Bioinformatics of genomic association mapping

Vaez Barzani, Ahmad

Chapter 8

General discussion

Bioinformatics of genomic association mapping

This dissertation aimed to provide a comprehensive walk-through of the classic steps of genomic association mapping using bioinformatics-based approaches. This walk-through started with a classic heritability study, continued with providing novel tools for genome-wide association studies (GWASs) of complex traits, and ended with an integrated post-GWAS pipeline for translating GWAS findings of any human trait or disease to biological knowledge (Figure 1). Our A-to-Z approach emphasized the importance of following the consecutive steps of genomic association mapping. To support this process we sought to 'develop' and/or 'apply' bioinformatics-based tools, such as technical algorithms, software packages, analysis pipelines, etc., for the analysis of high-throughput biological data.

Genomics of complex traits

In contrast to rare Mendelian disorders, common complex traits or diseases, also known as multifactorial or polygenic disorders, are controlled by a complex network of environmental and genetic factors. Examples are metabolic syndrome, hypertension, coronary heart disease, type 2 diabetes, autoimmune diseases, etc. For a long time genomic (linkage) mapping of common complex diseases or traits had been deeply disappointing, mainly because of the complex interplay between multiple genetic and environmental factors. It is only in recent years that new, more successful approaches of genomic association mapping have become available, promising gains in knowledge of biological mechanisms underlying common complex traits and diseases (1, 2).

To illustrate the walk-through of the classic steps of genomic association mapping, and as a running example of a typical human complex trait, we worked on serum levels of C-reactive protein (CRP), a well-known marker of systemic inflammation that has previously been associated with a variety of diseases or outcomes (3–9). Serum levels of CRP are controlled by different environmental and genetic factors (10, 11). However, the exact biological and genomic mechanisms that control serum levels of CRP, and more importantly, the causal contribution of CRP to the pathophysiology of associated diseases were still largely unknown (12–15).

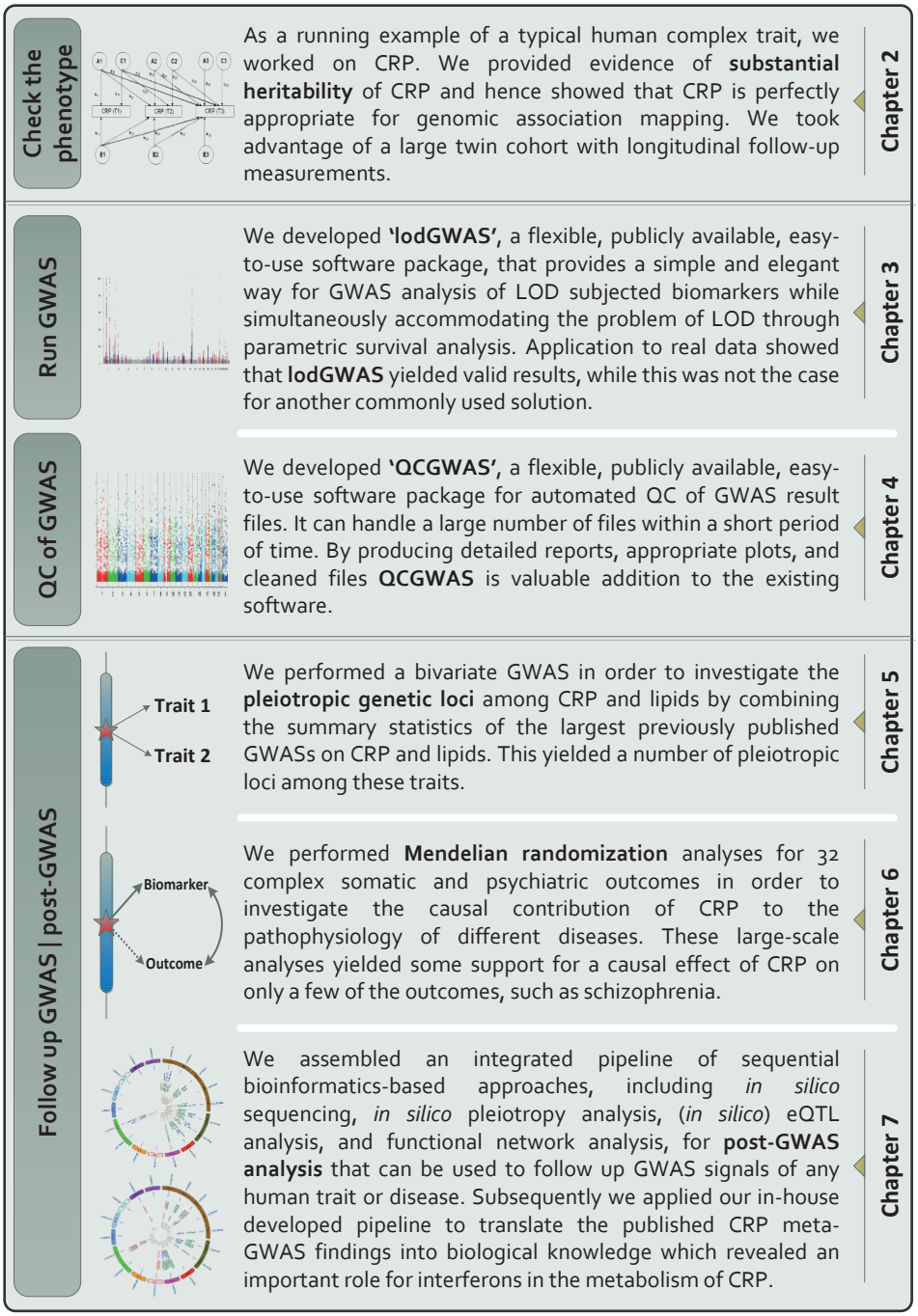

Figure 1 Results of this thesis at a glance. CRP indicates C-reactive protein; GWAS, genome-wide association study; LOD, limit of detection; and QC, quality control.

