Universidad Zaragoza

Trabajo Fin de Grado

Un estudio de asociación genómica basado en aprendizaje automático para la caracterización de la enfermedad de Alzheimer

A genomic-wide association study based in machine learning for the characterization of Alzheimer's Disease

Autor

Eduardo Alonso Monge

Directoras

Elvira Mayordomo Cámara
Mónica Hernández Giménez

Ingeniería Informática - Computación

Escuela de Ingeniería y Arquitectura
2020

Índice

Anexo 1: Software utilizado

1.1 PLINK

PLINK es un conjunto de herramientas de código abierto para análisis de asociación de genoma completo, diseñado para realizar una gama de análisis básicos a gran escala de una manera computacionalmente eficiente. El enfoque de PLINK es puramente el análisis de datos de genotipo/fenotipo, por lo que no hay soporte para los pasos previos a esto.

La opción básica para generar un nuevo conjunto de datos es –recode:

```
plink --file data --recode
```

que generará las etiquetas de los alelos tal como aparecen en el original; también, el código del genotipo que falte se conserva si es diferente de 0. Además, si se especifica –output-missing-genotype, este valor se usará en su lugar.

Para fusionar dos archivos PED / MAP se utiliza el siguiente comando:

```
plink --file f1 --merge f2.ped f2.map --recode --out merge
```

La opción –merge debe ir seguida de 2 argumentos: el nombre del segundo archivo PED y el nombre del segundo archivo MAP. Una opción –recode (o –make-bed, etc.) es necesaria para generar el archivo recién fusionado; en este caso, la opción –out creará los archivos merge.ped y merge.map.

La opción –merge también se puede usar con archivos PED binarios, ya sea como entrada o salida, pero no como segundo archivo, es decir

```
plink --bfile f1 --merge f2.ped f2.map --make-bed --out merge
```

creará merge.bed, merge.fam y merge.bim, ya que se usó la opción –make-bed en lugar de la opción –recode.

El siguiente formato a menudo es útil si se desea utilizar un paquete estadístico para analizar los datos, ya que aquí los genotipos se codifican como un número de alelo único. Para crear un archivo con genotipos SNP recodificados en términos de componentes aditivos y dominantes, se usa la opción:

```
plink --file data --recodeAD
```

que, suponiendo que C es el alelo menor, recodificará los genotipos de la siguiente manera:

SNP		SNP_A	,	SNP_HET
A A	–>	0	,	0
A C	–>	1	,	1
C C	–>	2	,	0
0 0	–>	NA	,	NA

El valor predeterminado para la recodificación aditiva es contar el número de alelos menores por persona. La opción –recodeAD produce una codificación aditiva y de dominación, se usa –recodeA en su lugar para omitir la codificación SNP_HET. La opción –recodeAD/–recodeA guarda los datos en un solo archivo plink.raw.

El comportamiento de los comandos –recodeA y –recodeAD se puede cambiar con el comando –recode-allele. Esto permite que el recuento 0, 1, 2 refleje el número de un tipo de alelo especificado previamente por SNP, en lugar del número del alelo menor. Este comando toma como argumento único el nombre de un archivo que enumera el nombre SNP y el alelo para informar.

1.2 ANNOVAR

ANNOVAR es una herramienta de software para anotar funcionalmente las variantes genéticas detectadas de diversos genomas (incluido el genoma humano hg18, hg19, hg38, así como ratón, gusano, mosca, levadura y muchos otros). Dada una lista de variantes con cromosoma, posición inicial, posición final, nucleótido de referencia y nucleótidos observados, ANNOVAR puede realizar:

