Bonjour à tous,

Je vous contacte une dernière fois pour faire un petit bilan (un peu tardif, je m'en excuse) de l'examen de CGH que vous avez passé fin janvier. Dans l'ensemble c'était plutôt bien, en attendant que vous récupériez vos copies corrigées, vous trouverez ci-dessous quelques explications sur ce que j'attendais de vous dans ces deux sujets et comment les points ont été attribués.

Vous trouverez également en pièces jointes les sujets des deux promotions, je vous invite à parcourir celui que vous n'avez pas passé et vous demander comment vous vous en seriez sorti, car toutes les facettes du cours n'ont pu être abordées dans un seul sujet. Si vous êtes pressé, je vous invite à privilégier la partie D dans laquelle j'apporte quelques éclaircissements précieux, et la partie A qui pourrait vous sauver un jour de concours.

Parties A des deux sujets

Pour les deux promotions, il s'agissait de décrire comment vous mèneriez une expérience de CGH array sur un problème donné, légèrement différent d'une promotion à l'autre. C'est une épreuve assez fastidieuse j'en conviens, mais c'est une question de concours ou d'entretien assez typique qui permet de cerner rapidement si vous disposez de connaissances pratiques dans un domaine, ou s'il vous faudra vous documenter longuement avant d'être opérationnel. Du point de vue de l'employeur, le choix est vite fait.

Une petit remarque avant de détailler : soyez prudents avec les termes techniques que vous employez ! Ça vous sauve peut être des points en examen d'apprendre uniquement leurs noms pour les replacer, mais dans la vraie vie évitez de lancer des termes que vous ne maitrisez pas, ça se voit assez vite et de mon point de vue c'est encore pire que de ne rien dire. A titre d'exemple j'ai lu beaucoup d'abus de « dérivation », « LOWESS », « normalisation » et « filtration », qui sont des outils génériques qui n'ont de sens que si vous expliquez comment vous les appliquez.

Il y avait donc 5 thèmes :

- Le choix des hybridations (1 point).

J'attendais que vous me parliez des ADN à hybrider sur chaque canal (Cy3 rouge et Cy5 vert), éventuellement de dye swap et de réplicats. Le choix des couleurs n'avait pas d'importance. Certains m'ont parlé d'ADN cot, attention à ne pas confondre la CGH sur puce à la CGH telle qu'on la pratiquait à ses débuts ! Dans la CGH originale effectivement on ajoutait un ADN correspondant aux régions répétées car l'hybridation avait lieu sur des chromosomes entiers, en CGH array on se contente de ne pas inclure de sondes dans les régions répétées. Je vous avais préparé quelques pièges, j'ai été assez déçu de voir tout le monde y sauter à pieds joints.

Pour les M2.1 il s'agissait d'un problème simple dans lequel on devait simplement hybrider l'ADN des cellules tumorales contre un ADN de référence, soit un pool d'ADN commercial soit l'ADN normal du patient correspondant. **Utiliser l'ADN sanguin comme référence était exclu** puisqu'il s'agit d'un cancer touchant des cellules sanguines circulantes. On aurait pu envisager un tri cellulaire sur des marqueurs tumoraux (assez hasardeux) ou de récupérer de l'ADN ailleurs (prélèvement buccal).

Pour les M2.2, on recherchait des anomalies **constitutionnelles, c'est à dire des polymorphismes présents avant la survenue du cancer !** Il s'agissait donc d'étudier l'ADN sain des patients (à partir du sang cette fois ci puisqu'il s'agit d'un cancer solide, dont les métastases circulantes éventuelles devraient être négligeables), hybridé contre une référence commune (pool de préférence).

