, and G. Rätsch (2008): "Transcript normalization and segmentation of tiling array data." *Pacific Symposium on Biocomputing*, 12, 527–538.

Zimmermann, P., M. Hirsch-Hoffmann, L. Hennig, and W. Gruissem (2004): "Genevestigator. arabidopsis microarray database and analysis toolbox." *Plant Physiology*, 136(1), 2621–2632.

DOI: 10.2202/1544-6115.1692

Bibliographie

Bibliographie

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control* **AC-19** 716-723.

Archer, G.E.B. and Titterington, D.M. (2002). Parameter estimation for hidden Markov chains. *Journal of Statistical Planning and Inference* **108** 365-390.

Aubourg, S. and Rouzé, P. (2001). Genome annotation. *Plant Physiology and Biochemistry* **39**, 181-193.

Banfield, J.D. and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803-821.

Baudry, J.P., Raftery, A.E., Celeux, G., Lo, K. and Gottardo, R. (2010). Combining Mixture Components for Clustering. *JCGS*, **9**(2) : 332-353.

Baudry, J.P. (2009). Sélection de Modèle pour la Classification Non Supervisée. Choix du Nombre de Classes. *Thèse de Doctorat*, Université Paris-Sud XI.

Baum, L. and Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics* **37** 1554-1563.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* **41** 164-171.

Bérard, C., Martin-Magniette, M-L. and Robin, S. (2011) Mixture model approach to compare two samples of tiling array data : ChIP-chip and Transcriptome. *Statistical Applications in Genetics and Molecular Biology* **10**, Iss. 1, Article 50.

Bérard, C., Martin-Magniette, M.-L., To, A., Roudier, F., Colot, V. and Robin, S. (2009). Mélanges gaussiens bidimensionnels pour la comparaison de deux échantillons de chromatine immunoprécipité. *La revue de MODULAD*, **40** 53-68.

Berger, S.L. (2007). The complex language of chromatin regulation during transcription. *Nature* **447**, 407-412.

Bernatavichute, Y.V., Zhang, X., Cokus, S., Pellegrini, M. and Jacobsen, S.E. (2008). Genome-Wide Association of Histone H3 Lysine Nine Methylation with CHG DNA Methylation in Arabidopsis thaliana. *PLoS ONE* **3**(9) :e3156.

Bernot, A., Choisne, N. and Salanoubat, M. (2001). Séquençage des génomes eucaryotes : Arabidopsis, le quatrième élément. *médecine/sciences* **17**, 829-835.

Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(7) 719-725.

Biernacki, C., Celeux, G. and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis* **41** 561-575.

Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based clustering and discriminant analysis with the MIXMOD software. *Computational Statistics and Data Analysis* **51/2** 587-600.

Bilmes, J.A. (1998). A gentle Tutorial of the EM algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. *International Computer Science Institute.*

Boulicaut, J.F. and Gandrillon, O. (2004). Informatique pour l'analyse du transcriptome. Chapitre Techniques statistiques pour l'analyse du transcriptome. *Hermès.*

Buck, M.J. and Lieb, J.D. (2004). Chip-chip : considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83**(3), 349-360.

Buck, M.J., Nobel, A.B. and Lieb, J.D. (2005). ChIPOTle : a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol.* **6**(11).

Cawley, S., Bekiranov, S., Ng, H. *et al.* (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. *Cell* **116**(4), 499-509.

Celeux, G. and Diebolt, J. (1985). The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comp. Statis. Quaterly* **2** 73-82.

Celeux, G. and Durand, J.B. (2008). Selecting Hidden Markov Model State Number with Cross-Validated Likelihood. *Computational Statistics* 541-564.

Celeux, G. and Govaert, G. (1992). A Classification EM Algorithm for Clustering and Two Stochastic versions. *Computational Statistics and Data Analysis* **14** 315-332.

Celeux, G. and Govaert, G. (1995). Gaussian Parsimonious Clustering Models. *Pattern Recognition* **28** 781-793.

Cenik, C., Derti, A., Mellor, J.C., Berriz, G.F. and Roth, F.P. (2010). Genome-wide functional analysis of human 5'untranslated region introns. *Genome Biology* **11** :R29.

Chatzis, S.P. (2010). Hidden Markov Models with Nonelliptically Contoured State Densities. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(12) 2297-2304.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorihtm (with discussion). *J. R. Statis. Soc. B* **39** 1-38.

Dèrozier, S., Samson, F., Tamby, J.P., Guichard, C., Brunaud, V., Grevet, P., Gagnot, S., Label, P., Leplé, J.C., Lecharny, A. and Aubourg, S. (2011). Exploration of plant genomes in the FLAGdb++ environment. *Plant Methods* **7**(1) :8.

Devijver, P.A. (1985). Baum's forward-backward algorithm revisited. *Pattern Recognition Letters* **3** 369-373.

Du, J., Rozowsky, J.S., Korbel, J.O., Zhang, Z.D., Royce, T.E., Schultz, M.H., Snyder, M. and Gerstein, M. (2006). A Supervised Hidden Markov Model Framework for Efficiently Segmenting Tiling Array Data in Transcriptional and ChIP-chip Experiments : Systematically Incorporating Validated Biological Knowlegdge. *Bioinformatics* **22**(24), 3016-3024.

