ON GENOTYPE-PHENOTYPE ASSOCIATION USING SAS

Jing Hua Zhao, Jian'an Luan, Ruth JF Loos, Nick Wareham
MRC Epidemiology Unit & Institute of Metabolic Science
Addenbrooke's Hospital Box 285
Hills Road
Cambridge CB2 0QQ, UK
email: [jinghua.zhao,jianan.luan,ruth.loos,nick.wareham]@mrc-epid.cam.ac.uk

ABSTRACT
We provide updates on various aspects that have been described in two previous papers concerning about the utility of general software systems in analysis of complex traits and the experiences from the European Prospective Investigation of Cancer (EPIC) Norfolk, a large population-based cohort study (`http://www.srl.cam.ac.uk/epic/`). Our current focus is restricted to SAS (`http://www.sas.com`) including data handling, foreign language calling, statistical modelling, and a brief comparison with other software such as R (`http://www.r-project.org`). We give examples on exact test of Hardy-Weinberg equilibrium, meta-analysis, and test of genettype-phenotype association involving family data including kinship calculation. Advantages, limitations and future work are also indicated. We believe these will be of general interest and practical use.

KEY WORDS
GENETICS AND GENOMICS, BIOINFORMATICS, GENOME-WIDE ASSOCIATION STUDIES, META-ANALYSIS, MIXED MODELS, PEDIGREE ANALYSIS

1 INTRODUCTION

Genome-wide association studies are routinely conducted (`http://www.genome.gov/gwastudies`) to identify genetic variants associated with human quantitative and disease traits[1]. They now involve millions of single nucleotide polymorphisms (SNPs), the most abundant genetic variants in the human genome. General software systems are appealing[2] for many statistical and computational challenges and shown to be feasible[3].

The current work has been motivated by our analysis of both population-based and family-based samples. For instance, in contribution to a consortium-wide analysis on lung function within Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium (`http://web.chargeconsortium.com/`), repeated measures of forced expiratory volume in 1 second (FEV1) were used via recommended coding distributed in SAS and R, while in the last two Genetic Analysis Workshops (GAWs, `http://gaworkshop.org`) family data were provided so it was natural to consider SNP-covariate interaction in these data. We realized that together with meta-analysis which is now commonly used in genetic association studies, both types of data can appropriately be put in the mixed-model framework, so can be meta-analysis. The consideration of various types of outcomes naturally invokes generalized linear mixed models. It is reminiscent of the earlier work in the literature[4,5] on meta-analysis, and [6] on analysis of genetic data. A further but somewhat different aspect in contribution to consortium from our data has involved calling C/C++ from general software systems such as SAS and Stata (`http://www.stata.com`). In the following we will describe our recent work on these following the latest development in SAS 9.2.

2 METHODS

In the following, we describe in more details the two datasets mentioned earlier and highlight some aspects of the data management and statistical analysis.

The EPIC-Norfolk study. The EPIC-Norfolk cohort study consists of 25,631 residents near Norwich, England[7]. Participants were 39-79 years old during the baseline health check between 1993 and 1997. A genome-wide association study was carried out in 2006 using a case-cohort design in which the subcohort was a random sample of the whole cohort at baseline and cases were the remaining individuals with body mass index (BMI) being over 30 kg/m^2. In contrast to a commonly used case-control design, the subcohort sample was an unselected sample of the population and allows for a variety of traits to be investigated. For the replication, approximately 20,000 individuals have been involved in the remaining EPIC-Norfolk cohort [8].

GAW 17. The workshop has distributed 200 replicates of simulated data based on 697 unrelated individuals and eight extended families. The former were from CEPH, Chinese, Japanese, Luhya, Tuscan and Yoruba populations in the 1000 genomes project, while the latter were founded by a random sample of 202 of those unrelated individuals. Both data had information on sex, age, smoking, three quantitative traits (Q1, Q2, Q4), an affection status (AFFECTED), and a common set of 24,487 SNPs from 3,205

genes for all 200 replicates. Besides the SNP data, identity-by-descent (IBD) information was also provided for these genes assuming fully informative markers.

