Published in final edited form as:

Transl Res. 2009 October ; 154(4): 161–164. doi:10.1016/j.trsl.2009.07.001.

Genome-Wide Association Studies: hypothesis-“free” or “engaged”?

Georgios Kitsios, MD, MSc^{1,2} and Elias Zintzaras, MSc, PhD.^{1,2}

¹ Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts Medical Center, Boston, MA

² Department of Biomathematics, University of Thessaly School of Medicine, Larissa, Greece

Despite extensive efforts and a vast amount of resources spent, the genetic basis of common, multifactorial disorders remains poorly understood. Following two decades of research with the candidate-gene approach and thousands of relevant publications, very few genetic regions and genes can reliably be considered as true positives; the majority of gene-disease associations claimed by genetic association studies have shown inconsistency and non-reproducibility.¹

More recently, the transition to genome-wide association studies (GWAS) has provided a new conceptual framework in the search for variants underlying common disorders.² Rather than focusing on biological candidate genes, the genome is screened without any prior predilection for specific regions, genes, or variants thereof. Based on this interrogation of genetic variation on a genome-wide basis, GWAS have been characterized as “hypothesis-free” or “agnostic” approaches.^{3,4} The “hypothesis-free” basis of GWAS offered the opportunity to overcome difficulties and obstacles imposed by the incomplete understanding of disease pathophysiology.

The initial successes of GWAS provided a proof of their concept and dissolved the skepticism about the viability of this approach. Previously unsuspected genes and pathways were uncovered whereas some surprising associations of intergenic regions with common disorders, (such as 9p21.3 locus with coronary artery disease or 8q24 locus with cancer), posed unprecedented questions regarding the mechanistic role of genetic variation.⁵ Variants in such gene-depleted areas would have never been considered by the candidate-gene approach.

Much of the excitement that GWAS brought to the scientific community was based on the expectation that since these studies are “hypothesis-free”, and thus independent of the preexisting bias of traditional biology, a comprehensive description of the genetic causes of complex disease would become feasible. The available output from these expensive experiments explains a small fraction of disease heritability and the modest genetic effects detected to date are of little predictive value.^{6,7} Discovered associated polymorphisms are often likely to represent tagging markers rather than the culprit functional genetic variation. Given these tough realizations, a series of methodological concerns regarding the design, analysis and interpretation of GWAS, as well as translation of their findings, have been raised and strategies to address these issues and accelerate progress in the field have been proposed.^{8–10}

However, it remains probably underappreciated that GWAS are dependent on underlying hypotheses despite their “hypothesis-free” label- which account for important limitations of this approach and could explain why the information derived from GWAS is incomplete.

Characterizing these experiments as “hypothesis-free” or “agnostic” can be misleading and disregards the fact that the output of any biological experiment is primarily determined by the extent to which the hypotheses tested truly hold. Although not explicitly stated, GWAS are based on a priori hypotheses, dictated by the design of genotyping platforms or the analysis methodologies:

1. Common Disease/Common Variant hypothesis

The design of current GWAS makes them suitable mainly for the discovery of common variants conferring low/moderate risks, in the context of the common disease/common variant hypothesis (CDCV).¹¹ Given the current near impossibility of reliably detecting effects of rare alleles (owing to sample size and sequencing constraints), most of GWAS have analyzed single nucleotide polymorphisms (SNPs) with minor allele frequencies of more than 5%. Rare variants, with frequencies lying somewhere between the limits of deleterious mutations and polymorphic variations (i.e. 0.1–1%) are not examined by current GWAS. According to an alternative hypothesis (common disease/rare variant hypothesis), complex traits are caused collectively by multiple rare variants with moderate to high penetrance.¹² Evolutionary theories, empirical evidence and data from the HapMap project support this hypothesis.¹³ Because of their low frequency and individually small contributions to the overall inherited disease susceptibility, rare variants will not be detectable by population association studies based on the use of linked polymorphic markers. Consequently, if the CDCV hypothesis does not hold, or at least holds only in some of the cases, then the “missing heritability” will not be tracked by GWAS. Rare variants are more likely to be detected by extensive resequencing of carefully selected candidate genes in relatively large numbers of carefully chosen cases, together with a thorough analysis of the functional effects of any suspected variants.¹⁴ Multiple rare pathogenic variants have been detected by this approach.^{15,16} However, candidate gene sequencing will probably not suffice to identify important rare variants and sequencing on a whole-genome basis will be required.⁹ It is anticipated that advances in genotyping technologies and novel genetic variation maps that capture rare variants (1,000 Genomes Project¹⁷) will make whole-genome searches for rare variants feasible.

