M1 SVS – UE8.305 – année 2014-2015
TP : Bases de Données Biologiques
MD Devignes (devignes@loria.fr)

TP : Utilisation de BD biologiques
Exemple de scénario : Recherche de gènes candidats pour une maladie

But du TP : Utilisation d'OMIM, Genome Browser, GOA, GeneCards…

La dégénérescence vitréorétinienne (en anglais : « vitreoretinal degeneration ») en flocons de neige (« snowflake type ») est une maladie génétique progressive qui touche plusieurs tissus de l'œil et se traduit par une dégénérescence de l'humeur vitrée, une cataracte, de légers dépôts cristallins dans la rétine neurosensorielle et un décollement de la rétine. Hejtmancik et coll.[1] ont séquencé 20 des 59 gènes du locus impliqué et ont identifié en 2008 une mutation hétérozygote du gène *KCNJ13* chez la première famille chez qui ce syndrome a été décrit. *KCNJ13* code un canal potassium exprimé dans la rétine et son épithélium pigmentaire.

Dans ce TP il s'agit de mettre en œuvre une approche bioinformatique pour réaliser l'inventaire des gènes présents au locus d'une maladie et pour classer par ordre de priorité d'étude les gènes candidats les plus plausibles. Cette approche permettra de se familiariser avec l'interrogation de plusieurs bases de données biologiques en ligne.

1. *Exploration d'OMIM : Online Mendelian Inheritance in Man*

1.1. Interrogation d'OMIM

- Connectez-vous à OMIM (http://www.omim.org/) et entrez l'expression :

"vitreoretinal degeneration"

(entre guillemets pour limiter la recherche à l'expression entière) dans le champ de recherche

- Combien trouvez-vous de réponses ? ……

- Quels préfixes remarquez-vous (en rouge) devant chaque numéro d'identifiant des réponses parmi les 10 premières réponses (triées par « relevance », cf option en haut à droite)?

……..…

- Cliquer sur le lien « Advanced Search » pour accéder aux explications sur les préfixes et reportez les dans le tableau ci-dessous

MIM Préfixe	Type d'entrée

[1] Hejtmancik, J. F., Jiao, X., Li, A., Sergeev, Y. V., Ding, X., Sharma, A. K., Chan, C.-C., Medina, I., Edwards, A. O. Mutations in KCNJ13 cause autosomal-dominant snowflake vitreoretinal degeneration. Am. J. Hum. Genet. 82: 174-180, 2008.

#	……..
%	……..
*	……..
+	……..

Les entrées OMIM sans préfixe correspondent aux entrées qui ne satisfont aucune de ces définitions en particulier des phénotypes dont l'origine génétique est suspectée (et non démontrée).

- Avec la page Advanced Search on peut limiter la recherche selon divers critères pour mieux cibler les réponses. Essayer par exemple limiter la recherche en ne recherchant l'expression « vitreoretinal degeneration » que dans les titres (Title). Combien de réponses obtenez-vous ? [……]

Quels préfixes sont représentés ? [……]

- Sur cette page on peut aussi savoir combien OMIM compte d'entrées de chaque type (pour chaque préfixe, sans rien saisir dans le champ texte de recherche). Recherchez par exemple combien de phénotypes OMIM restent encore aujourd'hui sans gène responsable identifié ? [……]

- Revenez à votre recherche avec l'expression « vitreoretinal degeneration » limitée aux titres en cliquant sur le lien Search History. C'est la recherche qui avait donné 3 résultats. Notez l'acronyme, l'identifiant et le locus de la dégénérescence vitréorétinienne en flocons de neige.

[……] , [……] , [……]

1.2. Analyse d'une notice OMIM

- Affichez l'entrée complète sur la dégénérescence vitréorétinienne en flocons de neige (= notice OMIM) en cliquant dessus.

- Observez la TABLE OF CONTENTS (sur la droite de la page) d'une notice OMIM et notez les différentes rubriques ci-dessous.

- ……..
- ……..
- ….....
  - ……
  - ……
  - ……
- .....
- ……

- ……..
- ……
- ……..

- Explorez les sous-sections Clinical Features et Mapping et notez quels gènes autres que le gène *KCNJ13* sont discutés puis exclus comme cause de la maladie.

……

- Dans la sous-section Mapping, repérez les marqueurs génétiques permettant la délimitation du locus de la maladie

…… , ……

- Que constatez-vous quand à la bande cytogénétique citée dans la notice OMIM (sous-section Mapping) par rapport à celle notée en 1.1. ?