- **Anotación basada en genes:** identifica si los SNP o CNV causan cambios en la codificación de proteínas y los aminoácidos que se ven afectados. Se pueden usar con flexibilidad los genes RefSeq, los genes UCSC, los genes ENSEMBL, los genes GENCODE, los genes AceView o muchos otros sistemas de definición de genes.
- **Anotación basada en regiones:** identifica variantes en regiones genómicas específicas, por ejemplo, regiones conservadas entre 44 especies, sitios de unión a factores de transcripción pronosticados, regiones de duplicación segmentaria, aciertos de GWAS, base de datos de variantes genómicas, sitios de hipersensibilidad de ADN, ENCODE H3K4Me1 / H3K4Me3 / H3K27Ac / Sitios de CTCF, picos de ChIP-Seq, picos de RNA-Seq o muchas otras anotaciones en intervalos genómicos.
- **Anotación basada en filtros:** identifica variantes que están documentadas en bases de datos específicas, por ejemplo, si se informa una variante en dbSNP, cuál es la frecuencia de alelos en el Proyecto 1000 Genoma, exomas NHLBI-ESP 6500 o el genoma base de datos de agregación (gnomAD).
- **Otras funcionalidades:** recupera la secuencia de nucleótidos en cualquier posición genómica específica, identifica una lista de genes candidatos para enfermedades mendelianas a partir de datos del exoma y otras utilidades.

En este trabajo se hará uso de la funcionalidad en anotación basada en genes. Antes de trabajar en ella, un archivo de definición de genes y el archivo FASTA asociado deben descargarse en un directorio. Llámese este directorio, por ejemplo, humandb/

```
$ annotate_variation.pl -downdb -buildver hg19
                        -webfrom annovar refGene humandb/
NOTICE: Web-based checking to see whether ANNOVAR new version
 is available ... Done
NOTICE: Downloading annotation database
 http://openbioinformatics.org/hg19_refGene.txt.gz ... OK
NOTICE: Downloading annotation database
 http://openbioinformatics.org/hg19_refLink.txt.gz ... OK
NOTICE: Downloading annotation database
 http://openbioinformatics.org/hg19_refGeneMrna.fa.gz ... OK
NOTICE: Uncompressing downloaded files
NOTICE: Finished downloading annotation files for hg19 build
 version, with files saved at the 'humandb' directory
```

A continuación se crea un fichero .txt con la lista de los SNPs que se pretende conocer el gen al que pertenecen:

```
$ cat example/snplist.txt
rs74487784
rs41534544
rs4308095
rs12345678
...
```

El siguiente paso es convertir el fichero .txt al formato .avinput con el que trabaja ANNOVAR con el siguiente comando:

```
$ convert2annovar.pl -format rsid example/snplist.txt
                     -dbsnpfile humandb/hg19_snp138.txt
                     > snnplist.avinput

NOTICE: Scanning dbSNP file humandb/hg19_snp138.txt...
NOTICE: input file contains 4 rs identifiers, output file
 contains information for 4 rs identifiers
WARNING: 1 rs identifiers have multiple records (due to
 multiple mapping) and they are all written to output
```

Lo que daría como resultado un fichero como el siguiente:

```
$ cat snplist.avinput
chr2 186229004 186229004 C T rs4308095
chr7 6026775 6026775 T C rs41534544
chr7 6777183 6777183 G A rs41534544
chr9 3901666 3901666 T C rs12345678
chr22 24325095 24325095 A G rs74487784
```

La anotación basada en genes puede realizarse mediante el siguiente comando (de forma predeterminada están activados los siguientes flags –geneanno -dbtype refGene) pasándole como entrada el fichero previamente creado.

```
$ annotate_variation.pl -build hg19 snplist.avinput humandb/
                        -out snplist

NOTICE: The --geneanno operation is set to ON by default
NOTICE: Reading gene annotation from humandb/hg19_refGene.txt
  ... Done with 48660 transcripts for 25588 unique genes
NOTICE: Reading FASTA sequences from humandb/hg19_refGene.fa
WARNING: A total of 333 sequences will be ignored.
NOTICE: Finished gene-based annotation on 15 genetic variants
NOTICE: Output files were written to snplist.variant_function,
 snplist.exonic_variant_function
```

Se generarán dos archivos de salida: snplist.variant_function y snplist.exonic_variant_function (para cambiar los nombres de los archivos de salida se usa el argumento –outfile).

```
UTR5 ISG15(NM_005101:c.-33T>C) 1 948921 948921 T C
UTR3 ATAD3C(NM_001039211:c.*91G>T) 1 1404001 1404001 G T
intronic DDR2 1 162736463 162736463 C T
intronic DNASE2B 1 84875173 84875173 C T
intergenic LOC6454(dist=11566),LOC33(dist=1102) 1 1393 1394 TC
intergenic UBD1(dist=5505),PHD2(dist=1399) 1 11596 11596 - AT
intergenic LOC108(dist=8738),NONE(dist=NONE) 1 121 131 A ATA
exonic IL23R 1 67705958 67705958 G A
exonic ATG16L1 2 234183368 234183368 A G
exonic NOD2 16 50745926 50745926 C T
exonic NOD2 16 50756540 50756540 G C
exonic NOD2 16 50763778 50763778 - C
exonic GJB2 13 20763686 20763686 G
```

