Bioinformatics of genomic association mapping

Vaez Barzani, Ahmad

Chapter 1

General introduction

Preface

The successful completion of the Human Genome Project marked a milestone for genomics research (1). Likewise, new microarray technologies have become widely available providing simple and elegant ways of rapidly delivering vast amounts of biological data. Moreover, three major, multi-country projects have been launched in the past few years to improve the understanding of genomic diversity: The International HapMap Project (2), the 1000 Genomes Project (3), and the Encyclopedia of DNA Elements (ENCODE) project (4). All of the data of these projects have been made publicly available to the scientific community (Box 1, (5)). However, generating and storing vast amounts of biological data only makes sense if data are creatively used for further research to produce novel insights into biological mechanisms (6). Bioinformatics is an interdisciplinary field to address this growing need. Hence, bioinformatics has been one of the most dynamically evolving fields of science in the past few years. The Wikipedia encyclopedia describes bioinformatics as an interdisciplinary field that 'develops' and 'applies' methods and software tools for understanding biological data. It combines computer science, statistics, mathematics, and engineering to analyze biological data. Bioinformatics is an umbrella term for both (i) methodological studies that 'develop' novel algorithms, software packages, or analysis pipelines, and (ii) biological studies that 'apply' those tools as part of their methodology (7).

Genetic epidemiology studies the role of genetic factors in health and disease in families and/or populations (8). It was defined by Morton as "a science which deals with the etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations" (9). The vast availability of biological data as well as handiness of modern bioinformatics-based methods have yielded rapid advances in discovering genetic variants underlying different human diseases or traits, so-called genomic association mapping (10, 11).

Unlike rare Mendelian disorders, common complex diseases or traits, such as ischemic heart disease, hypertension, or serum levels of inflammatory markers, are controlled by the complex interplay between multiple genetic and environmental factors (11). For that reason gene mapping of complex traits or diseases had been largely disappointing in the past. Genomic association mapping is now a systematic approach aiming to find genetic variants that are associated with the trait or disease of interest. Thanks to inexpensive availability of microarray data and also thanks to accessibility of bioinformatics pipelines for data analysis,

Box 1 Timeline of key events of major projects facilitating genomic association mapping.

1990	Human Genome Project was launched
2001	Draft of Human Genome Project was published
2002	International HapMap Project was launched
2003	Human Genome Project was completed
2003	International HapMap Project first major public data released
2003	ENCODE project was launched
2007	Pilot phase of the ENCODE project was finished
2008	1000 Genomes Project was launched
2010	International HapMap Project latest public data released
2010	Phase 1 of 1000 Genomes Project was completed
2012	ENCODE project initial results were published
2014	Phase 3 of 1000 Genomes Project data released

genomic mapping of complex traits or common diseases has become very successful in recent years (11).

Genomic association mapping consists of a number of classic steps as described in Figure 1. First we should check if the trait or disease of interest is appropriate for genomic mapping. This can be done by a typical heritability study, that is investigating the proportion of variation in a trait that is due to genetic differences between individuals. If there is evidence of heritability, the trait is (partially) controlled by genetic factors and genomic association mapping is a logical next step (12).

Once it has been established that the trait of interest is heritable, we should attempt to identify genetic variants that are significantly associated with the trait. Although a number of different methodologies can be used, the experimental design of the genome-wide association study (GWAS) has been very successful and played a key role in genomic association mapping during the past few years (11). GWAS analysis is typically aimed at detecting genetic variants associated with common complex traits or diseases without prior hypothesis of functionality.