Foreman, L.A. (1993). Generalization of the viterbi algorithm. *Journal of Mathematics Applied in Business and Industry* **4** 351-367.

Forney, G.D.Jr (1973). The Viterbi algorithm. *Proceedings of the IEEE.*

Fraley, C. and Raftery, A.E. (1999). Mclust : Software for Model-based Cluster Analysis. *Journal of Classfication* **16** 297-306.

Fraley, C. and Raftery, A. E. (2006). MCLUST version 3 for R : Normal mixture modeling and model-based clustering. *Tech. Rep. 504*, University of Washington, Department of Statistics, Seattle, WA.

Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G. and Jain, A.N. (2004). Hidden Markov models approach to the analysis of array CGH data. *J Multivariate Analysis* **90**, 132-153.

Ghosh, S., Hirsch, H.A., Sekinger, E.A., Kapranov, P., Struhl, K. and Gingeras, T.R. (2007). Differential analysis for high density tiling microarray data. *BMC Bioinformatics*, **8** :359.

Glaz, J., Pozdnyakov, V. and Wallenstein, S. (2009). Scan Statistics Methods and Applications. *Statistics for Industry and Technology, Birkhauser.*

Halasz, G., van Batenburg, M.F., Perusse, J., Hua, S., Lu, X.J., White, K.P. and Bussemaker, H. (2006). Detecting transcriptionally active regions using genomic tiling arrays. *Genome Biology* **7**.

Hanada, K., Zhang, X., Borevitz, J.O., Li, W.H. and Shiu, S.H. (2007). A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. *Genome Research* **17**, 632-640.

Hathaway, R.J. (1986). Another interpretation of the em algorithm for mixture distributions. *Statistics & Probability Letters* **4** 53-56.

He, K., Li, X., Zhou, J., Deng, X.W., Zhao, H. and Luo, J. (2009). NTAP : for NimbleGen tiling array ChIP-chip data analysis. *Bioinformatics* **25**, 1838-1840.

Hennig, C. (2010). Methods for merging Gaussian mixture components. *Adv Data Anal Classif* 3-34.

Huber, W., Toedling, J. and Steinmetz, L.M. (2006). Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22**(6), 1963-1970.

Humburg, P., Bulger, D. and Stone, G. (2008). Parameter estimation for robust HMM analysis of ChIP-chip data. *BMC Bioinformatics* **9** :343.

Hupé, P., Stransky, N., Thiery, J.P., Radvanyi, F. and Barillot, E. (2004). Analysis of array CGH data : from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**(18) :3413-3422.

Jarvis, K. and Robertson, M. (2011). The noncoding universe. *BMC Biology* **9** :52.

Ji, H. and Wong, W.H. (2005). TileMap : create chromosomal map of tiling array hybridizations. *Bioinformatics* **21**, 3629-3636.

Johannes, F., Wardenaar, R., Colomé-Tatché M. *et al.* (2010). Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. *Bioinformatics* **26** 1000-1006.

Johnson, W.E., Li, W., Meyer, C.A., Gottardo, R., Carroll, J.S., Brown, M. and Liu, X.S. (2006). Model-based analysis of tiling-arrays for ChIP-chip. *PNAS* **103**, 12457-12462.

Karlis, D. and Xekalaki,E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics and Data Analysis*, **41**(3) :577-590.

Kass, E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American statistical Association* **90** 773-795.

Keles, S., van de Laan, M., Dudoit, S. and Cawley, S.E. (2004). Multiple Testing Methods for ChIP-chip high density oligonucleotide array data. *University of California Berkeley Division of Biostatistics Working Paper Series* **147**.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A* **62**(1) :49-66.

Kerr, M.K., Afshari, C.A., Bennett, L., Bushel, P., Martinez, J., Walker, N.J. and Churchill, G.A. (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* **12** 203-217.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing* **85** 1501-1510.

Lebarbier, E. and Mary-Huard, T. (2004). Le critère BIC : fondements théoriques et interprétation. *Technical Report 5315*, INRIA.

Lebarbier, E. and Mary-Huard, T. (2011). Classification non supervisée. *Polycopié AgroParisTech.*

Li, J. (2005). Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics* 547-568.

Li, W., Meyer, A. and Liu, X.S. (2005). A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* **211**, 274-282.

Maitra, R. (2009). Initializing partition-optimization algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **6** 144-157.

Martin-Magniette, M.L., Mary-Huard, T., Bérard, C. and Robin, S. (2008). ChIPmix : mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics* **24**, 181-186.

Mary-Huard, T., Martin-Magniette, M.L., Bérard, C. and Robin, S. (2010). Statistical methodology for the analysis of multi-sample ChIP-chip experiments. *International Biometric Conference* Florianopolis (Brésil).

McLachlan, G.J. and Krishnan, T. (2008). The EM Algorithm and Extensions. *Wiley Series in Probability and Statistics.*

McLachlan, G.J. and Peel, D. (2000). Finite Mixture Models. *New York : John Wiley.*

Mockler, T.C. and Ecker, J.R. (2005). Applications of DNA tiling arrays for whole genome analysis. *Genomics* **85**, 1-15.

