How to analyze interSNP results of DB_TEST_Singlemarker.txt

The first goal is to be able of run a *clump* over the obtained results. Our first approach will be take a *Singlemarker* result from interSNP, parse it and write the data in plink results format.

- First, plink data must contain:

```
CHR     Chromosome
SNP     SNP ID
BP      Physical position (base-pair)
A1      Minor allele name (based on whole sample)
F_A     Frequency of this allele in cases
F_U     Frequency of this allele in controls
A2      Major allele name
CHISQ   Basic allelic test chi-square (1df)
P       Asymptotic p-value for this test
OR      Estimated odds ratio (for A1, i.e. A2 is reference)
```

- Second, interSNP file contains:

```
1 - No
2 - Chr
3 - rs_No
4 - Position
5 - Gene
6 - minor
7 - major
8 - MAF(All/Co/Ca)
9 - SNP_MR
10 - HWE_Ca
11 - HWE_Co
12 - P_Single-marker
13 - P_Corr
14 - A_Ca_N
15 - B_Ca_N
16 - A_Co_NB_Co_N
17 - A_Ca
18 - B_Ca
19 - A_Co
20 - B_Co
21 - OR_A
22 - LCL_A
23 - RCL_A
24 - OR_B
25 - LCL_B
26 - RCL_B
```

- So, the relation must be:

```
CHR --> Chr
SNP --> rs_No
```

```
   BP --> Position
   A1 --> minor
   F_A --> MAF(All/Co/Ca) --> Ca
   F_U --> MAF(All/Co/Ca) --> Co
   A2 --> major
   CHISQ --> Not available, lets try skipping it
   P --> P_Single-marker
   OR --> OR_A
```

Instead try writting a perl pattern to catch the desired variables I'm going to use sed & awk first to write a *csv* like temporary file

```
echo "CHR SNP BP A1 F_A F_U A2 P OR" > tmpfile; awk -F'\t' '{print
$2,$3,$4,$6,$8,$7,$12,$21}' myresults.file | sed 's/\(.*\) .*\/\(.*\)\/\([^
]*\) \(.*\)/\1 \3 \2 \4/' | tail -n +2 >> tmpfile
```

Yep, I know is awful but it save me from writing a nearly to infinity matching regexp in perl.

- Now, this file is ready to feed plink! Something like this,

  ```
   plink --bfile ~/data/Variomics/ADMURimpQC2 --clump test_clean.txt --
  clump-p1 0.01
  ```

  produces this output:

  ```
  @----------------------------------------------------------@
  |        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 [ plink.log ]
  Analysis started: Wed Mar  6 14:06:03 2013

  Options in effect:
          --bfile /home/osotolongo/data/Variomics/ADMURimpQC2
          --clump test_clean.txt
          --clump-p1 0.01

  Reading map (extended format) from [
  /home/osotolongo/data/Variomics/ADMURimpQC2.bim ]
  1034238 markers to be included from [
  /home/osotolongo/data/Variomics/ADMURimpQC2.bim ]
  Reading pedigree information from [
  /home/osotolongo/data/Variomics/ADMURimpQC2.fam ]
  ```

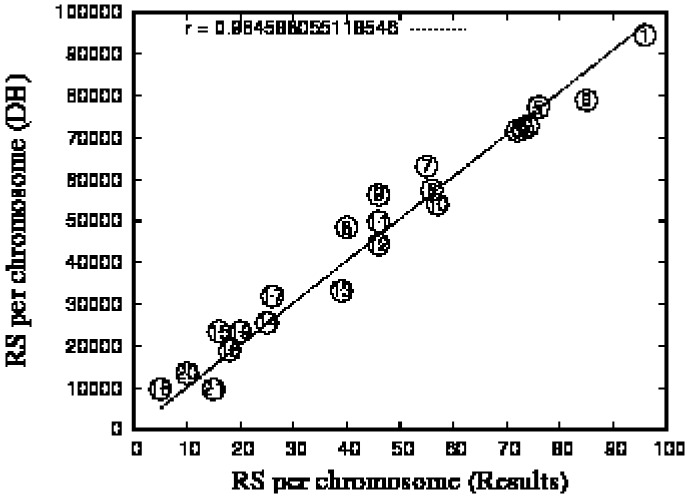
```
1088 individuals read from [
/home/osotolongo/data/Variomics/ADMURimpQC2.fam ]
1088 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
319 cases, 769 controls and 0 missing
511 males, 577 females, and 0 of unspecified sex
Reading genotype bitfile from [
/home/osotolongo/data/Variomics/ADMURimpQC2.bed ]
Detected that binary PED file is v1.00 SNP-major mode
Before frequency and genotyping pruning, there are 1034238 SNPs
1088 founders and 0 non-founders found
1130 heterozygous haploid genotypes; set to missing
Writing list of heterozygous haploid genotypes to [ plink.hh ]
Total genotyping rate in remaining individuals is 0.986847
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 1034238 SNPs
After filtering, 319 cases, 769 controls and 0 missing
After filtering, 511 males, 577 females, and 0 of unspecified sex

Parameters for --clump:
          p-value threshold for index SNPs = 0.01
      Physical (kb) threshold for clumping = 250
     LD (r-squared) threshold for clumping = 0.5
        p-value threshold for clumped SNPs = 0.01

Reading results for clumping from [ test_clean.txt ]
Extracting fields SNP and P
Indexing on all files
Writing clumped results file to [ plink.clumped ]

Analysis finished: Wed Mar  6 14:17:58 2013
```

This look fine 😉

What happens then with the obtained results?

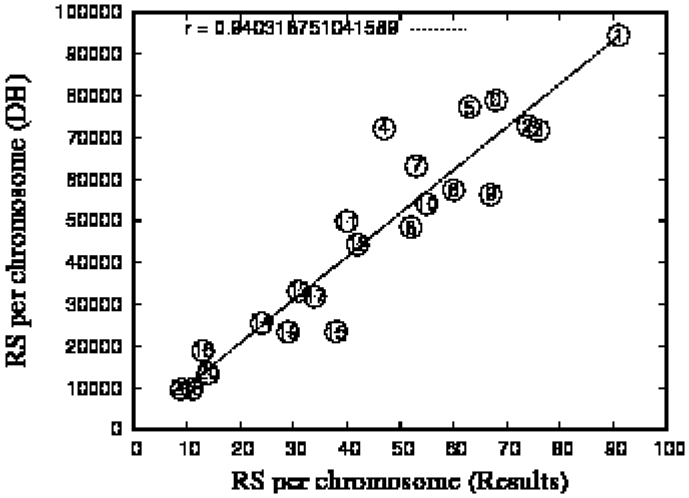
Well, we can make a bootstrap of the better results (e.g. 1k) and compare this with the frequency of markers by chromosome.

This first one is a self test. I mean, a bunch of random 1k SNPs of the DB against the SNPs DB frequency.



As expected, unless the RNG was rubbish, the linear regression is very good between both data. However if our results are half good the Pearson coefficient must be far from 0.9.

Look at this picture,



This was the above explained test over single marker results of interSNP test 5 against one of our DBs. The goodness of regression is self explanatory, results don't worth a dime.

Want to see what happens next? go to the DB_TEST_BestMarkerCombi2.txt example