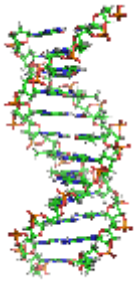


Imputando a 1000 Genome con plink

Preparando la DB

Espero que ya hayas pasado por esto: 1000 Genome a .ped. En caso contrario no vas a hacer mucho aqui.

I got fresh meat

El primer paso es garantizar que los marcadores tengan la misma posicion en nuestra DB y la de 1000Genome. Para ello voy a tener que convertir nuestra DB a *ped* y *map*.

```
$ plink --bfile admurcia --recode --out admurcia
```

y ahora ya puedo hacer un parser en el *bim* de 1000 Genome que cambie la posicion en el *map* nuestro.

refresh_map.pl

```
#!/usr/bin/perl

# Copyright 2013 O. Sotolongo <osotolongo@fundacioace.com>

use strict; use warnings;
use Parallel::ForkManager;
use File::Slurp qw(read_file);
use File::Remove 'remove';
use File::Basename qw(basename);
use Data::Dump qw(dump);

my $debug = 1;

#my $library =
'/home/data/GNlib/1000Genome/CEU_all/1000genome_CEU_merged.map';
my $library =
'/home/osotolongo/data/GNlib/1000Genome/EUR/1000genome_eur.bim';
```

```
my $db = shift;
die "Must supply database\n" unless $db;
my $odb = $db;
$odb =~ s/(.*)\.(.*)/$1_freshed\.$2/;

print "I will try to write results to $odb\n";

my %genmap;
my %data_line = reverse map {/^(.*\s+(rs\d+)\s+.*)$/} grep
{/^(.*\s+(rs\d+)\s+.*)$/} read_file $db;
print "Data stored, lets try to understand it\n";
foreach my $marker (sort keys %data_line){
    (@{$genmap{$marker}} {qw/chr data position/}) = $data_line{$marker}
=~ /^(\d+)\s+rs\d+\s+(.*)\s+(\d+)$/;
}

print "Data read, lets work now\n";
open(IN, "<$library") || die "can't open $library";
while(<IN>){
    if (my ($libmarker, $libpos) = /^\d+\s+(rs\d+)\s+.*\s+(\d+)\s+.*/){
        if(exists $genmap{$libmarker}) {
            $genmap{$libmarker}{position} = $libpos;
        }
    }
}
close IN;

print "Writing to $odb\n";
open ODF, ">$odb" || die "Could not open output file";
foreach my $marker (sort {($genmap{$a}->{chr} <=> $genmap{$b}->{chr})
or ($genmap{$a}->{position} <=> $genmap{$b}->{position})} keys
%genmap){
    print ODF
"$genmap{$marker}{chr}\t$marker\t$genmap{$marker}{data}\t$genmap{$marke
r}{position}\n";
}
close ODF;
```

```
$ refresh_map.pl admurcia.map
I will try to write results to admurcia_freshed.map
Data stored, lets try to understand it
Data read, lets work now
Writing to admurcia_freshed.map

$ mv admurcia.map admurcia_old.map
$ mv admurcia_freshed.map admurcia.map
```

y ahora lo llevo a binario de nuevo. Esto no deberia demorar demasiado.

```
$ plink --file admurcia --make-bed --out admurcia_freshed --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 )
Connecting to web...  OK, v1.07 is current

Writing this text to log file [ admurcia_freshed.log ]
Analysis started: Tue May  7 09:27:33 2013

Options in effect:
    --file admurcia
    --make-bed
    --out admurcia_freshed
    --allow-no-sex

198429 (of 198429) markers to be included from [ admurcia.map ]
1088 individuals read from [ admurcia.ped ]
1088 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
319 cases, 769 controls and 0 missing
511 males, 577 females, and 0 of unspecified sex
Before frequency and genotyping pruning, there are 198429 SNPs
1088 founders and 0 non-founders found
1130 heterozygous haploid genotypes; set to missing
Writing list of heterozygous haploid genotypes to [ admurcia_freshed.hh ]
Total genotyping rate in remaining individuals is 0.993951
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 198429 SNPs
After filtering, 319 cases, 769 controls and 0 missing
After filtering, 511 males, 577 females, and 0 of unspecified sex
Writing pedigree information to [ admurcia_freshed.fam ]
Writing map (extended format) information to [ admurcia_freshed.bim ]
Writing genotype bitfile to [ admurcia_freshed.bed ]
Using (default) SNP-major mode

Analysis finished: Tue May  7 09:29:11 2013
```

Merging

De aqui ya se puede intentar hacer un *merge*.

```
$ plink --bfile ~/data/GNlib/1000Genome/EUR/1000genome_eur --bmerge
admurcia_freshed.bed admurcia_freshed.bim admurcia_freshed.fam --make-bed --
out admurcia_merged
```

y nos dice que algunos no andan bien

```
Found 17054 SNPs that do not match in terms of allele codes
Might include strand flips, although flipped A/T and C/G SNPs will be
undetected)
Writing problem SNPs to [ admurcia_merged.missnp ]

ERROR: Stopping due to mis-matching SNPs -- check +/- strand?
```