- Les biais à surveiller (2 points)

Beaucoup ont oublié qu'on demandait également une brève explication et une description des moyens employés pour surveiller ces biais, et qu'on **ne** demandait **pas** de décrire la correction de ces biais. J'attendais donc que vous me parliez du **bruit de fond** (suffisamment évident pour se passer d'explication) que le DLRS permet assez bien de surveiller, du **biais d'intensités** (tendance des log-ratios à varier en fonction de l'intensité moyenne des deux canaux, du fait d'une différence d'efficacité entre les fluorochromes) qu'un MA-plot permet de surveiller, des **biais spatiaux** (variations du log-ratio dans des régions physiques de la puce, malgré qu'il n'y ai pas de lien entre coordonnées physiques et chromosomiques) qu'on peut surveiller par un « M-XY » ou « spatial » plot, du **phénomène de vagues** (variations du log-ratio en fonction de la composition en GC de la région chromosomique abritant la sonde) qu'on peut identifier visuellement avec une moyenne locale ou un scatter plot des log-ratios le long d'un chromosome.

Les M2.1 auraient du également me parler de **cellularité** (tendance des log-ratios à tendre vers 0 à cause d'une contamination de l'échantillon tumoral en cellules normales), qu'on peut identifier en analysant la distribution des log-ratios après segmentation, et éventuellement de **polymorphismes** (anomalies constitutionnelles) s'ils n'avaient pas choisi comme ADN de référence l'ADN du patient.

Les M2.2 quand à eux auraient du s'en abstenir, car ces biais ne les concernent pas ... En effet ils étudient justement des polymorphismes, et le génome est sensé être identique dans toutes les cellules (tumeur exclue) du patient.

- La comparabilité des puces (1 point)

A vrai dire j'espérais que vous me parleriez du calling, en me rappelant qu'attribuer un état ou un nombre de copies à chaque segment permet d'obtenir une donnée comparable d'une puce à l'autre, quels que puissent être les biais qui les touchent. Vu qu'aucun d'entre vous ne l'a fait, j'ai accepté qu'on me parle de normalisation des log-ratios en fonction de l'intensité, ou de la nécessite de suire un protocole identique pour toutes les puces.

- Les outils et leur fonctionnement (2 points)

La liste était longue et je n'attendais pas de vous une liste complète, mais au moins quelques noms qui prouvent que si jamais vous avez un jour une telle analyse à mener, vous savez où commencer vos recherches. Vous auriez donc pu me parler du **MA-plot** (log-ratio des sondes en fonction du log de l'intensité moyenne des deux canaux), de **R** et **Bioconductor** de façon générale, de **DNAcopy** ou **CBS** (pour la segmentation des chromosomes en régions de log-ratios constants) en particulier, des **suites logicielles** développées par les fabricants de puces (ca semble trivial mais c'est important de savoir qu'elles existent et la plupart du temps font correctement le travail). Vous auriez également pu me parler de **MANOR** pour surveiller les biais spatiaux, de **BioHMM** (reposant sur des chaines de Markov caché pour la segmentation) ou **CGHcall** (reposant sur une modélisation des log-ratios par un modèle de mélange gaussien) pour le copy calling.

- Le transfert (1 point)

La mise au point d'un outil diagnostic bon marché passait par la recherche d'une ou plusieurs anomalies intéressante pour l'étude, ce qui était l'occasion de parler de **MCR** (Minimal Common Region, la région minimale communément amplifiée ou délétée chez un groupede patients). Venait ensuite la précision de la localisation de cette anomalie, via la technique de marche moléculaire vue en cours. Enfin le test aurait reposé sur une **PCR quantitative** de la région en question, QMPSF ou autre. Pour rejoindre ma mise en garde initiale contre les termes techniques mal employés, certains m'ont parlé à tors de **RT**-PCR : la PCR après **R**everse **T**ranscription permet d'étudier de l'ARN, nous en avons parlé à l'occasion des transcrits de fusion. La CGH ne permet pas de détecter les translocations qui conduisent à la formation de transcrits de fusion, il est donc inutile de travailler sur l'ARN (moins stable et qui demanderait une phase humide supplémentaire). Certains m'ont parlé de **FISH**, ce que j'ai accepté malgré qu'il s'agisse d'une technique plus complexe et plus couteuse, qu'on préfère pour la détection de translocations là encore.

