Disease Association Studies

OA Genetics
Open Access

Critical review

Penalised regression methods in genetic research

H Mallick[1], M Li[2]*

Abstract

Introduction

Complex human diseases usually have multifactorial causes, and may develop as a result of the collective effects of multiple genetic variants, complex gene-gene/gene-environment interactions, rare sequence variants, copy number alterations, epigenetic modifications, etc. Understanding the genetic aetiology of complex human diseases require a comprehensive assessment of these causes. Recently, penalised regression methods have gained popularity in genetic research, aiming to detect genetic, epigenetic and environmental factors contributing to complex human diseases. In this article, we attempt to provide a brief overview of these methods in light of their applications in various contexts of genetic research.

Conclusion

These methods are built on the assumption that, given a genotype-phenotype association, the genetic similarity would contribute to the phenotype similarity and aggregate multiple rare and common variants through the genetic similarities between individuals.

Introduction

Recent advances in genotyping and sequencing technology have enabled researchers to rapidly collect an enormous amount of high-dimensional genotype data throughout the entire genome[1]. The first generation genome-wide association studies (GWAS) commonly test each variant separately. Although thousands of variants have been identified[2], for most complex diseases, the current identified variants explain only a small percentage of the disease heritability[3]. Recently, there has been an intensive effort dedicated to examine multiple variants in a single model for their association with complex diseases/traits. This multi-locus strategy has many advantages over the single-locus search. First, all the genetic variants are not independent, but form Linkage-Disequilibrium (LD) blocks. Fitting a single model for multiple variants allow one to test the effect of one variant while controlling the effect of others, increasing the power to detect weak signals by accounting for other causal effects, and remove false signals by including a stronger causal association[4,5]. Second, analyses of multiple variants simultaneously will reduce the burden of multiple testing; thus, improve the power to detect a causal association.

Penalised regression methods have become popular in genetic research as an attractive alternative for a single marker analysis. In a penalised regression framework, the genetic effects are shrunk by maximizing the log-likelihood function subject to a penalty term, which is a function of the coefficients indexed by one or more tuning parameters. The form of the penalty term determines the general behaviour of penalised methods. For example, various methods have utilised least absolute shrinkage and selection operator (LASSO), which assumes that only a small number of genetic variants are causal to the disease phenotypes and allows simultaneous selection of causal variants and estimation of their effect sizes. The selection of causal variants is achieved by shrinking the effects of non-causal variants to zero, and retaining only a small subset of genetic variants with non-zero effects[5]. On the other hand, ridge-based methods do not directly conduct model selection for causal variants. However, they are much easier to be computed than lasso-based methods, and enjoy stable estimation in the presence of multi-collinearity. In practice, a variety of penalty terms have been proposed, among which the most popular ones are ridge, lasso, adaptive lasso, fused lasso, group lasso and elastic net (mixture of lasso and ridge). The mathematical formulas of all these methods are summarised in Table 1. The turning parameter(s) control the degree of penalisation, and the 'best' value of the tuning parameter can be selected either by minimizing Akaike information criterion (AIC) or Bayesian information criterion (BIC) or by cross-validation[6]. Penalised regression methods are also closely related to Bayesian shrinkage methods, which achieve the same goal of variable selection by specifying shrinkage prior distributions (probability distributions that have high probability near zero) on the coefficients[7,8]. In Bayesian analysis, it is a standard procedure to fully explore the posterior distribution through Markov chain Monte Carlo (MCMC) simulations. Most of the coefficient estimates are expected to be close to zero through their posterior distributions[7].

Generally, these penalised regression methods are able to handle high-dimensional data, robust to the data

* Corresponding author
Email: MLi@uams.edu

[1] Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama, USA
[2] Division of Biostatistics, Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA

FOR CITATION PURPOSES: *Mallick H, Li M*. Penalised regression methods in genetic research. OA Genetics 2013 Jul 01;1(1):7.