Asi que ahi vamos, a hacer un *flip*

```
$ plink --bfile admurcia_freshed --flip admurcia_merged.missnp --make-bed --
out admurcia_fandf
```

y a ver que tal ha quedado,

```
$ plink --bfile ~/data/GNlib/1000Genome/EUR/1000genome_eur --bmerge
admurcia_fandf.bed admurcia_fandf.bim admurcia_fandf.fam --make-bed --out
admurcia_merged2
```

pos no ha mejorado mucho,

```
Found 11404 SNPs that do not match in terms of allele codes
Might include strand flips, although flipped A/T and C/G SNPs will be
undetected)
Writing problem SNPs to [ admurcia_merged2.missnp ]

ERROR: Stopping due to mis-matching SNPs -- check +/- strand?
```

Lo que voy a hacer es borrar estos ultimos

```
$ plink --bfile admurcia_fandf --exclude admurcia_merged2.missnp --make-bed
--out admurcia_ffex

@----------------------------------------------------------@
|        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 [ admurcia_ffex.log ]
Analysis started: Tue May  7 10:06:59 2013

Options in effect:
    --bfile admurcia_fandf
    --exclude admurcia_merged2.missnp
    --make-bed
    --out admurcia_ffex

Reading map (extended format) from [ admurcia_fandf.bim ]
198429 markers to be included from [ admurcia_fandf.bim ]
Reading pedigree information from [ admurcia_fandf.fam ]
1088 individuals read from [ admurcia_fandf.fam ]
1088 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
319 cases, 769 controls and 0 missing
511 males, 577 females, and 0 of unspecified sex
Reading genotype bitfile from [ admurcia_fandf.bed ]
Detected that binary PED file is v1.00 SNP-major mode
Reading list of SNPs to exclude [ admurcia_merged2.missnp ] ... 11404 read
Before frequency and genotyping pruning, there are 187025 SNPs
1088 founders and 0 non-founders found
1117 heterozygous haploid genotypes; set to missing
Writing list of heterozygous haploid genotypes to [ admurcia_ffex.hh ]
Total genotyping rate in remaining individuals is 0.99397
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 187025 SNPs
After filtering, 319 cases, 769 controls and 0 missing
After filtering, 511 males, 577 females, and 0 of unspecified sex
Writing pedigree information to [ admurcia_ffex.fam ]
Writing map (extended format) information to [ admurcia_ffex.bim ]
Writing genotype bitfile to [ admurcia_ffex.bed ]
Using (default) SNP-major mode

Analysis finished: Tue May  7 10:07:07 2013
```

y volver a intentarlo

```
$ plink --bfile ~/data/GNlib/1000Genome/EUR/1000genome_eur --bmerge
admurcia_ffex.bed admurcia_ffex.bim admurcia_ffex.fam --make-bed --out
admurcia_merged3
```

Ahora si que funciona,

```
@----------------------------------------------------------@
|        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 [ admurcia_merged3.log ]
Analysis started: Tue May  7 10:08:41 2013

Options in effect:
        --bfile /home/osotolongo/data/GNlib/1000Genome/EUR/1000genome_eur
        --bmerge admurcia_ffex.bed admurcia_ffex.bim admurcia_ffex.fam
        --make-bed
        --out admurcia_merged3

Reading map (extended format) from [
/home/osotolongo/data/GNlib/1000Genome/EUR/1000genome_eur.bim ]
39706712 markers to be included from [
/home/osotolongo/data/GNlib/1000Genome/EUR/1000genome_eur.bim ]
Reading pedigree information from [
/home/osotolongo/data/GNlib/1000Genome/EUR/1000genome_eur.fam ]
379 individuals read from [
/home/osotolongo/data/GNlib/1000Genome/EUR/1000genome_eur.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 379 missing
0 males, 0 females, and 379 of unspecified sex
Warning, found 379 individuals with ambiguous sex codes
Writing list of these individuals to [ admurcia_merged3.nosex ]
Reading genotype bitfile from [
/home/osotolongo/data/GNlib/1000Genome/EUR/1000genome_eur.bed ]
Detected that binary PED file is v1.00 SNP-major mode
Using merge mode 1 : consensus call (default)

187025 markers to be merged from [ admurcia_ffex.bim ]
Of these, 1218 are new, 185807 already exist in current data
1088 individuals merged from [ admurcia_ffex.fam ]
Of these, 1088 were new, 0 were already in current data

Detected that binary PED file is v1.00 SNP-major mode
1088 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
319 cases and 769 controls
Before frequency and genotyping pruning, there are 39707930 SNPs
1467 founders and 0 non-founders found
1117 heterozygous haploid genotypes; set to missing
```

