Analysis of Large Genomic Data *in Silico*: The EPIC-Norfolk Study of Obesity

Jing Hua Zhao[1], Jian'an Luan[1], Qihua Tan[2], Ruth Loos[1], and Nick Wareham[1]

[1] MRC Epidemiology Unit, The Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK
[2] Dept of Biochemistry, Pharmacology and Genetics, Odense University Hospital, Sdr. Boulevard 29, DK-5000, Odense C, Denmark
{Jing.Hua.Zhao,Jian'an.Luan,Qihua.Tan,Ruth.Loos, Nick.Wareham,jinghua.zhao}@mrc-epid.cam.ac.uk

Abstract. In human genetics, large-scale data are now available with advances in genotyping technologies and international collaborative projects. Our ongoing study of obesity involves Affymetrix 500k genechips on approximately 7000 individuals from the European Prospective Investigation of Cancer (EPIC) Norfolk study. Although the scale of our data is well beyond the ability of many software systems, we have successfully performed the analysis using the statistical analysis system (SAS) software. Our implementation trades memory with computing time and requires moderate hardware configuration. By using such an established system, it extends some earlier discussions in a more constructive and accessible way. We report our findings and give some recommendations with SAS. We also compare briefly with alternative implementations. Our work is relevant to researchers conducting analysis of large-scale data in general, and genomewide association studies in particular.

Keywords: Data mining, genomewide association, obesity, statistical analysis system.

1 Introduction

The problems associated with large data are common to many fields of research such as banking and marketing research, neuroimaging, remote sensing and weather forecasting. In human genetics, this is demonstrated by projects such as HAPMAP (http://www.hapmap.org) and genomewide association (GWA) studies of complex traits such as obesity and diabetes [1], [2], both involve many single nucleotide polymorphisms (SNPs)[3], the most abundant genetic variants in human genome. However, *in silico* approach to these data is far from adequate, as shown by calls for appropriate tools ([4], Genetic Analysis Workshops, http://www.gaworkshop.org); meanwhile the established systems and the practice remain to be separated entities and to some extent this has been accepted with a great deal of complacency among geneticists[5]. Reviews of computer software used in human genetics have been made

D.-S. Huang, L. Heutte, and M. Loog (Eds.): ICIC 2007, CCIS 2, pp. 781–790, 2007.

[6], [7], [8]. Other notable remarks are "most statisticians have their pet methods, which they are loath to give up"[4], "the number of haplotype analysis programs is equal to the number of statisticians" (Mark McCarthy, Personal communication).

Earlier, we have described the drawbacks of the current practice in relation to established systems [9], [10]. Here we use our GWA study of obesity as an example to illustrate this further. Our case-cohort design is based on the EPIC-Norfolk totaling about 25000 individuals, and the genomewide association study of obesity involves Affymetrix (http://www.affymetrix.com/index.affx) 500k genechips on about 7000 individuals, to be followed by Illumina (http://www.illumina.com) 317k genechips. Data of such a scale will be awkward to deal with by the current practice.

In the following sections, we describe pilot and full analyses of obesity in SAS (http://www.sas.com, http://en.wikipedia.org/wiki/SAS), a comprehensive software system with complete manuals available online, while referring to several other software systems. Although our target systems are SAS and Linux, their features are generic. All the computer programs are available from the authors. Our contributions, both as a workable example and as accessible tool to handle large scale data, along with solution to various problems, will be of considerable value to researchers facing similar problems. For ease of exposition, we give some background in Appendix, detailing the kind of analysis involving call rate calculation, Hardy-Weinberg equilibrium (HWE) tests, and regression analyses under several genetic models, with adjustment for multiple testing. SAS programs typically consist of a data step to prepare for data to be processed by other modules called procedures, in the following we will use capital letters to indicate SAS data step (DATA) and procedures (PRINT, SQL, etc). Occasionally, we also give codes.