……

2. *Récupération d'une liste de gènes localisés entre 2 marqueurs génétiques à partir du Genome Browser*

- Connectez-vous au **Génome Browser** (http://genome.ucsc.edu/). Ce site a été conçu et est maintenu depuis de nombreuses années à l'Université de Californie Santa Cruz. Il existe depuis peu un site miroir en Europe (http://genome-euro.ucsc.edu/). Deux autres sites permettent aussi d'explorer les génomes : MapView au NCBI (USA) et ENSEMBL à l'EBI (Europe, Hinxton, UK). L'ergonomie de chaque site est différente et demande un temps d'adaptation. Malheureusement nous n'avons pas le temps de les explorer tous les trois. Le Genome Browser est le plus abordable à prendre en main dans un temps limité (d'après mon expérience).

- Cliquer sur le menu Genomes en haut de la page et sélectionnez le génome humain, assemblage de Février 2009 (GRCh37/hg19).

- Dans le champ « Search Term » on peut rentrer différents types de requêtes : voir la liste énumérée dans le reste de la page sous les en-têtes : « Request » « Genome Browser Response ». On peut en particulier utiliser des marqueurs génétiques dans leur nomenclature D-number (D + numéro du chromosome + S + identifiant à 4 chiffres).

- Saisissez le D-number du 1er marqueur génétique trouvé sur OMIM et cliquez sur submit. Observez la réponse. L'affichage par défaut représente environ 200kb avec le marqueur au centre, surligné en noir et sous un autre nom que celui que vous avez entré !!! (AFM….). Il s'agit du nom donné au Genethon à l'époque des travaux sur la première carte génétique complète du génome humaine (travaux français, pilotés par Jean Weissenbach : AFM = Agence Française contre les Myopathies, organisatrice du Téléthon et financeur du Généthon). En dépit de la fierté de voir cette référence à des travaux français, on ne peut que déplorer le problème de l'hétérogénéité et de la redondance des nomenclatures en génomique (!...).

- Notez le nom AFM du marqueur et cliquez dessus pour obtenir plus d'information. Relever les positions de début et de fin, ainsi que la bande cytogénétique, et remplissez le tableau page suivante pour les deux marqueurs. Remarquez la multiplicité des alias de nomenclature (other names). Quelle en est la raison à votre avis ?

	Marqueur à gauche	Marqueur à droite
D-number trouvé sur OMIM		
Nom AFM		
Chromosome		
Position début		
Position fin		
Bande cytogénétique		

- En déduire les coordonnées précises du locus de la maladie étudiée (entre les deux marqueurs) dans la version du génome interrogée :

……..

- Ainsi que l'intervalle cytogénétique concerné :

…….

- Faire afficher cette région sur le Genome Browser selon la syntaxe proposée par le système (regarder l'exemple fourni en haut de la page sous le mot « position »). Quelle est la taille en Megabases de la région à explorer ? (regardez bien, le système la calcule pour vous et l'affiche)

……..

- Sur le Genome Browser l'affichage des différentes pistes peut être personnalisé très facilement. Chaque piste (track) est repérée par un titre centré sur la figure (ex : STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps), et peut apparaître en version complète (full) ou condensée sur une seule ligne (dense). Pour passer du complet au condensé il suffit de cliquer sur le titre de la piste.
Une grande variété de pistes peuvent être affichées : allez plus bas sur la page, après la figure. Elles sont regroupées par thématique : Mapping and Sequencing, Genes and Gene Prediction, Phenotype and Literature, mRNA and EST, Expression, Regulation, Comparative Genomics, etc. Pour chaque

groupe on peut choisir d'afficher ou non (hide) un certain nombre de pistes qui correspondent à des éléments alignés sur le génome.
- Essayez de simplifier votre figure de la région étudiée en supprimant toutes les pistes sauf Base Position, Chromosome Band, STS markers, UCSC Genes, CpG islands, Repeat Masker. Positionner toutes les pistes en mode condensé (dense).

- Observez la piste UCSC Genes. On peut aussi l'afficher en mode squish, pack, (en plus de full et de dense). Quel est l'intérêt de chaque mode ?

- Pour télécharger la liste des gènes de la région, il faut utiliser une fonctionnalité du Genome Browser le **Table Browser** dans le menu **Tools** (sur le bandeau du haut de la page).