Check the suitability of the phenotype

As a first classic step of genomic association mapping, we sought to check the suitability of our phenotype of interest, i.e. serum levels of CRP. Hence, **Chapter 2** was devoted to investigate relative impact of genetic and environmental factors underlying serum levels of CRP with advancing age. Although cross-sectional twin and family studies have already reported a moderate heritability of baseline levels of CRP ranging from 0.10 to 0.65 for different age ranges (16–19), the (in)stability of heritability of CRP with age was still unclear. To apply a longitudinal classical twin design, we took full advantage of using a large dataset of twins from the TwinsUK registry (20) through applying software specifically designed for the analysis of twin and family data (21). We included data of a maximum of 6,201 female twins with up to three CRP measurements over a 10 year follow up period. In **Chapter 2** we showed that although CRP levels itself is not stable with advancing age, its heritability is around 50% and very constant over time. We also showed, in spite of stability of heritability of CRP, the genomic factors underlying control of CRP are not stable over time, indicating emergence of new genetic effects on CRP levels with age. All this means that CRP is perfectly suitable for genomic association mapping, warranting next classic steps presented in this dissertation.

Run GWAS of the phenotype

As a second classic step of genomic association mapping, we sought to perform GWAS analysis on our phenotype of interest, i.e. serum levels of CRP. In **Chapter 3**, however, we faced the problem that the detection range of CRP, like for some other biomarkers, is restricted by the assay procedure, i.e. there exists a so-called Limit of Detection (LOD). An LOD is the floor or ceiling value of the biomarker at which the level of the biomarker can still be accurately detected by the assay. Any value of the biomarker outside the range determined by the LOD(s) cannot be determined precisely (22, 23). Those observations that fall outside the LOD, so-called non-detects (NDs), cannot be simply excluded from the GWAS analysis, because NDs cannot be considered as 'missing at random'. Hence, NDs should be included in GWAS analysis by applying an appropriate statistical technique (22–25).

Appropriate statistical solution

There are a number of suboptimal statistical approaches described to deal with the problem of LOD while including the NDs into the statistical analysis, such as: analyzing detect/non-detect dichotomies, substituting NDs with a random value

between zero and lower LOD, substituting NDs with a constant smaller than the lower LOD (e.g. zero, LOD/√(2), or LOD/2), substituting NDs with LOD itself, etc. (24). However, the information provided by NDs is not optimally used by these methods, particularly when a substantial proportion of observations fall below the LOD. We found a better solution by concluding that biomarker data constrained by LOD most closely fit the characteristics of 'censored' time-to-event data and hence, the statistical methods of survival analysis for censored data can be applied to any variable that is subject to LOD (26, 27).

Software implementation

To the best of our knowledge, the problem of LOD in GWAS analysis of biomarkers has been usually overlooked, which could cause biased results. On the other hand, applying a proper statistical approach for combining NDs with valid measured values has not been feasible for most GWAS analysts, mainly due to limited software availability (28–30). In view of these shortcomings, we developed a software package that is capable of performing GWAS analysis of biomarkers while simultaneously accommodating the problem of LOD.

'lodGWAS' software package

We developed 'lodGWAS', a software package for GWAS analysis of biomarkers with a limit of detection. It treats NDs as censored data, either left-censored or right-censored or both, and performs a parametric survival analysis by including both 'measured' and 'censored' values. As an example application of lodGWAS and to demonstrate the robustness of its results, we performed GWAS analyses on serum levels of CRP in the LifeLines cohort study (31, 32) by means of both lodGWAS and PLINK (29) software. The latter results were considered as *gold standard*. In **Chapter 3** we showed that GWAS results of analyses by PLINK and lodGWAS on the complete dataset were identical. Then we assumed an arbitrary lower LOD of 1.0 mg/L for the CRP assay. The samples below this cut-off, that is 39% of the whole dataset, were disregarded by PLINK, but appropriately treated by lodGWAS. We showed that when analyzing this LOD subjected dataset, GWAS results of lodGWAS were very similar to the *gold standard*, while PLINK results were biased. Moreover, some regions that were genome-wide significant when analyzing the complete dataset, disappeared in the latter PLINK analysis leading to false negative results.

Our in-house developed software package has been submitted to the Comprehensive R Archive Network (CRAN) (33) and is now publicly available to the scientific community at http://cran.r-project.org/web/packages/lodGWAS.

Check the quality of GWAS results

As a third classic step of genomic association mapping, we sought to check the quality of GWAS results. GWAS results are the ultimate product of multiple sequential sophisticated lab- and bioinformatics-based steps, including genotyping of study samples, genotype calling, genotype data cleaning, genotype imputation, phenotype assay, statistical analysis of phenotype, (sometimes) merging of different datasets of genotypes or phenotypes, and finally performing GWAS analysis on the genotype and phenotype datasets. In addition to the large number of steps to complete a GWAS analysis, a great variety of software programs and pipelines can be used. Furthermore, each GWAS result file comprises millions of lines of data. All this means that GWAS result files are highly prone to different types of errors and mix-ups (34). Therefore in **Chapter 4** we emphasize the importance of a thorough quality control (QC) of GWAS output files prior to further steps of genomic association mapping.

Stringent QC pipeline

We developed a stringent pipeline for a thorough quality control of GWAS output files, which consists of multiple stages. It starts with checking the dataset for missing and invalid data. Duplicated genetic markers or those markers with missing crucial variables should be removed. Then, the alleles and also the allele frequencies of single nucleotide polymorphisms (SNPs) should be checked against a given reference, such as HapMap (35), 1000 Genomes Project (36), GoNL (37), UK10K (38), etc. Any strand switch or allele flip should be treated appropriately. Next, several QC plots and summary statistics are generated, which altogether provide a detailed view of the quality of the contents of the GWAS result file. Subsequent to detecting and fixing erroneous lines, and after harmonizing the alleles against the given reference, a cleaned GWAS result file should be generated. Such a quality controlled GWAS file can then be used for further steps of genomic association mapping.