- Bonus (1 point)

La plupart d'entre vous m'ont parlé de la technique de CGH en elle même, ce que je n'avais pas prévu mais qui semblait assez logique en relisant le sujet. J'ai donc ajouté un demi-point pour une explication correcte du principe (hybridation compétitive entre un ADN cible et un ADN de référence, calcul du log-ratio) et un autre demi-point pour m'avoir décrit correctement le principe de l'analyse (calcul des log-ratios de chaque sonde, segmentation du génome et copy calling).

Partie B, M2.1

Pour la première question, la technique a reconnaitre était une FISH multiplexe, et plus précisément COBRA, qui était pourtant facilement reconnaissable par le marquage binaire caractéristique qui s'ajoute au marquage multiple de la M-FISH. Il y avait 0,5 points à gagner en me parlant de FISH, 0,5 supplémentaires pour m'expliquer le principe et 0,5 pour avoir reconnu COBRA.

D'un point de vue génomique on pouvait observer qu'une partie du chromosome 13 avait été dupliquée ou **transloquée** sur le chromosome 6, vu la qualité de l'image c'était assez difficile de savoir si les deux chromosomes 13 étaient encore intacts. J'ai accepté les deux, du moment que la conclusion n'était pas trop catégorique. A cette explication qui valait 0,5 points s'ajoutait un autre demi point pour m'avoir parlé de translocation, équilibrée ou non.

Pour les effets physiologiques, la t(6;13) n'était pas dans le cours et il fallait donc se contenter de suppositions. Pour 1,5 points on pouvait parler de la formation de gène de fusion, de la production d'une protéine ou d'un ARN tronqué et donc mal ou non fonctionnel, de substitution de promoteur pouvant avoir un impact sur le niveau d'expression d'un gène proche du point de cassure.

Partie B, M2.2

Dans ce sujet il s'agissait d'un caryotype classique, issu d'une expérience de chromosome banding. 0,5 points pour le terme, 0,5 points pour une explication rapide du principe.

Il s'agissait ici aussi d'une translocation entre les chromosomes 9 et 22, attention car rien n'indique dans la formule si la translocation est équilibrée ou non, ne soyez donc pas trop catégoriques ici non plus. Pour 1,5 points il fallait expliquer que la partie q34 du chromosome 9 a été échangée avec la partie q11 du chromosome 22, et on pouvait également expliquer brièvement les autres éléments de la formule (génome féminin diploïde de 46 chromosomes).

La translocation t(9;22) était dans le cours, pour 1,5 points il fallait réexpliquer le principe ou faire plusieurs hypothèses (voir le sujet des M2.1).

Partie C des deux sujets

On avait pour les deux sujets les 3 mêmes types de figures et 4 questions à 0,5 points chacune. C'était un peu fastidieux mais il fallait répondre à toutes les questions pour toutes les figures ! Bien évidemment les conclusions diffèrent d'un sujet à l'autre.

Le premier tableau donnait le DLRS (Derivative Log Ratio Spread) de chaque puce, avec un code couleur qui facilitait tout de même grandement la conclusion. Peu ont été capables de décrire le principe avec exactitude, je me suis donc contenté de l'acronyme pour attribuer le demi point de la question 1. Ce score permet d'avoir une idée de la qualité de chaque puce, mais surtout du **bruit de fond**, ce que peu d'entre vous ont noté. Les puces avec un DLRS en blanc étaient donc de très bonne qualité (faible DLRS), les puces avec un DLRS foncé de mauvaise qualité etc. D'un point de vue correction, j'ai accepté qu'on bote en touche en disant qu'il faudrait identifier le biais à la source du problème avec d'autres analyses, j'attendais surtout qu'on envisage l'exclusion pure et simple de l'analyse pour les puces de mauvaise qualité.