2.1 Data manipulation

In general, to use a database system or SAS managing genomic data, we could work on data in chunks before proceeding to later processing. As this typically does not have a big overhead, it is also viable with text files so that data partition can be done through Linux utilities such as awk while enables the usual setup in SAS, e.g., text file processing and data analysis in the familiar tabular format (one line per indidivual). However, it is also possible to work on all data in SAS. Specifically, we have revised our initial deposition of data before working on 30 partitions of each chromosome[3] to one in which information is stored by SNP genotypes per individual and optionally with map information (SNP name, position, allele labels) embedded. Data in this format has no redundancy of information and can be used to generate a long format with map information, family structures if any, as well as trait and covariate information and individual's SNP genotypes. We pragmatically treat the second allele to be effective so that to characterize genotypic effect in a desirable direction (e.g., increasing phenotypic value, recessive or dominant models), it is only required to change direction of additive effect or swap recessive and dominant models when appropriate.

The long format is appropriate for analysis as described earlier[3], where SNPs and individuals can be filtered through criteria such as Hardy-Weinberg equilibrium (HWE), call rates (proportions of successfully genotyped SNPs/individuals per individual/SNP) and frequency of the minor allele at each SNP. Subsequently regression analyses were conducted assuming additive, dominant or recessive models, whose results were meta-analyzed across studies. A macro has been implemented for the data generation and is available from the first author.

We have used the GAW17 dataset as provided by the workshop organizers. As the number of SNPs is relatively small, all the data mangement is furnished within R.

2.2 Exact test of HWE

Given the large number of variants involved, it is often necessary to perform test of HWE on genotype counts quickly. We can SAS 9.2 PROC PROTO and PROC FCMP to call the exact procedure in C/C++ [9] natively as a SAS function.

```
proc proto package = work.mathfun
           label = "SNPHWE function";
     double
     SNPHWE(int obs_hets,
            int obs_hom1,
            int obs_hom2);
     link 'snphwe.so';
run;
proc fcmp inlib=work
          outlib=work.mathfun.trial;
     function HWE(b,a,c);
        pHWE=SNPHWE(b,a,c);
        return(pHWE);
     endsub;
run;
```

where `snphwe.so` is a shared object compiled from the C/C++ program. It works by creating a prototype function of the C/C++ function `SNPHWE` and in turn formally defining as a SAS function `HWE`. An inline version, which is simpler to compile but more lengthy, is to include C/C++ source code directly within PROC PROTO; where the functional body of `SNPHWE` is embedded within *externc SNPHWE;* and *externcend;*. Both PROC PROTO and PROC FCMP are available in SAS 9.2.

2.3 Mixed models

As outlined earlier, meta-analysis and analysis of longitudinal and family data can all be cast in a mixed model framework in which linear mixed model is the simplest. The model has the form

$$y = X\beta + Z\gamma + \epsilon$$

which links outcome y, explanatory variable X, known design matrix Z, and error term ϵ, where β is the unknown fixed effects parameter vector, γ is the vector of unknown random-effects parameter. It is assumed that $(\gamma\ \epsilon)$ has a joint multivariate normal distribution with mean 0, and covariance matrix with blocks G, R.

Control for familial relationship can be achieved through specification of the G and R matrices via the *random* and *repeated* statements in PROC MIXED. In particular, given the kinship coefficients A_1, the associated polygenic variance θ_1 the appropriate covariance matrix is $\theta_1 A_1$ for family membership as a random effect. When the IBD matrix A_2 for a particular genomic location is also available, the polygenic variance and quantitative trait locus (QTL) variance θ_2 will be allowed through the linear combination $\theta_1 A_1 + \theta_2 A_2$.