Software implementation

We developed a software package that automates the abovementioned steps of our stringent QC pipeline. It automatically opens GWAS result files and checks their content against QC steps, provides plenty of plots and summary statistics regarding the quality of each of the GWAS result files, makes comparisons between files, and generates cleaned files. It is built as a package for R (33), ensuring compatibility with different operating systems and capability of handling large data files.

'QCGWAS' software package

'QCGWAS' is a flexible and comprehensive software package for automated QC of GWAS result files (39). As described in **Chapter 4**, its automated capability of multistage QC can be broadly summarized in two key features:

(i) It generates cleaned, quality controlled GWAS files by (a) standardizing column names, (b) matching alleles and also allele frequencies with a given reference, (c) removing SNPs with mismatching alleles, (d) removing SNPs with invalid variable values, (e) removing duplicated SNPs, etc.

(ii) It provides plenty of valuable information about the GWAS result files, including detailed summary statistics as well as a wide variety of QC plots. This information helps the user to strictly interpret the quality of the GWAS result file and hence, to precisely detect any hidden errors in the file, even very minor bugs.

The strength of QCGWAS is that it is geared towards producing reports, summary statistics and plots that are ultimately based on the cleaned quality controlled dataset, which is much more reliable than pre-QC file. We try to show this strength with two real examples, as follows:

Example 'A'. The allele frequency correlation plots of pre- and post-QC GWAS file 'A' against the given reference can be seen in Figure 2. There is a significant difference between the two panels of the figure because the pre-QC file is the GWAS output file as is, while the post-QC dataset has been appropriately treated by QCGWAS through matching alleles and allele frequencies with the given reference.

Example 'B'. The genomic inflation factor (lambda) of GWAS file 'B' is calculated to be 0.89 and 1.12 for pre- and post-QC GWAS files, respectively. Lambda

quantifies presence or absence of population stratification, a phenomenon that can induce false positive results. The expected value of lambda when there is no population stratification is 1.00. There is a significant difference between the two calculated lambdas because the pre-QC file is the GWAS output file as is, while the post-QC dataset has been appropriately treated by QCGWAS through removing ~9% of SNPs of the pre-QC file as they contained invalid summary statistics: for all ~9% of SNPs of the pre-QC file the effect sizes, standard errors, and p-values were reported as -1, -1, and +1, respectively. Obviously, the invalid p-values of the ~9% of SNPs would markedly shift the distribution of p-values of the pre-QC file to the right, yielding a smaller lambda. Removing the invalid p-values yielded a valid distribution of p-values and hence, an accurate estimate of lambda.

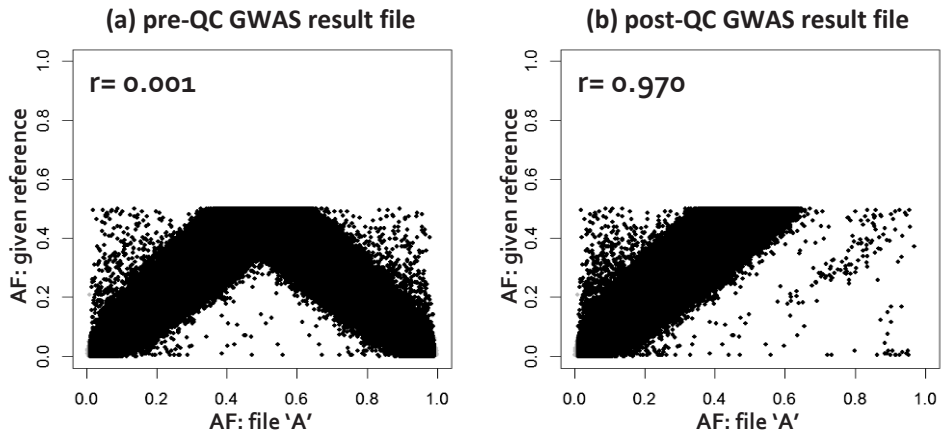

Figure 2 The allele frequency correlation plots of an illustrative (a) pre- and (b) post-QC GWAS result file 'A' against the given reference. QC indicates quality control; GWAS, genome-wide association study; and AF, allele frequency.

These examples emphasize the value of generating a cleaned, quality controlled dataset prior to interpretation of the quality of the file and any further use, e.g. in meta-analysis. They also emphasize how such a cleaned dataset can significantly change our judgment about the quality of a GWAS result file. We conclude that we can only rely on QC reports, summary statistics, and plots based on post-QC datasets. To the best of our knowledge, the only software package with such capabilities is QCGWAS. Where other software packages for QC of GWAS result files just produce some reports and plots, QCGWAS actually cleans the file and then appropriately produces reports and plots that are more informative. In

other words, QCGWAS is a valuable addition to the programs for the QC of GWAS result files.

Our in-house developed software package has been submitted to the Comprehensive R Archive Network (CRAN) (33) and is now publicly available to the scientific community at http://cran.r-project.org/web/packages/QCGWAS.

Follow-up of GWAS results: post-GWAS analyses

As the final classic step of genomic association mapping, we sought to follow up GWAS results by translating them into biological knowledge. GWAS results are typically provided as a list of SNPs that are significantly associated with the investigated phenotype (2). SNP names are presented as Reference SNP Cluster Identifier (rsID) numbers (i.e, rs<number>). Hence, the ultimate result of a successful GWAS is merely a list of rsID numbers without biological meaning. It means that further steps of genomic association mapping, so-called post-GWAS analyses, are needed to translate such a list of rsIDs into biological insights (40, 41).

As running examples of post-GWAS analyses, in **Chapters 5**, **6**, and **7** we followed up a large-scale meta-GWAS on serum levels of CRP (14) and combined these results with those from other large meta-GWASs on outcomes potentially associated with CRP. (42–67).

Pleiotropy analysis

Some of the identified genetic variants are associated with more than one phenotype, a phenomenon termed *genetic pleiotropy*. Several lines of evidence suggest an overlap between the biology of CRP and lipids (68–72). However, the exact mechanisms underlying this overlap are yet to be investigated. To identify genetic pleiotropy among CRP and lipids and hence, to uncover the genetic mechanisms underlying their overlap, **Chapter 5** was devoted to follow up of results from the largest meta-GWAS on CRP (14) in combination with the largest meta-GWASs on lipids, i.e. low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, triglycerides, and total cholesterol (66, 67).

