Analyse bioinformatique d'un jeu de données NGS
Rôle de la protéine G3BP dans le cerveau de la souris

*Formez des groupes de 2 à 3 personnes. Pour répondre aux questions posées dans l'énoncé, ouvrez un nouveau document CR-*Nom-Prenom *dans lequel vous indiquerez vos noms et prénoms et recopierez les titres des différentes parties. Puis dans chaque partie, écrivez le numéro des questions et vos réponses au fur et à mesure que vous les traitez.*

PHASE PREPARATOIRE - premières commandes Unix et découverte de Galaxy

Les données dont nous disposons correspondent aux clusters de tags obtenus dans le cadre d'une expérience CLIP-Seq (aussi appelé HIT-CLIPS), un type de manipulation permis par les NGS, indiquant quelles molécules (ARN ou protéines) se lient physiquement aux exemplaires de la protéine G3BP dans une cellule en activité (in vivo). Cette protéine est suspectée de se lier surtout avec des ARN, et ainsi de jouer un rôle dans l'expression d'autres protéines (en retenant les ARN de ces protéines, G3BP empêche momentanément la traduction de ces ARN en la protéine cible, sans compter que les ARN tenus par G3BP peuvent au final être "dégradées" (coupées en morceau), empêchant alors définitivement la création de la protéine cible depuis ces ARN.

Le séquenceur fournit des morceaux («**tags**») des séquences liées aux protéines de G3BP en activités. Ces **tags** ont été positionnés sur le génome de la souris pour savoir d'où ils proviennent et regroupés en **clusters** de tags se chevauchant (plusieurs tags peuvent venir d'une même région du génome). On dispose de fichiers textes indiquant, pour chaque cluster, le numéro du chromosome dont il provient, la position précise sur ce chromosome, ainsi que le brin (sens) sur lequel le cluster a été trouvé.

L'expérience de séquençage NGS de cellules de souris a été faite à deux endroits : Montpellier (MPL) et Edimbourg (ED). Dans les deux cas les souris concernées sont de type sauvages (*wild type*), mais d'un centre à l'autre elles sont légèrement différentes génétiquement. La personne ayant fait l'extraction des cellules n'est pas la même dans les deux cas, par contre la préparation de l'expérience de séquençage a été faite par la même personne, sur des séquenceurs Illumina. Entre la 1ère manip (ED) et la 2ème (MPL), la technologie de séquençage a évolué, aussi dans le 1er cas on a récupéré au départ 10 millions de tags de 36 nucléotides, quand 80 millions de tags de 100 nucléotides ont été obtenus dans la deuxième expérience. Toutes ces différences entre le protocole de ED et MPL expliquent qu'on ne retrouvera pas précisément les mêmes résultats dans les deux cas.

(1) Sous Linux, récupérez les fichiers du TP sur les pages web de votre enseignant et répartissez les dans les dossiers attendus.

(2) Parcourez le contenu des fichiers récupérés pour le TP et indiquez dans votre compte-rendu de TP pour chaque fichier la signification des colonnes qu'il contient (pour voir son contenu depuis un terminal de commandes, utilisez les commandes *cat, more, head, tail* ou bien utilisez un éditeur de texte). Indiquez suivant quel critère les lignes du fichier sont ordonnées dans le fichier actuellement.

(3) Modifiez les fichiers de façon à ne conserver que les colonnes qui nous serviront dans cette analyse : le numéro du chromosome, la position sur le chromosome, le sens de codage sur ce chromosome et le nombre de tags dans le cluster. Impossible à faire dans un éditeur de texte, mais facile en utilisant les **commandes unix** comme `cut` et `awk` et en redirigeant la sortie vers un fichier (symbole '>'). Utilisez ces commandes pour traiter les données de Montpellier uniquement.

Une façon alternative de réaliser ce traitement est d'utiliser le pipeline en ligne **Galaxy** (http://usegalaxy.org) où vous pouvez charger des fichiers dont les données sont au format texte et composées de colonnes différentes (pour le traitement à réaliser ici, regardez dans les outil du menu *Text Manipulation*). Utilisez cet outil pour traiter les données d'Edimbourg.

