Int J Mol Epidemiol Genet 2011;2(1):30-35
www.ijmeg.org /ISSN1948-1756/IJMEG1009006

Bioinformatics Application Note

Using Fisher's method with PLINK 'LD clumped' output to compare SNP effects across Genome-wide Association Study (GWAS) datasets

Hui Shi, Christopher Medway, Kristelle Brown, Noor Kalsheker and Kevin Morgan

Human Genetics, School of Molecular Medical Sciences, Queens Medical Centre, University of Nottingham, Nottingham NG7 2UH, UK.

Received September 29, 2010; accepted November 16, 2010; Epub November 25, 2010; published January 1, 2011

Abstract: As the number of publically available GWAS datasets continues to grow, bioinformatic tools which enable routine manipulation of data are becoming increasingly useful. Meta-analysis using multiple GWAS datasets has become essential to elucidate novel SNP associations which may not be readily discovered in each GWAS individually due to insufficient power. Replication of GWAS findings is critical and is the 'arbiter' of genuine SNP associations. We have developed an 'LD aware' bioinformatics application which allows efficient comparison of SNP effects across multiple GWAS datasets using Fisher's combined probability test from PLINK (v1.06) 'LD clumped' output. Availability: the application is freely available from the authors.

Keywords: Genome-wide Association Studies (GWAS), Fisher's method, bioinformatic tools

Motivation

Although genome wide association studies (GWAS) have given a valuable insight into the biological aetiology of many complex human diseases, insufficient power has meant a large number of these studies have failed to generate any significant 'hits' [1]. GWAS have often fallen below the numbers of cases and controls required to detect a modest effect (OR ~1.25) from a common variant [1-3]. However, despite these power issues, they are still valuable for meta-analysis purposes. Combining individually underpowered GWAS will propel genuine associations and spurious associations will diminish [4].

Unfortunately, there are a number of constraints that have limited the effectiveness of whole-genome meta-analysis to date; i) given that GWAS use different genotyping platforms (Illumina or Affymetrix), each assaying different panels of SNPs, the number of 'matched' SNPs available for meta-analysis is limited. This is often confounded by SNP dropout during quality control procedures ii) whole genome meta-analysis is labour intensive without suitable bioinformatics software.

To overcome these constraints, we have developed an application written in PERL programming language which allows simple and efficient meta-analysis of multiple GWAS. Unlike other whole-genome meta-analysis tools, our application is 'LD aware' - if a single SNP is present in one study but is not present directly in the other, the p-value of the SNP is inferred using a perfect proxy (LD, r^2=1).

We consider this application to have several advantages; i) maximize the number comparable SNPs (and therefore genome coverage) using LD (r^2=1), ii) provide summary p-value statistics using Fisher's method, iii) annotate SNPs which have an odds ratio (OR) in opposing directions from different studies ('flippers'), and iv) tabulate the results to facilitate simple visualisation by conventional tools such as Microsoft

Excel or SPSS.

Implementation

General

PLINK (v1.06) [5] provides an 'ld-clump' analysis which allows automatic calculation of 'clumps' (blocks of SNPs in LD) across genotyping chip platforms. The software we have developed uses the output file from this 'LD clump' analysis and performs the downstream analysis – meta-analysis of SNPs in each clump (taking one SNP/proxy (r^2=1) from each study and combining their p-values).

The application consists of two files '*perl.pl'*, '*module.pm*', where '*perl.pl*' is an executable file when PERL language is installed. PERL programming language can be downloaded freely from http://www.perl.org/.
'*perl.pl*' file can be edited using a conventional text file editor, and allows users to define two parameters; i) the location and the filename of the input ('*ld_clump*') file generated in PLINK, and ii) the type of the study - case/control (CC) analysis or quantitative trait (QT) analysis. Fields requiring modification have been annotated in the '*perl.pl*' file. The default input is case/control analysis. Failure to adjust the parameter for the correct type of analysis will cause incorrect output in the results file.

To execute the programme, simply double click the file '*perl.pl*'; alternatively, type in '*perl perl.pl*' into the command line. The command-line interface icon can be found in 'Start -> All programs -> Accessories' in Windows.

The programme was designed to handle an unlimited number of GWAS datasets and unlimited SNPs. However, currently we have only validated the application with a maximum of ten modelled datasets.

