Genomics software: The view from 10,000 feet

*Michael E. Weale**

Department of Medical and Molecular Genetics, King's College London, 8th Floor, Tower Wing, Guy's Hospital, London SE1 9RT, UK
**Correspondence to*: Tel: +44 (0)207 188 2601; Fax: +44 (0)207 188 2585; E-mail: michael.weale@kcl.ac.uk

Date received: 18th September 2009

Abstract

The rate of change in genomics, and 'omics generally, shows no signs of slowing down. Related analysis software is struggling to keep apace. This paper provides a brief review of the field.

Keywords: *genomics, software, systems genetics, 'omics, genome-wide association studies, SNP annotation, networks*

For the bioinformaticist, and still more so the traditional genetic epidemiologist, the big view on how to tackle genomic data analysis looks daunting. Only a few years ago, the genome-wide association study (GWAS) represented the overshadowing Everest on the landscape, and commentators fretted about the computational feasibility of analysing 500,000 or so single nucleotide polymorphisms (SNPs) against one phenotype variable. Now, the single-phenotype GWAS is a foothill from which to launch attacks on datasets of much greater scale and variety. A new term — systems genetics — has emerged to describe this expanded world view. Analytical tools for dealing with molecular data have always lagged behind the acceleration of high-throughput methods for generating them, and that seems especially true at the present time. Here, I provide an overview of the software currently available, and look ahead to future developments.

First, a look at more familiar territory. The software package PLINK[1] has become the favoured work-horse of GWAS analysis, thanks to the untiring efforts of Shaun Purcell to keep the software well documented, flexible, fast and compact in its use of data structures. Few other packages surpass PLINK as far as basic quality control and first-pass SNP-by-SNP analysis are concerned, and many other, more advanced features are available and are being expanded continuously. In addition to SNP probes, modern GWAS panels are equipped with additional probe sets for interrogating copy number variation (CNV). PennCNV[2] is a popular software for calling these. CNV call uncertainty poses downstream problems for association analysis, and software for dealing with this has been reviewed recently in this journal.[3] Another trend is towards imputation of SNPs that are not present in the GWAS panel but can be inferred via linkage disequilibrium (LD), also reviewed here recently.[4] Popular choices are Mach,[5] Impute[6] and Beagle.[7] A more specialist imputation problem, but one of general interest due to the role of the immune response system in many diseases, is to call classical human leukocyte antigen (HLA) genotypes from SNPs typed in the HLA region of chromosome 6. Recently improved software from Gil McVean and colleagues is available for this.[8]

SNP annotation tools provide the most straightforward window from GWAS hits and also sequence data into the wider 'omic universe. A recent review is by Rachel Karchin.[9] The SNP Function Portal[10] provides one of the more comprehensive lists of annotation for each SNP, including those arising via LD proxy or 'tagging'. Other options include FastSNP,[11] PupaSuite,[12] SNPnexus,[13] SNPinfo,[14] SNPselector,[15] F-SNP[16] and TAMAL.[17]

WGAviewer[18] is geared specifically towards the analysis of GWAS results, and has a nice visual interface. All these tools struggle to keep up with the rapidly expanding set of available annotations. For example, several different datasets are now publically available that combine GWAS SNP data with genome-wide gene expression data (so-called genetical genomics or expression quantitative trait locus [eQTL] data). Currently, however, no one tool integrates the ability to search all these datasets simultaneously. One option for the more proficient investigator is to keep one step ahead by using the Galaxy web tool[19] to design their own application for integrating different annotation tracks with their GWAS hits. SNAP[20] is a useful tool for feeding LD proxy information into such a custom-made Galaxy application.

Beyond SNP annotation, there are more formal attempts at linking genetic data into functional networks. These may be created from internal sources, such as *p*-values for SNP–SNP interactions, or extrinsic sources, such as protein–protein interactions (reviewed here recently[21]) and gene ontology categories. A repository of types of network data is available at http://www.pathwaycommons.org. While network visualisation tools were previously the domain of expensive commercial software, Cytoscape[22] has become an excellent freeware alternative. For formal statistical significance of coincident patterns within these networks, there is a rapidly expanding literature and no consensus yet on the best approach to take. Two examples are ALIGATOR[23] and gene-set enrichment analysis (GSEA). The latter has been adapted from gene expression studies and applied to GWAS *p*-values. Web-based implementations are available at http://bioinfo.vanderbilt.edu/webgestalt and http://www.broadinstitute.org/gsea.

How can one keep up to date in the rapidly changing world of genomic software? Certainly, review sections such as the one here in *Human Genomics* will help. *Nucleic Acids Research* publishes a useful annual review of web server applications,[24] now also available online (http://bioinformatics.ca/links_directory). The Applications Note section of the journal *Bioinformatics* provides the best, but by no means only, location for primary literature on new software. Looking ahead, software for handling high-throughput sequencing is an area where we can expect much development in the coming months. *Bioinformatics* has a useful online 'virtual issue' on tools for next generation sequencing which they are recurrently updating (http://www.oxfordjournals.org/our_journals/bioinformatics/nextgenerationsequencing.html). One wonders whether 10,000 feet will be high enough for a synoptic view in 12 months time.