I. REGROUPEMENT DE CLUSTERS

Avant de commencer les analyses des clusters en regard des zones connues (annotées) du génome de la souris, une première question de regroupement de clusters se pose : suivant la chance que l'on a eu lors du séquençage, il est tout à fait possible que plusieurs clusters apparemment séparés ne forment en fait qu'un seul cluster mais qu'on ne le sache pas faute d'avoir obtenu au séquençage un tag chevauchant à la fois ces deux clusters. On se pose donc la question de détecter les clusters séparés par peu de nucléotides, et de les regrouper à chaque fois en un seul cluster.

(4) Détaillez en quelque phrases le principe d'un algorithme naturel pour détecter les clusters séparés par moins de n nucléotides. Ecrivez l'algorithme en pseudo-code français et évaluez sa complexité en fonction du nombre k de clusters du fichier.

(5) Si l'algorithme que vous avez proposé ci-dessus a une complexité au moins quadratique en k (c-a-d en $O(k^2)$) alors demandez-vous si en profitant d'un ré-ordonnancement préalable des lignes de clusters vous ne pourriez obtenir une complexité plus faible. Une fois trouvée l'idée, décrivez le principe et la complexité obtenue, pré-traitement inclus.
Triez ensuite les lignes de chaque fichier de clusters (commande `sort`) de façon à ce que les clusters proches sur le génome soient dans des lignes consécutives (commande `sort` avec les options `-u -k` ; regardez le manuel de la commande : `man sort`). De façon similaire à la ligne de commande, Galaxy propose des outils de tris dans la partie *Filter and Sort* de la colonne de gauche, parvenez-vous à lui faire trier les données de façon satisfaisante ?

(6) Ecrivez un petit programme nommé **`fusion`** (par exemple en Python ou Perl) qui accepte en entrée une valeur d'écart `E` (en nb de nucléotides) et un nom de fichier de clusters. Ce programme lit le fichier et fusionne toute paire de lignes qui se suivent et dont les clusters sont à moins de `E` nucléotides l'un de l'autre (voir ce que l'on peut fusionner de ces lignes comme information, c-a-d ce qu'il faut garder pour la suite du TP).

(7) En re-dirigeant la sortie de ce programme vers un fichier, appliquez-le avec une valeur `E=10` à l'ensemble des fichiers de départ pour en obtenir de nouvelles versions. Notez dans le compte-rendu de TP pour chaque fichier de combien de clusters a été réduit le fichier (utilisez la commande `wc -l`). Notez aussi de combien de clusters seraient réduits les fichiers si l'on utilise un écart de 300 nucléotides plutôt que 10.

II. DECOUVERTE de l'UCSC Genome Browser

(8) Trouvez le site ***UCSC Genome Browser*** et ouvrez le dans votre navigateur. Choisissez l'entrée "Genome" et demandez à voir le génome de la souris dans sa dernière version.

Dans votre rapport de TP indiquez à quel clade (groupe d'espèces le plus précis) appartient cette espèce et les informations sur la version la plus récente de son génome.

Une fois la visualisation du génome obtenue (lignes de couleurs noires, bleus, etc sur un fond blanc), observez les champs au dessus vous permettant de zoomer, ou d'indiquer une zone que vous souhaitez visualiser.

(9) Dans les options sous la fenêtre de présentation du génome demandez à afficher les pistes ("track") correspondant aux gènes RefSeq et aux gènes UCSC (gènes connus par ces institutions) en demandant à ce que ces pistes soient affichées avec tout leur détails ("full"). Maintenant à l'aide du champ "gene" au dessus de la fenêtre du génome, localisez le gène correspondant à la protéine G3BP qui sera notre protéine d'intérêt pour cette étude. Indiquez dans votre rapport sur quel chromosome elle se situe. Utilisez le zoom et/ou l'intervalle de positions pour localiser le gène qui se situe juste avant, et celui qui se situe juste après sur le même chromosome. Indiquez leur nom dans votre compte-rendu de TP

Bien, nous allons maintenant importer nos données sur le site de l'UCSC pour visualiser ces données en regard du génome de la souris. Ces données correspondent aux clusters de tags obtenus dans le cadre de l'expérience CLIP-Seq (une expérience de type NGS), indiquant quelles molécules (ARN ou protéines) se lient physiquement à G3BP dans la cellule en activité. Cette protéine est suspectée de se lier surtout avec des ARN, et ainsi de jouer un rôle dans l'expression d'autres protéines (en retenant les ARN de ces protéines, G3BP empêche momentanément la traduction de ces ARN en la protéine cible, sans compter que les ARN tenus par G3BP peuvent au final être "dégradées" (coupées en morceau), empêchant alors définitivement la création de la protéine cible depuis ces ARN.