Competing interests: none declared. All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript. All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

OA Genetics
Open Access

Table 1 Different penalty functions (Assuming p coefficients in the model; for the group lasso we assume that there are K groups with m_k elements in each group, $k = 1 \ldots K$; for adaptive lasso the weights w_j's are assumed to be known)

Penalised method	Tuning parameters	Penalty function
Lasso	$\lambda > 0$	$\lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$
Ridge	$\lambda > 0$	$\lambda \sum_{j=1}^{p} \beta_j^2$
Bridge	$\lambda > 0, \alpha \leq 1$	$\lambda \sum_{j=1}^{p} \lvert\beta_j\rvert^{\alpha}$
Adaptive Lasso	$\lambda > 0$	$\sum_{j=1}^{p} w_j \lvert\beta_j\rvert$
Elastic Net	$\lambda_1 > 0, \lambda_2 > 0$	$\lambda_1 \sum_{j=1}^{p} \lvert\beta_j\rvert + \lambda_2 \sum_{j=1}^{p} \beta_j^2$
SCAD	$\lambda \geq 0, a > 2$	$\sum_{j=1}^{p} p_{\lambda,a}(\lvert\beta_j\rvert) \text{ where } p_{\lambda,a}(\lvert\beta_j\rvert) = \begin{cases} \lambda\lvert\beta_j\rvert, & \text{if } \lvert\beta_j\rvert \leq \lambda \\ -\dfrac{\beta_j^2 - 2a\lambda\lvert\beta_j\rvert + \lambda^2}{2(a-1)}, & \text{if } \lambda < \lvert\beta_j\rvert \leq a\lambda \\ \dfrac{(a+1)}{2}\lambda^2, & \text{if } \lvert\beta_j\rvert > a\lambda \end{cases}$
MCP	$\lambda \geq 0, \gamma > 1$	$\sum_{j=1}^{p} p_{\lambda,\gamma}(\lvert\beta_j\rvert) \text{ where } p_{\lambda,\gamma}(\lvert\beta_j\rvert) = \begin{cases} \lambda\lvert\beta_j\rvert - \dfrac{\beta_j^2}{2\gamma}, \text{if } \lvert\beta_j\rvert \leq \gamma\lambda \\ \dfrac{1}{2}\gamma\lambda^2, \text{if } \lvert\beta_j\rvert > \gamma\lambda \end{cases}$
Group Lasso	$\lambda > 0$	$\lambda \sum_{k=1}^{K} \sqrt{\sum_{j=1}^{m_k} \beta_{k_j}^2}$
Fused Lasso	$\lambda_1 > 0, \lambda_2 > 0$	$\lambda_1 \sum_{j=1}^{p} \lvert\beta_j\rvert + \lambda_2 \sum_{j=2}^{p} \lvert\beta_j - \beta_{j-1}\rvert$

multicollinearity due to highly linked variants and reduce the burden of multiple testing. Penalised regression methods attempt to attain a parsimonious model comprising only a small number of variants that are most important for accurate disease prediction, stable effect estimation and easy interpretation. For example, Ayers and Cordell[4] compared various penalised regression methods and single variant analysis under various simulation scenarios and found that penalised methods usually outperform single-SNP analysis, preventing correlated variants from entering the model and producing a sparse model with causal variants. Because of these appealing features, penalised methods and their Bayesian counterparts have been applied extensively in a wide variety of

FOR CITATION PURPOSES: *Mallick H, Li M*. Penalised regression methods in genetic research. OA Genetics 2013 Jul 01;1(1):7.

Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

OA Genetics
Open Access

topics in genetic research, such as, analyses of multiple genetic variants, complex gene-gene/gene-environment interactions, rare sequence variants, copy number variants, epigenetic modifications, etc. In the following sections, we briefly review some of their successful applications in various contexts of genetic research.