2.3.1 Meta-analysis

Suppose data from 15 studies are available, fixed and random effects models of meta-analysis can be furnished with the following code.

```
proc mixed method=reml;
     class studyid;
     model beta = / s cl;
     repeated / group=studyid;
     parms / parmsdata = g
             eqcons = 1 to 15;
run;
```

```
proc mixed method=reml covtest;
     class studyid;
     model beta = / s cl outp=predp;
     random studyid / g gdata = g s v;
     ods output CovParms=cp
                G=G V=V
                SolutionF=SF
                SolutionR=SR;
run;
```

The ODS statement clearly shows all the relevant statistics. In general, when *repeated / r* is specified, one can also include *R=R* in the statement to output the R matrix.

2.3.2 Association accounting for familial relationship

The following program performs linear and logistic regressions for family data via MIXED and GLIMMIX.

```
proc mixed method=ml covtest asycov
           noclprint
           noitprint
           noprofile;
     ods output FitStatistics=mixed1;
     class pid id;
     model q1=sex age smoke
              / noint notest;
     random int / type=lin(1) ldata=kmat
            sub=pid;
     parms (0.2) (0.1) / lowerb=0,0;
run;
proc glimmix method=mmpl asycov
            noclprint
            noitprint
            noprofile;
     ods output FitStatistics=glimmix1;
     class pid id;
     model affected(event='1')
           = sex age smoke
           / dist=binary link=logit;
     random int / type=lin(1)
                  ldata=kmat sub=pid;
     parms / lowerb=0,0;
     nloptions technique=newrap;
run;
```

To allow for IBD matrix, we have *type=lin(2)* and approriate change in the *parms* statement.

A critical procedure in SAS necessary for association analysis involving family data is PROC INBREED which can be used to obtain the kinship or relationship matrix between relatives in a pedigree. According to PROC INBREED documentation, individuals in a pedigree have to be ordered such that parents precede their children. As kinship calculation is readily available from the R `kinship` package, we have used an example pedigree shown in Figure 1 to examine this. The order of individuals is as follows: 2, 88, 8, 10, 20, 22, 24, 26, 18, 34, 12, 50, 56, 64, 66, 68, 70, 72, 1, 4, 6, 99, 28, 30, 36, 38, 40, 42, 44, 46, 52, 54, 48, 78, 60, 62, 80, 82, 3, 5, 7, 9, 11, 13, 15, 14, 16, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 73, 74, 75, 76, 77, 79, 59, 61, 63, 65, 32, 67, 69, 71, 58, 81, 83. Parental information can be read from Figure 1.

Table 1. Comparisons of SAS and R

Analysis	R functions	SAS procedures
Test of HWE	HWExact HWE.exact	PROC ALLELE
Meta-analysis	rma	PROC MIXED
Family data:		
Linear regression	`lmekin` `pedigreemm`	PROC MIXED PROC GLIMMIX
Logistic regression	`pedigreemm`	PROC GLIMMIX
Cox regression	`coxme`	

2.4 Comparison with R

Unlike SAS which has eased the C/C++ call until very recently in 9.2, R is more established with foreign language calls whose list includes Fortran, C/C++, Java, among others. It is perhaps also not surprising that appropriate functions can be found in R. For instance, function `HWE.exact` in package `genetics` is appropriate for exact test of HWE whereas function `HWExact` in package `GWASExactHW` is dedicated to test of HWE by calling the C/C++ program in [9]. For meta-analysis a close correspondence would be `metafor`, which implements the DerSimonian-Laird moment estimator in addition to the fixed and random effects models described here. More interestingly, for family data the function `lmekin` from the R package `kinship` and `pedigreemm` from package `pedigreemm` are appropriate for linear and logistic mixed models with the former can also utilize IBD information at a genomic location that is similar to PROC MIXED. A brief comparison is given in Table 1.