L'importation de nos clusters de tags sur le génome nous permettra de voir si les deux expériences différentes (MPL et ED) trouvent des molécules qui se lient à G3BP qui proviennent des mêmes endroits du génome. Ces clusters de tags ont été obtenus après une localisation individuelle des tags issus du CLIP-Seq sur le génome en cherchant une séquence qui leur correspond *exactement*.

(10) On a peu de chances de se tromper (ou d'hésiter) en procédant de la sorte pour des tags qui font en général de 50 à 70 bases de long. La souris a un génome de 3,4 milliards de base, sachant qu'un tag peut se trouver sur un brin d'ADN ou son complémentaire, quelle est la probabilité d'avoir une séquence sur le génome correspondant à un tag de 50 nucléotides si la probabilité de chaque base (A,T,G,C) est de 1/4 ?

(11) L'importation de nos données se fera dans les "Custom track" du site de l'UCSC. Trouvez sur le site sous quels formats on peut importer des données, et donnez le nom de trois de ces formats dans votre compte-rendu de TP. D'un point de vue informatique, indépendamment du nom donnés par UCSC à ces formats disponibles, quel est le type de format qui revient le plus : format texte ? format compressé ? format word ? autre ?

De cette constatation, on en déduit qu'Unix (dont Linux est une variante) sera le système le plus commode pour mettre en forme nos fichiers de données car il contient plein de commandes utiles pour jouer avec ce type de format.

(12) Pour la suite du TP, nous allons utiliser le format BED. Trouvez sa description et notez les champs qu'il nous faut préparer pour chaque "track" (piste) que nous voulons observer le long du génome : notez dans le compte-rendu de TP combien de champs (informations) sont obligatoires, quel est leur contenu (dans l'ordre).

(13) Ici pour ne pas perdre le brin sur lequel le cluster apparaît (+ ou -), nous devons donner un nom et un score à chaque cluster. Comme nom nous utiliserons «Clustx» où x sera remplacé par le numéro de cluster (on va utiliser `awk` pour ça, et sa variable `NR`), comme score nous utiliserons le nombre de tags du cluster. Allez-y, transformez le fichier correspondant aux clusters de tags se liant à la protéine G3BP (fichiers dont le nom contient «wt1») dans l'expérience de Montpellier (MPL) en fichier au format BED. Alternativement, utilisez l'outil *Galaxy*. Dans un cas comme dans l'autre, indiquez dans votre compte-rendu de TP la liste de manipulations nécessaires.

Importez maintenant ce fichier BED dans le site UCSC (bouton <*manage custom tracks*>). Une fois ceci fait, dans la même page vous pouvez cliquer sur le nom des données importées pour indiquer des précisions sur la façon d'afficher votre piste dans le *Genome Browser* : entrez la ligne suivante dans la boîte "Edit configuration:"

track name='WT1 MPL' description='WT1 MPL' colorByStrand='255,0,0 0,0,255' visibility=3

Cliquez ensuite sur <*submit*> puis <*go to genome browser*> pour voir votre piste (assurez-vous que les Custom tracks sont affichées et particulièrement celle qu'on vient d'importer).

(14) Localisez le cluster correspondant à un gène connu et apparaissant le plus au début du chromosome M (un petit chromosome). Indiquez dans votre compte-rendu de TP les caractéristiques de ce cluster : position de début, position de fin, sens de lecture (click droit sur le cluster dans le browser permet d'avoir les infos qui lui correspondent), et en allant voir dans le fichier d'origine indiquez le nombre de tags de ce cluster.