Multi-locus association test

Penalised regression methods have been effectively applied for multi-locus analyses in both candidate gene-based and genome wide association studies. In these applications, the joint effect of multiple variants is assumed to be additive without any interactions. For example, Li et al.[9] proposed a Bayesian lasso method for GWAS by assuming that the genetic effects have a double exponential distribution as prior. The method models each individual variant with an additive effect and a dominant effect, on which lasso penalties are imposed. The method used MCMC simulations to provide posterior median effect estimates for each variant, while adjusting for the effects of the other variants and covariates. Furthermore, based on posterior samples, a heritability value was estimated for each variant to guide variable selection for variants contributing significantly to phenotype[9]. Breheny and Huang[10] incorporated a grouping structure into the analysis and applied a group MCP method by grouping variants located in the same gene. The group minimaxconcave penalty (MCP) penalty achieves variable selection at both individual level and group level. Therefore, not only can it identify important genes but also can select important variants within those genes. Cho et al.[11] used an elastic net method to jointly analyse variants on a genome-wide scale. Each variant was assumed to have an additive effect. The elastic net penalty also takes advantages of regularization properties (i.e. automatic variable selection and stable estimation in presence of multicollinearity). Both Bayesian lasso and adaptive lasso methods have also been applied to detect quantitative trait loci in plant and animal studies[12,13].

Gene-gene interaction

Gene-gene interaction, or termed "epistasis", occurs when the effect of one genetic variant is influenced by the existence of others[8,14]. Accumulating evidence has suggested that genetic interactions exist pervasively in biological pathways[15], and is a major source accounting for the issue of "missing heritability" and the low replication power for the current positive findings[16,17]. Due to their capability of handling high-order interactions and differentiating interaction effects from main genetic effects, penalised regression methods are also widely used to detect epistasis interactions. Wu et al.[18] developed a two-stage lasso penalised logistic regression to handle genetic interactions in genome-wide association studies. In the first stage, the top variants with the most significant effects were identified. In the second stage, the two-way or higher-order interactions among the selected variants were examined. Yang et al.[19] proposed a group adaptive lasso method for GWAS analysis. All variants and their interactions were treated as multi-level factors, which were detected in a group manner.

A particular type of gene-gene interaction in maternal and prenatal research is the maternal-foetal genotype (MFG) interactions, which occurs when an MFG combination jointly alters the phenotype or risk of disease in the offspring. A well-known example of an MFG interaction is Rh incompatibility[20]. An Rh-negative mother may produce immune antibodies to the Rh antigens on the red blood cells of her Rh-positive foetus, causing Rh isoimmunization. Penalised regression methods have also been used to detect MFG interactions. Li et al.[21] defined a genetic conflict indicator if the baby has a different genotype from its mother. A ridge regression is used to address the data collinearity between maternal and foetal genomes. Alternatively, Li et al.[22] used adaptive lasso embedded within an EM algorithm (to simultaneously detect phased haplotype probabilities) to detect haplotype-haplotype interaction between maternal and foetal genomes, which also differentiate the genetic effects from maternal genotypes, foetal genotypes and MFG interactions.

Gene-environment (GXE) interaction

Similar to epistasis, gene-environment interaction plays a crucial role in understanding the genetic basis of complex diseases. Application of penalised methods to detect gene-environmental interaction is similar to those for detecting gene-gene interaction. The product terms between genetic variants and environmental factors can be incorporated in the model, the effects of which are further estimated subject to a penalty term. Park and Hastie[23] used ridge-penalised logistic regression followed by a forward selection strategy to detect epistasis and gene-environment interactions. Tanck et al.[24] implemented penalised regression, with ridge penalty on main effects and lasso penalty on epistasis and GXE interaction effects. Therefore, all the main effects are always included in the model (property of ridge), while irrelevant interactions are automatically removed (property of lasso). Due to their capability of handling correlated variables, these methods can handle a large number of variants and their GXG and GXE interactions.