Both `lmekin` and `coxme` here are available from the R `kinship` package and allow for linear combination of variance components. Function `pedigreemm` is available from the R `pedigreemm` package and allows for count outcome but cannot accommodate linear combination of variance components. PROC GLIMMIX is similar to `pedigreemm` in their design to handle various types of outcomes. In analogy to PROC INBREED, function `kinship` from the R `kinship` package is used to obtain kinship matrix.

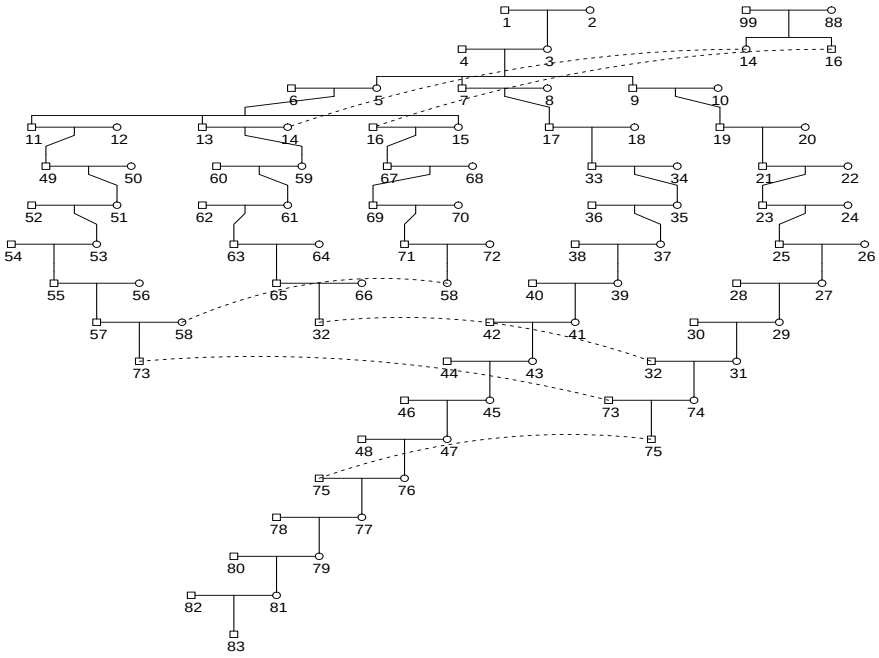

Figure 1. An example pedigree (dashed lines are used to connect the same individuals, figure generated by the R `kinship` package)

3 EXAMPLE APPLICATIONS

We here provide some timing information for the EPIC-Norfolk data. A full comparison has not been run as our recent analysis of the data has been centering around imputed data[10] which is beyond the focus of this paper.

For directly genotyped SNPs in the subcohort (2,417) individuals in the EPIC-Norfolk study using BMI as outcomes with additive coding adjusting for age took 1.5 hours on a single Linux node with 16G RAM for generating the long format file including allele coding, a few seconds for obtaining call rates, 1 hour for summary statistics, 15 minutes for linear regression.

3.1 Exact test of HWE

For rs1121980, whose three genotypes counts are 6615, 9953, 3774, the exact p value for test of HWE can be obtained as follows,

```
options cmplib=(work work.mathfun);
data abc;
     input a b c;
     pHWE=HWE(b,a,c);
     datalines;
     6615 9953 3774
run;
proc print;
     format pHWE 20.15;
run;
```

This gives an exact p value of 0.79.