(15) Pour savoir si les clusters correspondent à des ARN particuliers ou des protéines, quelles autres "tracks" nous intéressent ? (demandez leur visualisation dans le browser).

(16) Cherchez maintenant un de nos cluster tombant dans une région de type **exon** sur le chromosome 1, mais vous n'avez droit qu'à 5mn chrono pour y arriver (c'est l'expérience qui compte ici, pas le résultat). Décrivez dans le compte-rendu de TP quelle méthode vous avez adopté pour cette recherche (et si les 5mn vous ont suffis pour aboutir, si oui, indiquez le nom du gène contenant la région exon trouvée).

Ceci ne semble pas évident à réaliser pour un cluster, alors imaginez comment cela serait fastidieux à faire pour tous les clusters.

Pour automatiser ce travail pénible, des outils bioinformatiques existent (comme le programme *fjoin*). Nous allons utiliser dans un premier temps pour cela le **Table Browser** (dans le menu *Tools* de la barre horizontale de menu sur le site UCSC).

(17) Explorez cette page Table Browser et décrivez comment vous faîtes pour obtenir au format BED la liste des clusters se positionnant sur des gènes connus de RefSeq (indice : utiliser une intersection, attention l'ordre dans lequel vous indiquez les tables a une influence sur la forme du résultat : l'intersection est en réalité un filtre des tuples d'une table principale en regard d'une table secondaire). Notez que le résultat d'une requête effectuée dans le Table Browser peut soit être récupéré sous la forme d'un fichier texte (remplir le champ output file) dont on précise le format, soit sinon est incorporé à vos données UCSC comme une nouvelle «Custom Track».

(18) Sauvegardez le fichier BED obtenu et comptez à l'aide la commande Unix grep le nombre de clusters correspondant aux chromosomes 1, 2 et M. Quel est parmi ces chromosomes celui qui contient le plus de clusters tombant dans des gènes connus, proportionnellement à sa taille ?

III. REPRODUCTIBILITE EXPERIMENTALE

L'expérience de séquençage de cellules de souris a été faite à deux endroits : Montpellier (MPL) et Edimbourg (ED). Pour certaines analyses, on souhaite ne retenir que les clusters apparaissant dans les deux expériences à la fois, c-a-d chevauchant une même région génomique.

(19) Décrivez dans le compte-rendu de TP les étapes qui vont vous permettre d'obtenir cette liste par l'intermédiaire de l'UCSC ou de Galaxy, et réalisez-les au fur et à mesure.

(20) Quel pourcentage des clusters de l'expérience MPL conservez-vous ? Quel pourcentage des clusters de l'expérience ED ?

(21) Faîtes aussi la même manipulation sur les expériences concernant la petite protéine (CtBp désignée par wt2 dans les noms de fichiers), une protéine alternative à G3BP. Quels pourcentages obtenez vous respectivement ?

(22) Plus tard dans le TP nous aurons besoin des séquences d'ADN correspondant aux clusters. Décrivez comment vous pouvez obtenir les séquences d'ADN correspondant aux clusters d'un jeu de données par l'intermédiaire du Table Browser. Faîtes un essai et indiquez la séquence d'ADN obtenue pour le cluster de la question 8 (vérifiez que vous avez la bonne taille de séquence).

IV. DETERMINATION DES ENSEMBLES COMMUNS ET SPECIFIQUES A LA GRANDE PROTEINE (G3BP) ET A LA PETITE PROTEINE (CtBP)

Lors de la préparation expérimentale pour le séquenceur, outre la protéine d'intérêt (G3BP), une petite protéine (nommée CtBP) a été observée et les tags obtenus par séquençage concernent en fait des ARN liés soit à G3BP soit à CtBP. Les clusters de tags liés à G3BP sont dans des fichiers nommés «wt1...» tandis que ceux liés à CtBP sont dans des fichiers «wt2...».