Issues and problem-solving

It is worth noticing that, when generating the 'clump' files, PLINK requires a reference GWAS dataset (usually HapMap) to calculate LD values between two SNPs. This data is freely available at http://hapmap.ncbi.nlm.nih.gov/. Given that LD values vary between ethnic populations, it is imperative that the reference dataset and the GWAS datasets are from a similar population thus avoiding stratification issues.

It is common that a SNP in one GWAS will have multiple perfect SNP proxies (r^2= 1) in another independent GWAS. In this situation, although it is considered appropriate to use any pair of SNPs to perform meta-analysis, the application decides which proxy to use based on which proxy is closest to the index SNP. The distance between the SNP/proxy and the index SNP is annotated in the results file. It should be noted that if an index SNP is unique to one study and does not have a perfect proxy in any other studies, no meta-analysis results will be displayed for this SNP.

This application only uses perfect proxies (r^2 = 1). We acknowledge that this is a limitation as using imperfect proxies ($r^2 < 1$) will increase the number of comparable SNPs between studies. However, we do not recommend this as it will lead to inaccurate Fisher's combined probability test p-values. Currently, there is no weighting algorithm implemented therefore any SNP p-value inferred from a proxy with $r^2 < 1$ will be inaccurately treated as a perfect match. If the user wishes to use imperfect proxies, a minimal LD value of $r^2 = 0.9$ maybe a reasonable threshold. To use imperfect proxies, simply alter the (--clump-r2) parameter in PLINK, and run the application as usual. This may indeed be a valuable approach to increase coverage, analogous to imputation, but until the output is weighted accordingly the results will have to be interpreted with caution.

A 'flipper' refers to single SNP or SNP proxy in one dataset which has the opposite effect in the original study. Our application compares the OR (case/control analysis) or regression coefficient (BETA from quantitative trait analysis) of all SNP pairs from different studies, and annotates according to the following rules:

i) 'YES – flipper' – ORs are in opposite directions in two (or more) GWAS studies, irrespective of missing OR data in additional studies, ii) 'NO - non-flipper' - all SNPs OR in the same direction and OR data is present for all studies, and iii) 'NA - not applicable' - there is either no OR data, or datasets are missing OR data making it inappropriate to call a 'non-flipper'. In studies with missing OR or BETA, the field has to be encoded as '-9' in the '*.assoc*' file for subsequent 'ld-clumping' analysis in PLINK.

The application automatically recognizes the number of GWAS from PLINK 'ld-clump' output files and tabulates the results accordingly. Although the application was designed for GWAS meta-analysis the user can perform analysis on much smaller datasets.

We suggest that this approach is used prior to more formal meta-analysis. It is essential that any potential finding that emerges using the application is verified by further investigation in a rigorous manner. Adjusting the genotypic data for covariates and taking into account heterogeneity between studies/samples will verify if observations involving both 'flipping' and 'non-flipping' alleles are likely to be genuine and worthy of downstream study.

EXAMPLE: Meta-analysis of Carrasquillo et al. 2009 [6], Reiman et al. 2007 [7] and Beecham et al. 2008 [8] GWAS

We used two Late-onset Alzheimer's disease (LOAD) GWAS datasets - Carrasquillo el al. 2009 and Reiman et al. 2007 and the 'top hits' tabulated in Beecham et al. 2008 to test the performance of the application. The sample sizes and genotyping platforms for the three studies are - Carrasquillo et al. 2009, 799 LOAD cases and 1199 controls on Illumina 300 chip, Reiman et al. 2007, 859 LOAD cases and 552 controls on Affymetrix 500K chip and Beecham et al. 2008, 492 LOAD cases and 496 controls on Illumina 550 chip. The total sample size of all three GWAS is 4,397 (2150 LOAD cases and 2247 controls).

No apolipoprotein E (*APOE*) related SNPs were listed in Beecham et al. 2008 'top hits'.

Procedure

Before using the software a number of PLINK analyses have to be undertaken:

1) Subject-level genotype data was obtained from Carrasquillo et al. 2009 and Reiman et al. 2007. The files were converted into PLINK format where necessary, and the SNP identifiers were converted into dbSNP 'rs' number.