3.2 Meta-analysis

The data as reported on rs9939609 near *FTO* and BMI in [11] are reproduced below.

```
data g;
     input beta se studyid$;
     col=_n_;
     row=_n_;
     value=se*se;
cards;
 0.35 0.09 CoLaus
 0.16 0.13 SardiNIA
 0.19 0.11 EPIC-Obesity
-0.04 0.13 NHS
 0.20 0.11 PLCO
 0.08 0.16 KORA
 0.11 0.16 WTCCC-Controls
 0.11 0.17 BC58
 0.14 0.17 DGI-Controls
 0.24 0.15 FUSION-Controls
```

```
-0.06 0.13 WTCCC-HT
 0.24 0.14 WTCCC-CAD
 0.47 0.17 WTCCC-T2D
 0.32 0.18 DGI-T2D
 0.04 0.20 FUSION-T2D
;
```

The results from fixed and random effects model are fairly similar, with -2 restricted loglikelihood=-15.05, regression coefficient (standard error)= 0.1715 (0.03537), resulting a t=4.85 and p=0.003. The regression results are comparable to [10] but here it does not involve weighting by sample sizes. Potentially covariates are easily incorporated in both models, though in this example one can resort to the usual regression.

3.3 Test of association

The GAW17 pedigree data is contained in a SAS dataset called `pheno`, in which all individuals are uniquely labelled (`id`) and founders have missing information for father (`fa`) and mother (`mo`). PROC INBREED is called to set up the appropriate relationship matrix to be used by PROC MIXED and PROC GLIMMIX.

```
proc inbreed data=pheno covar
             outcov=amatrix;
     var id fa mo;
run;
data kmat;
     parm=1;
     row=_n_;
     set amatrix;
run;
```

We have used *VEGFC* and replicate one as an example, as it contains a causal variant for Q1. PROC MIXED has yielded a loglikelihood of -939.45 with kinship and similar results with both kinship and IBD matrices. For binary outcome (AFFECTED), the loglikelihood is -1708.00 with kinship and again similar results with both kinship and IBD matrices. It is not certain whether the latter result could be improved.

3.4 Comparison with R

As for association testing, function `lmekin` allows for the polygenic and QTL specific variances to be estimated, so that both linkage and association are allowed in a consistent framework. On the other hand, SAS has difficulty to obtain the QTL specific variance. For the *VEGFC* example `lmekin` has obtained a loglikelihood of -839.9436 with kinship alone compared to -832.0997 with additional IBD information, leading to a log-likelihood ratio statistic of 15.70 for a single degree of freedom. This shows that through testing nested models with and without IBD component it has successfully recovered the causal gene through linkage. The results on binary outcome (AFFECTED) are similar using PROC GLIMMIX, as with `pedigreemm` since the latter cannot accommodate IBD matrix at the same with kinship matrix.

Clearly, PROC INBREED can give wrong kinship coefficients if parents of pedigree founders are not coded as missing (.) but zero (0), which unfortunately is the most popular in human genetics. The order of individuals also matters a lot. On the other hand, the `kinship` function as in the R `kinship` package does not have such restriction.

4 DISCUSSION

Due to time restriction, we could not provide a full comparison using various strategies described here but only some snapshots of general software systems such as SAS and R. A complete picture could be given on all the timing data using the original long-formatted and data partition scheme, a pre-partitioned set of raw data as input to existing setup in SAS, or the informative wide-formatted with map information. At least practically, partitioning data of a whole study from a long format is laborious, it is natural to resort to a pre-partitioned version with the familiar setting. Examples as in [6] have involved iteration over SAS procedures, which would be less efficient compared to a single data step generating all the data to be used by analytical procedures later on (as can be seen from the macro we have written). We feel that there are a lot of advantages in data management and testing of association analysis with them. The drawback perhaps is that they do not have capacity for specific tasks in genetic analysis. In which case, standalone programs continue to have their place in the field of human genetics. For a general description of facilities in R for genetic analysis, the readers are advised to read through the comprehensive R archive network (CRAN, `http://cran.r-project.org`) task view (at the CRAN location /web/views/Genetics.html) where a C++ package `PLINK` has been referred. Note this is an up-to-date link than reported earlier [12].