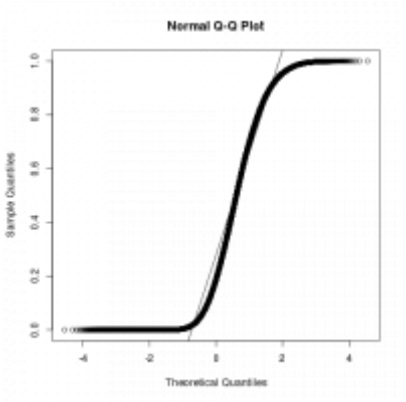
```
Writing list of heterozygous haploid genotypes to [ admurcia_merged3.hh ]
Total genotyping rate in remaining individuals is 0.261815
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 39707930 SNPs
After filtering, 319 cases, 769 controls and 379 missing
After filtering, 511 males, 577 females, and 379 of unspecified sex
Writing pedigree information to [ admurcia_merged3.fam ]
Writing map (extended format) information to [ admurcia_merged3.bim ]
Writing genotype bitfile to [ admurcia_merged3.bed ]
Using (default) SNP-major mode

Analysis finished: Tue May  7 13:52:56 2013
```

Assoc y seguimos cambiando

Pero no he acabado todavía. Necesito hacer un estudio de asociacion entre mi DB y la de 1000genome para detectar los *flips* que plink no ha sido capaz de encontrar.

```
$ sed 's/^\([^ ]* [^ ]* [^ ]* [^ ]* [^ ]*\) 1/\1 2/' admurcia_merged3.fam |
sed 's/^\([^ ]* [^ ]* [^ ]* [^ ]* [^ ]*\) -9/\1 1/' >
admurcia_merged3_case_control.fam
$ mv admurcia_merged3.fam admurcia_merged3_real.fam
$ mv admurcia_merged3_case_control.fam admurcia_merged3.fam
$ plink --bfile admurcia_merged3 --assoc --out flipthis --allow-no-sex
```

Esto nos da como resultado una lista de marcadores que tienen diferentes frecuencia alelica entre nuestra DB y la de 1000Genome. No tiene porque estar mal pero claramente hay algunos que son muy distintos.



El qq plot nos muestra que hay una serie de marcadores que se desvian de lo esperado (para $p \approx 0$). Esto es basicamente lo que hay que arreglar en la medida de lo posible.

Procesando la salida del assoc

Escogemos un punto de corte ($\chi^2 > 100$) y para estos marcadores seguimos dos

procedimientos diferentes. Todos los que sean **AT** o **CG** los ponemos aparte y hacemos un flip en la ultima DB buena. El resto hay que cambiarlos en el *.bim*

Primero los ordenamos y separamos,

```
$ awk {'if($9!="NA") print'} flipthis.assoc > flipthis_nona.txt
$ sort -k9,9 -g flipthis_nona.txt > flipthis_nona_sorted.txt
$ awk {'if(($4=="T" && $7=="A")||($4=="C" && $7=="G") || ($4=="A" &&
$7=="T")||($4=="G" && $7=="C")) print'} flipthis_nona_sorted.txt | awk
{'if($8>100) print'} > flipthis_nona_sorted_toflip.txt
$ awk {'if(!(($4=="T" && $7=="A")||($4=="C" && $7=="G") || ($4=="A" &&
$7=="T")||($4=="G" && $7=="C"))) print'} flipthis_nona_sorted.txt | awk
{'if($8>100) print'} > flipthis_nona_sorted_tochangeinDB.txt
```

Ahora vamos a tomar la primera lista y hacer *flip* en la DB,

```
$ awk {'print $2'} flipthis_nona_sorted_toflip.txt > secondflip.txt
$ plink --bfile admurcia_ffex --flip secondflip.txt --make-bed --out
admurcia_ffex_flip2
```

Lo segundo que necesitamos hacer es tomar este nuevo *.bim* y cambiar **A1** por **A2** para los resultados de la segunda lista. Exactamente asi,

flip_bim.pl

```
#!/usr/bin/perl

# Copyright 2013 O. Sotolongo <osotolongo@fundacioace.com>

use strict; use warnings;
use File::Slurp qw(read_file);
use Data::Dump qw(dump);

my $db = shift;
my $assoc = shift;
die "Must supply database\n" unless $db;
my $odb = $db;
$odb =~ s/(.*)\.(.*)/$1_bflipped\.$2/;

print "I will try to write results to $odb\n";

#first I will load .bim
print "Reading data from $db ...\n";
my %genmap;
my %data_line = reverse map {/^(.*\s+(rs\d+)\s+.*)$/} grep
{/^(.*\s+(rs\d+)\s+.*)$/} read_file $db;
print "Data stored, lets try to understand it\n";
foreach my $marker (sort keys %data_line){
    (@{$genmap{$marker}} {qw/chr gposition fposition allele1 allele2/})
= $data_line{$marker} =~
/^(\d+)\s+rs\d+\s+(.*)\s+(\d+)\s+([A,T,C,G])\s+([A,T,C,G])\s*$/;
```

```
}

#now I will read target markers from the assoc file
print "OK, let see what markers I have to change, reading $assoc
...\n";
my @targets = map {/^.*\s+(rs\d+)\s+.*$/} grep
{/^(.*\s+(rs\d+)\s+.*)$/} read_file $assoc;

# and then flip the allele information of targets in db info
print "I'm going to flips those markers now\n";
foreach my $marker (@targets){
    my $tallele = $genmap{$marker}{allele1};
    $genmap{$marker}{allele1} = $genmap{$marker}{allele2};
    $genmap{$marker}{allele2} = $tallele;
}

# go and write .bim now
print "Writing to $odb\n";
open ODF, ">$odb" || die "Could not open output file";
foreach my $marker (sort {($genmap{$a}->{chr} <=> $genmap{$b}->{chr})
or ($genmap{$a}->{fposition} <=> $genmap{$b}->{fposition})} keys
%genmap){
    print ODF
"$genmap{$marker}{chr}\t$marker\t$genmap{$marker}{gposition}\t$genmap{$
marker}{fposition}\t$genmap{$marker}{allele1}\t$genmap{$marker}{allele2
}\n";
}
close ODF;
```