2) A case/control allelic association test was undertaken using '--assoc' using PLINK (v1.06) for Carrasquillo et al. 2009 and Reiman et al. 2007 GWAS datasets. This generates two '*.assoc*' files '*Carrasquillo.assoc*' and '*Reiman.assoc*'. Since we do not have the Beecham et al. 2008 GWAS dataset, a file called '*Beecham.assoc*' was manually generated conforming to the format of an '*.assoc*' file [8]. Three compulsory columns are required in the '*.assoc*' file, and the headers have to be named exactly as 'SNP', 'OR' and 'P' without quotation marks. All other information such as 'BP', 'CHR', 'A1' in the standard '*.assoc*' file are not required, and can simply be ignored. As OR data was not included in the Beecham data, these values were set to '-9' in the 'OR' column.

3) The 'ld-clump' analysis was performed using PLINK (v1.06) with the three files we generated '*Carrasquillo.assoc*', '*Reiman.assoc*' and '*Beecham.assoc*'. We downloaded the filtered version of HapMap data (CEU population, release 23) in PLINK binary format (.bed, .bim and .fam) from the PLINK website. http://pngu.mgh.harvard.edu/~purcell/plink/res.shtml' [5]. The HapMap data downloaded contains 2.3 million SNPs.

The 'ld-clump' analysis was performed using the following PLINK command:

```
          plink     --bfile HapMapCEU23
                    -clump Carras-
quillo.assoc,Reiman.assoc,Beecham.assoc
                    --clump-verbose
                    --clump-annotate OR
                    --clump-p1 1
                    --clump-p2 1
                    --clump-r2 0.99
                    --out ld_clump
                    --noweb
```

The PLINK method reads '--bfile' (the HapMap data in PLINK format) and 'clumps' the three datasets based on HapMap LD r-squared values. '--clump-r2 0.99' ensures that only SNPs which are perfect proxies (r^2>0.99) are clumped ('--clump-r2 1' does not work). All SNPs were used to perform the 'ld-clump' analysis irrespective of p-values ('--clump-p1 1' and '--clump-p2 1'). The output of ORs was specified using '--clump-annotate OR' (Use '--clump-annotate BETA' for quantitative trait analysis). A single output file '*ld_clump.clumped*' was generated using '--out ld_clump'. All these commands listed are compulsory to the subsequent analysis except '--clump-p1' and '--clump-p2' which allow the user to control p-value threshold.

4) The '*perl.pl*' file was edited using a text file

Table 1. Results from Meta-analysis of Carrasquillo et al. 2009, Reiman et al. 2007 and Beecham et al. 2008. The table shows Single Nucleotide Polymorphisms (SNPs) with Fisher's combined probability test p-value less than 1E-05. 'F1', 'F2' and 'F3' refers to the studies which have been inputted to perform the meta-analysis. 'F1' indicated which study the 'SNP' is taken from ('1', '2' and '3' refers to Carrasquillo et al. 2009, Reiman et al. 2007 and Beecham et al. 2008, respectively). 'F2' and 'F3' indicated which study 'PROXY1' and 'PROXY2' is taken from ('1', '2' and '3' refers to the studies as above). 'KB1' and 'RSQ1' refer to the distance and LD between the index 'SNP' and 'PROXY1'. 'KB2' and 'RSQ2' refer to the distance and LD between the index 'SNP' and 'PROXY2'. The suffix (1,2 or 3) on column headers 'Pvalue' indicated p-value in each study individually as listed. 'ClumpNo' - index number ranked based on descending p-value in each study, 'CHR' - chromosome number, 'FISHER' - Fisher's combined probability test p-value and 'FLIPPER' - indicates whether the SNP is a 'flipper' (see main text for definition of a 'flipper').