Pour exécuter cette fonction, cliquer dessus et dans le formulaire qui s'affiche :

➢ Vérifier ou préciser les paramètres suivants

- Clade : Mammal,
- Genome : Human
- Assembly : Feb. 2009 (GRCh37/hg19)
- Group : Genes and Gene Predictions Tracks
- Track : UCSC Genes
- Table : knownGene
- Region : cocher ⊙ « Position » et vérifier que les coordonnées de la région sont correctement affichées
- Output format : choisir « selected fields from primary and related tables »
- File type return : s'assurer que ⊙ « plain text » est coché

➢Puis envoyer (avec le bouton « get output »). On a alors la possibilité de choisir les champs qui seront affichés.

- Sélectionner dans la 1ère table (« from hg19.knownGene »)
  - ✓ name
  - ✓ txStart
  - ✓ tsEnd
- Sélectionner dans la 2ème table (« hg19.kgXref fields »)
  - ✓ geneSymbol

Cliquer sur « get output » (en dessous de la première table). Copier le texte affiché sur le browser dans un fichier texte (ex : Wordpad ou OpenOffice à enregistrer en type txt) que vous nommerez LocusSVD-genelist.txt

- Ouvrir Excel et importer le fichier SVD-locus-genelist.txt en suivant les instructions d'import d'Excel[2]. Observer les différentes lignes : que remarquez-vous dans la colonne GeneSymbol ? A quoi correspondent les lignes différentes pour le même GeneSymbol ? Suggestion : pour répondre à cette question retourner sur le site du GenomeBrowser et interroger avec l'un de ces GeneSymbols qui apparaissent plus d'une fois.

[2] Sous OpenOffice, utiliser la fonction Insertion feuille à partir d'un fichier dans le menu Insertion, en faisant attention lors de la sélection du fichier de bien préciser le type de fichier qui doit être Texte CSV (*.csv ;*.txt). Dans la fenêtre qui s'ouvre, préciser que les données sont séparées par des tabulations.

……..

……....

……..

- Supprimez les doublons[3] (lignes pour lesquels le GeneSymbol est identique à celui d'une ligne antérieure). Enregistrez sous LocusSVD-genelistSansDoublons.xls.

- Combien obtenez-vous ainsi de gènes candidats distincts dans le locus ?

……

3. Tri et filtrage de la liste de la liste de gènes

Il apparaît ensuite bien difficile de distinguer parmi tous ces gènes les meilleurs candidats à séquencer en priorité ! Il faut formuler des hypothèses et collecter les annotations des gènes permettant de filtrer les gènes candidats sur la base de ces hypothèses.

IMPORTANT : Pour la suite on ne travaillera que sur une liste courte (« short list ») constituée d'une douzaine de gènes autour du gène KCNJ13 (copier les symboles des gènes entre TIGD1 et ATG16L1), les insérer dans une nouvelle feuille appelée « ShortList » dans le fichier Excel ou OpenOffice).

3.1. Elimination des gènes déjà liés à une maladie dans OMIM

- Revenir à OMIM et interroger successivement avec chacun des symboles des gènes de la liste courte. Ne retenez que les réponses préfixées par # qui indiquent que ce gène a été impliqué dans la maladie préfixée par #. Quelles autres réponses obtenez-vous ? Notez dans une colonne supplémentaire de la feuille ShortList, le ou les numéros OMIM des pathologies qui font référence à ce gène. Si on avait le temps, il faudrait naturellement aller voir dans les notices OMIM s'il s'agit effectivement du gène responsable de la maladie ou simplement d'un candidat mentionné dans la notice.

Bilan : En plus du gène KCNJ13 qui est bien sûr associé à SVD, combien de gènes de la liste courte sont déjà associés à une maladie dans OMIM ? ……

Combien sont inconnus dans OMIM ? ……

[3] La suppression des doublons est facile avec Excel 2007 (ou +). C'est une option du menu « Données ». Avec OpenOffice ou une version antérieur d'Excel, il faut faire la manip suivante : trier le tableau selon les GeneSymbol (colonne D), puis insérer dans la case E2 de la colonne E, à droite des GeneSymbol la fonction suivante: =SI(D2=D1 ; 1 ; 0) , puis recopier vers le bas pour toutes les lignes. Il s'affiche un 0 chaque fois que le GeneSymbol est différent du précédent et un 1 quand c'est le même. Recopier en colonne F seulement les valeurs (pas la formule) en utilisant collage special, et en cochant seulement nombres. Puis trier le tableau en 1er tri sur la colonne F et en 2ème tri sur la colonne B (Start Position). On en obtient en début de tableau les gènes sans leurs doublons (valeur 0 en colonne F) dans l'ordre de localisation sur le génome.

