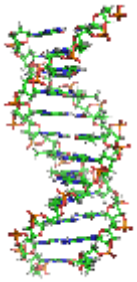

Convertir las bases de datos de 1000Genome a formato plink

Bajando los archivos

Primeramente hay que obtener las bases de datos del proyecto *1000Genome*. Estos archivos estan en formato VCF (Variant Call Format) en el ftp de *1000Genome*,

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/

Observese que esta es la última *release* de los archivos pero hay otras. La forma más fácil de bajar todos los archivos es hacer un **wget** para todos los ficheros. Hacemos la lista primero,

files2download.txt

```
ALL.chr1.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr1.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.tb
i
ALL.chr10.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr10.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.t
bi
ALL.chr11.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr11.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.t
bi
ALL.chr12.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr12.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.t
bi
ALL.chr13.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr13.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.t
bi
ALL.chr14.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr14.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.t
bi
ALL.chr15.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr15.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.t
bi
ALL.chr16.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
```

```
ALL.chr16.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.t
bi
ALL.chr17.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr17.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.t
bi
ALL.chr18.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr18.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.t
bi
ALL.chr19.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr19.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.t
bi
ALL.chr2.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr2.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.tb
i
ALL.chr20.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr20.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.t
bi
ALL.chr21.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr21.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.t
bi
ALL.chr22.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr22.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.t
bi
ALL.chr3.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr3.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.tb
i
ALL.chr4.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr4.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.tb
i
ALL.chr5.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr5.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.tb
i
ALL.chr6.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr6.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.tb
i
ALL.chr7.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr7.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.tb
i
ALL.chr8.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr8.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.tb
i
ALL.chr9.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz
ALL.chr9.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz.tb
i
```

Nota: Los archivos .tbi son los indices que usan las herramientas de procesamiento de VCF para poder leer mas rapidos las DBs. Los he puesto aqui para que no sorprenda su existencia.

y despues la bajamos,

```
alias fget='wget -U "Mozilla/5.0 (X11; Linux x86_64; rv:10.0.11)
Gecko/20121121 Firefox/10.0.11"';
for x in `cat files2download.txt`; do fget
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/$x ; done
```

Convirtiendo con vcf_to_ped_converter.pl

Una vez que tengamos los archivos vamos a buscar los conversores necesarios. La descripcion de como se convierte puede verse en http://www.1000genomes.org/faq/can-i-convert-vcf-files-plinkped-format

Hay dos scripts que son necesarios *vcf_to_ped_converter.pl* y *tabix*. El primero es un script de perl cuya ultima version puede obtenerse de, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/browser/vcf_to_ped_converter/

El segundo es un paquete que forma parte de las SAM Tools y que puede bajarse de http://sourceforge.net/projects/samtools/files/tabix/. La pagina de manual esa en http://samtools.sourceforge.net/tabix.shtml. Una vez bajado y **compilado** se debe configurar el *PATH* del usuario para que se tenga acceso a el. Todo esto está hecho en **detritus** y lo que faltaria por usuario es añadir el *PATH* al archivo *~/.bash_profile*. Algo así,

```
PATH=$PATH:/opt/tabix
```

Cuando se intenta convertir los archivos que se han bajado da un error,

```
$ /opt/gntics/vcf_to_ped_convert.pl -vcf
/media/1000Genome/ALL.chr1.phase1_release_v3.20101123.snps_indels_svs.genoty
pes.vcf.gz -sample_panel_file
/media/1000Genome/phase1_integrated_calls.20101123.ALL.panel -population CEU
-region 1:10583-249239465
[tabix] the index file either does not exist or is older than the vcf file.
Please reindex.
tabix exited with status 1 at /opt/gntics/vcf_to_ped_convert.pl line 198.
```

Lo que esta diciendo es que hay que reindexar el archivo. Esto se hace con,

```
tabix -p vcf
/media/1000Genome/ALL.chr1.phase1_release_v3.20101123.snps_indels_svs.genoty
pes.vcf.gz
```

Despues de esto ya es posible convertir el archivo a formato *ped*,

```
$ /opt/gntics/vcf_to_ped_convert.pl -vcf
/media/1000Genome/ALL.chr1.phase1_release_v3.20101123.snps_indels_svs.genoty
pes.vcf.gz -sample_panel_file
/media/1000Genome/phase1_integrated_calls.20101123.ALL.panel -population CEU
-population TSI -population IBS -region 1:10583-249239465
Created 1_10583-249239465.info and 1_10583-249239465.ped
```

Aqui ya casi hemos terminado. Falta sólo convertir el archivo *.info* a un archivo *.map*. Esto se hace,

```
awk {'print "1 "$1" 0 "$2'} 1_10583-249239465.info > 1_10583-249239465.map
```

batching

Reindex:

```
for x in `cat files2download.txt | grep -v tbi`; do tabix -f -p vcf $x; done
```