Entonces viramos esos marcadores,

```
$ flip_bim.pl admurcia_ffex_flip2.bim flipthis_nona_sorted_tochangeinDB.txt
```

Merge again :-P

y volvemos a hacer el merge,

```
$ plink --bfile ~/data/GNlib/1000Genome/EUR/1000genome_eur --bmerge
admurcia_ffex_flip2.bed admurcia_ffex_flip2_bflipped.bim
admurcia_ffex_flip2.fam --make-bed --out admurcia_merged4
```

Repetimos, el estudio de asociacion a ver que pasa,

```
$ sed 's/^\([^ ]* [^ ]* [^ ]* [^ ]* [^ ]*\) 1/\1 2/' admurcia_merged4.fam |
sed 's/^\([^ ]* [^ ]* [^ ]* [^ ]* [^ ]*\) -9/\1 1/' >
admurcia_merged4_case_control.
$ mv admurcia_merged4.fam admurcia_merged4_real.fam
$ mv admurcia_merged4_case_control.fam admurcia_merged4.fam
```

```
$ plink --bfile admurcia_merged4 --assoc --out flipthis4 --allow-no-sex
$ awk {'if($9!="NA") print'} flipthis4.assoc > flipthis4_nona.txt
$ sort -k9,9 -g flipthis4_nona.txt > flipthis4_nona_sorted.txt
```

y en este ultimo archivo nos quedan los marcadores resultantes.

Analizando el segundo Assoc

Lo que se observa tras este segundo estudio de asociacion es que los marcadores intercambiados tienen distinto valor de *p-value*. Aquellos que se han intercambiado correctamente tienen un *p-value* mayor (o un menor χ^2) pero los que eran correctos y se han intercambiado incorrectamente muestran ahora un menor *p-value* (o un χ^2 mayor).

Lo que debe hacerse para corregir la situacion es volver a intercambiar los marcadores, pero solo aquellos que hayan disminuido su *p-value* (o aumentado el χ^2). Para ello tomamos el resultado del primer estudio de asociacion y lo comparamos con este ultimo.

Primero limpiamos todo,

```
$ awk {'if($9!="NA") print'} flipthis4.assoc | sort -k9,9 -g >
flipthis4_nona.txt
$ awk {'if(($4=="T" && $7=="A")||($4=="C" && $7=="G") || ($4=="A" &&
$7=="T")||($4=="G" && $7=="C")) print'} flipthis4_nona.txt | awk
{'if($8>100) print'} > flipthis4_toflip.txt
$ awk {'if(!(($4=="T" && $7=="A")||($4=="C" && $7=="G") || ($4=="A" &&
$7=="T")||($4=="G" && $7=="C"))) print'} flipthis4_nona.txt | awk
{'if($8>100) print'} > flipthis4_tochangeinDB.txt
```

Ahora, siguiendo este razonamiento, tenemos que comparar dos pares de archivos diferentes. Primero los marcadores a los que hay que hacer un *flip* con plink, comparando *flipthis_nona_sorted_toflip.txt* y *flipthis4_toflip.txt*. el resultado ha de guardarse en un archivo aparte. Luego hay que comparar los archivos *flipthis_nona_sorted_tochangeinDB.txt* y *flipthis4_tochangeinDB.txt* para decidir cuales hay que volver a cambiar en el archivo **.bim**. Ahi va un sencillo script que debe hacer esto,