La figure du milieu était un MA-plot, c'était bien de le dire mais c'était encore mieux d'être capables de décrire à quoi correspondent M (le log-ratio de chaque sonde) et A (le « log-mean » de chaque sonde, c'est à dire la moyenne des logarithmes des intensités rouges et vertes pour chaque sonde). Cette figure permet donc d'identifier des **biais d'intensité**, c'est à dire une tendance des M à varier en fonction de A, ce qui n'a aucune raison biologique mais peut s'expliquer par une différence d'efficacité des fluorochromes. Chez les M2.2 le biais d'intensité était bien présent, les sondes de faible intensité avaient tendance à avoir un log-ratio négatif. On pouvait donc envisager une régression locale de type LOWESS des M de chaque sonde en fonction de A, de manière à « redresser » ce graphique. Chez les M2.1 les données avaient déjà été normalisées et il n'y avait en théorie rien à voir, j'ai cependant accepté qu'on demande une centralisation (soustraire la moyenne des M au M de chaque sonde de manière à ramener la moyenne des M à 0), bien que cela repose sur l'hypothèse assez discutable qu'on ait autant d'amplifications que de délétions. On pouvait aussi tout simplement déclarer qu'il n'y avait pas besoin de corriger quoi que ce soit.

La dernière figure était un « M-XY » ou « spatial plot », qui consistait à colorer chaque position physique de la puce en fonction du log-ratio des sondes qui s'y trouvent (j'ai appliqué un lissage, d'où l'aspect en plages de couleurs qui change légèrement de ce qu'on a pu voir en TD). Vu qu'il n'est pas sensé y avoir de lien entre la position physique et la position chromosomique de chaque sonde, aucune région ne devrait se faire remarquer. Chez les M2.1 on avait un très beau gradient vertical, les log-ratios étant plus négatifs dans les premières lignes que les dernières. C'était un cas d'école, qu'on pouvait donc corriger facilement en appliquant une régression LOWESS aux log-ratios en se basant sur le numéro de ligne de chaque sonde, ou plus simplement grâce au package R MANOR. On m'a proposé de supprimer les sondes des 100 premières lignes, pour un gradient ça reste en général une mauvaise idée car le principe du gradient est qu'il touche toute la puce, il vaut donc mieux essayer d'en annuler les effets que d'éliminer la partie la plus touchée. Chez les M2.2 seul l'angle inférieur droit posait problème, c'était un « artefact local » assez typique. On pouvait décider d'éliminer de la puce les sondes touchées, soit manuellement soit grâce à MANOR.

Partie D, M2.1

L'exercice D était le plus difficile, mais malheureusement le plus proche de ce qu'on peut attendre de vous lors de l'analyse des résultats. C'est bien compréhensible vu le peu de temps qu'on a passé sur le sujet, mais la plupart d'entre vous ont sauté à pieds joints dans les pièges qu'une analyse classique de CGH peut comporter.

Pour les M2.1 on cherchait à conclure sur l'intérêt que pouvaient présenter deux régions spécifiques dans les résultats d'une série de puces hybridant ADN tumoral et normal. La première région n'était en fait pas intéressante pour l'étude, car il y a de fortes chances qu'il s'agisse d'un **polymorphisme**, c'est à dire une altération du nombre de copies qui ne soit pas spécifique à l'ADN de la tumeur mais en fait déjà présente chez le patient à sa naissance (et donc que d'autres individus sains ont de grandes chances de présenter aussi). Deux arguments à cela :

- **L'extrême similitude des points de cassure**. Ne perdez pas de vue qu'une translocation ou une amplification est un phénomène qui doit pas mal au hasard, et il est extrêmement rare que deux patients voient leurs chromosomes casser au même endroit, à la base près. Généralement les cassures ont lieu dans les mêmes introns ou régions intergéniques, ce qui donne au final les mêmes effets. Plus la résolution des puces augmente, plus on s'aperçoit de ces différences, et c'est bien toute la difficulté des analyses de CGH array que de conclure avec des points de cassure différents. Les régions répétées en revanche correspondent à des motifs précis, leurs bornes sont bien définies.

- La présence simultanée **d'amplifications** (bleu) et de **délétions** homozygotes (rouge foncé), qui suggère une grande variabilité du nombre de copie (de +1 minimum à -2 donc). Les amplifications et délétions tumorales ont rarement de telles amplitudes, on s'estime déjà heureux de gagner ou perdre une copie. De plus on espère en analysant une série de patients atteints de la même maladie mettre en évidence des similarités, c'est à dire le même comportement (amplification ou délétion) chez la majorité des patients, une région aussi instable est donc très suspecte (peut elle contenir un gène suppresseur de tumeur si certains patients la sur-exprime et présentent pourtant un cancer ? Et vice versa ...).