La primera columna dice si la variante golpeó a los exones o las regiones intergénicas, o los intrones o los genes de ARN no codificantes. Si la variante es exónica / intrónica / ncRNA, la segunda columna da el nombre del gen (si se intersecta con múltiples genes, se agregará una coma entre los nombres de los genes)

Los posibles valores de la primera columna se resumen a continuación:

Valor		**Explicación**
exonic	1	La variante se superpone a una codificación
splicing	1	La variante se encuentra a 2 pb de una unión de empalme (use -splicing_threshold para cambiar esto)
ncRNA	2	La variante se superpone a una transcripción sin anotación de codificación en la definición del gen
UTR5	3	La variante se superpone a una región no traducida de 5'
UTR3	3	La variante se superpone a una región no traducida de 3'
intronic	4	La variante se superpone a un intrón
upstream	5	La variante se superpone a la región de 1 kb por encima del lugar de inicio de la transcripción
downstream	5	La variante se superpone a la región de 1 kb por debajo del lugar de inicio de la transcripción
intergenic	6	La variante está en región intergénica

Anexo 2: Formato de ficheros

.fam (PLINK sample information file)

Archivo de información de muestra que acompaña a una tabla de genotipos binarios .bed. (–make-just-fam se puede usar para actualizar solo este archivo). Un archivo de texto sin línea de encabezado y una línea por muestra con los siguientes seis campos:

- ID de familia ('FID')
- ID dentro de la familia ('IID'; no puede ser '0')
- ID dentro de la familia del padre ('0' si el padre no está en el dataset)
- ID dentro de la familia de la madre ('0' si la madre no está en el dataset)
- Código de sexo ('1'=masculino, '2'=femenino, '0'=desconocido)
- Valor de fenotipo ('1'=control, '2'=caso, '-9'/ '0'/no numérico=falta de datos)

.bim (PLINK extended MAP file)

Archivo de información de variante extendido que acompaña a una tabla de genotipos binarios .bed. (–make-just-bim se puede usar para actualizar solo este archivo). Un archivo de texto sin línea de encabezado y una línea por variante con los siguientes seis campos:

- Código de cromosoma (ya sea un entero o 'X'/'Y'/ 'XY'/ 'MT'; '0' indica desconocido) o nombre
- Identificador de variante
- Posición en morgans o centimorgans (opcional o '0')
- Coordenada del par base (basado en 1; limitado a 231-2)
- Alelo 1 (correspondiente a bits claros en .bed; generalmente menor)
- Alelo 2 (correspondiente a los bits establecidos en .bed; generalmente mayor)

.bed (PLINK binary biallelic genotype table)

Representación primaria de llamadas de genotipo en variantes bialélicas. Debe ir acompañado de archivos .bim y .fam. Cargado con –bfile; generado en muchas situaciones, especialmente cuando se usa el comando –make-bed. No confundir con el formato BED de UCSC Genome Browser, que es totalmente diferente.

.map (PLINK text fileset variant information file)

Archivo de información de variantes que acompaña una tabla de pedigrí + genotipo de texto .ped. También generado por –recode rlist. Un archivo de texto sin

archivo de encabezado y una línea por variante con los siguientes 3-4 campos:

- Código de cromosomas.
- Identificador de variante
- Posición en morgans o centimorgans (opcional o '0')
- Coordenada del par base

.ped (PLINK text pedigree + genotype table)

Formato de texto estándar original para información de pedigrí de muestra y llamadas de genotipo. Normalmente debe ir acompañado de un archivo .map; Cargado con –file, y producido por –recode.

No contiene encabezado, contiene una línea por muestra con 2V + 6 campos donde V es el número de variantes. Los primeros seis campos son los mismos que los de un archivo .fam. Los campos séptimo y octavo son llamadas alélicas para la primera variante en el archivo .map ('0' = sin llamada); el noveno y el décimo son alelos para la segunda variante; y así.