To efficiently perform a bivariate GWAS on serum levels of CRP and lipids, we applied a new method and bioinformatics software package that uses GWAS summary statistics data rather than individual level data (73). Following up the largest meta-GWASs on CRP and lipids, our method eventually revealed a number of pleiotropic loci among these two traits. Three identified SNPs, which

were novel for CRP, were then successfully replicated in a large sample of independent individuals. The novel replicated SNPs were in or near *FTO*, *CTSB/FDFT1*, and *STAG1/PCCB*. In other words, in **Chapter 5** we showed that this bioinformatics-based method could be efficiently applied as a step in genomic association mapping not only to identify novel and pleiotropic loci, but also to provide further insight in the genetic interrelation between linked phenotypes.

Mendelian randomization analysis

To investigate the causal contribution of CRP to the pathophysiology of different diseases, **Chapter 6** was devoted to follow up the largest meta-GWAS on CRP (14) in combination with large meta-GWASs on a wide range of different disease outcomes potentially associated with CRP (42–65).

Serum levels of CRP have been associated with a great variety of complex somatic and psychiatric outcomes including coronary heart disease (6), hypertension (5), stroke (7), type 2 diabetes (4), bipolar disorder (8), and overall mortality (9). The association of CRP with each outcome can be due to real causation, reverse causation, confounding, or various types of bias (74). Hence, the causal involvement of CRP to the pathophysiology of different outcomes remains extremely controversial (75, 76).

Mendelian randomization (MR) is an increasingly popular method that uses genetically informed instrumental variables (IVs) to generate more reliable evidence whether the association of a biomarker, such as CRP level, with a disease outcome is due to real causation. To build appropriate IVs based on GWAS data, we used established genetic variants associated with CRP levels, and pooled them into two CRP genetic risk scores (GRSs): the first GRS consisted of four SNPs in the *CRP* gene (15), and the other GRS consisted of all eighteen genome-wide significant CRP SNPs (14). To optimize power, we applied a bioinformatics software package that uses GWAS summary statistics data rather than individual level data (77).

By collecting summary statistics of association results of several meta-GWAS consortia, we tested the association of CRP GRSs against 32 complex somatic and psychiatric outcomes including celiac disease (42), inflammatory bowel disease (i.e. Crohn's disease and ulcerative colitis) (43, 44), psoriasis (all types) (45, 46), rheumatoid arthritis (47), systemic lupus erythematosus (48), systemic sclerosis (49), type 1 diabetes (50), knee osteoarthritis (51), coronary artery disease (52), systolic and diastolic blood pressure (53), ischemic stroke (all types) (54), body mass index (55), type 2 diabetes (56), chronic kidney disease (57), eGFR for creatinine

(57), serum albumin and protein levels (58), amyotrophic lateral sclerosis (59), Alzheimer's disease (60), Parkinson's disease (61), autism (62), bipolar disorder (63), major depressive disorder (64), and schizophrenia (65).

In **Chapter 6**, we provide evidence for a potentially causal (protective) effect of elevated CRP level on schizophrenia. We also provide nominal (not Bonferroni corrected) significant evidence for a potentially causal association of elevated CRP level with knee osteoarthritis, bipolar disorder, and systolic and diastolic blood pressure. However, we found no further causal association of CRP with any of the other 27 complex diseases investigated in this large-scale effort.

A comprehensive solution for post-GWAS analysis

A number of post-GWAS analysis approaches have already been introduced in the literature (41). However, in the past post-GWAS analysis typically only covered one or more data domains in search of functional evidence, such as translating SNP rsID numbers into names of closest genes (simple annotation), identifying SNP associations with gene expression (expression quantitative trait loci [eQTL] analysis), or translating gene names into biological processes (functional network analysis). Hence, the results of each data domain were usually reported separately and not easily combined with those of other data domains impeding translation to biological insights. To the best of our knowledge there was no established pipeline integrating individual post-GWAS approaches and offering a comprehensive solution for post-GWAS analysis of GWAS findings (40, 41).

In **Chapter 7**, we assembled an integrated pipeline of sequential bioinformatics-based approaches for post-GWAS analysis that can be used to follow up GWAS signals of any human trait or disease. The components of our '*in silico*' pipeline are all based on publicly available tools or data resources (36, 78–91), which means it can easily be applied by other researchers who wish to follow up GWAS findings. The pipeline accepts a list of GWAS SNPs (gSNPs) as input, then performs a number of sequential bioinformatics-based steps including *in silico* sequencing, *in silico* pleiotropy analysis, (*in silico*) eQTL analysis, functional interaction network analysis, and finally, functional enrichment analysis.

The core concept of this pipeline is that GWAS SNPs merely tag the surrounding genomic region and are not necessarily causally related to the phenotype of interest (40). Hence, our pipeline emphasizes the genomic context in the vicinity of the gSNPs and, by using different resources from different data domains, provides a thorough description of the genomic context of the

surrounding regions. It not only integrates the results of different data domains, but also integrates and uses the results of all gSNPs combined. This holistic approach is consistent with the polygenic model of disease susceptibility (2).

As an example application of this pipeline and to demonstrate the robustness of its results, in **Chapter 7** we followed up the largest meta-GWAS on serum levels of CRP (14). That is, we submitted the 18 genome-wide significant CRP SNPs to the pipeline to translate these CRP meta-GWAS findings into biological knowledge. This strategy yielded a range of enriched biological processes that provided new information on mechanisms involved in CRP metabolism, as follows:

(i) The previously known overlap between the biology of CRP and lipids (68–71) was confirmed.

(ii) The results of the pipeline suggested an important role for interferons in the metabolism of CRP (92).