compare_merges.pl

```
#!/usr/bin/perl

# Copyright 2013 O. Sotolongo <osotolongo@fundacioace.com>

# This program is free software; you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation; either version 2 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.
```

```
#
use strict; use warnings;
use File::Slurp qw(read_file);
use Data::Dump qw(dump);

my $flip = 0;

while (@ARGV and $ARGV[0] =~ /^-/) {
    $_ = shift;
    last if /^--$/;
    if (/^-flip/) { $flip = 1;}
}
my $db;
my $odb;
unless ($flip) {
    $db = shift;
    die "Must supply database\n" unless $db;
    $odb = $db;
    $odb =~ s/(.*)\.(.*)/$1_bcomp_flipped\.$2/;
}else{
    $odb = 'just_flip_these.txt';
}
my $assoc_previous = shift;
my $assoc_last = shift;

print "I will try to write results to $odb\n";

my %genmap_previous;
my %genmap_last;
my %genmap;
my %data_line;
#first I store the files
unless($flip){
    %data_line = reverse map {/^(.*\s+(rs\d+)\s+.*)$/} grep
{/^(.*\s+(rs\d+)\s+.*)$/} read_file $db;
    foreach my $marker (sort keys %data_line){
        (@{$genmap{$marker}} {qw/chr gposition fposition allele1
allele2/}) = $data_line{$marker} =~
/^(\d+)\s+rs\d+\s+(.*)\s+(\d+)\s+([A,T,C,G])\s+([A,T,C,G])\s*$/;
    }
}

#now I will read target markers from the assoc files
print "OK, let see what markers I have to change, reading
$assoc_previous and $assoc_last ...\n";
%data_line = reverse map {/^(.*\s+(rs\d+)\s+.*)$/} grep
{/^(.*\s+(rs\d+)\s+.*)$/} read_file $assoc_previous;
foreach my $marker (sort keys %data_line){
    (@{$genmap_previous{$marker}} {qw/chr position allele1 freq1 freq2
allele2 chi2 pvalue oddratio/}) = $data_line{$marker} =~
/^\s*(\d+)\s+rs\d+\s+(\d+)\s+([A,T,C,G]+)\s+(\d+\.*\d*e*-
```

```
*\d*)\s+(\d+\.*\d*e*-*\d*)\s+([A,T,C,G]+)\s+(\d+\.*\d*e*-
*\d*)\s+(\d+\.*\d*e*-*\d*)\s+(\d+\.*\d*e*-*\d*|NA)\s*$/;
}

%data_line = reverse map {/^(.*\s+(rs\d+)\s+.*)$/} grep
{/^(.*\s+(rs\d+)\s+.*)$/} read_file $assoc_last;
foreach my $marker (sort keys %data_line){
    (@{$genmap_last{$marker}} {qw/chr position allele1 freq1 freq2
allele2 chi2 pvalue oddratio/}) = $data_line{$marker} =~
/^\s*(\d+)\s+rs\d+\s+(\d+)\s+([A,T,C,G]+)\s+(\d+\.*\d*e*-
*\d*)\s+(\d+\.*\d*e*-*\d*)\s+([A,T,C,G]+)\s+(\d+\.*\d*e*-
*\d*)\s+(\d+\.*\d*e*-*\d*)\s+(\d+\.*\d*e*-*\d*|NA)\s*$/;
}

#dump %genmap_last; exit;
my @outmarkers;
# and then flip the allele information of targets in db info
if($flip){
    foreach my $marker (keys %genmap_previous){
        if(exists($genmap_previous{$marker}) &&
exists($genmap_last{$marker})){
            if($genmap_previous{$marker}{chi2} <
$genmap_last{$marker}{chi2}){
                push @outmarkers, $marker;
            }
        }
    }
}else{
    print "I'm going to flips those markers now\n";
    foreach my $marker (keys %genmap_previous){
        if(exists($genmap_previous{$marker}) &&
exists($genmap_last{$marker})){
            if($genmap_previous{$marker}{chi2} <
$genmap_last{$marker}{chi2}){
                my $tallele = $genmap{$marker}{allele1};
                $genmap{$marker}{allele1} = $genmap{$marker}{allele2};
                $genmap{$marker}{allele2} = $tallele;
            }
        }
    }
}

# go and write .bim or markers list now
print "Writing to $odb\n";
open ODF, ">$odb" || die "Could not open output file";

if($flip){
    foreach my $marker (@outmarkers){
        print ODF "$marker\n";
    }
}else{
```

```
    foreach my $marker (sort {($genmap{$a}->{chr} <=>
$genmap{$b}->{chr}) or ($genmap{$a}->{fposition} <=>
$genmap{$b}->{fposition})} keys %genmap){
        print ODF
"$genmap{$marker}{chr}\t$marker\t$genmap{$marker}{gposition}\t$genmap{$
marker}{fposition}\t$genmap{$marker}{allele1}\t$genmap{$marker}{allele2
}\n";
    }
}
close ODF;
```