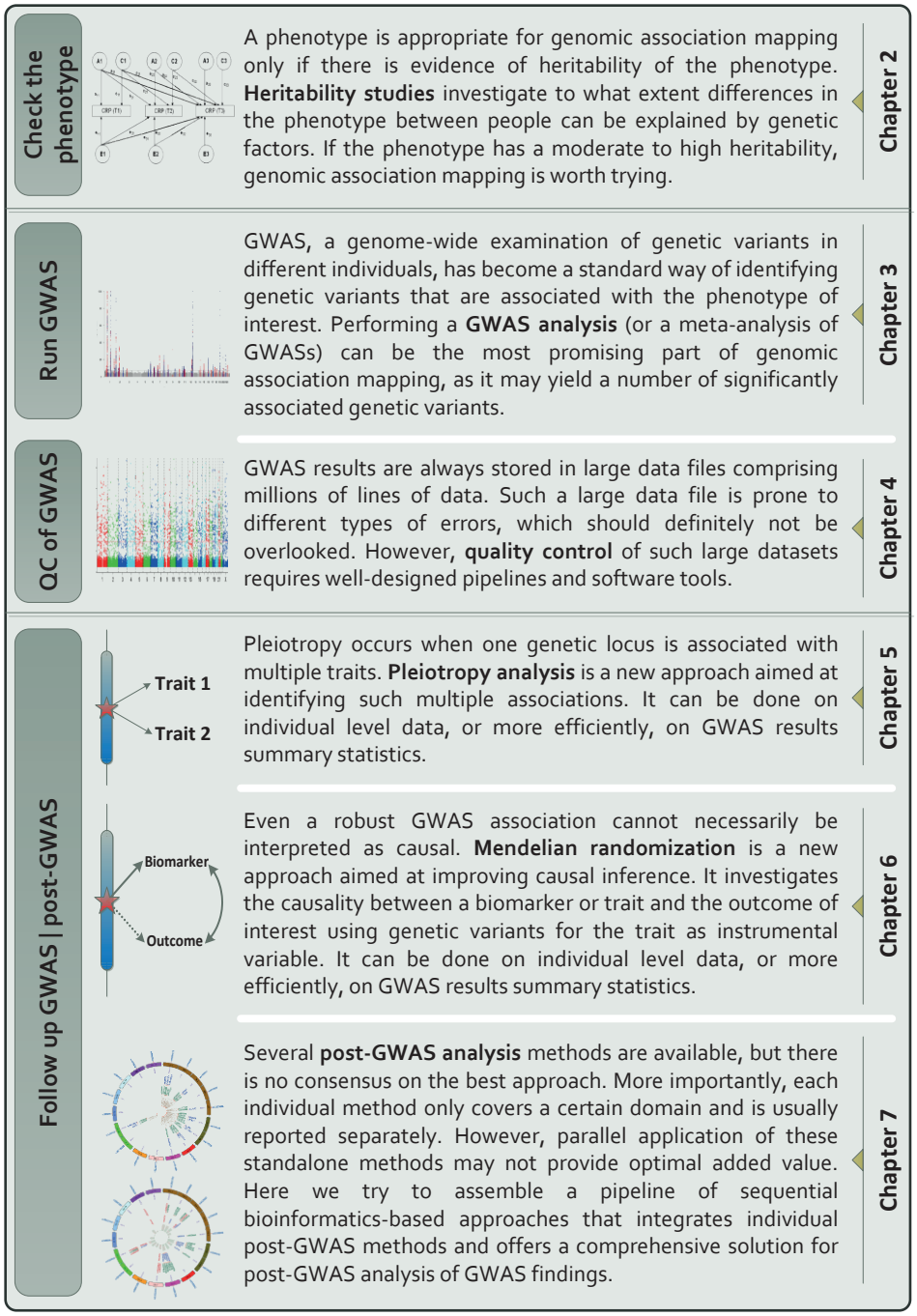

Figure 1 Flow diagram of sequential bioinformatics-based genetic epidemiological approaches for genomic association mapping. GWAS indicates genome-wide association study; and QC, quality control.

GWAS analysis is always done on millions of single-nucleotide polymorphisms (SNPs) per individual for thousands of individuals, and the statistical analyses may differ by genotyping array, imputation platform (13) data management pipeline, analysis design, and GWAS software package (14–16). The output of such sophisticated analysis is usually stored in large data files with millions of lines of data. Such a large data file may contain different kinds of (minor or major) errors, which should be carefully identified and fixed.

Once GWAS analysis is done and its quality is stringently checked and possible errors are fixed, we will have a list of significantly associated genetic variants, which merely flag the associated genomic loci (17). Hence, extra steps of genomic association mapping are inevitable to translate those GWAS findings to biological knowledge, so-called post-GWAS analyses (18). There is a great variety of post-GWAS methods for following up GWAS results. These methods can be broadly categorized into two major groups: (i) dry lab (*in silico*) approaches, which are based on reanalyzing (publicly) available biomedical datasets in order to unravel functional mechanisms underlying GWAS signals (6), such as functional network analysis, and (ii) wet lab (*in vitro* or *in vivo*) approaches, which are based on classic laboratory methods of working on cell lines, animals models, or human samples. Obviously, dry lab (*in silico*) approaches are faster, easier, and cheaper compared to wet lab approaches and hence should be tried first. The findings of dry lab studies can subsequently be backed up by further wet lab experiments (18).