2 Methods

To gain a solid ground about the kind of hardware and software systems required, and build our solution iteratively, we started with a pilot experiment on our current Linux system. This is based on a sample of 400 controls from the European Prospective Investigation of Cancer (EPIC) Norfolk study (http://www.srl.cam.ac.uk/epic/) each with SNPs from Perlegen (http://www.perlegen.com) 250k genechips (the EPIC 400 study). This is part of a screening sample from a multistage study of breast cancer in EPIC. We then extended this experiment to the full-scale of our obesity study which involves about 7000 individuals, 50% of which has SNP data on Affymetrix 500k at stage one.

2.1 The EPIC400 Study

The flowchart of our implementation is shown in Fig. 1. The input data have three sources of information, i.e., genotype, map and phenotype. The genotypic data contain actual genotypes for all individuals in the so-called long format (individual ID, SNP ID, and genotype). Map information show position of each SNP by chromosome. Phenotypic information is in the usual tabular format (individual ID, sex, body mass index and other measurements). These three sources of information are merged into a combined dataset for analysis. For the purpose of screening SNPs

of interest, only single point analysis was conducted. Call rates were obtained, as with HWE tests for all SNPs, the results of which were used as an inclusion/exclusion filter of SNPs in the regression analysis. For instance, SNP information including HWE tests can be obtained by chromosome and positions, as input to ALLELE procedure which accepts genotype and outputs summary information of SNPs, allele frequencies and genotype frequencies. The outputs are stored in ODS (output delivery system) databases by chromosomes and SNP positions, and all outputs for individual SNPs are suppressed. The raw genotype data and map information can be used to construct input files for HAPLOVIEW (http://www.broad.mit.edu/mpg/haploview/) for visualization. The SNPs involved can be submitted to ENSEMBL (http://www.ensembl.org/index.html) to obtain gene annotations.

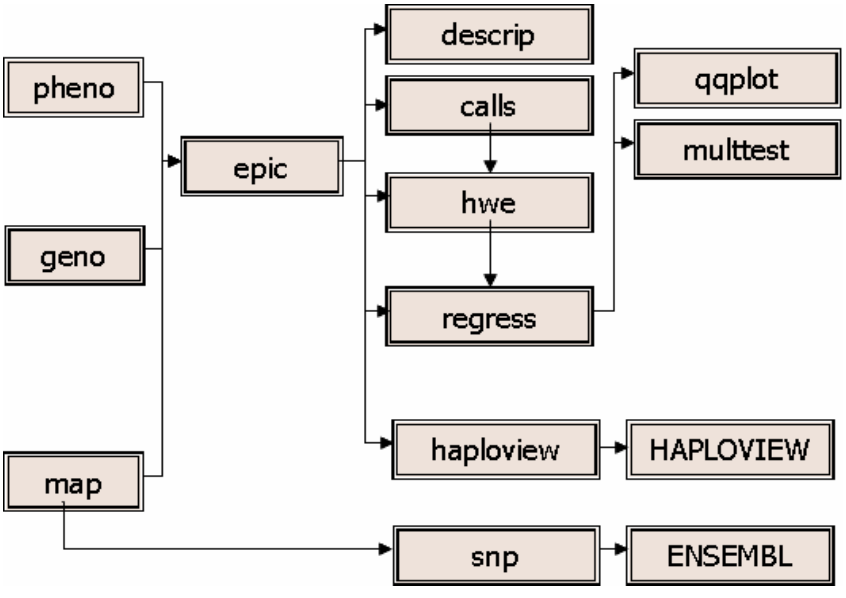

Fig. 1. A flowchart of the EPIC 400 analysis, with modules in brackets. Genotypes (geno) and phenotypes (pheno) are merged (epic) for descriptive statistics (descrip) call rates (calls), HWE (hwe), regression (regress) with adjustment for multiple testing (multtest) and comparison with theoretical distribution (qqplot). The raw data together with map information (map) can also be reformatted (haploview) into HAPLOVIEW input files so that specific region in the genome can be visualized, with annotation information from ENSEMBL according to SNPs (snp).

In a traditional statistical analysis, the data usually takes the so-called wide format, where rows indicate sample and columns variables. Since the number of SNPs is quite large, it is sensible to organize genotype data into the so-called long format. Although this requires larger amount of storage but the analysis is considerable simpler, for one can perform analysis for each SNP in sequence and store the results in a systematic fashion. We can take advantage of SAS/GENETICS module for HWE and haplotype analysis. All the outputs are available as databases for re-use and it is possible to generate data for external software programs such as HAPLOVIEW. The facility

of result database is possible with ODS. We have kept the comma-separated data in compressed format, to be readily processed by pipe mechanism in SAS.