On veut ici savoir si les catégories de tags associés à G3BP et CtBP sont similaires ou pas. Plutôt que de partir des fichiers de cluster bruts, on va partir des fichiers de clusters auxquels on a assigné des **régions spécifiques** (intron, exon, nom de transcrit (NM_...) ou de non-codant (NR_...)) — fichiers dont le nom comporte «RefGene» et situés dans le dossier REFSEQ des fichiers à télécharger). Ces informations ont été obtenues par un programme (*fjoin*) détectant les chevauchements entre les clusters et un fichier décrivant le contenu du génome de la souris. Ce programme réalise l'équivalent de ce que l'on peut faire par intersection de tables dans le *Table Browser* (mais ce dernier ne propose pas de table détaillant la nature des régions du génome : intron, exon, 3'UTR, etc), ce que l'on pourrait faire par programme, mais bon disposant d'un temps limité, nous partirons des fichiers «RefGene» fournis sur les pages webs de votre encadrant de TP (donc que vous avez déjà récupérés ;)

Soit WT1= clusters de tags se liant à la grande protéine (G3BP), soit 5184 clusters.
Soit WT2= clusters de tags se liant à la petite protéine (CtBP), c-a-d 57604 clusters.

Afin de cibler au mieux la protéine G3BP, on veut déterminer ce qui est spécifique de cette protéine, c-a-d se lie à cette protéine et pas à la petite protéine (CtBP). On appellera **A** cet ensemble :

A = WT1 - WT2

De même on nommera **B** l'ensemble = WT1 inter WT2 (c-à-d, tout cluster de WT1 chevauchant en partie un cluster de WT2) :

B = WT1 inter WT2 (c-a-d les clusters de WT1 ayant une intersection avec ceux de WT2 **plus** ceux de WT2 ayant une intersection avec des clusters de WT1 — précision donnée en regard de l'utilisation des sites UCSC et Galaxy).

(23) Chargez les pistes (si pas déjà fait) correspondant à WT1 et WT2 sur le site UCSC et utilisez le Table Browser pour obtenir les ensembles A et B sous formes de fichiers BED que vous enregistrerez sur votre ordinateur (indices : obtenir B se fait par des manipulations similaires à ce que nous avons fait précédemment ; A s'obtient en changeant une option dans le page où on précise les détails d'une intersection entre tables). Quelles sont les tailles de ces fichiers respectivement ?

(24) On veut aussi calculer l'ensemble **C** = WT2 - WT1. Quelle taille a cet ensemble (sauvez le fichier sur disque lui aussi) ?

V. CLASSIFICATION DES CLUSTERS (CORRESPONDANT A DES GENES, c-a-d associés à un RefSeq) EN TYPES DE SEQUENCES

On utilise ici les fichiers des ensembles A,B et C obtenus dans la partie précédente du TP.

Note : G3Bp et CtBp peuvent se lier dans le noyau avec des ARN prématures exportés ensuite dans le cytoplasme avec la protéine. Dans le cytoplasme la protéine peut aussi se lier à des ARN messagers.

Chaque cluster tombant dans une région génomique correspondant à un gène est intéressant car on a des infos supplémentaires (partie annotée du génome). Par exemple on peut savoir si ce cluster tombe dans la région «**3UTR**», «**5UTR**» (rôle dans la stabilisation de l'ARN) ou «**CDS**» (traduit en protéine) d'un transcrit ou bien en amont («**2k_up**») ou en aval («**1k_dn**») d'un gène (rôle de régulation de l'expression du gène : promoteur, ...).

Précisions sur le sens des annotations dans ces fichiers :

- Attention : les **annotations sont spécifiques de l'ARN considéré** : un même cluster peut être *exon* d'un ARN (transcrit (repéré par un numéro NM_...) ou non-codant (NR_...)) et *intron* dans un autre, en raison de l'épissage alternatif. Dans ce cas là, le même cluster donnera lieu à des lignes du fichier pour le 1er ARN et à des lignes pour le 2ème ARN. Attention ces lignes seront parfois entrelacées, d'où l'importance de trier le fichier avant analyse.
- «**exon**» désigne un morceau d'ADN qui sera présent dans un ARN (codant ou non-codant). Pour un transcrit donné, on peut donc avoir des «exons» dans le CDS, mais aussi dans une région 3UTR et 5UTR.
- «**CDS**» est ici un exon qui se retrouvera dans l'ARNm d'un transcrit pour être ensuite traduit en protéine (au contraire d'autres exons donnant un morceau de l'ARNm sans être traduits ensuite). Un tel exon donnera lieu à deux lignes dans le fichier d'annotations.
- Quand un cluster tombe dans les régions 3UTR ou 5UTR, on précise s'il s'agit d'un «**exon**» ou d'un «**intron**», c-a-d d'un morceau qui ne sera pas transcrit dans l'ARNm (pour un certain NR_...). Dans ce cas, deux lignes seront indiquées pour le même cluster. Attention, ces lignes ne sont pas forcément successives. Quand un cluster tombe dans un intron entre deux exons CDS, alors seule la ligne «intron» sera présente dans le fichier pour ce cluster.