Scope of the thesis

In this thesis we sought to work on bioinformatics-based genetic epidemiological approaches for genomic association mapping, with emphasis on human quantitative traits and their contribution to complex diseases. This thesis had two principal aims as indicated in Figure 2 and Table 1:

(i) To 'develop' new bioinformatics-based tools (software packages and/or analysis pipelines) supporting appropriate analysis of high-throughput biological data for genomic association mapping.

(ii) To 'apply' bioinformatics-based tools on biological data to gain knowledge of underlying mechanisms controlling complex traits or diseases.

Considering aim ii, and as a running example of a typical human complex trait, we worked on serum levels of C-reactive protein (CRP). CRP is a marker of

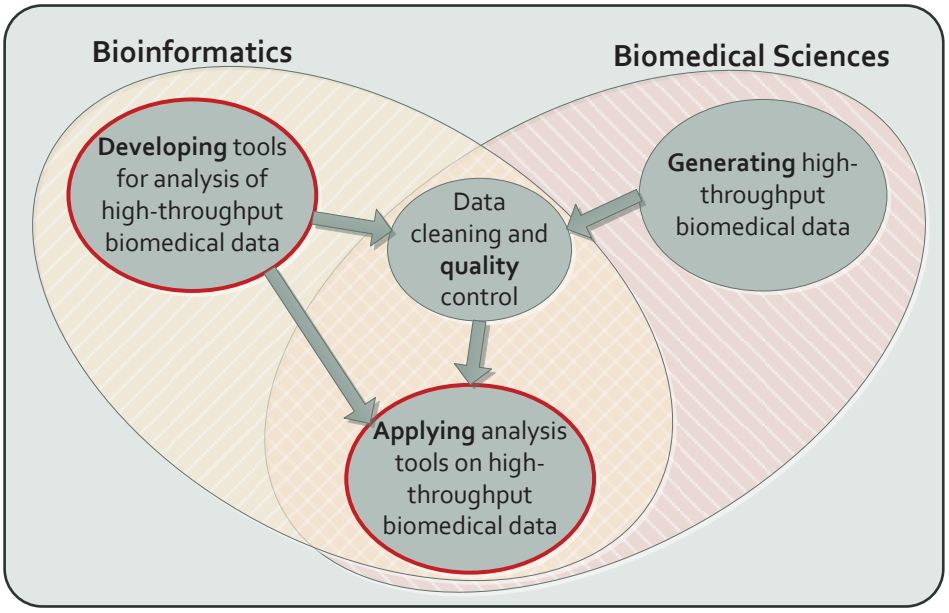

Figure 2 Bioinformatics is an umbrella term for either methodological studies that 'develop' new tools, or biomedical studies that 'apply' those tools in order to gain knowledge. The two red circles indicate the two principal aims of the current thesis.

systemic inflammation and its levels, like other complex traits, are controlled by a variety of environmental and genetic factors. So far, a number of environmental and genetic factors are recognized to influence CRP levels. However, the extent to which these factors account for the total variation in CRP levels, and more importantly, the exact mechanisms that underlie the regulation of CRP levels are still unknown (19). The heritability of CRP is estimated to range from 10% to 65% (20–23) and GWAS analyses have successfully identified several genetic variants associated with serum levels of CRP (24–26). However, although those GWASs have included tens of thousands of individuals, the explained variance in CRP levels by all identified variants is not more than five percent (26). Likewise, although CRP levels are already associated with many diseases or outcomes (27–33), its causal contribution to the pathophysiology of those diseases remains highly controversial (34–36), warranting investigation of the underlying mechanisms that control serum levels of CRP.

Outline of the thesis

This thesis is organized in four distinct parts and includes eight chapters (including this general introduction). The next six chapters address the typical sequential bioinformatics-based genetic epidemiological approaches for genomic association mapping (Figure 1). The final chapter provides a general discussion on the results and implications of this dissertation.