2.2 The GWA Study

The EPIC400 analysis described above provides us a skeleton of the implementation to be described here. Most importantly, with our initial success we can stick to the same hardware and software systems. However, we soon realized that it is not feasible to run all the data, even individuals chromosomes using the simple code just we had.

Although our original plan was to use grid computing via SAS/CONNECT, it takes some efforts to tune the grid, we also try to resolve this by partitioning the data evenly within each chromosome *i, i=1,…22, X*, to be analyzed one at a time with the individual results later being assembled. The data partition works as follows, we obtain the number of records in the dataset and store into a macro variable in the SAS system environment (*N*), a group indicator is then created dynamically according to *ceil(s/N*_n_)*, where *s* is the number of partitions and *_n_* is the running record number from SAS. Partition or particular SNPs can be obtained via a *where* clause in the SQL procedure. We can simply call the macro a number of times each with a different chromosome number. With four nodes available during the experiment, we split chromosomes into four groups run separately on each node. The first few chromosomes are fairly long and 30 partitions were used, while most other chromosomes 20 partitions were sufficient. Similarly, we have used *where* clause to exclude individuals who failed quality control criteria in the analysis. The quality controls are with respect to call rate (>94%), discordance in SNP pairs with r^2=1 in HAPMAP (>5%), heterozygosity (<23%, or > 30%), concordance with another DNA (related or duplicated), the *where* clause to a SQL statement is simply *where id not in (select id from exclude);* where *exclude* contains the list of individuals to be excluded. The simplicity is remarkable. The analysis was performed while the data from Affymetrix were processed and examined. After initial results were obtained, additional data arrived. However, it is easy to incorporate the extra individuals. Since these data have exactly the same structure, we used pipe command such as *filename case pipe "gunzip –c data.gz extradata.gz"* to be used by SAS data step statement *infile case dlm='09'x*. No further code alteration is necessary. This feature is generic; for instance it will allow for reading data on all chromosomes at once.

To show further the complexity one might get into as an end-user of the genomic data (more details is available from the SAS program *rotate.sas* provided), the finding from the extracted data has been reported[11]. The Affymetrix 500k experiment was designed such that case and cohort samples are put in the same plate to reduce potential bias due to experimental conditions. Incidentally, at one time this remained to be the case when the data were distributed to us, so that case and cohort samples in these plates have been rotated 180 degrees according to some pre-specified rules, the genotype of a case actually corresponds to the genotype of another individual in the cohort. The information has been given in three files containing the plate number for cases, for cohort controls and rotation rules. The problem was solved as follows. First, replace the case dataset and cohort dataset each with an additional key variable called *code*, then merge the cases with the data containing rotation rules, each rule specifying the labels of source of plate and target plate. The key word *code* here is

also the label of source plate. Now rename *code* in the data just generated as *source*, and the target label to *code*. Then merge the combined data with the cohort sample, using key word *code,* with the genotypes in both cases and cohort renamed. We then merge the combined data with raw genotypes data using subject ID. We then swap the genotypes from cases and cohort. We finally append parts of the combined file involving cases and controls, with appropriate re-labeling. Clearly, this is a less daunting task for database management than using high-level language such as C/C++.

3 Results

The study results are largely standard concerning quality control including call rates, HWE, minor allele frequencies and results from regression analysis on additive models, some of which has been reported[11]. Independent case-control analyses have been performed within the MRC Epidemiology Unit, the Wellcome Trust Sanger Institute and the University of Cambridge and gave comparable results, through which 149 SNPs have been chosen for our second stage genotyping. We give some technical results here as it is prohibitive at this stage to compare the timing with alternative implementations. Our collaborators at Sanger Institute used Stata (http://www.stata.com), while colleagues at the University of Cambridge used customized standalone programs. Both were grid computing environments with Perl scripts and C programs to extract the raw data and feed them into Stata. Their implementations are thus more segmented and likely to be more difficult to use by others. Grid computing from Stata is recently available but the system is still lack of supported routines for genetic analysis. During our analysis, another independent work has been carried out, but our experiences showed the package requires[12] database management systems for allele coding and our experiences showed that it is yet to tune for actual analyses. In contrast, the software we developed is applicable to most research groups with moderate resources and for both Linux and Windows systems. The approach is also generic and not limited to genetic data.

