Gearing up software systems for genome data: a case study with SAS

Jing Hua Zhao[1,*], Qihua Tan[2], Jian'an Luan[1], Shengxu Li[1], Fuzhong Xue[3], Wendi Qian[4], Ruth J. F. Loos[1], Nicholas J. Wareham[1]

[1] MRC Epidemiology Unit, Institute of Metabolic Science, Box 285, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, United Kingdom

[2] Dept of Biochemistry, Pharmacology and Genetics, Odense University Hospital, Sdr Boulevard 29, Odense C, DK-5000, Denmark.

[3] Institute of Public Health, Shandong University, 44 Wen Hua Xi Lu, Jinan 250012, PR China

[4] MRC Clinical Trials Unit, 222 Euston Road, London, NW1 2DA, United Kingdom

ABSTRACT

Summary: We described implementation of analysis for genome-wide association studies (GWASs) followed by a brief comparison with alternative implementation and software. We expect they will be useful for GWASs and studies with many variables.

Availability: The software and supplementary information are available at http://www.mrc-epid.cam.ac.uk/~jinghua.zhao.

Contact: jinghua.zhao@mrc-epid.cam.ac.uk

Supplementary information: Supplementary materials are available at *Bioinformatics* online.

Large-scale genome data have underlain the recent discoveries of genes or single nucleotide polymorphisms (SNPs) associated with diseases or other traits in humans. While the major scientific aspects in these genome-wide association studies (GWASs) are well recognized (Couzin and Kaiser, 2007; Altshuler, et al., 2008; Donnelly, 2008; Pearson and Manolio 2008), we believe that many practical issues are wide open. We previously argued the case for analysis of genomic data in general software systems such as SAS or R for data management, programming, validation and development of analytical tools (Zhao and Tan, 2006a). SAS has been widely used by both the industries and academic institutions and R is a free, open source statistical and programming environment that is increasingly popular (Vance, 2009a,b). Here we will focus on SAS, for we have used it for genetic data before most of the recent tools for GWASs became available.

Despite its advantage, our experience (Zhao, *et al.*, 2007; Loos, *et al.*, 2008) with a study of ~4000 population-based individuals each with Affymetrix 500K GeneChip data has shown that additional challenges present themselves when tackling genomic analysis. In particular, we found that data partition (i.e., into smaller subsets of SNPs) was necessary for calculation of per SNP summary statistics and association testing. We here present procedures that can be used to overcome this problem along with other developments. These include 1. converting from an individual-by-SNP format (one column per SNP) to a more desirable SNP-by-individual format (one column per individual), with the option to include allele labels from a map file for direct coding of SNPs as will be detailed below, 2. placing the phenotype and covariates in a separate file for greater flexibility, 3. using established procedures for analysis involving haplotypes, imputed genotypes, meta-analysis, and 4. obtaining individual data or summary statistics efficiently obtained from HapMap (The International HapMap Consortium, 2007). These procedures simplify the implementation of analyses a great deal and allow researchers to undertake a wide range of analysis effectively.

We will elaborate the data organization first. With many SNPs and the prospect of involving the whole genome, the SNP-by-individual format is preferable but nevertheless non-standard for most general statistical systems. This is because per SNP association model would involve specification of a particular SNP name, it would be laborious to literally lay out all models from data with an individual-by-SNP format. To get around this we have implemented an algorithm such that for each SNP, we read map information and iteratively proceed through individuals' genotype column-wise together with a record in the trait file containing phenotype and covariates corresponding to the particular column in the genotype file, and all three sources of information are stored in a long format file. We can keep track of successful genotype calls per SNP. The long file can be processed by SAS/GENETICS with a BY statement plus the *notsorted* option, which really instructs a per SNP analysis. With additional indicators in the genotype and phenotype files, one can focus on a subset of SNPs and individuals without much overhead.