Part 1 | How to check if the phenotype is appropriate for genomic association mapping

In **Chapter 2**, we sought to identify the proportion of variation in serum levels of CRP that is explained by genetic factors. To conduct a heritability analysis, we used data from TwinsUK, a well-known adult twin registry (37). By including the data of CRP levels on more than 6,000 twins, our study was one of the largest heritability studies of CRP. Furthermore, and for the first time, we used longitudinal follow-up measurements of CRP levels. Taking advantage of multiple within-subject CRP measurements over time, we were able to estimate the (in)stability of the effects of genetic and environmental factors with advancing age.

Part 2 | How to perform a GWAS analysis

In **Chapter 3**, we sought to provide an appropriate solution for GWAS analysis of those biomarkers whose measurement assays are restricted by the problem of limit of detection (LOD). Although several GWAS software packages are already available (14–16), GWAS analysis of those biomarkers with the problem of LOD is poorly addressed. Here we provide a statistical solution as well as a software package for GWAS analysis of such biomarkers accounting for LOD.

Table 1 List of chapters of this thesis, indicating whether they 'develop' or 'apply' bioinformatics-based approaches for genomic association mapping.

Chapters	To 'develop' bioinformatics-based approaches	To 'apply' bioinformatics-based approaches
Chapter 2		√
Chapter 3	√	√
Chapter 4	√	
Chapter 5		√
Chapter 6		√
Chapter 7	√	√

In **Chapter 4**, we sought to provide an appropriate solution for quality control (QC) of GWAS result files. GWAS results are prone to different types of errors, which should be detected and fixed. Here we provide a QC pipeline as well as a software package for stringent quality control of GWAS result output files.

Part 3 | How to follow up GWAS results: post-GWAS analyses

In **Chapter 5**, we sought to investigate the pleiotropic genetic loci among CRP and lipids. Considering the previously known link between the biology of CRP and lipids (31, 38–41), we followed up the GWAS results of large meta-GWASs on serum levels of CRP (26) and lipids (42). We applied a new method and software package that combines the summary statistics of those meta-GWASs aiming to find pleiotropic genetic loci among two phenotypes.

In **Chapter 6**, we sought to investigate the causality between CRP and somatic and psychiatric complex outcomes. To this end, we again followed up the results of the large scale meta-GWAS on serum levels of CRP (26) and used them to investigate GWAS results of 32 complex outcomes. To perform efficient Mendelian randomization (MR) analyses, we applied a new method and software package that combines the summary statistics of those meta-GWASs aiming to test the causality between CRP and those outcomes (43).

In **Chapter 7**, we sought to build an integrated pipeline of sequential bioinformatics-based approaches that can be used for a systematic follow-up of GWAS results of any human trait or disease. Then we applied our pipeline to the GWAS results of the aforementioned large-scale meta-GWAS on serum levels of CRP (26). In this chapter, we provide the details of our in-house developed pipeline along with this example application, which illustrates how GWAS results of any trait or disease can be followed up and translated to biological knowledge.

Part 4 | Bioinformatics of genomic association mapping: an A to Z walk-through

Finally, **Chapter 8** is devoted to a general discussion of the previous chapters by providing an A to Z walk-through of classic steps of genomic association mapping. In this chapter, we summarize and discuss the main findings of the dissertation, as well as elaborate on implications and prospects for further research.

References

1. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

2. Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al. (2003) The International HapMap Project. *Nature*, **426**, 789–796.

3. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E. and McVean, G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

4. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

5. Fox, C.S., Hall, J.L., Arnett, D.K., Ashley, E.A., Delles, C., Engler, M.B., Freeman, M.W., Johnson, J.A., Lanfear, D.E., Liggett, S.B., et al. (2015) Future Translational Applications From the Contemporary Genomics Era A Scientific Statement From the American Heart Association. *Circulation*, 10.1161/CIR.0000000000000211.

6. Call for data analysis papers (2014) *Nat. Genet.*, **46**, 213–213.

7. Bioinformatics, Wikipedia Free Encycl. Available: http://en.wikipedia.org/wiki/Bioinformatics. Accessed 29 April 2015.

8. Genetic epidemiology, Wikipedia Free Encycl. Available: http://en.wikipedia.org/wiki/Genetic_epidemiology. Accessed 29 April 2015.

9. Morton, N.E. (1982) Outline of Genetic Epidemiology S Karger Pub, Basel u.a.

10. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci.*, **106**, 9362–9367.