References

1. Purcell, S. *et al.* (2007), 'PLINK: A tool set for whole-genome association and population-based linkage analyses', *Am. J. Hum. Genet.* Vol. 81, pp. 559–575.
2. Wang, K. *et al.* (2007), 'PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data', *Genome Res.* Vol. 17, pp. 1665–1674.
3. Plagnol, V. (2009), 'Association tests and software for copy number variant data', *Hum. Genomics* Vol. 3, pp. 191–194.
4. Ellinghaus, D., Schreiber, S., Franke, A. and Nothnagel, M. (2009), 'Current software for genotype imputation', *Hum. Genomics* Vol. 3, Vol. 371–380.
5. Li, Y. and Abecasis, G.R. (2006), 'Mach 1.0: Rapid haplotype reconstruction and missing genotype inference', *Am. J. Hum. Genet.* Vol. S79, p. 2290.
6. Marchini, J., Howie, B., Myers, S., McVean, G. *et al.* (2007), 'A new multipoint method for genome-wide association studies by imputation of genotypes', *Nat. Genet.* Vol. 39, pp. 906–913.
7. Browning, B.L. and Browning, S.R. (2009), 'A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals', *Am. J. Hum. Genet.* Vol. 84, pp. 210–223.
8. Leslie, S., Donnelly, P. and McVean, G. (2008), 'A statistical method for predicting classical HLA alleles from SNP data', *Am. J. Hum. Genet.* Vol. 82, pp. 48–56.
9. Karchin, R. (2009), 'Next generation tools for the annotation of human SNPs', *Brief. Bioinform.* Vol. 10, pp. 35–52.
10. Wang, P. *et al.* (2006), 'SNP Function Portal: A web database for exploring the function implication of SNP alleles', *Bioinformatics* Vol. 22, pp. e523–529.
11. Yuan, H.Y. *et al.* (2006), 'FASTSNP: An always up-to-date and extendable service for SNP function analysis and prioritization', *Nucleic Acids Res.* Vol. 34, pp. W635–W641.
12. Conde, L. *et al.* (2006), 'PupaSuite: Finding functional single nucleotide polymorphisms for large-scale genotyping purposes', *Nucleic Acids Res.* Vol. 34 (Web Server issue), pp. W621–W625.
13. Chelala, C., Khan, A. and Lemoine, N.R. (2009), 'SNPnexus: A web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms', *Bioinformatics* Vol. 25, pp. 655–661.
14. Xu, Z. and Taylor, J.A. (2009), 'SNPinfo: Integrating GWAS and candidate gene information into functional SNP selection for genetic association studies', *Nucleic Acids Res.* Vol. 37 (Web Server issue), pp. W600–W605.
15. Xu, H. *et al.* (2005), 'SNPselector: A web tool for selecting SNPs for genetic association studies', *Bioinformatics* Vol. 21, pp. 4181–4186.
16. Lee, P.H. and Shatkay, H. (2008), 'F-SNP: Computationally predicted functional SNPs for disease association studies', *Nucleic Acids Res.* Vol. 36 (Database issue), pp. D820–D824.

17. Hemminger, B.M., Saelim, B. and Sullivan, P.F. (2006), 'TAMAL: An integrated approach to choosing SNPs for genetic studies of human complex traits', *Bioinformatics* Vol. 22, pp. 626–627.
18. Ge, D. *et al*. (2008), 'WGAViewer: Software for genomic annotation of whole genome association studies', *Genome Res.* Vol. 18, pp. 640–643.
19. Giardine, B. *et al*. (2005), 'Galaxy: A platform for interactive large-scale genome analysis. *Genome Res.* Vol. 15, pp. 1451–1455.
20. Johnson, A.D. *et al*. (2008), 'SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap', *Bioinformatics* Vol. 24, pp. 2938–2939.
21. Lehne, B. and Schlitt, T. (2009), 'Protein-protein interaction databases: Keeping up with growing interactomes. *Hum. Genomics* Vol. 3, pp. 291–297.
22. Shannon, P. *et al*. (2003), 'Cytoscape: A software environment for integrated models of biomolecular interaction networks', *Genome Res.* Vol. 13, pp. 2498–2504.
23. Holmans, P. *et al*. (2009), 'Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder', *Am. J. Hum. Genet*. Vol. 85, pp. 13–24.
24. Benson, G. (2009), 'Nucleic Acids Research Annual Web Server Issue in 2009', *Nucleic Acids Res.* Vol. 37, pp. W1–W2.