2. SNPs are the responsible genetic variants or can serve as proxies for the causal variants

SNPs are the markers that have been selected to serve as the tools for pinpointing the genetic determinants of complex traits. This choice was mainly based on their abundance (over 12 million SNPs), and the technological facilities for their high-throughput analysis. Based on the current design of GWAS, other important sources of variation are discarded, such as structural variants, noncoding RNAs or epigenetic changes. The most common types of variants are gains and losses of DNA, called copy number variants (CNVs), which likely exert important phenotypic effects on gene expression and function.¹⁸ Several CNVs have been shown to be implicated in common disorders, as both rare and common genomic changes. Most of the associations between CNVs and complex disorders reported so far have been unveiled through candidate gene or candidate region approaches.¹⁸ Despite some successful examples of SNP-tagged CNV association detections,^{19,20} the first wave of GWAS has largely missed the contribution of CNVs. This occurs because CNVs are not easily tagged by SNPs (influencing Hardy-Weinberg Equilibrium and thus being eliminated during genotype quality-control checks), CNVs might have a wide range of copy number variability, and often fall in genomic regions not well covered by genotyping platforms or not genotyped by the HapMap project.²¹ Recently developed genotyping chips allow the simultaneous interrogation of SNPs and CNVs across the genome with improved coverage.²² The genome-wide detection of rare CNVs, as well as the discrimination capacity for variants with wide range of variability in copy number, represent important challenges for genotyping technology research. Since the global

inventory of CNVs remains incomplete, and with limited empirical data currently available, the extent to which current arrays capture the structural variome and enable a more comprehensive elucidation of the spectrum of genomic variability remains unclear.^{18,21} Even under optimal design, genome-wide SNP screens will still not be informative about the role of other components of the regulatory architecture of the genome, such as noncoding RNAs or epigenetic modifications (somatically-acquired or trans-generationally inherited); complementary techniques, such as high-throughput methylome analysis, will be ultimately needed.^{23,24}

3. Genetic predisposition to complex disorders is conferred by independent (main) effects of SNPs

Most of GWAS to date have used a single-locus analysis strategy, in which each variant is tested individually for association with a specific phenotype, assuming independent effects, despite the axiomatic consideration that complex disorders are caused by the complex interplay between multiple genetic and environmental factors.²⁵ A genetic variant that functions primarily through such a complex mechanism might be missed in a GWAS, if the variant is examined in isolation without allowing for its potential interactions with other unknown factors. Many SNPs have been shown to have small individual effect sizes, but their combined effect may be much larger.²⁶ Unfortunately, these useful signals are embedded in a genome-wide sea of noise. Given the large multiple testing, very few signals exceed the genome-wide significance threshold and those that do not exceed this stringent statistical requirement are generally neglected. Exploration of interactions has been the exception rather than the rule in GWAS analyses, and only a few small-scale examples are available.²⁷ A search for epistatic effects is computationally demanding, as several billions of SNP pairs exist for typical genotyping chips and gene-environment interactions pose formidable methodological challenges. Recently, several methods and software packages have been developed for the consideration of the statistical interactions between loci^{26,27} and between genes and environmental factors.²⁸ The application of these methodologies is expected to enhance the power to detect associations and to provide further insight in the biological and biochemical pathways of disease.²⁶ Although the detection and interpretation of these interactions will not be straightforward, genome-wide interaction testing is the next natural step after the single-locus univariate analysis techniques.

In conclusion, GWAS represent one forward step towards a more complete understanding of the genetic architecture of complex disorders. Unbiased by prior pathophysiological assumptions, more than 350 GWAS have substantially changed the landscape of genetic associations for more than 80 traits and provided new insights into disease mechanisms.^{4,5,29} The independence of GWAS from the need of biological candidates has led to their characterization as “hypothesis-free” or “agnostic” experiments. However, these terms can be misleading; in essence, GWAS are based on prior hypotheses, implicitly determined by the statistical analysis or the design of the genotyping array. GWAS to date have largely focused on the role of common SNPs and probably most of the heritability of complex disorders captured by these variants has been accounted for.⁶ The remaining variance may be due to rare SNPs, structural variants, epigenetic effects, other unsuspected genomic mechanisms, gene-gene or gene-environment interactions that have not been adequately modeled.^{9,27}

In the recent past, other venues of research (genome-wide linkage scans and microarray expression profiling experiments)^{30,31} had been celebrated as “hypothesis-free” and unbiased approaches, considered to be liberated from the need of restrictive presumptions about the causative molecular mechanisms of disease. Nevertheless, these techniques fell short from providing a comprehensive picture of complex disorders genetics. It is important to acknowledge the underlying assumptions of GWAS, since in fact all biological research is

based on some hypotheses, although they may not be always explicitly stated.³² More hypotheses of these experiments regarding what is actually interrogated in genome-wide screens may also reveal themselves as the understanding of the molecular mechanisms progresses. Unless the inherent limitations of GWAS are addressed, and complementary genetic analysis methods, such as targeted or whole-genome sequencing, are implemented, the full advantage of genomic scans of human variation will not be realized.