The SAS datasets are approximately 30GB for the pilot study if some intermediate results are included; the running time is about a day or two on our Intel Linux systems with 2GB RAM. It was therefore estimated that the full obesity project is approximately 30 times larger but if we spread the task across chromosomes on a Linux cluster of 30 nodes the task can be furnished at similar speed. The GWA data used about 65GB disk space, while the long format with some phenotypic information (age, sex, BMI, case/control label and cohort indicator) required several folds larger, in total, just over 380GB. The major difference between the pilot study and the full GWA obesity case-cohort study was the need for data partitioning between and within chromosomes. As expected, the first batch (chromosomes 1-5) took the longest time whereas the fourth batch (chromosomes 18-22, X) was the fastest. Altogether, the whole analysis took about three days from data management, allele coding, phenotype merging, and statistical analyses.

Although SQL procedure provides functions such as *left join* but it can generate files of considerable size and it is much more efficient to use the in-line query given earlier, i.e. *select * from a where rsn in (select rsn from b)*, where *a,b* are datasets

containing the SNP name to be intersected. We have used this for selecting data for HAPLOVIEW or gene-specific analysis. We also found the performance of SAS procedures is heterogeneous. DATA step and procedures such as PRINT were less memory-efficient than SQL procedure. To extract the unique SNP IDs from the raw data, both SORT procedure with NODUPKEY option and SQL procedure with *unique()* function in the SELECT statement failed to work. However, one may use FREQ procedure since it can produce frequency tables for SNPs with non-missing data. We further noted some caveats associated with SAS. While the system was designed to handle large data, the implementation of its procedures is heterogeneous. For instance, we noted in general, SQL procedure is better than DATA step and PRINT procedure and can perform more sophisticated data management tasks. We do not necessarily need to segment the data in order to use MEANS and FREQ procedures. On the other hand, the SAS/GENETICS procedures often run into memory problems, but fortunately SAS has many alternative ways to do the same task. As the usage of disk space is quite heavy, it would be more useful to enable SAS/GENETICS procedures to read phenotype data separately. Over years the SAS language has been enriched but remains very stable and its powerful macro facility is also an extra advantage to many other software systems. We found that SAS uses a default value of 10 for the width of its character variable, which is insufficient to contain the full SNP names. When appending the segmented datasets the width of the variable *rsn* containing SNP names may be indicated as "*Variable rsn has different lengths on BASE and DATA files (BASE 13 DATA 10)*" from SAS's log, and SAS exits with warnings. However, we found that there is no loss of information despite this. When analyses are done by SNP names (*rsn*), SAS would change the SNP names containing dash to underscore (e.g. SNP_A-1969580 to SNP_A_1969580) so we can use *tranwrd(rsn, "A_","A-")* to convert it back in order to link with external data, e.g. gene annotation data.

4 Discussion

We have been able to furnish a timely analysis of our GWA analysis. Significant features of our implementation are its integrity, simplicity and generality. Although SAS can work with other database management systems its own SQL procedure is very powerful and comprehensive statistical analysis can be performed. Most tasks can be performed with small numbers of lines of coding. The programs we developed are modular and can run without change under Windows, with an MS-DOS batch file to call SAS from MS-DOS prompt. Besides being useful as it is, our implementation can potentially be used to prepare data and benchmark for standalone programs and software which require coded input.

We have focused on single-point analysis, and the problem can be more complex when multipoint analysis is involved[13]. We are yet to develop fully into other types of analysis available from SAS, e.g., principal component analysis and partial least squares method for structured association, cluster analysis for study of relatedness and outliers, covariance structure modeling for pathway analysis, just to name a few. As the SAS system is widely available, our work will be welcome.