The overlap in the biology of CRP and lipids is not novel, but its consistency with previous knowledge satisfactorily provided confidence in the results of the pipeline. Consequently it strengthened the main finding of this study, i.e. the link between CRP and interferons, which has not been generally accepted as a potential mechanism underlying the biology of CRP. Our extended literature review however revealed a number of indirect clinical observations, which supported this finding. Furthermore, an *in vitro* study showed that interferon-α is an inhibitor of CRP secretion. The suppression of CRP secretion was shown to be dose dependent and to act through the type I interferon receptor (93). Additionally, although the associations of CRP levels with most inflammatory diseases are well established, its elevated levels poorly correlate with severe inflammatory conditions that are typically known to have high levels of interferons, i.e. systemic lupus and viral infections (94–99).

Some odd observations in lupus patients like a dramatically increased risk of myocardial infarction (100, 101), but no association between cardiovascular risk factors and CRP (98), as well as poor correlation between Interleukin-6 and CRP in these patients (102) might very well be explained by suppression of CRP by interferons. Likewise, mild and poorly correlated increases of CRP in viral infections (99), making CRP a widely used biomarker in distinguishing bacterial from viral infections, can be explained by the suppression of CRP by interferons (93).

Box 1 Concluding remarks.

This dissertation succeeded to provide an A-to-Z coverage of classic steps of genomic association mapping using bioinformatics-based approaches.

We showed how bioinformatics tools can facilitate and support analysis of high-throughput biological data. In **Chapters 2, 3, 5, 6,** and **7** we 'applied' a number of already available tools, whereas in **Chapters 3, 4,** and **7** we 'developed' novel bioinformatics tools supporting appropriate analysis of big data for genomic association mapping.

Our in-house developed bioinformatics tools alongside with appropriate documentations are now freely available to the scientific community for any further application.

Furthermore, and as a running example of genomic association mapping of a typical human complex trait, we strictly adhered to serum levels of CRP. Using appropriate bioinformatics-based tools, either already available or our in-house developed ones, we succeeded to gain in knowledge of biological mechanisms controlling serum levels of CRP as well as its (causal) contribution to the pathophysiology of human diseases.

We also applied a preliminary version of our post-GWAS pipeline to GWAS results of a large meta-GWAS on serum protein concentrations, yielding new insights into underlying mechanisms controlling serum levels of proteins (58). Furthermore, we applied (part of) this pipeline to the results of an epigenome-wide association (EWAS) study of the correlation between maternal smoking and methylation and its relation to the birth weight of the offspring, which revealed a role in activating cell-mediated immunity (103). The results of a number of other meta-GWAS or meta-EWAS projects are also being followed up by this pipeline (manuscripts in preparation).

Future perspectives

Quality control of GWAS results

In **Chapter 4** we introduced QCGWAS, our in-house developed software package for quality control of GWAS result files (39). It was originally motivated by the need for QC of GWAS results of immunoproteins, such as CRP. Hence, it is currently geared towards quantitative traits. Further adaptations of the package will be needed to enable QC of GWAS results of case-control studies. Moreover, the availability of data of non-SNP variants, such as insertion-deletions or copy-number variants in e.g. the 1000 Genomes data (36), highlights the need and potential for improved versions of the package to process all variant types. Moreover, close resemblance of the framework of GWAS and EWAS data, warrants developing a new software tool for QC of EWAS results.

Identifying further genetic risk variants

In **Chapters 5**, **6**, and **7** we followed up the SNPs that were significantly associated with CRP levels in the largest meta-GWAS to date. Although this large scale meta-GWAS is based on >80,000 subjects, the explained variance in CRP level by all identified SNPs is only around five percent (14). To further unravel the underlying mechanisms controlling CRP level, an even larger meta-GWAS on CRP is highly needed. In parallel to this thesis, we have designed and launched such a mega meta-GWAS project with a sample size of more than 200,000 individuals aiming at identifying more genetic variants associated with serum levels of CRP. All the results of this thesis, i.e. expertise and bioinformatics tools, have been essential in conducting such an international, mega-scale project.

From SNP association to function

The variety, complexity, and size of publicly available data resources have dramatically increased in the last few years. For example, the initial results of the Encyclopedia of DNA Elements (ENCODE) Project were published in 2012, making more than 1,600 datasets of genome-wide functional elements publicly available (104). Additionally, there are currently more than 3,800 data sets available comprising data on more than 1.4 million samples of gene expression deposited in the NCBI Gene Expression Omnibus (GEO) database (accessed May 29, 2015). Likewise, other large-scale public repositories of biological data are becoming available, including (but not limited to): the International HapMap Project (35), the 1000 Genomes Project (36), the Roadmap Epigenomics Mapping Consortium

(105), the Genotype-Tissue Expression (GTEx) (106), The Cancer Genome Atlas (TCGA) (107), the Library of Integrated Network-Based Cellular Signatures (LINCS) (108), etc.

The effective mining of high throughput biological data, so-called 'big data', is extremely challenging. Although the data are publicly available, data collection, curation, calibration, and processing requires labor-intensive work together with high expertise. At any rate, bioinformatics tools and pipelines are needed to integrate different data types obtained from diverse biological domains. A variety of algorithms and tools for biological big data mining have become available, including (but not limited to): Cistrome (109), Oncomine (110), PAthway Recognition Algorithm using Data Integration on Genomic Models (Paradigm) (111), Multiple Association Network Integration Algorithm (GeneMANIA) (87), Data-driven Expression Prioritized Integration for Complex Traits (DEPICT) (112), Probabilistic Identification of Causal SNPs (PICS) (113), etc.

In **Chapter 7** we presented an integrated pipeline of sequential bioinformatics-based approaches for post-GWAS analysis of GWAS signals of human traits or diseases (92). It requires appropriate information from different data resources and treats the obtained information in a broader context by integrating them to other data domains (36, 78–91). We applied a bioinformatics algorithm based on the 'guilt-by association' approach to translate the vast amounts of diverse biological data to biological insights. However, we will be particularly keen to extend this approach by:

(i) Obtaining biological data from a wider variety of data domains, such as the Roadmap Epigenomics Mapping Consortium, ENCODE project, etc.