Primero lo corremos para sacar los marcadores a reintercambiar con plink,

```
$ compare_merges.pl -flip flipthis_nona_sorted_toflip.txt
flipthis4_toflip.txt
$ mv admurcia_ffex_flip2.bim admurcia_ffex_flip2_old.bim
$ mv admurcia_ffex_flip2_bflipped.bim admurcia_ffex_flip2.bim
$ plink --bfile admurcia_ffex_flip2 --flip just_flip_these.txt --make-bed --
out admurcia_flip3

@----------------------------------------------------------@
|        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 [ admurcia_flip3.log ]
Analysis started: Fri May 10 11:36:16 2013

Options in effect:
        --bfile admurcia_ffex_flip2
        --flip just_flip_these.txt
        --make-bed
        --out admurcia_flip3

Reading map (extended format) from [ admurcia_ffex_flip2.bim ]
187025 markers to be included from [ admurcia_ffex_flip2.bim ]
Reading pedigree information from [ admurcia_ffex_flip2.fam ]
1088 individuals read from [ admurcia_ffex_flip2.fam ]
1088 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
319 cases, 769 controls and 0 missing
511 males, 577 females, and 0 of unspecified sex
```

```
Reading genotype bitfile from [ admurcia_ffex_flip2.bed ]
Detected that binary PED file is v1.00 SNP-major mode
Reading SNPs to flip strand from [ just_flip_these.txt ]
Flipped strand of 56 SNPs
Before frequency and genotyping pruning, there are 187025 SNPs
1088 founders and 0 non-founders found
1117 heterozygous haploid genotypes; set to missing
Writing list of heterozygous haploid genotypes to [ admurcia_flip3.hh ]
Total genotyping rate in remaining individuals is 0.99397
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 187025 SNPs
After filtering, 319 cases, 769 controls and 0 missing
After filtering, 511 males, 577 females, and 0 of unspecified sex
Writing pedigree information to [ admurcia_flip3.fam ]
Writing map (extended format) information to [ admurcia_flip3.bim ]
Writing genotype bitfile to [ admurcia_flip3.bed ]
Using (default) SNP-major mode

Analysis finished: Fri May 10 11:36:25 2013
```

y ahora intercambiamos los marcadores en el **.bim**

```
$ compare_merges.pl admurcia_flip3.bim flipthis_nona_sorted_tochangeinDB.txt
flipthis4_tochangeinDB.txt
$ mv admurcia_flip3.bim admurcia_beforeflip4.bim
$ mv admurcia_flip3_bcomp_flipped.bim admurcia_flip3.bim
```

Merge (hasta cuando es esto?)

Y ahora hago el ultimo merge. Voy a suponer que ya el resultado va a estar bien o no termino nunca.

```
$ plink --bfile ~/data/GNlib/1000Genome/EUR/1000genome_eur --bmerge
admurcia_flip3.bed admurcia_flip3.bim admurcia_flip3.fam --make-bed --out
admurcia_merged5 --allow-no-sex
```

Imputando con plink

Primero dividimos la DB en chromosomas independientes,

```
$ for x in {1..22..1}; do plink --bfile admurcia_merged5 --chr $x --make-bed
--allow-no-sex --out admurcia_merged5_chr$x; done
```

La imputacion es **lenta**.

Ni se te ocurra hacer algo así,

```
$ parallel echo "plink --bfile admurcia_merged5_chr{} --proxy-impute all --
allow-no-sex --make-bed --out admurcia_merged5_imputed_chr{}" ::: {1..22..1}
```

Pero la idea básica es ir lanzando los chromosomas en grupos (4 o 5 a la vez). Para uno solo la orden es,

```
$ plink --bfile admurcia_merged5_chr22 --proxy-impute all --make-bed --
allow-no-sex --out admurcia_merged5_plink_imputed_chr22
```

Esto se puede hacer con algo asi,

plink_impute.pl

```
#!/usr/bin/perl

# Copyright 2013 O. Sotolongo <osotolongo@fundacioace.com>

use strict; use warnings;
use Parallel::ForkManager;
use GNTools qw(achtung);
use File::Basename qw(basename);
use constant CHROMOSOMES => 22;

my $max_processes = 4;
my $impute = '/usr/local/bin/plink --noweb --proxy-impute all --make-
bed --allow-no-sex';
my $db = shift;
die "Must supply database\n" unless $db;
my $pm = new Parallel::ForkManager($max_processes);
my $logfile = "/tmp/.parallel_impute_by_db_output.log";
my $ordfile = "/tmp/.parallel_impute_by_db_orders.log";
open ORDOUT, ">$ordfile" or die "Can't open file, $!";
open STDOUT, ">$logfile" or die "Can't redirect stdout";
open STDERR, ">&STDOUT" or die "Can't dup stdout";
my $bndb = basename($db);

for(my $chr = 1; $chr <= CHROMOSOMES; $chr++){
        my $order = $impute.' --bfile '.$db.'_chr'.$chr.' --out
'.$bndb.'_chr'.$chr.'_plink_imputed';
        $pm->start() and next;
        print ORDOUT "$order\n";
        system($order);
        $pm->finish;
}
$pm->wait_all_children;

close ORDOUT;
close STDERR;
```

```
close STDOUT;
achtung "plink impute", "La imputacion de plink ha terminado\n";
print "Have a nice day! ;-)\n";
```

Imputando con impute2

Para pasar a formato de *impute2* necesito los *.ped* y *.map*. Voy a convertir los chromosomas de uno en uno para no volverme loco. Ejemplo,

```
$ plink --bfile admurcia_merged5_chr22 --recode --allow-no-sex --out
admurcia_merged5_chr22
```

despues uso gtool para cambiar de formato,

```
$ gtool -P --ped admurcia_merged5_chr22.ped --map admurcia_merged5_chr22.map
--og admurcia_2imp2_chr22.gen --os admurcia_2imp2_chr22.sample
```

