

## CGH-array - Travaux Dirigés

*L'objet de ce TD est de vous mettre en situation réelle devant un résultat de CGH-array à analyser, sans logiciel « clé en main » ou protocole à disposition. La plupart des opérations réalisées durant ce TD sont automatisées dans des logiciels ou packages R tiers, pour une approche plus pédagogique et dans un souci d'interopérabilité nous les implémenterons nous même. A noter qu'un certain nombre de fonctions complexes vous sont fournies, n'hésitez pas à essayer d'en comprendre le fonctionnement malgré tout.*

Le fichier « CHB-04159.gz » dont vous disposez contient les résultats d'une expérience de CGH-array réalisée sur un ADN de biopsie ganglionnaire de lymphome B diffus à grandes cellules (fluorescence rouge). L'ADN de référence utilisé lors de cette expérience est l'ADN normal du patient, extrait du sang périphérique (fluorescence verte). Après hybridation, cette puce Agilent 44k a été scannée grâce au logiciel *Feature Extraction* fourni par le fabricant, qui a produit ce fichier.

- 1.1. Observez le contenu du fichier, en vous aidant de [la documentation](#) si nécessaire.**
- 1.2. Donnez les coordonnées spatiales d'une sonde de positionnement de la grille de lecture.**
- 1.3. Donnez les coordonnées spatiales d'une sonde saturée.**
- 1.4. Donnez le log-ratio d'une sonde située sur le gène CDKN2A.**

Face à ce fichier typique des puces à ADN, nous choisirons d'utiliser le package R *limma*.

- 2.1. Retrouvez et installez le package LIMMA.**
- 2.2. En vous aidant de la documentation, importez ce fichier dans R.**
- 2.3. Calculez les log-ratio des sondes en utilisant LIMMA (sans normalisation).**

Pour simplifier la manipulation des données, nous n'irons pas plus loin avec les formats de LIMMA, utiles pour la normalisation mais plus adaptés à la transcriptomique. Pour une étude plus poussée, le package Bioconductor « *snapCGH* » propose des formats de données simples à manipuler. Vue sa complexité d'installation, nous nous limiterons à un format personnalisé.

- 3.1. Sourcez le fichier de fonctions « Fonctions CGH.r » dans votre session, puis utilisez la fonction « *limma2df* » pour convertir votre objet LIMMA dans un format plus simple.**
- 3.2. Ajoutez les coordonnées génomiques des sondes à ce tableau (fonction fournie).**
- 3.3. Résolvez les répliquats (fonction fournie).**

Une fois les données importées, nous allons nous intéresser aux sondes.

- 4.1. Calculez le DLRS de la puce (fonction fournie). Commentez.**
- 4.2. Réalisez un « MA-plot ». Commentez.**
- 4.3. Réalisez un « spatial-plot » (fonction fournie). Commentez.**
- 4.4. Observez les fluctuations du log-ratio le long du chromosome 4 dans un graphique.**
- 4.5. Ajoutez la moyenne locale à ce graphique (fonction fournie). Observez les changements induits par la modification du paramètre 'w' (largeur de la fenêtre).**

Pour conclure avec plus de précision, la segmentation est nécessaire. Nous choisirons d'utiliser l'algorithme CBS (Circular Binary Segmentation), le standard de fait en ce domaine.

- 5.1. Retrouvez et installez le package R qui implémente CBS.**
- 5.2. En vous aidant de la documentation, formatez vos données pour leur prise en charge.**
- 5.3. Appliquez le lissage et la segmentation.**
- 5.4. Ajoutez les segments au graphique réalisé en 4.4 (rectangles).**
- 5.5. En vous basant sur l'explication de l'algorithme donnée en cours et la documentation de DNACopy, augmentez le nombre de segments. Commentez les différences observées sur le chromosome 4.**
- 5.6. Observez la segmentation du chromosome 6. Que remarquez-vous ? Corrigez le biais incriminé en utilisant les fonctions fournies.**

La segmentation terminée, il est nécessaire d'interpréter les log-ratios. Nous choisirons pour ce TD une approche simple basée sur des seuils manuels. Pour des puces moins évidentes à analyser, les packages Bioconductor « CGHcall », « aCGH » ou « snapCGH » peuvent être envisagés.

- 6.1. Observez la densité de distribution des log-ratios des sondes. Commentez.**
- 6.2. Observez la densité de distribution des log-ratios des segments. Dans un second temps, pondérez cette densité par la taille des segments. Commentez.**
- 6.3. Définissez un seuil d'amplification et un seuil de délétion grâce au graphique précédent.**
- 6.4. Attribuez un état (délété, normal ou amplifié) aux segments.**
- 6.5. Ajoutez cette information au graphique réalisé en 5.4 (couleur des segments).**
- 6.6. Donnez une estimation de la proportion de cellules normales présentes dans l'échantillon.**

Quelle que soit la question biologique posée sur ces données, y répondre passe par l'annotation des régions altérées obtenues (gènes ...). Nous choisirons d'utiliser le *R Genome Browser* dans cette optique (<http://bioinformatics.ovsa.fr/Rgb>).

- 7.1. Installez le package Rgb.**
- 7.2. En vous aidant de la vignette, créez une piste de segments, colorés selon l'état prédit précédemment. Utilisez l'interface graphique pour parcourir ces données.**
- 7.3. Observez les gènes amplifiés et délétés sur le chromosome 4. Donnez quelques exemples.**
- 7.4. Grâce au *genome browser*, recherchez le statut du gène CDKN2A.**
- 7.5. A l'aide du graphique réalisé en 6.2, précisez votre réponse à la question précédente.**

Observons maintenant les effets de la normalisation.

- 8.1. Ajoutez l'état obtenu par segmentation (délété, normal ou amplifié) aux sondes.**
- 8.2. Ajoutez cette information au « MA-plot » (couleur des points). Commentez.**
- 8.3. Reprenez l'analyse dans son ensemble en choisissant cette fois-ci une normalisation « loess ». Observez et commentez les différences avec l'analyse précédente concernant le DLRS (4.1), la distribution des segments (6.2) et le MA-plot modifié (8.2).**

Pour terminer, nous illustrerons le biais spatial avec des données publiques. Nous retrouverons les données à la main, sachez que le package « GEOquery » est capable d'automatiser ce processus.

- 9.1. Importez les données du patient GSM287077 sur la base Gene Expression Omnibus.**
- 9.2. Associez-y les coordonnées spatiales des sondes, que vous trouverez ailleurs sur GEO.**
- 9.3. Réalisez un « spatial-plot » de cette puce. Commentez.**