Next generation sequencing data and rare variants

Recent evidence has shown rare variants, though individually rare,

FOR CITATION PURPOSES: *Mallick H, Li M*. Penalised regression methods in genetic research. OA Genetics 2013 Jul 01;1(1):7.

Competing interests: none declared. All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript. All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

OA Genetics
Open Access

may have a stronger effect and collectively have a significant impact on disease phenotypes[25,26]. Though rare variants are suggested to be a potential source of 'missing' heritability'[16], detecting rare sequence variants associated with complex diseases remains to be a challenge. A single rare variant contains little variation owing to low minor allele frequency (<0.5% or 1%); testing these variants individually lack reasonable statistical power[5]. Many researchers have proposed ways of collapsing information across genes or across other regions so that the combined exposure becomes less rare[27–29]. A number of penalised regression methods have also been proposed to handle rare variants, most of which follows various collapsing strategies.

Recently, Zhou et al.[30] used a mixture of group lasso and lasso penalty to jointly test common and rare variants for association with disease phenotypes. The method collapsed rare variants with minor allele frequencies less than 1%, and introduced a group structure for all variants by genes and pathways. It was suggested that using a mixture of group lasso and lasso penalties outperformed using lasso penalties alone, especially when both common and rare variants are present. More recently, Ayers and Cordell[5] developed a method that groups SNPs by genes and collapses the rare variants in the gene into a single "super" variant. The penalty term imposed in their method allows an individual regression coefficient to be estimated for each common variant, effectively allowing individual common variants to be selected, while the grouping penalty allows a borrowing of strength between common and rare variants within the same gene. When applied to real and simulated datasets, their approach showed improved performance compared to its predecessors.

Copy number variants

In the past few years, solid evidence has shown that structural variations, due to insertions, deletions and inversions of the DNA, also contribute considerably to the diversity of the human genome[31–33]. These structural changes will cause copy number differences in particular genomic regions, ranging from one KB to complete chromosome arm. CNVs may contribute considerably to the development of complex diseases, like cancers[34] and are a major source of the "missing heritability" of complex human diseases[16].

Penalised methods have also been proposed to detect CNVs. Huang et al.[35] proposed to use a least squares regression model and penalised the difference between the relative copy numbers of the neighbouring markers. By using this lasso-type penalty, the change points of CNVs can be detected. Gao et al.[36] later considered the sparsity of CNVs in the genome, and proposed a robust penalised LAD regression model with the adaptive fused lasso penalty. The method was shown to be robust to outliers and correctly detected the numbers and locations of the true breakpoints.

Epigenetics

Epigenetic refers to the modifications of DNA or associated proteins, other than DNA sequence variation itself, which carry information content during cell division[37]. Two major molecular mechanisms for epigenetic inheritance are, DNA methylation and histone modification, both of which may lead to heritable changes in gene expression or cellular phenotypes[38]. Epigenetic alteration have long been linked to complex human diseases, such as cancers[37], disorders of genomic imprinting[39], neuropsychiatric diseases[40], autism[41], etc. Detecting epigenetic alterations contributing to complex human disease would also help to account for the issue of "missing heritability"[16].

Penalised regression methods are also used to investigate DNA methylation data. Sun and Wang[42] applied a penalised conditional logistic regression model for matched case-control studies. The method used a network-based penalty to favour selection of Cytosine-phosphate-Guanine (CpG) sites within a gene or genetic pathway. Liu et al.[43] applied a bridge-penalised logistic regression method. Compared to lasso that usually selects independent variants, the proposed sparse logistic regression was able to select highly correlated variants simultaneously. Application of the method to methylation data selected 6 out of 7 CpG regions, which are known to be predictive of lung cancer subtype.