Lo primero que pasa es que me da un *segfault*,

```
gtool -P --ped admurcia_merged5_chr22.ped --map admurcia_merged5_chr22.map -
-og admurcia_2imp2_chr22.gen --os admurcia_2imp2_chr22.sample
MAP...
Number of SNPs: 494342
Count Samples...
Number of samples: 1467
memory...2175599142
PED...
Segmentation fault (core dumped)
```

No da más información así que aquí estamos jodidos. Felizmente existe FCgen que permite convertir la información en múltiples formatos. Así que lo intentamos,

```
$ fcgene --map admurcia_merged5_chr22.map --ped admurcia_merged5_chr22.ped -
-oformat impute --out admurcia_2imp2_chr22

*->The analysis has started at Fri May 24 14:32:35 2013

||============================================================================
==||
*-> Copyright:  GNU General Public License
*-> Program Developer:
         Nab Raj Roshyara
         email: roshyara@yahoo.com
         Universitaet Leipzig
         Leipzig Research Center for Civilization Diseases(LIFE)
         Institute for Medical Informatics, Statistics and Epidemiology
(IMISE)
```

```
         Group: Genetical Statistics and Systems Biology
         Group Leader: Dr. rer. nat. Markus Scholz
||============================================================================
==||

*->Used command options:

         --map admurcia_merged5_chr22.map
         --ped admurcia_merged5_chr22.ped
         --oformat impute
         --out admurcia_2imp2_chr22_a

*->Reading map file "admurcia_merged5_chr22.map":

*-> Reading  ped file: "admurcia_merged5_chr22.ped":
 EROOR: Problem in reading  SNP esv2659493 in line1. The SNP has genotype [
CTAATTTTTTG ] and
[CAGAAATAATGTGTAACCAGATATGTGTGCATCCAGTGGCCCAGACAAGTTGACATATAGAATTAACCATCACAA
GTCCACCCCTTGTCAACCTGGTACCTACACACATCTCCTTAAACCATACTTAACCTCCAAATAAACACAATAACAA
AGTCAGTGTTAGCAGTGGAAGGTGTCTGAGTTAGTGGCAGCGAATCCGTATGGGTCTGCAGCAACCTCAATTCTTG
CCTCCTCAGAAAAAAGAATTCAACTGGCCGGGCATGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCAAG
GCGGGCAGACCACGAGGTCAGGAGATCGAGACCATCCTGGCTAACACGGTGAAACCTCGTCTTTACTAAAAATACA
AAAATTAGCCGGGCATGGTGGTGGGTGCCTGTTGTCCCAGCTACATGAGAGACTGAGGCAGGAGAATGGCAGGAAC
CCAGGAGGCGGAGCTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCAAGACTCCGTCT
CAAAAAAAAAAAGAATTCAACTGAGGGGCATAAGGCAGAAAAAGAGACCGAGGCAAGTTTTGGAGGAGGAGTGAAA
GTTTATTTAAAAAAGCTTTACAACAGGAAAGAAAGTATTCTTGGAAGAGACCTAAACAGGCACATAAAGGTCAAGT
GCGGTGTTTAACCTTGATTCTAGGACTTTATAGACTGACCCCTTTCCCAAGATTCTTCCCCTAGGGTGGGCTGCCT
ACATGCACAGTGTCCTCCTTACCCTTGGGAGATGAGCACACGCAGTGTGTTTAGGAAACTGTACGCATGCCCATCT
GAAGATTTCTTCCGTCCCAGAAGGTCATACTCTGCCATTGTCTCTTAATGCACATGGCCGGGGAAATTGCTTCTCC
CTGGTGCCTGCATTCAATTAACACTTTAATGAAACAGGTGTGACCCATCAGGAACTGGCCTCTCCCTGATGCCAGC
TGCCAATTTATCACTTTCATTTTTAATTTATTTTTTAGAAGGAGTCTCACTCTGTCACCCAGACTGGAGAGCAGTG
GCATGATCTCGGCTCACTGTAACCTCTGCCTCCCAGGTTCAAGTGATTCTCCTGCCTCAGCCTCCCGAGTGGCTGG