Although it is a common practice to designate alleles to be minor or major according to their population frequencies, and to code alleles of SNPs according the type or number of effect alleles, we found this to be a bottleneck for our previous implementation because we needed to count the number of alleles, pick up the minor allele and compare against the original genotypes. We now used an operational rule of keeping allele labels (i.e., A, C, G, T) in alphabetical order and treating the second allele to be effective so that additive or other types of coding can be done per genotype at runtime when the long file is generated. It is easily seen that for interpretation one only needs to swap the roles of minor and major alleles whenever appropriate, although minor allele is more an indication of the amount of information in our sample concerned about a particular SNP. This also allows for analysis via standard procedures other than those available from SAS/GENETICS, such

*To whom correspondence should be addressed.

as REG and LOGISTIC for linear and logistic regression analysis. We note that a typical SAS program divides tasks into data preparation (DATA step) and analysis (PROC step). Although it is possible to mix them in order to avoid the extra storage, the programming would be more involved and less efficient than what we have just described.

While the implementation is surprisingly short for most per SNP analyses, it is also computationally fast, memory efficient and SAS/BASE alone can furnish the necessary step for following analysis. Example using our EPIC-Norfolk data is provided as **Supplementary materials**. This suggested that it is viable to script tools such as *awk* yet more transparent and flexible. The long file uses more disk space when many variables are included in the analysis but a temporary use is often acceptable, especially for a subset of SNPs of interest. In addition, our example extends to family data by keeping identifiers for pedigrees, individuals, and parents in the phenotype file as key components in pedigree data analysis.

The remaining aspects we would like to mention are in regard to haplotypes, imputed genotypes, meta-analysis and publicly available data. Through SURVEYREG and SURVEYLOGISTIC procedures we can use posterior probability of probabilities of haplotype assignments, imputed genotypes, while meta-analysis can be facilitated by procedures such as MIXED, whether or not to account for heterogeneity or covariates. Data as available from HapMap in compressed format can be obtained through a short DATA step.

On the other hand, it would not be so difficult though not as straightforward to wrap up our algorithms via SAS/IML, an interactive matrix language, to process data without apparent use of the extra storage. We also use SAS in conjunction with independent programs as are listed at the Rockefeller linkage server (http://linkage.rockefeller.edu), for it provides the much needed validation and it would generally be a slow process to tune the comprehensive procedures for large data but to fast implement specific tasks, e.g., PLINK (Purcell, *et al.*, 2007) in a similar spirit and the C++ programming language, although it might be unnecessary or a matter of convenience with its feature to split tasks such as frequency calculation and association modeling. In an ongoing study of lung function for a consortium involves 38 analyses at the first phase running SNPTEST on imputed genotypes produced by IMPUTE (Marchini, *et al.*, 2007), all input files were orchestrated by SAS.

A collection of R packages is under development (Zhao and Tan, 2006b, and for GWASs these include SNPassoc, GenABEL, pbatR and snpMatrix (Gonzalez, *et al.*, 2007; Aulchenko, *et al.*, 2007; Clayton and Leung, 2007), all available from the Comprehensive R Archive Network (http:///cran.r-project.org) and/or BioConductor (http://www.bioconductor.org). They customarily couple R language with calls to C/C++/Fortran routines or independent programs. Since a data object in R can often be directly treated as a matrix, it renders flexible specification of data in statistical modeling. Moreover, R has been our software of choice for graphical presentation.

Epidemiological cohorts eventually play a big role in gene characterization and discoveries (Longholz, *et al.*, 1999; Bodmer and Bonilla, 2008; Manolio, 2009). Our work is a good example of how to reframe a problem in order to arrive at a better solution. We have only covered elementary analysis with a lot of non-genetic variables and a naïve method for data compression, but for sequence data it would be worthwhile to consider better approaches (e.g., Christley, *et al.*, 2009) or faster algorithms (e.g., Wigginton, *et al.*, 2005). If general software systems themselves could go beyond comprehensiveness and correctness to be fast, they would be more in line with the pace of GWASs or studies with data of similar kind.

ACKNOWLEDGEMENTS

We wish to thank Dr Mike Weale for many helpful suggestions and Prof Pak Sham for comment.

REFERENCES

Altshuler, D., *et al*. (2008). Genetic mapping in human disease. *Science* **322**, 881-888.

Aulchenko,Y.S., *et al.* (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, **23**, 1294-1296.

Bodmer, W., and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet*, **40**, 695-701.

Christley, S., *et al.* Human genome as email attachments (2009). *Bioinformatics*, **25**, 274-275.

Clayton,D., and Leung, H.T. (2007) An R package for analysis of whole-genome association studies. *Hum. Hered.* **64**, 45-51.

Couzin, J., and Kaiser, J. (2007). Closing the net on common disease genes. *Science*, **316**, 820-822.