Discussion

With the advent of genomic era, it is now feasible to investigate the influence of the entire spectrum of human genetic variations on complex human diseases. Quite often we will need to examine a large number of genetic variants far exceeding the number of individuals in the study population. Traditionally regression-based methods would be overwhelmed to jointly consider all variants simultaneously. Over the years, penalised methods have emerged as a powerful tool in genetic research, covering a wide variety of topics, such as multivariant analysis, gene-gene/gene-environment interactions, rare sequence variants, copy number variants, epigenetics, etc. These research areas represent various sources that may account for the "missing heritability" of complex human diseases. Penalised regression methods have shown a number of advantages, such as facilitating model selection in high-dimensional data analysis, achieving stable effect estimation in the presence of multicollinearity, and reducing the burden of multiple testing. Application of penalised regression methods in various research topics also successfully

FOR CITATION PURPOSES: *Mallick H, Li M*. Penalised regression methods in genetic research. OA Genetics 2013 Jul 01;1(1):7.

Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

OA Genetics
Open Access

identified genetic/epigenetic/copy number variants associated with complex human diseases, differentiated causal variants from non-causal ones, and estimated their effects while adjusting for the effect of others. In this article, we attempt to provide a survey of the application of penalised regression methods in various contexts of genetic research. The application of penalised methods is not limited to these topics. Adaptation of these methods to transcriptomics, proteomics and metabolomics data can be straightforward and have been investigated[44–46].

It should also be noted that penalised regression methods might also have a few limitations. First, despite their success in model selection, their performance has been unsatisfactory in hypothesis testing and interval estimation[47]. In particular, in genetic studies with large number of variants, lasso and related methods may also produce false positive results[47]. To overcome this limitation, one might combine multiple test statistics by averaging over multiple tuning parameters, rather than building a single 'best' model[47]. Second, although penalised methods provide robust performances for detecting causal variants under various disease models[48], they might not be the 'best' in all the situations. Other non-penalised approaches can be more advantageous depending on the underlying mechanism and allele frequency of the disease model[49]. For example, compared to lasso and group lasso method, the multifactor-dimensionality reduction (MDR) method was shown to have a higher power to detect pure epistatic interactions among common variants[49]. Therefore, in practice, optimised performance of these methods would be both model and context-dependent. Third, although penalised methods have been applied to various research areas, their performance requires further improvement, especially for detecting rare sequence variants. The current available penalised methods also have a few limitations due to the collapsing strategy of the rare variants due to the presence of: (i) both disease causing and disease-protective variants and (ii) both functional and non-functional variants within a region. There is a huge scope of improvement and further assessment in association testing with rare variants[50]. Recently, a number of similarity-based methods, such as SIMreg and SKAT, have been shown to be robust to the bi-direction of genetic effect in sequencing data analysis[51–55].

Conclusion

These methods are built on the assumption that, given a genotype-phenotype association, the genetic similarity would contribute to the phenotype similarity and aggregate multiple rare variants and common variants through the genetic similarities between individuals. It would be interesting to see if penalised methods can be incorporated to conduct selection among the similarity of genes, environmental factors and their interactions.

Acknowledgements

This work is supported in part by a cooperative agreement grant U01 NS041588 from the National Institute of Neurological Disorders and Stroke, National Institutes of Health and Department of Health and Human Services.

Abbreviations list

GWAS generation genome-wide association studies; LD, Linkage-Disequilibrium; LASSO, least absolute shrinkage and selection operator; AIC, Akaike information criterion; BIC Bayesian information criterion; MCMC, Markov chain Monte Carlo; MCP minimaxconcave penalty; MFG, maternal-foetal genotype; CpG Cytosine-phosphate-Guanine; MDR, multifactor-dimensionality reduction.