3.2 Recherche des annotations GO : Gene Ontology

Le vocabulaire contrôlé GO (Gene Ontology) rassemble dans trois hiérarchies (Molecular Function, Biological Process et Cellular Component) des termes qui sont utilisés par les biologistes pour annoter les gènes et les protéines codées par ces gènes.
La base de données GOA (pour GO Annotations) collecte et donne accès à toutes ces annotations. On peut l'interroger sur le site de l'EBI.

http://www.ebi.ac.uk/QuickGO/

Pour chacun des gènes de la shortlist, non associés à une maladie, dans la page de recherche rapide (Quick Search) taper le symbole du gène. Suivre les liens pour arriver au tableau listant les annotations GO associées à ce gène chez l'homme.

Observez toutes les informations données dans chaque ligne du tableau. N'hésitez pas à poser des questions si vous ne comprenez pas.

Compléter le tableau Excel précédant en recopiant au moins un terme GO de la branche Biological Process associé à chacun des gènes étudiés.

Réfléchir aux annotations les plus favorables pour suggérer un gène candidat pour la pathologie étudiée.

3.3 Recherche du lieu d'expression dans l'organisme

Pour discuter de la plausibilité d'un gène candidat on peut s'appuyer sur l'hypothèse selon laquelle le gène candidat doit être exprimé dans le tissu où s'exprime la maladie. Ici la rétine ou le corps vitré ou de façon plus grossière dans l'œil ou le cerveau.

Il existe différents sites que l'on peut interroger pour cela.

- Dans le Génome Browser il est possible de faire afficher les sites d'expression des gènes à partir des données GNF
- Dans les fiches de la base GENE du NCBI (notée Entrez-GENE)
- Dans les fiches Gene Cards, rubrique Expression
- Etc.

On explorera ici GeneCards http://www.genecards.org , une ressource intégrée très complète et bien mise à jour (on peut vérifier les dates de mise à jour en bas de la fiche du gène interrogé)

Entrer le symbole du gene et sélectionner l'option Search by« symbol only ».

Après affichage de la page, sélectionner à gauche dans « Jump to section » la rubrique « Expression». Regarder les histogrammes et repérer s'il y a une expression dans la rétine pour l'un ou l'autre des types d'expérience.

Compléter le tableau Excel en ajoutant trois colonnes pour « Expression dans la rétine (selon GeneCards) », selon Microarray, RNAseq ou SAGE, et en notant de façon schématique les résultats rapportés par GeneCards (par exemple : - , +, ++, etc.).

- Que penser des résultats obtenus pour le gène KCNJ13 ?

……

3.4 Recherche des interactants et des pathways

La protéine codée par le gène candidat peut interagir avec une autre protéine au sein d'un pathway donné. Dans ce cas on pourra discuter de l'effet d'une mutation dans le gène candidat en supposant que la mutation empêche l'interaction et perturbe le pathway. Pour retrouver ces informations on peut aussi consulter Genecards dans la rubrique « Pathways and interactions ».

Compléter le tableau Excel en ajoutant une colonne « Pathway KEGG » et en notant le nom du ou des pathways de la base de données KEGG qui apparait dans la fiche GeneCards. Remarquez qu'il existe de nombreuses autres sources d'informations selon les gènes considérés.

Observer les réseaux d'interaction proposés par la base de données STRING : cliquer sur le lien vers STRING Interaction Network pour comprendre les informations fournies par cette base de données.

3.5 Autres Outils en ligne : calculs de similarité

Fonction Gene Sorter du Genome Browser : Cette fonction (accessible par le bandeau supérieur du génome browser) recherche dans le génome des gènes similaires à un gène ou à un critère que l'on saisit en entrée. On peut limiter la recherche à une région du génome. La mesure de similarité peut concerner les lieux d'expression, les annotations GO, la séquence etc.

Endeavour du groupe de Yves Moreau http://homes.esat.kuleuven.be/~bioiuser/endeavour/endeavour.php
Cet outil est le plus complet. Il permet de classer par ordre de priorité une liste de gènes en fonction de divers critères de similarité par rapport à une liste de gènes donnée en paramètre comme gènes témoins. Il faudrait ici réfléchir à une liste de gènes témoins impliqués dans d'autres dégénérescences rétiniennes.