Donnelly, P. (2008). Progress and challenges in genome-wide association studies in humans. *Nature* **456**, 728-731.

Gonzalez, J.R., *et al.* (2007) SNPassoc: an R package to perform whole genome association studies. *Bioinformatics*, **23**, 644-645.

Langholz, B, *et al.*(1999): Cohort studies for characterizing measured genes. *J. Natl. Cancer Inst. Monogr.*, 39-42.

Loos, R., *et al.* (2008). Common variants near *MC4R* are associated with fat mass, weight and risk of obesity. *Nat. Genet.*, 40, 768-75.

Manolio, T. A. (2009) Cohort studies and the genetics of complex disease. *Nat. Genet.* **41**, 5-6.

Marchini, *et al.* (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. *Nat. Genet.* **39**, 906-913.

Pearson, T. A., and Manolio, T.A. (2008). How to interpret a genome-wide association study. *JAMA* 299:1335-134.

Purcell, S., *et al.* (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.*, **81**, 559-575.

Vance, A. (2009a). Data analysts captivated by R's power. *New York Times*, 6th January.

Vance, A. (2009b). R you ready for R? *New York Times*, 8th January.

Wigginton, J. E., *et al.* A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.*, **76**, 887-893.

Zhao, J.H., *et al.*(2007). Analysis of Large Genomic Data *in Silico*: The EPIC-Norfolk Study of Obesity. In DS Huang, L Heutte, and M Loog (Eds). *ICIC2007, CCIS 2 Advanced Intelligent Computing Theories and Applications with Aspects of Contemporary Intelligent Computing Technologies* 781-790.

Zhao, J.H., and Tan, Q. (2006a). Genetic dissection of complex traits *in silico*: approaches, problems and solutions. *Curr Bioinformatics*, **1**, 359-369.

Zhao, J.H., and Tan, Q. (2006b): Integrated analysis of genetic data with R. *Hum. Genomics*, **2,** 258-265.

Supplementary materials

Code listing and timing for the EPIC-Norfolk GWAS

```
%macro wtl(data, trait, map, snpid=rsn, vlist=, inc=one);
data out.long (keep=&snpid id &vlist a1a2 add n);
     set &data;
     fid=open("&data");
     length id $11. add 3. a1a2 $3.;
     format add 1.;
     set &map point=_n_;
     n=0;
     do col=2 to attrn(fid,"nvars");
        iid=col-1;
        set &trait (keep=&vlist &inc) point=iid;
        if &inc=1 then do;
           id=varname(fid,col);
           a1a2=vvaluex(id);
           add=.;
           if a1a2 ne "   " then do;
              a1=substr(a1a2,1,1);
              a2=substr(a1a2,3,1);
              add=(a1=b)+(a2=b);
              n+1;
           end;
           output;
        end;
     end;
     rc=close(fid);
run;
%mend wtl;
libname in ("." "/genetics/data/GWA/EPIC/6-5-7/wide2");
libname out '/tmp';
options compress=yes mprint ps=max;
proc sql;
     create table ind as select 1 as one, id from in.id order by id;
     create table bmi as select id, bmi from in.trait order by id;
     create table map as select rsn, chr, b from in.map order by chr, rsn;
quit;
data trait;
     merge bmi ind;
     format bmi 5.2 obesity 1.;
     if bmi ne . then obesity=(bmi>=30);
     by id;
run;
%wtl(in.a1a2, map, trait, snpid=rsn, vlist=bmi obesity);
ods select none;
data out.cr (keep=rsn n);
     set out.long;
     by rsn notsorted;
     if last.rsn;
run;
proc allele data=out.long genocol;
     by rsn notsorted;
     var a1a2;
     ods output markersumm=out.ms allelefreq=out.af;
run;
proc reg data=out.long;
     by rsn notsorted;
     ods output parameterestimates=out.bmipm;
     model bmi = add / b stb;
quit;
proc logistic data=out.long descending;
     by rsn notsorted;
     ods output parameterestimates=out.obpm CLOddsPL=out.obclpm;
     model obesity = add / expb clodds=pl;
run;
ods select all;
```

The listing consists of a SAS macro (Lines 1-26) to transform the SNP-by-individual to long format, followed by preparation of phenotypic information (Lines 31-41), summary statistics and association testing (Lines 44-63). The macro accepts three sources of information in separate files as follows (Table S1).