11. Visscher, P.M., Brown, M.A., McCarthy, M.I. and Yang, J. (2012) Five Years of GWAS Discovery. *Am. J. Hum. Genet.*, **90**, 7–24.

12. van Dongen, J., Slagboom, P.E., Draisma, H.H.M., Martin, N.G. and Boomsma, D.I. (2012) The continuing value of twin studies in the omics era. *Nat. Rev. Genet.*, **13**, 640–653.

13. Marchini, J. and Howie, B. (2010) Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, **11**, 499–511.

14. Aulchenko, Y.S., Ripke, S., Isaacs, A. and van Duijn, C.M. (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, **23**, 1294–1296.

15. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

16. Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913.

17. Wang, X., Prins, B.P., Sõber, S., Laan, M. and Snieder, H. (2011) Beyond Genome-Wide Association Studies: New Strategies for Identifying Genetic Determinants of Hypertension. *Curr. Hypertens. Rep.*, **13**, 442–451.

18. Freedman, M.L., Monteiro, A.N.A., Gayther, S.A., Coetzee, G.A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D., et al. (2011) Principles for the post-GWAS functional characterization of cancer risk loci. *Nat. Genet.*, **43**, 513–518.

19. Kathiresan, S., Larson, M.G., Vasan, R.S., Guo, C.-Y., Gona, P., Keaney, J.F., Wilson, P.W.F., Newton-Cheh, C., Musone, S.L., Camargo, A.L., et al. (2006) Contribution of Clinical Correlates and 13 C-Reactive Protein Gene Polymorphisms to Interindividual Variability in Serum C-Reactive Protein Level. *Circulation*, **113**, 1415–1423.

20. Neijts, M., van Dongen, J., Kluft, C., Boomsma, D.I., Willemsen, G. and de Geus, E.J.C. (2013) Genetic architecture of the pro-inflammatory state in an extended twin-family design. *Twin Res. Hum. Genet.*, **16**, 931–940.

21. Saunders, C.L. and Gulliford, M.C. (2006) Heritabilities and shared environmental effects were estimated from household clustering in national health survey data. *J. Clin. Epidemiol.*, **59**, 1191–1198.

22. Rahman, I., Bennet, A.M., Pedersen, N.L., de Faire, U., Svensson, P. and Magnusson, P.K.E. (2009) Genetic Dominance Influences Blood Biomarker Levels in a Sample of 12,000 Swedish Elderly Twins. *Twin Res. Hum. Genet.*, **12**, 286–294.

23. Su, S., Miller, A.H., Snieder, H., Bremner, J.D., Ritchie, J., Maisano, C., Jones, L., Murrah, N.V., Goldberg, J. and Vaccarino, V. (2009) Common Genetic Contributions to Depressive Symptoms and Inflammatory Markers in Middle-Aged Men: The Twins Heart Study. *Psychosom. Med.*, **71**, 152–158.

24. Ridker, P.M., Pare, G., Parker, A., Zee, R.Y.L., Danik, J.S., Buring, J.E., Kwiatkowski, D., Cook, N.R., Miletich, J.P. and Chasman, D.I. (2008) Loci Related to Metabolic-Syndrome Pathways Including LEPR,HNF1A, IL6R, and GCKR Associate with Plasma C-Reactive Protein: The Women's Genome Health Study. *Am. J. Hum. Genet.*, **82**, 1185–1192.

25. Elliott, P., Chambers, J.C., Zhang, W., Clarke, R., Hopewell, J.C., Peden, J.F., Erdmann, J., Braund, P., Engert, J.C., Bennett, D., et al. (2009) Genetic Loci

associated with C-reactive protein levels and risk of coronary heart disease. *JAMA J. Am. Med. Assoc.*, **302**, 37–48.

26. Dehghan, A., Dupuis, J., Barbalic, M., Bis, J.C., Eiriksdottir, G., Lu, C., Pellikka, N., Wallaschofski, H., Kettunen, J., Henneman, P., et al. (2011) Meta-Analysis of Genome-Wide Association Studies in >80 000 Subjects Identifies Multiple Loci for C-Reactive Protein Levels. *Circulation*, **123**, 731–738.

27. Allin, K.H., Bojesen, S.E. and Nordestgaard, B.G. (2009) Baseline C-Reactive Protein Is Associated With Incident Cancer and Survival in Patients With Cancer. *J. Clin. Oncol.*, **27**, 2217–2224.