References

1. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods. 2008 Jan;5(1):16–8.
2. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA. 2009 Jun;106(23):9362–7.
3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461 (7265):747–53.
4. Ayers KL, Cordell HJ. SNP Selection in genome-wide and candidate gene studies via penalised logistic regression. Genet Epidemiol. 2010 Dec;34(8):879–91.
5. Ayers KL, Cordell HJ. Identification of grouped rare and common variants via penalised logistic regression. Genet Epidemiol. 2013 Sep;37(6):592–602.
6. Li Z, Sillanpää MJ. Overview of LASSO-related penalised regression methods for quantitative trait mapping and genomic selection. Theor Appl Genet. 2012 Aug;125(3):419–35.
7. Mallick H, N Yi. Bayesian Methods for High Dimensional Linear Models. J Biomet Biostat. 2013 Jun;1:005.
8. YI N. Statistical analysis of genetic interactions. Genet Res Camb. 2010 Dec;92(5–6):443–59.
9. Li J, Das K, Fu G, Li R, Wu R. The Bayesian lasso for genome-wide association studies. Bioinformatics. 2011 Feb; 27(4):516–23.
10. Breheny P, Huang J. Penalised methods for bi-level variable selection. Stat Interface. 2009 Jul;2(3):369–80.
11. Cho S, Kim K, Kim YJ, Lee JK, Cho YS, Lee JY, et al. Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. Ann Hum Genet. 2010;74(5): 416–28.
12. Yi N, S Xu. Bayesian LASSO for quantitative trait loci mapping. Genetics. 2008 Jun;179(2):1045–55.
13. Sun W, Ibrahim JG, Zou F. Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalised regression. Genetics. 2010 May; 185(1):349–59.
14. Cordell HJ. Detecting gene–gene interactions that underlie human diseases. Nat Rev Genet. 2009 Jun;10(6): 392–404.

FOR CITATION PURPOSES: *Mallick H, Li M*. Penalised regression methods in genetic research. OA Genetics 2013 Jul 01;1(1):7.

Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

OA Genetics
Open Access

15. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered. 2003;56(1–3):73–82.
16. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010 Jun;11(6):446–50.
17. Maher B. The case of the missing heritability. Nature. 2008 Nov;456(7218):18–21.
18. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalised logistic regression. Bioinformatics. 2009 Mar;25(6):714–21.
19. Yang C, Wan X, Yang Q, Xue H, Yu W. Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso. BMC Bioinformatics. 2010 Jan;11(Suppl 1):S18.
20. Kulich V, Kout M. Hemolytic disease of a newborn caused by anti-k antibody. Cesk Pediatr. 1967 Sep;22(9):823–6.Czech.
21. Li S, Lu Q, Fu W, Romero R, Cui Y. A regularized regression approach for dissecting genetic conflicts that increase disease risk in pregnancy. Stat Appl Genet Mol Biol. 2009;8.
22. Li M, Romero R, Fu WJ, Cui Y. Mapping haplotype-haplotype interactions with adaptive LASSO. BMC Genet. 2010 Aug;11:79.
23. Park MY, Hastie T. Penalised logistic regression for detecting gene interactions. Biostatistics. 2008 Jan;9(1):30–50.
24. Tanck MW, Jukema JW, Zwinderman AH. Simultaneous estimation of gene-gene and gene-environment interactions for numerous loci using double penalised log–likelihood. Genet Epidemiol. 2006 Dec;30(8):645–51.
25. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008 Jun;40(6):695–701.
26. McClellan J, King MC. Genetic heterogeneity in human disease. Cell. 2010 Apr; 141(2):210–17.
27. Li B, SM Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008 Sep;83(3):311–21.
28. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009 Feb;5(2):e1000384.
29. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010;34(2):188–93.
30. Zhou H, Sehl ME, Sinsheimer JS, Lange K. Association screening of common and rare genetic variants by penalised regression. Bioinformatics. 2010 Oct;26(19):2375–82.
31. Bhangale TR, Stephens M, Nickerson DA. Automating resequencing-based detection of insertion-deletion polymorphisms. Nat Genet. 2006 Dec;38(12):1457–62.
32. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet. 2006;38(1):75–81.
33. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, et al. Common deletion polymorphisms in the human genome. Nat Genet. 2005 Jan;38(1):86–92.
34. Frank B, Bermejo JL, Hemminki K, Sutter C, Wappenschmidt B, Meindl A, et al. Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk. Carcinogenesis. 2007 Jul;28(7):1442–45.
35. Huang T, Wu B, Lizardi P, Zhao H. Detection of DNA copy number alterations using penalised least squares regression. Bioinformatics. 2005 Oct;21(20):3811–7.
36. Gao X, J Huang. A robust penalised method for the analysis of noisy DNA copy number data. BMC Genomics. 2010 Sep;11:517.
37. Feinberg AP, Tycko B. The history of cancer epigenetics. Nat Rev Cancer. 2004 Feb;4(2):143–53.
38. Hammoud SS, Nix DA, Zhang H, Purwar J, Carrell DT, Cairns BR. Distinctive chromatin in human sperm packages genes for embryo development. Nature. 2009 Jul;460(7254):473–78.
39. Horsthemke B, Buiting K. Genomic imprinting and imprinting defects in humans. Adv genet. 2008;61:225–46.
40. Petronis A, Gottesman II, Crow TJ, Delisi LE, Klar AJ, Macciardi F, et al. Psychiatric epigenetics: a new focus for the new century. Mol Psychiatry. 2000 Jul;5(4):342–6.
41. Bakkaloglu B, O'Roak BJ, Louvi A, Gupta AR, Abelson JF, Morgan TM, et al. Molecular Cytogenetic Analysis and Resequencing of Contactin Associated Protein-Like 2 in Autism Spectrum Disorders. Am J Hum Genet. 2008;82(1):165–73.
42. Sun H S Wang. Network-based regularization for matched case-control analysis of high-dimensional DNA methylation data. Stat Med. 20133 May;32(12):2127–39.
43. Liu Z, Jiang F, Tian G, Wang S, Sato F, Meltzer SJ, et al. Sparse logistic regression with Lp penalty for biomarker identification. Stat Appl Genet Mol Biol. 2007;6(1).
44. Gevaert O, Villalobos V, Sikic BI, Plevritis SK. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. Interface Focus. 2013 Aug;3(4):20130013.
45. Lagani V, G Kortas, I Tsamardinos. Biomarker signature identification in "omics" data with multi-class outcome. Comput Struct Biotechnol J. 2013;6.
46. Rohart F. Paris A, Laurent B, Canlet C, Molina J, Mercat MJ et al. Phenotypic prediction based on metabolomic data: for growing pigs from three main European breeds. 2012 Dec;90(13):4729–40.
47. Basu S, Pan W, Shen X, Oetting WS. Multilocus association testing with penalised regression. Genet Epidemiol. 2011 Dec;35(8):755–65.
48. Abraham G, Kowalczyk A, Zobel J, Inouye M. Performance and robustness of penalised and unpenalised methods for genetic prediction of complex human disease. Genet Epidemiol. 2013 Feb; 37(2):184–95.
49. Winham S, C Wang, AA Motsinger-Reif. A comparison of multifactor dimensionality reduction and L1-penalised regression to identify gene-gene interactions in genetic association studies. Stat Appl Genet Mol Biol. 2011;10(1).
50. Yi N, Zhi D. Bayesian analysis of rare variants in genetic association studies. Genet Epidemiol. 2011 Jan;35(1):57–69.
51. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012 Sep;13(4):762–75.
52. Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, et al. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. Am J Hum Genet. 2011 Aug;89(2):277–88.
53. Tzeng JY, Zhang D, Chang SM, Thomas DC, Davidian M. Gene-trait similarity regression for multimarker-based association analysis. Biometrics. 2009 Sep;65(3):822–32.
54. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010 Jun;86(6):929–42.
55. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011 Jul;89(1):82–93.

Competing interests: none declared. Conflict of interests: none declared.
All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.