Part of our next experiment will be for grid computing using clusters, which is now the state-of-the-art alternative to supercomputers. SAS/CONNECT is sufficient and it can maintain communications between nodes of a cluster, and between personal computers and remote server system. The functions include the data transfer between local and remote computers via data transfer, remote directory read/write and then task scheduling.

Acknowledgments. We wish to thank Wendi Qian for comments and colleagues in MRC Epidemiology Unit, the University of Cambridge, and Sanger Institute for computing and genotyping support, data quality control and many helpful discussions.

References

1. Grant, S. F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A. *et al.*: Variant of Transcription Factor 7-Like 2 (TCF7L2) Gene Confers Risk of Type 2 Diabetes. Nat Genet 38 (2006) 320-323
2. Herbert, A., Gerry, N. P., McQueen, M. B., Heid, I. M., Pfeufer, A., Illig, T., Wichmann, H. E., Meitinger, T., Hunter, D., Hu, F. B. *et al.*: A Common Genetic Variant is Associated with Adult and Childhood Obesity. Science 312 (2006) 279-283
3. Thomas, D. C., Haile, R. W., Duggan, D.: Recent Developments in Genomewide Association Scans: a Workshop Summary and Review. Am J Hum Genet 77 (2005) 337-345
4. Guo, S. W., Lange, K.: Genetic Mapping of Complex Traits: Promises, Problems, and Prospects. Theor Popul Biol 57 (2000) 1-11
5. Excoffier, L., Heckel, G.: Computer Programs for Population Genetics Data Analysis: A Survival Guide. Nat Rev Genet 7 (2006) 745-758
6. Dudbridge, F.: A Survey of Current Software for Linkage Analysis. Hum Genomics 1 (2003) 63-65
7. Weale, M. E.: A Survey of Current Software for Haplotype Phase Inference. Hum Genomics 1 (2004) 141-144
8. Salem, R. M., Wessel, J., Schork, N. J.: A Comprehensive Literature Review of Haplotyping Software and Methods for Use with Unrelated Individuals. Hum Genomics 2 (2005) 39-66
9. Zhao, J. H., Tan, Q.: Integrated Analysis of Genetic Data with R. Hum Genomics 2 (2006) 258-265
10. Zhao, J. H., Tan, Q.: Genetic Dissection of Complex Traits *in Silico*: Approaches, Problems and Solutions. Curr Bioinformatics 1 (2006) 359-369
11. Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Prry, J. R. B., Elliott, K. S., Lango, H., Rayner, N. W. *et al.*: A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. Science online (2007)
12. Clayton, D., Leung, H.-T.: An R Package for Analysis of Whole-Genome Association Studies. Hum Hered 64 (2007) 45-51
13. Zhao, J. H., Sham, P. C.: Faster Haplotype Frequency Estimation Using Unrelated Subjects. Hum Hered 53 (2002) 36-41
14. Olson, J. M., Witte, J. S., Elston, R. C.: Genetic Mapping of Complex Traits. Stat Med 18 (1999) 2961-2981