ClumpNo	SNP	F1	CHR	PROXY1	F2	KB1	RSQ1	PROXY2	F3	KB2	RSQ2	FISHER	FLIPPER	Pvalue1	Pvalue2	Pvalue3
1	rs1114832	1	19	rs1114832	2	0	1	-	-	-	-	1.09E-09	NO	1.37E-06	0.000799	-
4	rs10402271	1	19	rs10402271	2	0	1	-	-	-	-	1.13E-07	NO	4.54E-06	0.0248	-
2	rs2318144	1	8	rs6982990	2	0.532	1	-	-	-	-	3.57E-07	NO	2.22E-06	0.161	-
12	rs929156	3	6	rs929156	1	0	1	rs9261519	2	-22.6	1	4.32E-07	NA	0.0887	0.288	1.69E-05
20	rs3746319	3	19	rs3746319	1	0	1	rs2061332	2	1.43	1	4.34E-07	NA	0.985	0.0149	2.96E-05
7	rs11205641	3	1	rs11205641	1	0	1	rs11205641	2	0	1	1.10E-06	YES	0.385	0.34	8.41E-06
107	rs468345	2	21	rs468345	1	0	1	-	-	-	-	1.15E-06	NO	0.00353	0.000326	-
42	rs7679738	1	4	rs510115	2	2.67	1	-	-	-	-	1.63E-06	NO	9.72E-05	0.0168	-
5	rs9659092	3	1	rs12022125	2	-83.5	1	-	-	-	-	1.83E-06	NA	-	0.404	4.54E-06
3	rs249153	2	12	rs249153	1	0	1	-	-	-	-	1.91E-06	NO	0.717	2.66E-06	-
68	rs9474661	2	6	rs4486000	1	-2.44	1	-	-	-	-	2.19E-06	NO	0.0142	0.000154	-
16	rs8039031	1	15	rs8039031	2	0	1	-	-	-	-	2.24E-06	NO	2.26E-05	0.0992	-
86	rs4693305	2	4	rs4693305	1	0	1	-	-	-	-	2.64E-06	NO	0.0119	0.000222	-
10	rs7318037	1	13	rs4456389	2	11.5	1	-	-	-	-	2.75E-06	YES	1.15E-05	0.239	-
6	rs3007421	1	1	rs3007421	2	0	1	-	-	-	-	3.06E-06	NO	6.54E-06	0.468	-
74	rs385771	1	5	rs385771	2	0	1	-	-	-	-	3.78E-06	YES	0.000163	0.0232	-
76	rs10501120	1	11	rs10501120	2	0	1	-	-	-	-	4.00E-06	NO	0.000171	0.0234	-
9	rs4313171	2	8	rs359819	1	-169	1	-	-	-	-	5.01E-06	NO	0.501	1.00E-05	-
144	rs6695249	1	1	rs17113051	2	4.38	1	-	-	-	-	5.40E-06	NO	0.00045	0.012	-
11	rs3807031	3	6	rs3807031	1	0	1	-	-	-	-	5.73E-06	NA	0.494	-	1.16E-05
25	rs2119067	3	2	rs2119067	1	0	1	-	-	-	-	6.92E-06	NA	0.158	-	4.38E-05
35	rs11033712	2	11	rs12271660	1	26.9	1	-	-	-	-	7.85E-06	YES	0.127	6.18E-05	-
8	rs6546452	1	2	rs17680828	2	9.37	1	-	-	-	-	8.28E-06	NO	8.55E-06	0.968	-
14	rs4759173	2	12	rs10876820	1	-22.7	1	-	-	-	-	8.86E-06	NO	0.445	1.99E-05	-
22	rs2387100	3	13	rs2387100	1	0	1	rs9551404	2	-12.5	1	9.40E-06	NA	0.644	0.382	3.82E-05
31	rs7537266	2	1	rs7537266	1	0	1	-	-	-	-	9.52E-06	NO	0.186	5.12E-05	-
17	rs13213247	2	6	rs16892136	1	-115	1	-	-	-	-	9.55E-06	NO	0.417	2.29E-05	-
79	rs4904864	1	14	rs10484035	2	14	1	-	-	-	-	9.58E-06	YES	0.000186	0.0515	-

editor to ensure it contains the correct PLINK 'ld-clump' output filename and correct type of analysis (either 'CC' for case/control analysis or 'QT' for quantitative trait analysis) as discussed earlier.

5) The application was executed by double clicking the '*perl.pl*' icon in Windows. Results file named '*results.txt*' was generated automatically.

Results

The top 7 results in **Table 1** illustrate the utility of this application. The top 2 SNPs are in LD with the *APOE* locus and demonstrate highly significant p-values as expected (rs1114832 Fisher's method p-value 1.09 x 10^{-9} and rs-10402271 Fisher's method p-value 1.13x10^{-7}); the first SNP met genome-wide significance (p = 1.67x10^{-7}) after correcting for the number of independent tests as described previously [4] and the second SNP approached this value.