Dernière remarque : dans le cas d'un ARN non codant (NR_...) la nomenclature utilisée est un peu plus flou.

Pour un ensemble donné (A, B ou C), on peut établir une sorte de profil en regardant comment se classent ses clusters. On peut faire se relever pour tous les ensembles, et ensuite se demander si ces ensembles ont globalement des profils similaires ou pas. Dans la négative, on peut conclure que la petite protéine et la grande protéine n'ont pas le même rôle vis à vis de la liaison des ARN de la cellule.

Dans un premier temps, on veut d'abord comparer les ensembles A et B pour voir si les clusters spécifiques de G3BP (la grande protéine) ont un profil différent des clusters qui ne lui sont pas spécifiques (ie ceux de B qui se lient indifféremment avec la petite ou la grande protéine).

On retient les **catégories** suivantes :
a) 3'UTR (peu importe que intron ou exon)
b) 5'UTR (peu importe que intron ou exon)
c) Exons dans les CDS
d) Intron (dans une région de CDS)
e) clusters tombant dans une «cassette», c-a-d dans une région génomique qui est exon pour un ARN et intron pour un autre (on parle aussi d'»exons alternatifs»).

On veut compter les clusters d'un ensemble (A ou B ou C) qui tombent dans ces catégories. Attention, les catégories sont complexes à extraire depuis le fichier de cet ensemble contenant le refSeq de chaque cluster.

(25) Faîtes un relevé des situations possibles en explorant le fichier contenant les RefSeq et notez vos observations dans le fichier de compte-rendu de TP.

(26) Décrivez un algorithme (une façon de s'y prendre) pour compter de façon automatique le nombre de clusters appartenant aux catégories a) b) c) d) énoncées précédemment et **implémentez cet algorithme** dans un langage de votre connaissance. Puis utilisez cet algorithme sur les données pour obtenir le nombre de clusters de chaque catégorie pour chaque grand ensemble A, B et C. Notez dans votre compte-rendu de TP les valeurs obtenues.

(27) Soit C la *catégorie* qui regroupe le plus de clusters dans les calculs ci-dessus (parmi a) b) c) et d)), cherchez sur internet (wikipedia, google, etc) des informations pour savoir s'il arrive que des protéines se lient à ces parties d'ARN dans le cytoplasme ou le noyau.

(28) A l'aide d'un tableur d'une suite bureautique, pour chaque grand *ensemble* (A, B et C) obtenez une représentation circulaire (en "camembert") des effectifs des différentes catégories de cluster. A votre avis, ces trois ensembles ont-ils des profils globalement identiques ou pas ?

VI. TEST STATISTIQUE

Pour confirmer (ou infirmer) notre intuition sur les différences de profils entre ensembles A, B et C, nous allons recourir à un test statistique afin de faire cela de façon rigoureuse.

Nous partons des données nominales (qualitatives) que nous avons : pour chaque cluster, nous connaissons l'ensemble auquel il appartient (A,B ou C) et la région génomique dans laquelle il a été trouvé (catégorie a,b,c,d ou e). Ces deux informations s'apparentent à des variables aléatoires que l'on observe sur un grand échantillon de clusters.

Nous voulons **tester l'indépendance de ces deux variables** pour les clusters (et ainsi savoir si oui ou non les ensembles A, B et C peuvent être jugés comme ayant des profils similaires ou pas).

