Minireview

Using prior knowledge and genome-wide association to identify pathways involved in multiple sclerosis

Marylyn D Ritchie

Address: Department of Molecular Physiology & Biophysics, Center for Human Genetics Research, Vanderbilt University, 519 Light Hall, Nashville, TN 37232-0700, USA. Email: ritchie@chgr.mc.vanderbilt.edu

Published: 29 June 2009

Genome Medicine 2009, 1:65 (doi:10.1186/gm65)

The electronic version of this article is the complete one and can be found online at http://genomemedicine.com/content/1/6/65

Abstract

The efforts of the Human Genome Project are beginning to provide important findings for human health. Technological advances in the laboratory, particularly in characterizing human genomic variation, have created new approaches for studying the human genome - genome-wide association studies (GWAS). However, current statistical and computational strategies are taking only partial advantage of this wealth of information. In the quest for susceptibility genes for complex diseases in GWAS data, several different analytic strategies are being pursued. In a recent report, Baranzini and colleagues used a pathway- and network-based analysis to explore potentially interesting single locus association signals in a GWAS of multiple sclerosis. This and other pathway-based approaches are likely to continue to emerge in the GWAS literature, as they provide a powerful strategy to detect important modest single-locus effects and gene-gene interaction effects.

Current approaches in genome-wide association studies

In the search for susceptibility genes for common complex diseases, we are faced with enormous challenges. The past decade's paradigm of focusing a study on just one or a few candidate genes limits our ability to identify novel genetic effects associated with disease. In addition, many susceptibility genes can show effects that are partially or solely dependent on interactions with other genes and/or the environment. Genome-wide association studies (GWAS) have been proposed as a solution to these problems; however, the analysis of whole-genome data is problematic because we must separate the one or few true but modest signals from the extensive background noise. Moreover, with GWAS data alone, the ability to elucidate gene-environment interactions is limited. GWAS researchers must embrace the abundant clinical and environmental data available to complement the rich genotypic data, with the ultimate goal of revealing the genetic and environmental factors that are important for disease risk. So far, GWAS have taken a simplistic, 'one SNP at a time' analysis approach. This approach is ignoring the complexity of common complex diseases.

Recent technological advances enable the genotyping of hundreds of thousands of human single-nucleotide polymorphisms (SNPs) on thousands of samples. We are hindered in exploiting these laboratory advances because strategies for analyzing the data have not kept pace with technological progress. Even with these challenges, successful reports of GWAS have emerged in the literature [1-6]. In fact, the National Human Genome Research Institute (NHGRI) keeps an updated GWAS catalog on their website [7], which lists over 273 published GWAS so far. Unfortunately, as expected, only the strongest associations can be detected using these traditional approaches, and there are many more genes still to be found [8,9].

The majority of these studies analyzed one SNP at a time, meaning that they have barely scratched the surface of interesting information within these datasets. Ultimately, supplementary data, replication datasets, or multiple analytical approaches must be used to filter the results down to a manageable number of the 'most likely' genes. In their recent report in *Human Molecular Genetics,* Baranzini *et al.* [10] developed and applied a pathway- and network-based analysis to exploit interesting association signals in the SNPs that fell between the thresholds of $P = 0.05$ and $P = 10^{-8}$ in the original single-SNP association analysis. It is methods such as this that are likely to allow us to better characterize and exploit GWAS signals in the so-called 'gray region' between genome-wide significance ($P = 10^{-8}$) and the typical $P = 0.05$. In the field, genome-wide significance has been determined to be at $P = 10^{-8}$ because that is the Bonferroni correction for $P = 0.05$ for 1 million tests.

Pathway-based results in multiple sclerosis

Because some of the replicating, positive results in GWAS fell below the level of genome-wide significance (that is, had *P*-values over 10^{-8}), Baranzini and colleagues [10] propose a protein interaction and network-based analysis (PINBPA) for the study of a multiple sclerosis (MS) dataset. This approach is similar to those in microarray studies in which gene ontologies are used for analysis [11]. The idea of using prior knowledge for GWAS has been used successfully in studies of diseases such as Parkinson's disease, age-related macular degeneration, bipolar disorder, rheumatoid arthritis, and Crohn's disease [12-14].