Sub-significant hits may prove to be genuine when more datasets are included. rs2318144 is located 200kb upstream of the inositol monophosphatase domain containing 1 gene, *IMPAD1*; as of October 2010 this gene has yet to figure as an AD candidate in the AlzGene forum [3].

rs929156 (p = 4.32x10^{-7}) is located in an exon of tripartite motif containing protein 15 (*TRIM15*) gene. There is evidence that *TRIM15* is involved in the innate immune system however, its association with Alzheimer's disease has yet to be sufficiently investigated [9-11].

rs3746319 (p = 4.34x10^{-7}) was found to be located in an exon of the zinc finger protein (*ZNF224*). Although this SNP is close to the *APOE* region, the effect has been suggested to be independent of *APOE* status [8].

rs11205641 demonstrates a dichotomy of effect i.e. it's OR is not compatible between datasets (shows opposing effects) and is thus indicated as a 'flipper'.

rs468345 (p = 1.15x10^{-6}) is located ~120kb upstream of Amyloid-beta precursor protein (*APP*) gene. It is well known that *APP* proteolysis generates beta amyloid (Aβ), and extracellular deposition of Aβ plaques is a hallmark of Alzheimer's disease [12]. Aβ is the central player in amyloid cascade hypothesis in AD [11, 13].

Acknowledgement

This work was supported by the Alzheimer's Research Trust and the Big Lottery Fund. We would like to express our sincere thanks to Dr. Minerva Carrasquillo and Professor Steven Younkin (Mayo Clinic, Jacksonville, Florida, USA) for kindly providing us with the data from their GWAS. Conflict of Interest: none declared.

Please address correspondence to: Professor Kevin Morgan, Professor of Human Genomics and Molecular Genetics, Institute of Genetics, School of Molecular Medical Sciences, A Floor, West Block, Room 1306, Queens Medical Centre, Nottingham NG7 2UH, United Kingdom, Tel: 0115 8230724, E-mail: kevin.morgan@nottingham.ac.uk

References

[1] McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP and Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008; 9: 356-369.

[2] Welcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007; 447: 661-678.

[3] Bertram L, McQueen MB, Mullin K, Blacker D and Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet 2007; 39: 17-23.

[4] Shi H, Medway C, Bullock J, Brown KS, Kalsheker N and Morgan K. Analysis of Genome-Wide Association Study (GWAS) data looking for replicating signals in Alzheimer's disease (AD). Int J Mol Epidemiol Genet 2010; 1: 53-66.

[5] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007; 81: 559-575.

[6] Carrasquillo MM, Zou F, Pankratz VS, Wilcox SL, Ma L, Walker LP, Younkin SG, Younkin CS, Younkin LH, Bisceglio GD, Ertekin-Taner N, Crook JE, Dickson DW, Petersen RC and Graff-Radford NR. Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. Nat Genet 2009; 41: 192-198.

[7] Reiman EM, Webster JA, Myers AJ, Hardy J, Dunckley T, Zismann VL, Joshipura KD, Pearson JV, Hu-Lince D, Huentelman MJ, Craig DW, Coon KD, Liang WS, Herbert RH, Beach T,

Rohrer KC, Zhao AS, Leung D, Bryden L, Marlowe L, Kaleem M, Mastroeni D, Grover A, Heward CB, Ravid R, Rogers J, Hutton ML, Melquist S, Petersen RC, Alexander GE, Caselli RJ, Kukull W, Papassotiropoulos A and Stephan DA. GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. Neuron 2007; 54: 713-720.

[8] Beecham GW, Martin ER, Li YJ, Slifer MA, Gilbert JR, Haines JL and Pericak-Vance MA. Genome-wide association study implicates a chromosome 12 risk locus for late-onset Alzheimer disease. Am J Hum Genet 2009; 84: 35-43.

[9] Ozato K, Shin DM, Chang TH and Morse HC, 3rd. TRIM family proteins and their emerging roles in innate immunity. Nat Rev Immunol 2008; 8: 849-860.

[10] Uchil PD, Quinlan BD, Chan WT, Luna JM and Mothes W. TRIM E3 ligases interfere with early and late stages of the retroviral life cycle. PLoS Pathog 2008; 4: e16.

[11] Sleegers K, Lambert JC, Bertram L, Cruts M, Amouyel P and Van Broeckhoven C. The pursuit of susceptibility genes for Alzheimer's disease: progress and prospects. Trends Genet

[12] Turner PR, O'Connor K, Tate WP and Abraham WC. Roles of amyloid precursor protein and its fragments in regulating neural activity, plasticity and memory. Prog Neurobiol 2003; 70: 1-32.

[13] Hardy J, Duff K, Hardy KG, Perez-Tur J and Hutton M. Genetic dissection of Alzheimer's disease and related dementias: amyloid and its relationship to tau. Nat Neurosci 1998; 1: 355-358.