Convert:

```
for x in `cat files2download.txt | grep -v tbi`;
do
chr=$(echo $x | sed 's/.*chr\(.*\)\.phase.*/\1/');
p0=$(zcat $x | grep "^$chr" | head -n 1 | awk {'print $2'});
pf=$(zcat $x | tail -n 1 | awk {'print $2'});
/opt/gntics/vcf_to_ped_convert.pl -vcf $x -sample_panel_file
phase1_integrated_calls.20101123.ALL.panel -population CEU -population TSI -
population IBS -region $chr:$p0-$pf;
done;
```

info2map:

```
for x in *.info;
do
chr=$(echo $x | awk -F"_" {'print $1'});
awk -v chr="$chr" {'print chr" "$1" 0 "$2'} $x > ${x%.info}.map;
done;
```

Convirtiendo con VCFtools

VCFtools es una herramienta para manipular los archivos VCF. Entre otras cosas permite convertir a otros formatos tal y como se indica en la documentacion. Tras bajar y **compilar** este paquete ha de añadirse al *PATH* del usuario en *~/.bash_profile*,

```
PATH=$PATH:/opt/vcftools_0.1.10/bin
```

Despues ejecutamos la conversión a *tped*, (la conversion a *ped* da problemas)

```
vcftools --gzvcf
ALL.chr9.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz --
plink-tped --out chr9
```

y hacemos la traspuesta al mismo tiempo que el binario

```
plink --tped chr9.tped --tfam chr9.tfam --alleleACGT --make-bed --out
```

```
chr9_ok
```

Ventajas

- Puede convertirse **todo** el archivo vcf, no hay que especificar la poblacion.
- Tampoco hay que especificar las regiones a convertir.

Desventajas

- No hay una forma clara de especificar la poblacion *target* por lo que en principio hay que procesar todo junto (La cosa va a tardar bastante mas)

batching

Para correr esto por todos los archivos,

```
for x in `cat files2download.txt | grep -v tbi`;
do
chr=$(echo $x | sed 's/.*chr\(.*\)\.phase.*/\1/');
vcftools --gzvcf $x --plink-tped --out chr${chr}_tmp;
plink --tped chr${chr}_tmp.tped --tfam chr${chr}_tmp.tfam --alleleACGT --
make-bed --out all_chr${chr};
rm -rf chr${chr}_tmp.*;
done;
```

Preprocesando con plink

Primero hay que recodificar los alelos a ACGT y de paso los pasamos a binario (solo es necesario si hemos converitdo con el primer metodo)

Cuidado: Hay que garantizar que el orden de los archivos sea el correcto para que el archivo resultante empiece por el cromosoma 1

```
for x in *.ped;
do
plink --file ${x%.ped} --out ${x%.ped} --alleleACGT --make-bed;
done;
```

Vamos a intentar juntarlos todos

```
afr=(`ls *.bed`);
for (( i=1; i<${#afr[@]}; i++ ));
do
x=${afr[$i]};
echo "${x} ${x%.bed}.bim ${x%.bed}.fam" >> allfiles.txt;
done;
x=${afr[0]};
plink --bfile ${x%.bed} --merge-list allfiles.txt --make-bed --out
```

```
1000genome_CEU_merged
```

y ya ta

Ahora hay que seguir fundiendo la DB que tenemos con nuestros datos: para poder imputar

porqueria que puede pasar

DUPLICATE MARKERS FOUND

```
$ vcftools --gzvcf
/media/1000Genome/ALL.chr22.phase1_release_v3.20101123.snps_indels_svs.genot
ypes.vcf.gz --plink-tped --out chr22_tmp && plink --tped chr22_tmp.tped --
tfam chr22_tmp.tfam --alleleACGT --make-bed --out all_chr22 && rm -rf
chr22_tmp.*

VCFtools - v0.1.10
(C) Adam Auton 2009

Parameters as interpreted:
    --gzvcf
/media/1000Genome/ALL.chr22.phase1_release_v3.20101123.snps_indels_svs.genot
ypes.vcf.gz
    --out chr22_tmp
    --plink-tped

Using zlib version: 1.2.3
Versions of zlib >= 1.2.4 will be *much* faster when reading zipped VCF
files.
Reading Index file.
File contains 494328 entries and 1092 individuals.
Applying Required Filters.
After filtering, kept 1092 out of 1092 Individuals
After filtering, kept 494328 out of a possible 494328 Sites
Writing PLINK TPED file ... Writing PLINK TFAM file ... Done.
Run Time = 1231.00 seconds

@----------------------------------------------------------@
|        PLINK!       |     v1.07      |   10/Aug/2009     |
|----------------------------------------------------------|
|  (C) 2009 Shaun Purcell, GNU General Public License, v2  |
|----------------------------------------------------------|
|  For documentation, citation & bug-report instructions:  |
|        http://pngu.mgh.harvard.edu/purcell/plink/        |
@----------------------------------------------------------@

Web-based version check ( --noweb to skip )
```

```
Connecting to web...  OK, v1.07 is current

Writing this text to log file [ all_chr22.log ]
Analysis started: Thu May  2 09:37:16 2013

Options in effect:
    --tped chr22_tmp.tped
    --tfam chr22_tmp.tfam
    --alleleACGT
    --make-bed
    --out all_chr22

Reading pedigree information from [ chr22_tmp.tfam ]
1092 individuals read from [ chr22_tmp.tfam ]
0 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
0 cases, 0 controls and 1092 missing
0 males, 0 females, and 1092 of unspecified sex
Warning, found 1092 individuals with ambiguous sex codes
Writing list of these individuals to [ all_chr22.nosex ]
494328 (of 494328) markers to be included from [ chr22_tmp.tped ]
 *** WARNING *** DUPLICATE MARKERS FOUND ***
Duplicate marker name found: [ rs11457237 ]
Duplicate marker name found: [ rs113940759 ]
Duplicate marker name found: [ rs71904485 ]

Before frequency and genotyping pruning, there are 494328 SNPs
1092 founders and 0 non-founders found
Total genotyping rate in remaining individuals is 1
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 494328 SNPs
After filtering, 0 cases, 0 controls and 1092 missing
After filtering, 0 males, 0 females, and 1092 of unspecified sex
Writing pedigree information to [ all_chr22.fam ]
Writing map (extended format) information to [ all_chr22.bim ]
Writing genotype bitfile to [ all_chr22.bed ]
Using (default) SNP-major mode

Analysis finished: Thu May  2 09:41:48 2013
```

