Recherche d'interactions sur données de type trio - Manuel d'utilisation

Langage : R

3 juillet 2014

1 Contraintes et pré-formatage

1.1 Fichier principal.

Le fichier principal contenant les informations sur les facteurs de risque, les informations génotypiques sur les SNP et les informations relatives aux individus telles que les identifiants, les identifiants de famille ou le sexe au format .csv doit être formaté selon les consignes suivantes :

- La colonne représentant le sexe de l'individu doit être nommée "Sex".
- La colonne représentant la famille de l'individu doit être nommée "Fam".
- La colonne représentant le statut de l'individu vis-à-vis de la maladie doit être nommée "Status".
- La colonne représentant le set de l'individu doit être nommée "set_complet".
- Les facteurs de risque (facteurs cliniques, environnementaux et génétiques qui ne sont pas les génotypes de SNP) doivent être représentés par des colonnes adjacentes.
- Les génotypes de SNP doivent être représentés par des colonnes rassemblées à la droite du tableau.
- Le séparateur de ce fichier doit être un caractère de tabulation.

1.2 Fichier d'informations sur les SNP.

Ce fichier doit contenir les informations suivantes :

- En première colonne le numéro du SNP (dans l'ordre des SNP sur le fichier contenant les informations génotypiques)
- En deuxième colonne le nom du SNP
- En cinquième et sixième colonne les allèles du SNP
- Le séparateur de ce fichier doit être ";".

Ces deux fichiers doivent être placés dans l'arborescence suivante : .../*nom de l'étude*/Donnees/

2 Créer et lancer une étude

2.1 Création du dossier de l'étude.

1. Lancer le programme "dir_creator.py".
2. Rentrer le chemin absolu du dossier où seront déposés tous les fichiers créés dans le cadre de l'étude.
3. Toute l'arborescence nécessaire à l'étude est maintenant créée.

2.2 Lancer l'étude.

2.2.1 Utilisation.

1. Lancer R en mode console.

2. Compiler le fichier .../*nom de l'étude*/Fonctions/**pipeline_generale.r**. Possible en chemin relatif en se rendant dans le dossier de pipeline_generale.r à partir de la commande setwd() de R : source("pipeline_generale.r"), ou en chemin absolu (rentrer le chemin absolu dans la fonction source()).

> **Attention**
> Pour R sous windows, les dossiers dans les arborescences sont séparés par \\

3. Lancer la fonction "pipeline_generale()" avec les arguments souhaités (voir partie suivante).

2.2.2 Description des arguments.

La fonction pipeline_generale() possède les arguments suivants : (chemin, chemEtude, nomDonnees, nomDonneesSNP, begClin, endClin, colSNP, initial, replication, seuilReglog, seuilGestionNA, kMDR).

- **chemin** *chaîne de caractères* : le chemin absolu du dossier dans lequel est contenu le dossier créé pour l'étude par "dir_creator.py".
- **chemEtude** *chaîne de caractères* : le nom du dossier créé pour l'étude par "dir_creator.py".
- **nomDonnees** *chaîne de caractères* : le nom du fichier contenant les données sur les familles de type trio.
- **nomDonneesSNP** *chaîne de caractères* : le nom du fichier contenant les données sur les SNP.
- **begClin** *entier* : le numéro de la première colonne des informations sur les facteurs de risque.
- **endClin** *entier* : le numéro de la dernière colonne des informations sur les facteurs de risque.
- **colSNP** *entier* : le numéro de la première colonne des informations sur les données génotypiques des SNP.
- **initial** *vecteur* : les numéros des sets compris dans le set initial.
- **replication** *vecteur* : les numéros des sets compris dans le set de réplication.
- **seuilReglog** *décimal* : le seuil de p-value à partir duquel sont séléctionnées pour les modèles multivariés d'hétérogénéité les variables sortant dans les modèles bivariés comprenant l'effet d'un SNP et d'une interaction entre ce SNP et un facteur de risque.
- **seuilGestionNA** *décimal* : la proportion maximale de valeurs manquantes autorisée pour que les données sur un SNP soient prises en compte dans les analyses dans lesquelles ont souhaite que les individus aient toutes les données renseignées. Valeur comprise entre 0 et 1.
- **kMDR** *entier* : le nombre maximal de facteurs considérés dans un même temps pour déterminer le meilleur modèle.

> **Attention**
> Sous R, un vecteur s'écrit de la façon suivante : c(a,b) où ce vecteur contient l'élément a et l'élément b. Par exemple, si on veut définir un vecteur contenant les chiffres 2 et 3, on écrira : c(2,3).
> De plus, la notation anglaise est utilisée pour écrire les nombres décimaux. Ainsi la valeur entière et la valeur décimale d'un nombre seront séparées par "." et non par ",".

2.2.3 Exemple.

```
1 pipeline_generale("C:\\Users\\Genhotel-Stat\\Desktop\\Etude_interactions",
2 "Etude_1", "donnees_trio.csv", "donnees_SNP.csv",
3 11, 19, 23,
4 c(1,2), c(3,4,5),
5 0.05, 0.2, 2)
```

Dans cet exemple, le nom de l'étude est "Etude_1" (ligne 2). Ce dossier contiendra tous les fichiers de sortie des analyses et sera compris dans le chemin "C :\\Users\\Genhotel-Stat\\Desktop\\Etude_interactions" (ligne 1). Le fichier contenant les données génotypiques est nommé "donnees_trio.csv". Le fichier contenant les données sur les SNP est nommé "donnees_trio.csv" (ligne 2).
Les données sur les facteurs de risques sont inclues dans les colonnes 11 à 19. Les données génotypiques sur les SNP sont contenues dans les colonnes 23 à la dernière colonne du fichier (ligne 3).
On souhaite rassembler les sets 1 et 2 en un set initial et les sets 3, 4 et 5 en un set de réplication (ligne 4).
On choisit un seuil de p-value égal à 0.05 pour tester les SNP trouvés dans des modèles bivariés, dans des modèles multivariés. On choisit que pour les analyses pour lesquelles les individus doivent posséder toutes les données ne soient pris en compte que les SNP pour lesquels on dispose d'au moins 80% des données. Enfin, on a choisi de considérer des modèles ayant au plus 3 variables dans le cadre des analyses MDR (ligne 5).