(29) Cherchez sur Internet quel test il faut employer pour décider de l'indépendance de deux variables aléatoires nominales et notez-le dans votre compte-rendu de TP. Cherchez une page qui

vous décrit comment effectuer ce test (choix de l'hypothèse H0, etc). Citez vos sources dans le compte-rendu de TP.

Pour implémenter ce test, nous allons nous appuyer sur le **logiciel statistique R** (un logiciel libre).

(30) Localisez ce logiciel qui doit être installé sur votre machine, ou installez-le (dans la mesure du possible).

(31) Créez un fichier de données ayant le format suivant dans un fichier regions.txt :

"3UTR" "5UTR" "CDS" "intron ds CDS" "cassette"
"A" 201 14 204 4884 34
"B" 3 2 1 54 0
"C" 398 166 1120 14418 154

où les valeurs numériques (séparées par des espaces) sont bien sûr celles que vous avez trouvées à l'exercice précédent, et où la dernière ligne doit finir par un caractère de fin de ligne (c-a-d tapez sur la touche «return» à la fin de cette ligne).

(32) Chargez ce jeu de données dans R de la façon suivante :

```
regions <- read.table("/chemin/vers/le/fichier/regions.txt")
```

où ce qui précède le nom du fichier est le chemin (linux) indiquant son emplacement dans les dossiers de l'ordinateur (exemple de chemin sous Windows : *C:\Documents and Settings\Toto \TPbioinfo*).

Pour être sûr que le fichier est bien chargé, tapez juste le nom du tableau contenant les données et son contenu devrait vous apparaître :

```
regions
```

Vous pouvez aussi avoir une idée de la répartition de vos clusters dans les différentes catégories par la commande suivante :

```
mosaicplot(regions)
```

Lancez le test statistique que nous avons choisi. Sachant que l'hypothèse nulle (H0) est "les deux variables sont indépendantes" et en prenant un un risque de type I de 5% (probabilité de rejeter H0 alors qu'elle est vraie) quelle conclusion vous dicte la p-value obtenue lors du test ? (c-a-d indépendance ou pas)

Que concluez-vous globalement de cette décision quant aux profils des régions A, B et C ? Ecrivez ces conclusions et les valeurs sur lesquelles elles sont basées dans votre rapport de TP ?

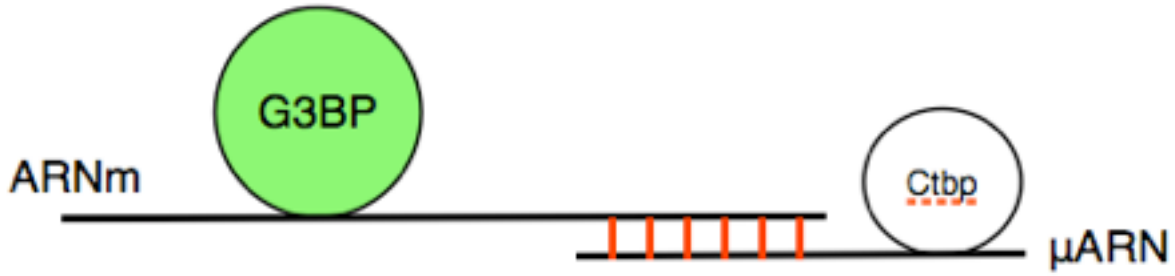
En installant le sous-package **vcd** du logiciel R, on peut obtenir des précisions supplémentaires :

```
assocx(regions)
```

Suivant ce schéma, quelles sont les classes qui s'écartent le plus de l'hypothèse d'indépendance (voir livre sur R ou Internet pour des explications complémentaires) ?

VII. DECOUVERTE D'ASSOCIATIONS INDIRECTES ENTRE G3BP et CtBP

La petite protéine (CtBP) est suspectée de s'associer essentiellement avec des micro-ARN (notés miRNA en anglais) tandis que la grande protéine (G3BP) à des ARN messagers (notés mRNA). Une hypothèse est que les deux protéines sont reliées dans la cellule par le biais de leurs ARN respectifs (voir schéma).



Pour tester cette hypothèse il nous faut trouver si les ARN se liant respectivement aux deux protéines peuvent s'apparier entre eux, ce qui nécessite qu'ils aient des bases complémentaires.