15. Elston, R. C., Anne Spence, M.: Advances in Statistical Human Genetics Over the Last 25 Years. Stat Med 25 (2006) 3049-3080
16. Balding, D. J.: A Tutorial on Statistical Methods for Population Association Studies. Nat Rev Genet 7 (2006) 781-791
17. Lander, E. S., Schork, N. J.: Genetic Dissection of Complex Traits. Science 265 (1994) 2037-2048
18. Risch, N., Merikangas, K.: The Future of Genetic Studies of Complex Human Diseases. Science 273 (1996) 1516-1517
19. Long, A. D., Grote, M. N., Langley, C. H.: Genetic Analysis of Complex Diseases. Science 275 (1997) 1328-1330
20. Kruglyak, L.: Prospects for Whole-Genome Linkage Disequilibrium Mapping of Common Disease Genes. Nat Genet 22 (1999) 139-144
21. Breslow, N. E.: Statistics in Epidemiology: the Case-control Study. J Am Stat Assoc 91 (1996) 14-28
22. Carlson, C. S., Eberle, M. A., Kruglyak, L., Nickerson, D. A.: Mapping Complex Disease Loci in Whole-Genome Association Studies. Nature 429 (2004) 446-452
23. Hirschhorn, J. N., Daly, M. J.: Genome-Wide Association Studies for Common Diseases and Complex Traits. Nat Rev Genet 6 (2005) 95-108
24. Wang, W. Y., Barratt, B. J., Clayton, D. G., Todd, J. A.: Genome-Wide Association Studies: Theoretical and Practical Concerns. Nat Rev Genet 6 (2005) 109-118
25. Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni J. P., Mane S. M., Mayne S. T. *et al.*: Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science 308 (2005) 385-389
26. Elston, R. C., Guo, X., Williams, L. V.: Two-Stage Global Search Designs for Linkage Analysis Using Pairs of Affected Relatives. Genet Epidemiol 13 (1996) 535-558
27. Holmans, P., Craddock, N.: Efficient Strategies for Genome Scanning Using Maximum-Likelihood Affected Sib-Pair Analysis. Am J Hum Genet 60 (1997) 657-666
28. Sham, P. C., Zhao, J. H.: The Power of Genome-Wide Sib Pair Linkage Scans for Quantitative Trait Loci Using the New Haseman-Elston Regression Method. Gene Screen 1 (2000) 103-106
29. Guo, X., Elston, R. C.: One-Stage Versus Two-Stage Strategies for Genome Scans. Adv Genet 42 (2001) 459-471
30. Satagopan, J. M., Verbel, D. A., Venkatraman, E. S., Offit, K. E., Begg, C. B.: Two-Stage Designs for Gene-Disease Association Studies. Biometrics 58 (2002) 163-170
31. Satagopan, J. M., Elston, R. C.: Optimal Two-Stage Genotyping in Population-Based Association Studies. Genet Epidemiol 25 (2003) 149-157
32. Satagopan, J. M., Venkatraman, E. S., Begg, C. B.: Two-Stage Designs for Gene-Disease Association Studies with Sample Size Constraints. Biometrics 60 (2004) 589-597
33. Thomas, D., Xie, R., Gebregziabher, M.: Two-Stage Sampling Designs for Gene Association Studies. Genet Epidemiol 27 (2004) 401-414
34. Skol, A. D., Scott, L. J., Abecasis, G. R., Boehnke, M.: Joint Analysis Is More Efficient Than Replication-Based Analysis for Two-Stage Genome-Wide Association Studies. Nat Genet 38 (2006) 209-213
35. Lin, D. Y.: Evaluating Statistical Significance in Two-Stage Genomewide Association Studies. Am J Hum Genet 78 (2006) 505-509
36. Wang, H., Thomas, D. C., Pe'er, I., Stram, D. O.: Optimal Two-Stage Genotyping Designs for Genome-Wide Association Scans. Genet Epidemiol 30 (2006) 356-368
37. Clerget-Darpoux, F., Bonaiti-Pellie, C., Hochez, J.: Effects of Misspecifying Genetic Parameters in LOD Score Analysis. Biometrics 42 (1986) 393-399

38. Curtis, D., Sham, P. C.: Model-Free Linkage Analysis Using Likelihoods. Am J Hum Genet 57 (1995) 703-716
39. Zhao, J. H., Curtis, D., Sham, P. C.: Model-Free Analysis and Permutation Tests for Allelic Associations. Hum Hered 50 (2000) 133-139
40. Hodge, S. E., Abreu, P. C., Greenberg D. A.: Magnitude of Type I Error When Single-Locus Linkage Analysis Is Maximized Over Models: A Simulation Study. Am J Hum Genet 60 (1997) 217-227
41. Nielsen, D. M., Ehm, M. G., Weir, B. S.: Detecting Marker-Disease Association by Testing for Hardy-Weinberg Disequilibrium at a Marker Locus. Am J Hum Genet 63 (1998) 1531-1540
42. Zou, G. Y., Donner, A.: The merits of testing Hardy-Weinberg equilibrium in the analysis of unmatched case-control data: a cautionary note. Ann Hum Genet 70 (2006) 923-933
43. Xu, J., Turner, A., Little, J., Bleecker, E. R., Meyers, D. A.: Positive Results in Association Studies Are Associated with Departure from Hardy-Weinberg Equilibrium: Hint for Genotyping Error? Hum Genet 111 (2002) 573-574
44. Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J., Gauderman, W. J.: Exploiting Gene-Environment Interaction to Detect Genetic Associations. Hum Hered 63 (2007) 111-119
45. Langholz, B., Rothman, N., Wacholder, S., Thomas, D. C.: Cohort Studies for Characterizing Measured Genes. J Natl Cancer Inst Monogr 26 (1999) 39-42
46. Manolio, T. A., Bailey-Wilson, J. E., Collins, F. S.: Genes, Environment and the Value of Prospective Cohort Studies. Nat Rev Genet 7 (2006) 812-820
47. Cai, J., Zeng, D.: Sample Size/Power Calculation for Case-Cohort Studies. Biometrics 60 (2004) 1015-1024