~~asi que voy y lo convierto a ascii. <code bash> $ plink --bfile all_chr22 --recode --out all_chr22 </code>~~

Edito a mano los marcadores dobles (**En el mismo .bim**) y pongo una *b* al segundo. Ejemplo:

```
22  rs11457237    0    34030843
22  rs11457237b    0    34030846
```

~~Ahora deja ver si son el mismo, <code> $ plink --file all_chr22 --twolocus rs11457237a rs11457237b~~

~~-allow-no-sex </code>~~

Bueno esta parte no la entendi asi que voy a borrar uno de los duplicados utilizando el criterio loreal (*Because I'm worth it*).

```
$ cat rmsnps.txt
rs11457237b
rs113940759b
rs71904485b
```

```
$ plink --bfile all_chr10 --merge-list allfiles.txt --make-bed --out
1000genome_all_merged --exclude rmsnps.txt --allow-no-sex

@----------------------------------------------------------@
|        PLINK!       |     v1.07      |   10/Aug/2009     |
|----------------------------------------------------------|
|  (C) 2009 Shaun Purcell, GNU General Public License, v2  |
|----------------------------------------------------------|
|  For documentation, citation & bug-report instructions:  |
|        http://pngu.mgh.harvard.edu/purcell/plink/        |
@----------------------------------------------------------@

Web-based version check ( --noweb to skip )
Recent cached web-check found... OK, v1.07 is current

Writing this text to log file [ 1000genome_all_merged.log ]
Analysis started: Fri May  3 12:56:36 2013

Options in effect:
    --bfile all_chr10
    --merge-list allfiles.txt
    --make-bed
    --out 1000genome_all_merged
    --exclude rmsnps.txt
    --allow-no-sex

Reading map (extended format) from [ all_chr10.bim ]
1882663 markers to be included from [ all_chr10.bim ]
Reading pedigree information from [ all_chr10.fam ]
1092 individuals read from [ all_chr10.fam ]
0 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
0 cases, 0 controls and 1092 missing
0 males, 0 females, and 1092 of unspecified sex
Warning, found 1092 individuals with ambiguous sex codes
Writing list of these individuals to [ 1000genome_all_merged.nosex ]
Reading genotype bitfile from [ all_chr10.bed ]
Detected that binary PED file is v1.00 SNP-major mode
Using merge mode 1 : consensus call (default)
Detected that binary PED file is v1.00 SNP-major mode
```

```
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode
Detected that binary PED file is v1.00 SNP-major mode

Merging 22 samples, final sample contains 1092 individuals and 39706715
markers
Reading list of SNPs to exclude [ rmsnps.txt ] ... 3 read
Before frequency and genotyping pruning, there are 39706712 SNPs
1092 founders and 0 non-founders found
Total genotyping rate in remaining individuals is 1
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 39706712 SNPs
After filtering, 0 cases, 0 controls and 1092 missing
After filtering, 0 males, 0 females, and 1092 of unspecified sex
Writing pedigree information to [ 1000genome_all_merged.fam ]
Writing map (extended format) information to [ 1000genome_all_merged.bim ]
Writing genotype bitfile to [ 1000genome_all_merged.bed ]
Using (default) SNP-major mode

Analysis finished: Fri May  3 17:11:00 2013
```

From:
http://detritus.fundacioace.com/wiki/ - **Detritus Wiki**

Permanent link:
http://detritus.fundacioace.com/wiki/doku.php?id=genetica:pre_1kg&rev=1367834176

Last update: **2020/08/04 10:48**