GATTAGAGGCATGCACCAACATGCCTGGCTAATTTTTGTATTTTTAGTAGAGATGGGCTTTCACCATGTTGGCCAG
GCTGGTCTCGAACCCCTGATGTCAAGTGACCCACCCACCTTGACCTCACAAAGTGCTAGGATTACAGGCATGAGAC
ACCATGCCCAGCCCTGGCTTTTTTTTTTTTTTTTTTTTTTTTGAGACGGAGTCTCGCTCTGTGGCCCAGGCTGTA
GTGCAGTGGCATGATCTCGCCTCGCTGCAAGCTCCGCCTCCTGGGTTCACGCCATTCTCCTGCCTCAGCCTCCAGA
GCAGCTGGGACCACAGGCGCCCACCACCACGCCCAGCTAATTTTTTGTATTTTTAGTAGAGACGGGATTTCACTGT
GTTAGCCAGGATGGTCTCGATCTGACCTCATGATCCGCCCGCCTTGGCCTCCCAAAGTGCTGGGATTACAGGCGTG
AGCCACTGTGCCCAGCCGCTAATTTTTTATTTTTTATTGTTTGTAGAGAGGGGGTCTCCCTATGTTGCCTAGGCTG
TTTTGTTTTTTTTTTTAAACAGAGACAGGGTCTTGCTATGTTGCCCAGGCTGGTCTTGAACTCCTGGGCTCAAGCA
GTCCTCCCATCTTGGTTTTCCAAAGTGCTAGGATTACAGGCATGAACCACCAGGCCCGGTCTCCACTGTGAATCTT
TTTTTTTTTTTTTTGAGACTGAGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGCACCATCTCGGCTCACCGC
AACCTCCGCCTCCTGGGTTCAAGCAATTCTGCTGCCTCAGCCTCCTGAGTAGCTCAGATTACAGGCATGCACCACT
ACGCCCAGCTAATTTTTTGTATTTTTAATAGATACAGGTTTTCACCATGCTACCCAGGATAGTCTTGATCTCCTGA
CCTCATGATCCACCTGCCTCAGCCTCCCAAAGTGCTGGGATTACAGATGTGAGCCACCGCACCCGGCCGCTAATGT
TTTATTTTTTATTGTTTGTAGAGACAGGGTCTCCCCATGTTGCCTAGGCTGTTTTTTGTTTTGTTTTGTTTTGTTT
AACAGAGACAGGATCTTGCTATGTTGCCAAGGCTGGTCTTGAACTCCTGGGTTCAAGCAGTCCTCCCATCTTGGTC
TCCCACCCAAAATGCTAGGATAGGCGTGAACCGCCAGGCCCGGTCTCCATTGTGAATCTTTTTTTTTTCTTTTTTT
TTTTTTGAGACTGTGTTTTGCTCTTGTTGCCCAGGCGGGAGTACAGTGGTGCGATCTCGGCTCACCGCAACCTCCG
TCCGCCTCCTGGGTTCAAGCGATTCTCCTGCCTCAGCCTTCCTGAGTAGCTGGGATTACAGGCATGCGCCACCATG
CCTGGCTAATTTTGTATTTTTAGTAGAGATGGGGTTTCCCCATGTTGGTCAAGCTGATCACGAACTCCCGACCTCA
```

```
GATGATCCACCCGCCTCGGCCTCCCAAAGTGCTGGGATTACAGGCATGAGCTACCGCGCCCGGTTTACCATTTTGA
ATCTTAAGAGGACAAAGTCTGGTTCTCTAGAAGGCCCGAGTAGCTTTTCCCCTGAGCATCTTGAGAAAAAAGTGCT
CTGAGCACTCCTTGGGAAGTCCGAGACAGCACATGCAGCCAGGTGCTCCCTGGCTGCTCTCCCAGAGGTCAGCACA
GCTGGAGGCAGCTGGAGGGAGGAGGCTGCAGAGGCGCTGAAGCCAGGAAGACCCAACAGTGGAGAGAGAAGTTGCC
TGGTGACTGACCCTCGGCTTGACCCTGTGCTCCCCTGGGCTGAGCTGGGTGTCAAAAGGCCTCCACCCTCCAAAGT
GCCAGCCTCCTGCTGTGGCTGGAAAGCATTTGCTCCTCCTCCTCTTCTCACTTCTTTGCATTCTTCTGATCTCTTA
AGGGCGTACAAAGATTCACATGGATATTAAAGTCAATGGTGATTTTCATAGCTGGAATACTCGTTGGGAGAAAGTA
GGAATTCTTTTTTTTTTTTTTTTCTCCCAAGCTCAAGCGATCCTCCCACCTCAGCCTTCTGAGTAGCTGGGACCAC
AGGTGTGAGCCTCCATGATGGCCTGGCTAATTTTTTATATTTTTAGT],  each should be only one
characters long
```

Lo que vemos es que los marcadores **esv** no se pueden convertir porque tienen mas de un caracter. Lo que hacemos es eliminar estos marcadores,

```
$ grep esv admurcia_merged5_chr22.map | awk '{print $2}' > esv_chr22.txt
$ plink --file admurcia_merged5_chr22 --exclude esv_chr22.txt --recode --
allow-no-sex --out admurcia_merged5_noesv_chr22
```

y ahora si,

```
$ fcgene --map admurcia_merged5_noesv_chr22.map --ped
admurcia_merged5_noesv_chr22.ped --oformat impute --out admurcia_2imp2_chr22
```

Una vez convertido, el mismo FCgene nos brinda un script que podemos correr o las ordenes individuales. El script hay que tocarlo un poco pero ya nos da el trabajo hecho