3 de mayo de 2016
Consuelo Dayzu Quinto

Cambio de alelos de acuerdo a la cadena FORWARD y REVERSE en los SNPs del microarreglo MEGA

Usando el manifiesto Multi-EthnicGlobal_A1.bpm y los datos crudos de la primera corrida de Maria Avila (maavp1v1), se obtuvo un reporte final (maavp1v1.raw_FinalReport.txt) que contiene la informacion de los alelos en la cadena Forward para cada SNP en el microarreglo MEGA.

Con la informacion de este reporte mas informacion de la base de datos dbSNP y 1KG, hice una lista `list_snps_to_change_alleles.txt`. Esta lista contiene el identificador del SNP, los alelos detectados por GenomeStudio, los alelos de dbSNP y 1KG, y la posicion fisica para 664,465 SNPs.

Archivo list_snps_to_change_alleles.txt:

```
JHU_8.51371909 A G T C 51371910
JHU_9.117096517 A G T C 0
rs55873141 C G G C 113845176
JHU_2.83159627 A G T C 83159628
rs10806671 A G T C 170270028
2:630995-T-G A C T G 630995
rs6079035 A C T G 13368741
rs2088629 A G T C 133287048
rs2066705 A G T C 25937004
```

Escribi un script en perl `convert_MEGA_alleles.pl` para procesar los datos crudos provenientes de GenomeStudio que realiza las siguientes cosas:

- Remueve SNPs mapeados en el cromosoma cero ($\sim$ 16,749 sin cluster; $\sim$ 10,791 con cluster).
- Remueve SNPs duplicados por ID y por posicion fisica.
- Cambia los alelos de acuerdo a la lista `list_snps_to_change_alleles.txt` (664,465 SNPs).
- Actualiza la posicion fisica de dos SNPs: rs9522257 y rs9480186 `new_positions_snps.txt`
- Renombra los SNPs con la lista `MEGA_Consortium_v2_15070954_A1_b138_rsids.txt`

Para correr este script solo se necesita un archivo PED:

```
perl convert_MEGA_alleles.pl archivo.ped
```

Este script genera:

- archivo.flip.ped y archivo.flip.map
- duplicate_snps_archivo.txt: ID de los SNPs duplicados
- duplicate_positions_entrenamiento_archivo.txt: ID de los SNPs que tienen la misma posicion
- to_remove.txt: ID de los SNPs duplicados que se excluyen del archivo map original

NOTA IMPORTANTE: Los archivos `list_snps_to_change_alleles.txt; new_positions_snps.txt; MEGA_Consortium_v2_15070954_A1_b138_rsids.txt` tienen que estar en el mismo directorio donde se encuentren los archivos de plink, y se requiere de la ultima version de plink (plink 1.9).

Aparte de los archivos post-procesados, hay dos archivos que se pueden dar a los usuarios:

- multiallelic_SNPs.txt tiene la lista de los SNPs que no son bialelicos (9,583 SNPs).
- complementary_snps.txt tiene la lista de SNPs cuyos alelos son C/G o A/T (2,036 SNPs).