Table S1. Details of the example files as required by the SAS macro

Filename	Variable name		Comments
data	rsn	SNP name at column 1	Any legitimate SAS names
	WTCCC139236---WTCCC147447	Individual IDs from columns "start"	Any legitimate SAS names
map	chr	Chromosome	Read through _n_
	rsn	SNP name	Any legitimate SAS names
	pos	Position	
	a	Label for the 1st allele	Baseline allele
	b	Label for the 2nd allele	Effect allele
trait	age	Individual's age	Read through vlist
	bmi	Body mass index (weight/height^2, kg/m^2)	
	obesity	Obesity status, 0=non-obese, 1=obese	

The genotypes are stored in *data* from column two and alphabetically coded, e.g., C/C, C/T, T/T, or blank if they are missing. The macro works on each record of data, iteratively reading individual genotypes, extracting the two alleles, and obtaining the additive coding. For both our map and genotype files, we have ordered the allele labels and genotypes alphabetically. The argument *inc* is an flag of inclusion and individuals with value 1 are kept in the analysis. The macro returns the desired long file containing SNP name, individual's ID, un-coded and coded genotypes, and the number of successful genotypes calls at each SNP. After execution of the macro, a DATA step obtains number of successful calls at each SNP (Lines 44-48) followed by summary statistics such as allele and genotype frequencies, Hardy-Weinberg equilibrium (Lines 49-53), linear (Lines 54-58) or logistic (Lines 59-63) association testing on obesity status. We have used the output delivery system (ODS) feature to direct relevant outputs into datasets.

We ran our analysis under Linux using BMI/obesity as outcomes with additive coding adjusting for age. It took ~1.5 hours for generating the long file including allele coding, a few seconds for obtaining call rates, ~1 hour for summary statistics, ~15 minutes for linear regression, and ~4 hours for logistic regression. We used individuals in the sub-cohort (2417) out of total 3850 individuals in the study.

Allelic coding by alphabetical order

We consider additive, dominant and recessive coding because these are the most widely used. It can be seen that one only needs to change direction of additive effect or swap recessive and dominant models when appropriate, but meta-analysis of studies following the same rule is straightforward (Table S2).

Table S2. Allelic coding when the minor allele A is coded as B by alphabetical order

Correct Model	Genotype coding			Coded Model	Genotype coding			Change direction of effect
	A/A	A/B	B/B		A/A	A/B	B/B	
Additive	2	1	0	Additive	0	1	2	Yes
Dominant	1	1	0	Recessive	0	0	1	Yes
Recessive	1	0	0	Dominant	0	1	1	Yes

Example programs for haplotype analysis, meta-analysis and use of Internet

Haplotype analysis. The SURVEYREG procedure is used with posterior haplotype assignment as probability weight (*weight p*) for individuals (*cluster id*), for the association between BMI, sex, age and haplotypes. Out of eight possible haplotypes from four SNPs of interest, six of them were observed. The haplotypes h1—h5 are against a common haplotype h6 can be modeled as follows.

```
proc surveyreg;
```

```
     ods output parameterestimates=bmipm (where=(parameter^="Intercept"));
     cluster id;
     bmi=sex age h1--h5 / clparm;
     weight p;
run;
```

The results are available as SAS datasets *bmipm*. It is possible to perform permutation tests (Zhao, *et al*., 2000) or with all haplotypes present in the model (Huang, *et al*., 2007). Single marker allelic analysis based on log-likelihoods is possible by adding a dummy marker (Zhao, 2004).

Meta-analysis. The following program performs meta-analysis on 15 results in a consortium in a SAS dataset called *giant* (with variable names *lor* and *se* as log(OR) and standard error from logistic regressions).

```
data giant;
     input studyid lor se;
     col=_n_; row=_n_;est=se*se;value=se*se;invse=1/se ;
cards;
…  data for 15 studies …
run;
proc mixed method=ml data=giant;
     class studyid;
     model lor = / s cl;
     repeated / group=studyid;
     parms / parmsdata = giant eqcons = 1 to 15;
run;
proc mixed data=giant covtest;
     class studyid;
     model lor = / s cl;
     random int / subject=studyid type=un;
     parms (0.1) (1) / hold=(2);
run;
data giant2;
     Set giant (obs=1) giant;
run;
proc mixed data=giant covtest;
     class studyid;
     model lor = / s cl outp=predp;
     repeated / group=studyid r;
     random int / g gdata = giant s v;
     parms / parmsdata=giant2 eqncons=2 to 16;
     ods output CovParms=cp G=G R=R V=V SolutionF=SF SolutionR=SR;
run;
```