The first step of PINBPA is to compute a gene-wise *P*-value by choosing the lowest *P*-value of all SNPs mapping to a given gene. These genes are then mapped onto a curated protein interaction network. Any markers that do not map to genes or unannotated genes are eliminated from this analysis. Next, using a plug-in for the Cytoscape [15] software, searches are conducted to extract potentially meaningful sub-networks associated with the phenotype of interest. Finally, a test is performed to determine the extent to which significant network modules could be obtained by chance. Baranzini and colleagues [10] applied PINBPA to two MS datasets: (i) the International Multiple Sclerosis Genetics Consortium (IMSGC) GWAS [16], consisting of 334,923 SNPs passing quality control from the Affymetrix Human Mapping 500K Array in 931 family trios, and (ii) the GeneMSA study [17], with 551,642 SNPs passing quality control from the Illumina HumanHap550 bead chip in 978 cases and 883 controls. After single-locus analysis using logistic regression, 78 and 87 SNPs had a *P*-value of less than 1×10^{-4} in the IMSGC and GeneMSA datasets, respectively.

Using PINBPA analysis, 346 significant modules were identified on the basis of their aggregate degree of association with MS. Because of the nature of the algorithm, many modules overlap extensively; thus, the modules with the highest scores were selected. Module I included several human leukocyte antigen (HLA) genes, including the known risk factor for MS *HLA-DRB1*. Interestingly, this module shows *HLA-DRA* as the most significant node. *HLA-DRB1* and *HLA-DRA* are in high linkage disequilibrium and some SNPs in *HLA-DRA* serve as proxies for *HLA-DRB1* with high sensitivity. Module II includes an extensive pattern of immunity-related genes, including several HLA genes: *CD4*, *CD82*, *ITGB2*, *IL2Ra*, and *CD58*. Finally, modules III and IV suggest a neural component, including genes expressed in neurons and glia, such as *NCK2*, *EPHA3*, *EPHA4* (module III) and glutamate receptor genes (module IV) and many more. The results of this study [10] provide insights into the role of several immunological pathways, including cell adhesion, signaling, and communication, and, more importantly, neural pathways in MS. In particular, signals for axon guidance and synaptic potentiation were overrepresented in MS. This is very exciting, as it is one of the first reports demonstrating genetic associations in a neural pathway contributing to the susceptibility of MS. Because the pathophysiology of MS suggests that neural pathways are likely to have a role, these results provide enormous potential for follow-up research.

The future of GWAS using prior knowledge and pathway-based approaches

Baranzini *et al.* [10] demonstrate the utility of protein interaction network information in the analysis of MS data. Several GWAS [12-14,18,19] have proposed the use of prior knowledge in the form of pathway databases, such as the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) and Biocarta, or gene ontology databases. Baranzini *et al.* [10] suggest that the network-based approach not only reduces the number of relevant interactions found but also increases the likelihood that proteins that interact are part of the same biological pathway [10]. This approach was certainly successful for MS.

Consistent with this line of thought, Bush *et al.* [20] have constructed the Biofilter as another alternative approach for detecting interactions in GWAS data. The Biofilter combines six sources of disease-independent information (information that is not related to the phenotype of interest) from the public domain: KEGG, Reactome, Gene Ontology, Database of Interacting Proteins (DIP), Protein Families Database (PFAM), and Netpath. It also includes disease-dependent information in the nature of previous linkage regions, association studies, and microarray expression results. All these sources are combined specifically to prioritize the search for gene-gene interactions in GWAS data [20].

Pathway-based approaches are continuing to emerge in the literature as a more comprehensive approach to the analysis of GWAS data. This trend is likely to continue as we learn more about the optimal strategies for incorporating prior knowledge into analyses. In fact, as we move to using next-generation sequencing data, such approaches may also expand into the next-generation sequencing arena: looking for rare variants in a particular pathway that are present in a higher proportion of disease cases than healthy controls. As more biological knowledge and genomic data become publicly available and more easily accessible, we will continue to see methodological developments exploit this information to better dissect the genetic architecture of common, complex disease.

Abbreviations

GWAS, genome-wide association studies; IMSGC, International Multiple Sclerosis Genetics Consortium; KEGG, *Kyoto Encyclopedia of Genes and Genomes;* MS, multiple sclerosis; PINBPA, protein interaction and network-based analysis; SNP, single-nucleotide polymorphism.

Competing interests

MDR is currently analyzing the IMSGC GWAS data, along with a subset of the GeneMSA data, in a Biofilter analysis to explore gene-gene interactions in MS.