Moghaddam, A.M.B., Roudier, F., Seifert, M., Bérard, C., Martin-Magniette, M.L., Ashtiyani, R.K., Houben, A., Colot, V. and Mette, M.F. (2011). Additive inheritance of histone modifications in Arabidopsis thaliana intra-specific hybrids. *The Plant Journal.*

Munch, K., Gardner, P.P., Arctander, P. and Krogh, A. (2006). A hidden Markov model approach for determining expression from genomic tiling micro arrays. *BMC Bioinformatics*, **7** :239.

Nicolas, P., Leduc, A., Robin, S., Rasmussen, S., Jarmer, H. and Bessières, P. (2009). Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. *Bioinformatics* **25**(18), 2341-2347.

Oh, S., Park, S. and van Nocker, S. (2008). Genic and global functions for Paf1C in chromatin modification and gene expression in Arabidopsis. *PLoS Genet* **4** :e1000077.

Oshlack, A., Robinson, M.D. and Young, M.D. (2010). From RNA-seq reads to differential expression results. *Genome Biology* **11** :220.

Penterman, J., Zilberman, D., Huh, J.H., Ballinger, T., Henikoff, S. and Fischer R.L. (2007). DNA demethylation in the Arabidopsis genome. *PNAS* **104** 6752-6757.

Picard, F. (2007). An introduction to mixture models. *SSB Research Report* **7**.

Picard, F., Robin, S., Lavielle, M., Vaisse, C. and Daudin, JJ. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics* **6** :27.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C. and Zhai, Y. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat.Genet.* **20**, 207-211.

Rabiner, L.R. (1989). A tutorial on Hidden Markov Models and selected applications in Speech Recognition. *Proceedings of the IEEE.*

Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T. *et al.* (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO Journal* **30** 1928-1938.

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

Schiex, T., Moisan, A. and Rouzé, P. (2001). EuGene : An Eucaryotic Gene Finder that combines several sources of evidence. *Computational Biology*, Eds. O. Gascuel and M-F. Sagot, LNCS 2066, 111-125.

Schwarz, G. (1977). Estimating the number of components in a finite mixture model. *Annals of Statistics* **6** 461-464.

Seifert, M., Banaei, A., Keilwagen, J., Mette, M.F., Houben, A., Roudier, F., Colot, V., Grosse, I. and Strickert, M. (2009). Array-based Genome comparison of Arabidopsis ecotypes using Hidden Markov models. *Biosignals*, Porto (Portugal).

Smyth, G.K., Yang, Y.H. and Speed, T.P. (2003). Statistical issues in cDNA microarray data analysis. *Methods in Molecular Biology* **224** :111-136.

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S. *et al.* (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat.Genet.* **29**, 263-264.

Sun, W. and Cai, T.T. (2009) Large-scale multiple testing under dependence. *J.R. Statist.Soc. B*, **71** Part2 393-424.

Sun, W., Buck, M.J., Patel, M. and Davis, I.J. (2009). Improved ChIP-chip analysis by a mixture model approach. *BMC Bioinformatics* **10** :173.

Tantrum, J. and Murua, A. (2003). Assessment and pruning of hierarchical model based clustering. *Pattern Recognition, Clustering, ACM Press.*

Thareau, V., Déhais, P., Serizet, C., Hilson, P., Rouzé, P. and Aubourg, S. (2003). Automatic design of gene-specific sequence tags for genome-wide functional studies. *Bioinformatics* **19(17)** :2191-8.

Turck, F., Roudier, F., Farrona, S., Martin-Magniette, M-L., Guillaume, E., Buisine, N., Gagnot, S., Martienssen, R., Coupland, G. and Colot, V. (2007). Arabidopsis TFL2/LHP1 Specifically Associates with Genes Marked by Trimethylation of Histone H3 Lysine 27. *PLoS Genet.v* **3(6)**.

Turner, T.R. (2000). Estimating the Propagation Rate of a Viral Infection of Potato Plants via Mixtures of Regressions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **49**(3) 371-384.

Velculescu, V.E, Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995). Serial analysis of gene expression. *Science* **270**, 484-487.

Wei, G.C.G, and Tanner, M.A. (1991). A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85** 699-704.

Wilkinson, J.Q., Lanahan, M.B., Clark, D.G., Bleecker, A.B., Chang, C., Meyerowitz, E.M. and Klee, H.J. (1997). A dominant mutant receptor from Arabidopsis confers ethylene insensitivity in heterologous plants. *Nature Biotechnology* **15**, 444-447.

Yamada, K., Lim, J., Dale, J.M. *et al.* (2003). Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* **302**, 842-846.

Zeller, G., Henz, S.R., Laubinger, S., Weigel, D. and Rätsch, G. (2008). Transcript normalization and segmentation of tiling array data. *Pacific Symposium on Biocomputing* **12**, 527-538.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L. and Gruissem, W. (2004). GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiology* **136**(1), 2621-2632.