28. Dehghan, A., Kardys, I., Maat, M.P.M. de, Uitterlinden, A.G., Sijbrands, E.J.G., Bootsma, A.H., Stijnen, T., Hofman, A., Schram, M.T. and Witteman, J.C.M. (2007) Genetic Variation, C-Reactive Protein Levels, and Incidence of Diabetes. *Diabetes*, **56**, 872–878.

29. Sesso, H.D., Buring, J.E., Rifai, N., Blake, G.J., Gaziano, J.M. and Ridker, P.M. (2003) C-reactive protein and the risk of developing hypertension. *JAMA J. Am. Med. Assoc.*, **290**, 2945–2951.

30. Danesh, J., Wheeler, J.G., Hirschfield, G.M., Eda, S., Eiriksdottir, G., Rumley, A., Lowe, G.D.O., Pepys, M.B. and Gudnason, V. (2004) C-Reactive Protein and Other Circulating Markers of Inflammation in the Prediction of Coronary Heart Disease. *N. Engl. J. Med.*, **350**, 1387–1397.

31. Kaptoge, S., Di Angelantonio, E., Lowe, G., Pepys, M.B., Thompson, S.G., Collins, R. and Danesh, J. (2010) C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. *Lancet*, **375**, 132–140.

32. De Berardis, D., Conti, C.M., Campanella, D., Carano, A., Scali, M., Valchera, A., Serroni, N., Pizzorno, A.M., D'Albenzio, A., Fulcheri, M., et al. (2008) Evaluation of C-reactive protein and total serum cholesterol in adult patients with bipolar disorder. *Int. J. Immunopathol. Pharmacol.*, **21**, 319–324.

33. Harris, T.B., Ferrucci, L., Tracy, R.P., Corti, M.C., Wacholder, S., Ettinger Jr, W.H., Heimovitz, H., Cohen, H.J. and Wallace, R. (1999) Associations of elevated Interleukin-6 and C-Reactive protein levels with mortality in the elderly. *Am. J. Med.*, **106**, 506–512.

34. Tremblay, J. (2007) Genetic determinants of C-reactive protein levels in metabolic syndrome: a role for the adrenergic system? *J. Hypertens.*, **25**, 281–283.

35. Ummarino, D. and Zeng, L. (2013) Is C reactive protein expression affected by local microenvironment? *Heart*, **99**, 514–515.

36. Danik, J.S. and Ridker, P.M. (2007) Genetic determinants of C-reactive protein. *Curr. Atheroscler. Rep.*, **9**, 195–203.

37. Moayyeri, A., Hammond, C.J., Hart, D.J. and Spector, T.D. (2013) The UK Adult Twin Registry (TwinsUK Resource). *Twin Res. Hum. Genet.*, **16**, 144–149.

38. Mendall, M.A., Patel, P., Ballam, L., Strachan, D. and Northfield, T.C. (1996) C Reactive protein and its relation to cardiovascular risk factors: a population based cross sectional study. *BMJ*, **312**, 1061–1065.

39. Kraja, A.T., Province, M.A., Arnett, D., Wagenknecht, L., Tang, W., Hopkins, P.N., Djoussé, L. and Borecki, I.B. (2007) Do inflammation and procoagulation biomarkers contribute to the metabolic syndrome cluster? *Nutr. Metab.*, **4**, 28.

40. Sakkinen, P.A., Wahl, P., Cushman, M., Lewis, M.R. and Tracy, R.P. (2000) Clustering of Procoagulation, Inflammation, and Fibrinolysis Variables with Metabolic Factors in Insulin Resistance Syndrome. *Am. J. Epidemiol.*, **152**, 897–907.

41. Ridker, P.M., Rifai, N., Rose, L., Buring, J.E. and Cook, N.R. (2002) Comparison of C-Reactive Protein and Low-Density Lipoprotein Cholesterol Levels in the Prediction of First Cardiovascular Events. *N. Engl. J. Med.*, **347**, 1557–1565.

42. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–713.

43. Johnson, T. (2012) Efficient calculation for multi-SNP genetic risk scores; (Abstract 1400W). Presented at the 62th Annual Meeting of The American Society of Human Genetics, Nov 7, 2012 in San Francisco, CA.

Part 1 | How to check if the phenotype is appropriate for genomic association mapping