The first MIXED call fits a fixed effects model through *model* and *repeat* statements with the variances (*est*) being fixed (*eqcons*) and regression coefficient and confidence interval (*s, cl*) given. The second call also fits random effects model with heterogeneity test using *covtest* options, where the *parms* constrains the random parameter. The third call uses G matrix as specified by *row, col and value.* The *repeated, random* and ODS statements output V, G, R matrices as with solutions (*s*) for fixed and random effects. Significance levels of the predicted values can be obtained through *probnorm(resid/stderrpred)* and in dataset *predp* by the option *outp*, which replaces the *P* option in SAS version 6 (Normand, 1999; van Houwelingen, *et al.* 2002). The MIXED procedure produces maximum likelihood rather than moment estimates (Dersimonian and Laird, 1986).

Downloading individual genotype and Linkage disequilibrium (LD) data from HapMap. The following code obtains genotypic data of Caucasian sample on chromosome 22, build 36 and release 23a.

```
%let dir=http://www.hapmap.org/genotypes/latest/rs_strand/non-redundant;
%let file=genotypes_chr22_CEU_r23a_nr.b36.txt;
data _null_;
     call system("wget &dir/&file..gz");
     call system("gzip –dfq &file..gz");
run;
proc import datafile=”&file” out=test dbms=dlm replace;
     getnames=yes;
     datarow=2;
     guessingrows=32767;
run;
```

The *call* statements download and decompress the file via *wget* and *gzip*. The IMPORT procedure treats the data as delimited format (*dlm*) whose first lines contains variable names and allows for 32767 lines of data to be read for deciding variable types. One can also avoid *wget.* For instance, the latest LD information for the Caucasian sample on chromosome 22 can be obtained using the following program.

```
%let url=http://www.hapmap.org/downloads/ld_data/latest;
%let file=ld_chr22_CEU.txt.gz;
```

```
filename ldin url "&url/&file" recfm=S;
filename ldout "&file";
data _null_;
     infile ldin;
     file ldout recfm=F;
     input;
     put _infile_;
run;
filename in2 pipe "gzip -dcq &file";
data ld;
     infile in2 dlm=' ';
     format pop $3. rs1 $15. rs2 $15.;
     input pos1 pos2 pop rs1 rs2 dprime r2 lod fbin;
run;
```

The *filename* engine seeds file download (*infile*, *input*), and storage (*file*, *put*), pipe (through *filename*) into DATA step and read (*infile*, *input*) as if it was an ordinary file. The dataset *ld* contains chromosomal position, population identifier, SNPs and their LD measures (*D'*, r^2, lod score). As *ld_chr22_CEU.txt.gz* is quite large (about 100MB), it is compressed. The latest version of HaploView (Barrett, *et al.*, 2005) has similar function for this purpose. One may be benefited from the ease of implementation for other types of data, e.g., piping in the *filename* statement to access ASCII files from the Internet if they are not compressed.

REFERENCES

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**:263-265.

DerSimonian, R. and Laird, N.(1986). Meta-analysis in clinical trials. *Control Clin. Trials* **7**:177-188.

Huang, B.E., Amos, C.I., and Lin, D.Y. (2007). Detecting haplotype effects in genomewide association studies. *Genet Epidemiol.* **31**:803-812.

Normand, S. (1999). Tutorial in Biostatistics: Meta-analysis: formulating, evaluating combining, and reporting. *Stat. Med.* **18**:312-359.

van Houwelingen, H.C., Arends, L.R., and Stijnen, T. (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat. Med.* **21**:589-624.

Zhao, J.H., Curtis, D., and Sham, P.C. (2000). Model-free and permutation tests for allelic associations. *Hum. Hered.* **50**:133-139.

Zhao, J.H. (2004). 2LD, GENECOUNTING and HAP: Computer programs for linkage disequilibrium analysis. *Bioinformatics* **20**:1325-1326.