Certains d'entre vous ont bien reconnu le polymorphisme, attention toute fois à ne pas être trop catégorique, vous risqueriez de passer à coté d'une région extrêmement intéressante si ce phénomène était effectivement propre à la tumeur. J'attendais donc qu'on cherche à vérifier quel ADN de référence a été utilisé (s'il s'agit d'un pool ca milite pour le polymorphisme, si c'est l'ADN du patient c'est tout de suite plus intéressant), ou si la région contenait des polymorphismes connus, grâce à la Database of Genomic Variants par exemple.

La deuxième région était la plus intéressante, même si certains se sont étonnés de trouver si peu de patients concernés (une quinzaine de délétions spécifiques à ce locus, plus une trentaine de délétions plus larges, probablement des pertes du chromosome entier). 15/80, c'est déjà une récurrence très intéressante lorsqu'on parle de cancers ! A part une poignée d'anomalies déjà bien connues, les découvertes se font aujourd'hui sur des anomalies récurrentes dans 10 % voir 5 % des cas, et qui permettent éventuellement de mettre en évidence des sous-types dans le cancer étudié.

Partie D, M2.2

L'exercice D était le plus difficile, mais malheureusement le plus proche de ce qu'on peut attendre de vous lors de l'analyse des résultats. C'est bien compréhensible vu le peu de temps qu'on a passé sur le sujet, mais la plupart d'entre vous ont sauté à pieds joints dans les pièges qu'une analyse classique de CGH peut comporter.

Pour les M2.2, on s'intéressait à une région chromosomique particulière pour un patient donné. Là encore beaucoup sont passés à coté d'un piège majeur lorsqu'on s'intéresse à des résultats de CGH : le biais de composition en GC ou l'effet « vagues » qui sautait pourtant aux yeux. A vrai dire j'ai ajouté moi même le biais à ces données pour qu'il soit évident, dans la réalité il ne l'est pas toujours autant. Beaucoup on donc considéré la partie droite (25 à 40 Mb) comme amplifiée et la partie centrale (10 à 20 Mb) comme délétée, ce que j'ai accepté malgré tout lorsque le raisonnement était bien posé. En réalité c'est justement ce qu'il faut éviter, passer à coté de ce biais c'est considérer des régions comme altérées alors qu'elles ne le sont pas, et rendre des résultats erronés, typiquement ce que votre expertise de bioinformaticien est sensé éviter par rapport au biologiste moyen. J'attendais donc que vous observiez cet aspect « sinusoïdal » et que vous proposiez l'explication d'un biais de composition en GC, en parlant éventuellement de solutions correctives (WACA et autres).

La véritable région d'intérêt dans ces données se situait vers 22 Mb, ce que peu d'entre vous ont remarqué. S'y trouvent au moins 6 sondes avec un log-ratio proche de -1, qui suggère donc fortement une délétion (pour ceux qui ont de la mémoire il s'agit de données réelles dans lesquelles on retrouve la délétion de CDKN2A observée en TD). Compte tenu du grossissement de l'image et du peu de sondes impliqué cela a pu vous sembler négligeable, là encore il est primordial de garder à l'esprit que les altérations les plus larges ne sont pas forcément les plus intéressantes, plus le locus est restreint plus la liste des gènes suspects est courte. Attention toutefois au nombre de sondes définissant le locus, vu le bruit de fond qu'on observe généralement avec cette technique il n'est pas raisonnable de miser sur un segment de 2 ou 3 sondes, mais avec 6 sondes on est déjà plus à l'aise.

Voila, j'espère qu'avec tout ca vous serez parés à toute éventualité concernant la CGH. Je sais bien que cela doit vous sembler un peu archaïque à l'heure des NGS, mais ne négligez pas cette technique qui reste encore très utilisée pour le screening de grandes cohortes, tout simplement pour des raisons de coûts.

Sylvain Mareschal