Acknowledgements

This work was supported by NIH grants HL065962, HG004608, and HG004603.

References

1. Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447:**661-678.
2. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R; SEARCH collaborators, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, *et al.*: **Genome-wide association study identifies novel breast cancer susceptibility loci.** *Nature* 2007, **447:**1087-1093.
3. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Thomas G, Chanock SJ: **A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer.** *Nat Genet* 2007, **39:**870-874.
4. Lyon HN, Emilsson V, Hinney A, Heid IM, Lasky-Su J, Zhu X, Thorleifsson G, Gunnarsdottir S, Walters GB, Thorsteinsdottir U, Kong A, Gulcher J, Nguyen TT, Scherag A, Pfeufer A, Meitinger T, Brönner G, Rief W, Soto-Quiros ME, Avila L, Klanderman B, Raby BA, Silverman EK, Weiss ST, Laird N, Ding X, Groop L, Tuomi T, Isomaa B, Bengtsson K, *et al.*: **The association of a SNP upstream of INSIG2 with body mass index is reproduced in several but not all cohorts.** *PLoS Genet* 2007, **3:**e61.
5. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Boström K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, *et al.*: **Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.** *Science* 2007, **316:**1331-1336.
6. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, *et al.*: **A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants.** *Science* 2007, **316:**1341-1345.
7. **Office of Population Genomics: A catalog of published genome-wide association studies** [http://www.genome.gov/26525384]
8. Couzin J, Kaiser J: **Genome-wide association. Closing the net on common disease genes.** *Science* 2007, **316:**820-822.
9. Williams SM, Canter JA, Crawford DC, Moore JH, Ritchie MD, Haines JL: **Problems with genome-wide association studies.** *Science* 2007, **316:**1840-1842.
10. Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BM, Kappos L; GeneMSA Consortium, Polman CH, Matthews PM, Hauser SL, Gibson RA, Oksenberg JR, Barnes MR: **Pathway and network-based analysis of genome-wide association studies in multiple sclerosis.** *Hum Mol Genet* 2009, **18:**2078-2090.
11. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
12. Torkamani A, Topol EJ, Schork NJ: **Pathway analysis of seven common diseases assessed by genome-wide association.** *Genomics* 2008, **92:**265-272.
13. Wang K, Li M, Bucan M: **Pathway-based approaches for analysis of genomewide association studies.** *Am J Hum Genet* 2007, **81:**1278-1283.
14. Lesnick TG, Papapetropoulos S, Mash DC, Ffrench-Mullen J, Shehadeh L, de Andrade M, Henley JR, Rocca WA, Ahlskog JE, Maraganore DM: **A genomic pathway approach to a complex disease: axon guidance and Parkinson disease.** *PLoS Genet* 2007, **3:**e98.
15. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
16. International Multiple Sclerosis Genetics Consortium, Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, De Jager PL, de Bakker PI, Gabriel SB, Mirel DB, Ivinson AJ, Pericak-Vance MA, Gregory SG, Rioux JD, McCauley JL, Haines JL, Barcellos LF, Cree B, Oksenberg JR, Hauser SL: **Risk alleles for multiple sclerosis identified by a genomewide study.** *N Engl J Med* 2007, **357:**851-862.
17. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, Barkhof F, Radue EW, Lindberg RL, Uitdehaag BM, Johnson MR, Angelakopoulou A, Hall L, Richardson JC, Prinjha RK, Gass A, Geurts JJ, Kragt J, Sombekke M, Vrenken H, Qualley P, Lincoln RR, Gomez R, Caillier SJ, George MF, Mousavi H, Guerrero R, Okuda DT, Cree BA, Green AJ, Waubant E, *et al.*: **Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis.** *Hum Mol Genet* 2009, **18:**767-778.
18. Pattin KA, Moore JH: **Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases.** *Hum Genet* 2008, **124:**19-29.
19. Wilke RA, Mareedu RK, Moore JH: **The pathway less traveled: moving from candidate genes to candidate pathways in the analysis of genome-wide data from large scale pharmacogenetic association studies.** *Curr Pharmacogenomics Person Med* 2008, **6:**150-159.
20. Bush WS, Dudek SM, Ritchie MD: **Biofilter: a knowledge integration system for the multi-locus analysis of genome-wide association studies.** *Pac Symp Biocomput* 2009:368-379.