Identifiez les miRNA qui peuvent se lier à la petite protéine en utilisant les services de l'UCSC pour filtrer les clusters de tags se liant à CtBP. Pour la façon de procéder, pensez que l'on veut la portion complète du génome qui correspond à ces miRNA et non juste le tag se liant à CtBP. Sauvegardez les références que vous obtenez dans un fichier **miRNA_wt2.bed**.

Vérifiez sur internet ce qui est dit sur la longueur des miRNA. A quelle longueur faut-il s'attendre environ (en nombre de bases) ? Comparez-le fichier obtenu au fichier wt2_miR.txt qui se trouve dans le dossier ED. Quelles remarques pouvez-vous faire sur la composition de ces deux fichiers ?

Extrayez de l'UCSC ou d'une autre base les séquences complètes de ces miRNA au format FASTA. Nommez **miRNA_wt2.fasta** le fichier résultat. Décrivez quelles manipulations permettent d'obtenir ce fichier et combien de séquences il contient.

Extrayez maintenant le fichier **mRNA_wt1.bed** des ARN messagers se liant à la protéine G3BP (identifiez la bonne piste / table sur l'UCSC et procédez maintenant comme vous savez faire). Puis obtenez le fichier **mRNA_wt1.fasta** des séquences de ces ARN. Décrivez quelles manipulations permettent d'obtenir ces fichier et combien de séquences vous obtenez.

Sachant que les clusters de tags contiennent une section de 10 nucléotides environ qui lie les ARN aux protéines, effectuez un schéma de la situation de liaison de G3BP à un mRNA, lui même lié à un miRNA, lui même lié à CtBP et déduisez-en la configuration d'appariement des séquences entre miRNA et des mRNA qui en résulterait, c-a-d la situation à chercher quand on va tester les appariements possibles entre ces deux types d'ARN.

Dans tous les programmes de la famille BLAST cherchez celui qui s'apparente le mieux à la recherche de parties se liant entre deux ensembles de séquences ADN et comment exploiter un tel programme. Vous pouvez par exemple en avoir une bonne description en français sur le site http://genomenews.free.fr/bioinfo.html . Installez ensuite une version de BLAST sur votre linux (adresse indiquée sur le site cité précédemment à une petite modification près peut-être).

Construisez ensuite deux banques de données locales avec la commande formatdb en partant d'une part des séquences de miRNA et d'autre part de celle des mRNA.

Réalisez les recherches d'appariement entre ces deux catégories d'ARN : attention, les appariements sont forcéments sur des bases complémentaires. Indiquez combien d'appariements vous obtenez. Stockez les alignements fournis par BLAST dans deux fichiers différents.

Constituez deux fichiers **mRNA_wt1_lien.fasta** et **miRNA_wt2.fasta** qui contiennent uniquement les séquences (et leurs noms) impliquées dans un appariement proposé par BLAST.

Sur la base des positions des sites de liaison, reprenez les fichiers d'alignements fournis par BLAST et mettez en majuscule seulement les parties correspondant aux sites de liaisons aux protéines respectivement, et le reste en minuscule (pour chaque alignement vous savez quelles séquences sont impliquées. Cette manipulation peut être faîte à la main ou par programme, comme vous le souhaitez. Appelez **align_laisons_wt1** et **align_laisons_wt1** les fichiers d'alignement ainsi modifiés.

Par combien de nucléotides en moyenne sont séparés dans ces alignements les deux sites de liaison aux protéine (le site liant à G3BP pour un ARN et le site liant à CtBP pour l'autre ARN) ? Dressez la liste des couples d'ARN pour lesquels les sites de liaisons aux protéines sont séparés par au moins 10 nucléotides.

VIII. Identification de la fonction des ARN liés à G3BP

Une question importante est de savoir quelle est la fonction des protéines dont G3BP bloque l'expression en retenant leurs ARN avant la traduction en protéine. Pour se faire explorer les liens avec la *Gene Ontology* et proposez une méthode pour récupérer les termes qui sont le plus souvent associés avec les clusters se liant à G3BP.