Appendix: Some Background of Design and Analysis of GWA

Reviews on statistical genetics and genetic association are available[14], [15], [16]. Due to the availability of large number of genetic variants and particularly SNPs, linkage design of family data examining co-segregation of genetic markers and putative disease loci has shifted to association designs using both family-based and population-based data, as foreseen over ten years ago[17], [18], [19], [20] when comparison[18] of affected sib-pair linkage, transmission/disequilibrium association tests showed that association tests are far more powerful. Case-control design has been the most established, with substantial contributions from biostatisticians[21] and is advantageous over family-based designs with its ease to implement. Association studies can be readily carried out and/or become an integral part in many established epidemiological cohorts. To reduce cost without compromising statistical efficiency, staged design is increasingly used due to the ever increasing scale of studies[22], [23], [24], which typically involve tens of thousands of SNPs, believed to hold the key to common diseases and population history; several recent papers provided such evidence [1], [2], [25]. The early work on staged design was in line with the development in genetic epidemiology in general, e.g. linkage studies [26], [27], [28], [29], followed by association studies including WGA [30], [31], [32], [33], [34], [35], [36].

A widely discussed topic in genetic epidemiology or statistical genetics is the so-called model-based or model-free method [37], [38], [39], [40], indicating the mode of inheritance of the disease locus, e.g. recessive, dominant and additive models, where one codes the three genotypes according to the number of minor or less

frequent allele of a SNP. Under additive model, the number of minor alleles range from 0 to 2. Under a dominant model, any genotype containing at least one minor allele is coded into 1 otherwise 0. Under a recessive model, a genotype is coded as 1 only if both alleles are minor. Moreover, HWE (http://en.wikipedia.org/wiki/Hardy-Weinberg_principle) test is a customary task to furnish[41], [42], [43], [44]. Indeed, experiences of many researchers including the authors have indicated deviation from HWE may suggest genotyping error, although this is less clear with large-scale genome data. It is useful to consider HWE in conjunction with call rate, the proportion of successful SNPs on a particular individual. Correlation analysis can be done between HWE and call rates to indicate genotype errors. It is worthwhile to note that for chromosome X, we have implemented HWE tests in women only, and women and men combined, considering men contribute only one copy of the alleles on chromosome X. A number of issues need to be considered for further analysis, including multi-locus modeling such as haplotype analysis[39], and meta-analysis across studies since not individual data are available from all studies. As noted earlier, comprehensive systems such as SAS will prove to be more useful in these settings.

Seeing that case-control design is easy to implement but with the drawback of control samples being highly selected, which can potentially lead to bias when used for any other purposes, our GWA study uses a case-cohort design, where a random sample of individuals is selected as controls. The sub-cohort is representative and can be used to compare against a wider range of phenotypes. The merits of cohort design in genetic association has been recognized recently[45], [46]. In addition, case-control studies can be nested within the cohort while a case-cohort design uses a randomly selected subset of the cohort. The case-cohort design is comparable to a two-stage case-control study in which each stage consisting of approximately 1700 cases and controls according to recent published work[47], giving larger cohort sample of about 2500 at each stage but allowing for some cases to be included. Reports[34], [35] suggested that 50:50 split of study samples between stages 1 and 2 achieves optimal power. However, compared to many studies our stage one has considerable more power to justify our SNP selection for the second stage genotyping.