Acknowledgments

Georgios Kitsios is a Pfizer Post-Doctoral Fellow in Clinical Research. Scientific support for this project was provided through the Tufts Clinical and Translational Science Institute (Tufts CTSI) under funding from the National Institute of Health/National Center for Research Resources (UL1 RR025752). Points of view or opinions in this paper are those of the authors and do not necessarily represent the official position or policies of the Tufts CTSI.

Abbreviations

GWAS	genome-wide association studies
CDCV	common disease/common variant
SNPs	single nucleotide polymorphisms
CNVs	copy number variants

References

1. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003;33:177–82. [PubMed: 12524541]
2. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005;6:95–108. [PubMed: 15716906]
3. Hunter DJ, Altshuler D, Rader DJ. From Darwin's finches to canaries in the coal mine--mining the genome for new biology. *N Engl J Med* 2008;358(26):2760–2763. [PubMed: 18579810]
4. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA* 2008;299:1335–44. [PubMed: 18349094]
5. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008;118:1590–605. [PubMed: 18451988]
6. Goldstein DB. Common genetic variation and human traits. *N Engl J Med* 2009;23(360):1696–8. [PubMed: 19369660]
7. Kraft P, Hunter DJ. Genetic risk prediction--are we there yet? *N Engl J Med* 2009;360:1701–3. [PubMed: 19369656]
8. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008;9:356–69. [PubMed: 18398418]
9. Hardy J, Singleton A. Genomewide association studies and human disease. *N Engl J Med* 2009;360:1759–68. [PubMed: 19369657]
10. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet* 2008;17:R156–65. [PubMed: 18852205]
11. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet* 2001;17:502–10. [PubMed: 11525833]
12. Iyengar SK, Elston RC. The genetic basis of complex traits: Rare variants or “common gene, common disease”? *Methods Mol Biol* 2007;376:71–84. [PubMed: 17984539]
13. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008;83:311–21. [PubMed: 18691683]
14. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008;40:695–701. [PubMed: 18509313]

15. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 2004;305:869–72. [PubMed: 15297675]
16. Ji W, Foo JN, O’Roak BJ, et al. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* 2008;40:592–9. [PubMed: 18391953]
17. 1000 Genomes Project. <http://www.1000genomes.org>
18. Estivill X, Armengol L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet* 2007;3:1787–1799. [PubMed: 17953491]
19. McCarroll SA, Huett A, Kuballa P, et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn’s disease. *Nat Genet* 2008;40:1107–12. [PubMed: 19165925]
20. Willer CJ, Speliotes EK, Loos RJ, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* 2009;41:25–34. [PubMed: 19079261]
21. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. *Nat Genet* 2007;39:S37–S42. [PubMed: 17597780]
22. McCarroll SA, Kuruvilla FG, Korn JM, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* 2008;40:1166–74. [PubMed: 18776908]
23. Butcher LM, Beck S. Future impact of integrated high-throughput methylome analyses on human health and disease. *J Genet Genomics* 2008;35:391–401. [PubMed: 18640619]
24. Mattick JS. The genetic signatures of noncoding RNAs. *PLoS Genet* 2009;5:e1000459. [PubMed: 19390609]
25. Cordell HJ. Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans. *Hum Mol Genet* 2002;11:2463–8. [PubMed: 12351582]
26. Cordell HJ. Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*. 2009 [In press].
27. Vineis P, Brennan P, Canzian F, et al. Expectations and challenges stemming from genome-wide association studies. *Mutagenesis* 2008;23:439–44. [PubMed: 18765424]
28. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol* 2009;169:219–226. [PubMed: 19022827]
29. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009;106:9362–7. [PubMed: 19474294]
30. Hoh J, Ott J. Genetic dissection of diseases: design and methods. *Curr Opin Genet Dev* 2004;14:229–32. [PubMed: 15172663]
31. Mir KU. The hypothesis is there is no hypothesis. The Microarray Meeting, Scottsdale, Arizona, USA, 22–25 September 1999. *Trends Genet* 2000;16:63–4. [PubMed: 10712160]
32. Wiley S. Hypothesis-free? No such thing. *The Scientist* 2008;22:31.