(ii) More effective integration and mining of the obtained data through improving algorithms and software tools. We are willing to work towards developing a software tool that will take full advantage of the strengths of current tools, such as GeneMANIA or DEPICT, but at the same time, will address their (e.g. technical) shortcomings.

Such a bioinformatics-based approach will yield an even deeper knowledge of biological processes promising novel biological discoveries, better understanding of disease mechanisms, more efficient diagnostic and prognostic tools, and last but not least, introducing individually tailored therapies based on the genomic context of each patient, so-called personalized medicine (114, 115).

References

1. 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.

2. 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.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

9. 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.

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

11. 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.

12. 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.

13. 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.

14. 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.

15. C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC), Wensley, F., Gao, P., Burgess, S., Kaptoge, S., Di Angelantonio, E., Shah, T., Engert, J.C., Clarke, R., Davey-Smith, G., et al. (2011) Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ*, **342**, d548.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. Neale, M.C., Boker, S.M., Xie, G. and Maes, H.H. (2003) Mx: Statistical Modeling. VCU Box 900126, Richmond, VA 23298: Department of Psychiatry.

22. Lubin, J.H., Colt, J.S., Camann, D., Davis, S., Cerhan, J.R., Severson, R.K., Bernstein, L. and Hartge, P. (2004) Epidemiologic Evaluation of Measurement Data in the Presence of Detection Limits. *Environ. Health Perspect.*, **112**, 1691–1696.

23. Armbruster, D.A. and Pry, T. (2008) Limit of Blank, Limit of Detection and Limit of Quantitation. *Clin. Biochem. Rev.*, **29**, S49–S52.

24. Uh, H.-W., Hartgers, F.C., Yazdanbakhsh, M. and Houwing-Duistermaat, J.J. (2008) Evaluation of regression methods when immunological measurements are constrained by detection limits. *BMC Immunol.*, **9**, 59.

25. Sattar, A., Sinha, S.K. and Morris, N.J. (2012) A Parametric Survival Model When a Covariate is Subject to Left-Censoring. *J. Biom. Biostat.*, **Suppl 3**.

26. Gillespie, B.W., Chen, Q., Reichert, H., Franzblau, A., Hedgeman, E., Lepkowski, J., Adriaens, P., Demond, A., Luksemburg, W. and Garabrant, D.H. (2010) Estimating Population Distributions When Some Data Are Below a Limit of Detection by Using a Reverse Kaplan-Meier Estimator. *Epidemiol. July 2010*, **21**.

27. Dinse, G.E., Jusko, T.A., Ho, L.A., Annam, K., Graubard, B.I., Hertz-Picciotto, I., Miller, F.W., Gillespie, B.W. and Weinberg, C.R. (2014) Accommodating Measurements Below a Limit of Detection: A Novel Application of Cox Regression. *Am. J. Epidemiol.*, **179**, 1018–1024.

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

29. 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.

30. 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.

31. Stolk, R.P., Rosmalen, J.G.M., Postma, D.S., Boer, R.A. de, Navis, G., Slaets, J.P.J., Ormel, J. and Wolffenbuttel, B.H.R. (2007) Universal risk factors for multifactorial diseases. *Eur. J. Epidemiol.*, **23**, 67–74.

32. Scholtens, S., Smidt, N., Swertz, M.A., Bakker, S.J., Dotinga, A., Vonk, J.M., Dijk, F. van, Zon, S.K. van, Wijmenga, C., Wolffenbuttel, B.H., et al. (2014) Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.*, 10.1093/ije/dyu229.

33. R Core Team (2014) R: A Language and Environment for Statistical Computing.

34. Bakker, P.I.W. de, Ferreira, M.A.R., Jia, X., Neale, B.M., Raychaudhuri, S. and Voight, B.F. (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.*, **17**, R122–R128.

35. 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.

36. 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.

37. Genome of the Netherlands Consortium (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.*, **46**, 818–825.

38. Muddyman, D., Smee, C., Griffin, H. and Kaye, J. (2013) Implementing a successful data-management framework: the UK10K managed access model. *Genome Med.*, **5**, 100.

39. Van der Most, P.J., Vaez, A., Prins, B.P., Munoz, M.L., Snieder, H., Alizadeh, B.Z. and Nolte, I.M. (2014) QCGWAS: A flexible R package for automated quality control of genome-wide association results. *Bioinformatics*, **30**, 1185–1186.

40. 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.

41. 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.

42. Dubois, P.C.A., Trynka, G., Franke, L., Hunt, K.A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G.A.R., Adány, R., Aromaa, A., et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.*, **42**, 295–302.

43. Franke, A., McGovern, D.P.B., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.*, **42**, 1118–1125.

44. Anderson, C.A., Boucher, G., Lees, C.W., Franke, A., D'Amato, M., Taylor, K.D., Lee, J.C., Goyette, P., Imielinski, M., Latiano, A., et al. (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.*, **43**, 246–252.

45. Nair, R.P., Duffin, K.C., Helms, C., Ding, J., Stuart, P.E., Goldgar, D., Gudjonsson, J.E., Li, Y., Tejasvi, T., Feng, B.-J., et al. (2009) Genome-wide scan

reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nat. Genet.*, **41**, 199–204.

46. Ellinghaus, E., Ellinghaus, D., Stuart, P.E., Nair, R.P., Debrus, S., Raelson, J.V., Belouchi, M., Fournier, H., Reinhard, C., Ding, J., et al. (2010) Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2. *Nat. Genet.*, **42**, 991–995.

47. Stahl, E.A., Raychaudhuri, S., Remmers, E.F., Xie, G., Eyre, S., Thomson, B.P., Li, Y., Kurreeman, F.A.S., Zhernakova, A., Hinks, A., et al. (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.*, **42**, 508–514.

48. Hom, G., Graham, R.R., Modrek, B., Taylor, K.E., Ortmann, W., Garnier, S., Lee, A.T., Chung, S.A., Ferreira, R.C., Pant, P.V.K., et al. (2008) Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. *N. Engl. J. Med.*, **358**, 900–909.