Our work has enabled faster analysis including using program in C/C++ as well as a recent implementation of meta-analysis and analysis of family data. We have also indicated drawback of procedure such as PROC INBREED, making us tempting to calculate the required kinship information from R. This should add to the list of improvements needed to be made such as exchanging dash (-) and underscore (_) when sorting data [3] while in Stata there is no such a problem. It is worthwhile to mention another apparent defect which is associated with PROC IMPORT, that SAS would truncate character variables wider than those in any of the first 32,767 records. In fact, the current number of genetic variants is in the order of millions which effectively means that PROC IMPORT would definitely go wrong for a simple text file reading.

We should mention that it remains to explore other aspects such as imputed genotypes and haplotype analysis which require probability weighting, and longitu-

dinal family data which requires finer specification of the variance structure. To meet many demands from the consortium contribution within the limited time, we have wrapped purpose-written analytical software SNPTEST[10] in Stata, which has probability weighting as a invaluable facility and a close counterpart in R would be package survey whereas for SAS, procedures such as SURVEYREG will be of interest. In all of general packages we have reviewed earlier[2] here is no generic routine for IBD calculation. A reviewer has pointed to us SAS JMP Genomics, but we do not have experience with it.

5 CONCLUSION

The updates we present here will add to the advantages of general software systems including both SAS and R that have previously been described. At the same time, there remain ample opportunities to translate our scientific thinking by fully integrating a variety of data and models efficiently. In general, SAS offers a broader range of options from more models to detailed implementation, which would be critical for validity check and further development in R. However, this is not necessarily the case, as is seen from the GAW17 data where the comparison between the two environments has not favored SAS so much. As we have not been aware of any previous work in the literature involving family data as is discussed here, our experiences presented here will be rather limited. Further work is necessary to consolidate our findings.

ACKNOWLEDGEMENTS

We are grateful of Dr Giovanni Montana, Conference Chair of Computational Bioscience 2011, and Karen Lee, the IASTED Secretariat, for their help on various technical issues, as with three anonymous referees for their suggestions leading to improvement of the paper.

The EPIC-Norfolk study is supported by research programme grant funding from Cancer Research UK and the Medical Research Council. We wish to thank colleagues during EPIC-Norfolk GWAS in the past few years which are bullt on tremendous contributions from the study participants. We wish to than Prof Terry Therneau for his work on the S-PLUS `kinship` package, Lukas Keller for the example pedigree in Figure 1.

REFERENCES

[1] L.A. Hindorff, P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 106, 2009, 9362-9367.

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

[3] J.H. Zhao, J.A. Luan, Q. Tan, R.J.F. Loos and N. Wareham. Analysis of large genomic data in silico: The EPIC-Norfolk study of obesity. *D.S. Huang, L. Heutte and M. Loog (Eds)*. ICIC2007, CCIS 2, 781-790.

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

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

[6] A. M. Saxton (Ed.) *Genetic Analysis of Complex Traits*, Cary NC, USA, SAS Institute Inc. 2004.

[7] N. Day, S. Oakes, R. Luben, K.-T. Khaw, S. Bingham, et al. EPIC-Norfolk: study design and characteristics of cohort. European Prospective Investigation of Cancer. *Br J Cancer* 80 suppl 1, 1999, 95-103

[8] R.J.F. Loos, C.W. Lindgren, S. Li, E. Wheeler, J.H. Zhao, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet* 40(6), 2008, 768-775.

[9] J.E. Wigginton. D.J. Cutler and G.R. Abecasis. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* 76, 2005, 887-893

[10] J. Marchini, B. Howie, S. Myers, G. McVean, P. Donnelly. A new multipoint method for genome-wide association studies via imputation of genotypes. *Nat Genet*, 39, 2007, 906-913

[11] C.J. Willer, E.K. Speliotes, R.J.F. Loos, S. Li, C.W. Lindgren, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet*, 41, 2009, 25-34.

[12] J.H. Zhao, Q. Tan. Integrated analysis of genetic data with R. *Hum Genomics* 2(40, 2006, 258-265.