49. Radstake, T.R.D.J., Gorlova, O., Rueda, B., Martin, J.-E., Alizadeh, B.Z., Palomino-Morales, R., Coenen, M.J., Vonk, M.C., Voskuyl, A.E., Schuerwegh, A.J., et al. (2010) Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. *Nat. Genet.*, **42**, 426–429.

50. Bradfield, J.P., Qu, H.-Q., Wang, K., Zhang, H., Sleiman, P.M., Kim, C.E., Mentch, F.D., Qiu, H., Glessner, J.T., Thomas, K.A., et al. (2011) A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet.*, **7**, e1002293.

51. Evangelou, E., Valdes, A.M., Kerkhof, H.J.M., Styrkarsdottir, U., Zhu, Y., Meulenbelt, I., Lories, R.J., Karassa, F.B., Tylzanowski, P., Bos, S.D., et al. (2011) Meta-analysis of genome-wide association studies confirms a susceptibility locus for knee osteoarthritis on chromosome 7q22. *Ann. Rheum. Dis.*, **70**, 349–355.

52. Schunkert, H., König, I.R., Kathiresan, S., Reilly, M.P., Assimes, T.L., Holm, H., Preuss, M., Stewart, A.F.R., Barbalic, M., Gieger, C., et al. (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.*, **43**, 333–338.

53. International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., Chasman, D.I., Smith, A.V., Tobin, M.D., Verwoert, G.C., et al. (2011) Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, **478**, 103–109.

54. International Stroke Genetics Consortium (ISGC), Wellcome Trust Case Control Consortium 2 (WTCCC2), Bellenguez, C., Bevan, S., Gschwendtner, A., Spencer, C.C.A., Burgess, A.I., Pirinen, M., Jackson, C.A., Traylor, M., et al.

(2012) Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nat. Genet.*, **44**, 328–333.

55. Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Lango Allen, H., Lindgren, C.M., Luan, J., Mägi, R., et al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.*, **42**, 937–948.

56. Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E., Huth, C., Aulchenko, Y.S., Thorleifsson, G., et al. (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.*, **42**, 579–589.

57. Köttgen, A., Pattaro, C., Böger, C.A., Fuchsberger, C., Olden, M., Glazer, N.L., Parsa, A., Gao, X., Yang, Q., Smith, A.V., et al. (2010) New loci associated with kidney function and chronic kidney disease. *Nat. Genet.*, **42**, 376–384.

58. Franceschini, N., van Rooij, F.J.A., Prins, B.P., Feitosa, M.F., Karakas, M., Eckfeldt, J.H., Folsom, A.R., Kopp, J., Vaez, A., Andrews, J.S., et al. (2012) Discovery and Fine Mapping of Serum Protein Loci through Transethnic Meta-analysis. *Am. J. Hum. Genet.*, **91**, 744–753.

59. Diekstra, F.P., Saris, C.G.J., van Rheenen, W., Franke, L., Jansen, R.C., van Es, M.A., van Vught, P.W.J., Blauw, H.M., Groen, E.J.N., Horvath, S., et al. (2012) Mapping of gene expression reveals CYP27A1 as a susceptibility gene for sporadic ALS. *PloS One*, **7**, e35333.

60. Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J.-C., Carrasquillo, M.M., Abraham, R., Hamshere, M.L., Pahwa, J.S., Moskvina, V., et al. (2011) Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nat. Genet.*, **43**, 429–435.

61. International Parkinson Disease Genomics Consortium, Nalls, M.A., Plagnol, V., Hernandez, D.G., Sharma, M., Sheerin, U.-M., Saad, M., Simón-Sánchez, J., Schulte, C., Lesage, S., et al. (2011) Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet*, **377**, 641–649.

62. Weiss, L.A., Arking, D.E., Gene Discovery Project of Johns Hopkins & the Autism Consortium, Daly, M.J. and Chakravarti, A. (2009) A genome-wide linkage and association scan reveals novel loci for autism. *Nature*, **461**, 802–808.

63. Psychiatric GWAS Consortium Bipolar Disorder Working Group (2011) Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.*, **43**, 977–983.

64. Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium, Ripke, S., Wray, N.R., Lewis, C.M., Hamilton, S.P., Weissman, M.M., Breen, G., Byrne, E.M., Blackwood, D.H.R., Boomsma, D.I., et al. (2013)

A mega-analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatry*, **18**, 497–511.

65. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.

66. 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.

67. Global Lipids Genetics Consortium, Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., et al. (2013) Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, **45**, 1274–1283.

68. 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.

69. 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.

70. 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.

71. 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.

72. Kraja, A.T., Chasman, D.I., North, K.E., Reiner, A.P., Yanek, L.R., Kilpeläinen, T.O., Smith, J.A., Dehghan, A., Dupuis, J., Johnson, A.D., et al. (2014) Pleiotropic genes for metabolic syndrome and inflammation. *Mol. Genet. Metab.*, **112**, 317–338.

73. Hsu, Y. and Chen, X. (2011) A Multivariate Approach on Genome-Wide Association Studies (GWAS) by Modeling multiple Traits Simultaneously to Identify Pleiotropic Genetic Effects/Plenary Talk. The Joint Statistical Meetings (American Statistics Association), Miami Beach, FL, USA, 2011.

74. Smith, G.D. and Hemani, G. (2014) Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.*, **23**, R89–R98.

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

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

77. 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.

78. Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164–e164.

79. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.

80. Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2011) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.

81. Kumar, P., Henikoff, S. and Ng, P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.

82. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

83. Li, H. (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, **27**, 718–719.

84. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

85. Saito, R., Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L., Lotia, S., Pico, A.R., Bader, G.D. and Ideker, T. (2012) A travel guide to Cytoscape plugins. *Nat. Methods*, **9**, 1069–1076.

86. Mostafavi, S. and Morris, Q. (2012) Combining many interaction networks to predict gene function and analyze gene lists. *PROTEOMICS*, **12**, 1687–1696.

87. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. and Morris, Q. (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.*, **9**, S4.

88. Montojo, J., Zuberi, K., Rodriguez, H., Kazi, F., Wright, G., Donaldson, S.L., Morris, Q. and Bader, G.D. (2010) GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics*, **26**, 2927–2928.

89. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

90. Schröder, M.S., Gusenleitner, D., Quackenbush, J., Culhane, A.C. and Haibe-Kains, B. (2013) RamiGO: an R/Bioconductor package providing an AmiGO Visualize interface. *Bioinformatics*, **29**, 666–668.

91. Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C.T., et al. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W220.

92. Vaez, A., Jansen, R., Prins, B.P., Hottenga, J.-J., Geus, E.J.C. de, Boomsma, D.I., Penninx, B.W.J.H., Nolte, I.M., Snieder, H. and Alizadeh, B.Z. (2015) An in silico Post-GWAS Analysis of C-Reactive Protein Loci Suggests an Important Role for Interferons. *Circ. Cardiovasc. Genet.*, 10.1161/CIRCGENETICS.114.000714.

93. Enocsson, H., Sjöwall, C., Skogh, T., Eloranta, M.-L., Rönnblom, L. and Wetterö, J. (2009) Interferon-α mediates suppression of C-reactive protein: Explanation for muted C-reactive protein response in lupus flares? *Arthritis Rheum.*, **60**, 3755–3760.

94. Theofilopoulos, A.N., Baccala, R., Beutler, B. and Kono, D.H. (2005) TYPE I INTERFERONS (α/β) IN IMMUNITY AND AUTOIMMUNITY. *Annu. Rev. Immunol.*, **23**, 307–335.

95. Lech, M., Rommele, C. and Anders, H.-J. (2013) Pentraxins in nephrology: C-reactive protein, serum amyloid P and pentraxin-3. *Nephrol. Dial. Transplant.*, **28**, 803–811.

96. Becker, G.J., Waldburger, M., Hughes, G.R. and Pepys, M.B. (1980) Value of serum C-reactive protein measurement in the investigation of fever in systemic lupus erythematosus. *Ann. Rheum. Dis.*, **39**, 50–52.

97. Honig, S., Gorevic, P. and Weissmann, G. (1977) C-Reactive Protein in Systemic Lupus Erythematosus. *Arthritis Rheum.*, **20**, 1065–1070.

98. Nikpour, M., Gladman, D.D., Ibañez, D. and Urowitz, M.B. (2009) Variability and correlates of high sensitivity C-reactive protein in systemic lupus erythematosus. *Lupus*, **18**, 966–973.

99. Sasaki, K., Fujita, I., Hamasaki, Y. and Miyazaki, S. (2002) Differentiating between bacterial and viral infection by measuring both C-reactive protein and 2'-5'-oligoadenylate synthetase as inflammatory markers. *J. Infect. Chemother.*, **8**, 76–80.

100. Manzi, S., Meilahn, E.N., Rairie, J.E., Conte, C.G., Medsger, T.A., Jansen-McWilliams, L., D'Agostino, R.B. and Kuller, L.H. (1997) Age-specific

Incidence Rates of Myocardial Infarction and Angina in Women with Systemic Lupus Erythematosus: Comparison with the Framingham Study. *Am. J. Epidemiol.*, **145**, 408–415.

101. Esdaile, J.M., Abrahamowicz, M., Grodzicky, T., Li, Y., Panaritis, C., Berger, R.D., Côté, R., Grover, S.A., Fortin, P.R., Clarke, A.E., et al. (2001) Traditional Framingham risk factors fail to fully account for accelerated atherosclerosis in systemic lupus erythematosus. *Arthritis Rheum.*, **44**, 2331–2337.

102. Gabay, C., Roux-Lombard, P., de Moerloose, P., Dayer, J.M., Vischer, T. and Guerne, P.A. (1993) Absence of correlation between interleukin 6 and C-reactive protein blood levels in systemic lupus erythematosus compared with rheumatoid arthritis. *J. Rheumatol.*, **20**, 815–821.

103. Küpers, L.K., Xu, X., Jankipersadsing, S.A., Vaez, A., Gemert, S. la B., Scholtens, S., Nolte, I.M., Richmond, R.C., Relton, C.L., Felix, J.F., et al. (2015) DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. *Int. J. Epidemiol.*, 10.1093/ije/dyv048.

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

105. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, **28**, 1045–1048.

106. The GTEx Consortium, Ardlie, K.G., Deluca, D.S., Segrè, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., et al. (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, **348**, 648–660.

107. The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C. and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.

108. Vempati, U.D., Chung, C., Mader, C., Koleti, A., Datar, N., Vidović, D., Wrobel, D., Erickson, S., Muhlich, J.L., Berriz, G., et al. (2014) Metadata Standard and Data Exchange Specifications to Describe, Model, and Integrate Complex and Diverse High-Throughput Screening Data from the Library of Integrated Network-based Cellular Signatures (LINCS). *J. Biomol. Screen.*, **19**, 803–816.

109. Liu, T., Ortiz, J.A., Taing, L., Meyer, C.A., Lee, B., Zhang, Y., Shin, H., Wong, S.S., Ma, J., Lei, Y., et al. (2011) Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.*, **12**, R83.

110. Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A. and Chinnaiyan, A.M. (2004) ONCOMINE: a cancer

microarray database and integrated data-mining platform. *Neoplasia N. Y. N*, **6**, 1–6.

111. Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D. and Stuart, J.M. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinforma. Oxf. Engl.*, **26**, i237–245.

112. Pers, T.H., Karjalainen, J.M., Chan, Y., Westra, H.-J., Wood, A.R., Yang, J., Lui, J.C., Vedantam, S., Gustafsson, S., Esko, T., et al. (2015) Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.*, **6**.

113. Farh, K.K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A., et al. (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, **518**, 337–343.

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

115. Jiang, P. and Liu, X.S. (2015) Big data mining yields novel insights on cancer. *Nat. Genet.*, **47**, 103–104.