Published in final edited form as:
Genet Epidemiol. 2012 April ; 36(3): 179–182. doi:10.1002/gepi.21612.

Genetic Epidemiology with a Capital E: Where Will We Be in Another 10 Years?

Duncan C. Thomas
Department of Preventive Medicine University of Southern California 2001 N. Soto St., SSB-220F
Los Angeles, CA 90089-9234 Phone: 323-442-1218 Fax: 323-442-2349

Abstract

In a commentary on the evolution of the field of genetic epidemiology over the past 10 years in this issue, Khoury et al. highlight several important developments, including the emergence of evaluation of genetic discoveries for their translational utility and of standards for reporting genetic findings. In this companion to their article, I reflect on some of these trends and speculate about the direction of the field in the future. In particular, I emphasize the opportunities posed by novel technologies like next-generation sequencing and the biological insights emerging from integrative genomics, but I also question the utility of large consortia. The basic principles of population-based research and the importance of taking account of the environment remain important to the field.

Keywords

environment; genome-wide association studies; consortia; next-generation sequencing; epigenetics; genetic testing; translational research; integrative genomics

I was pleasantly surprised to see my 1999 “Genetic Epidemiology with a Capital E” address to the International Genetic Epidemiology Society [Thomas 2000] featured in Muin Khoury’s plenary lecture at the this year’s North American Congress of Epidemiology and subsequently in the article in the current issue [Khoury et al. 2011]. Reviewing their manuscript inspired me more to ponder the direction of the field on my own, perhaps more than to critically review their paper. Hopefully, their thought-provoking piece will inspire others to reflect in a similar manner. In what follows, I thought I’d share some of my own reflections.

Genome-wide association studies

(GWAS) have become routine, as we move into the “post-GWAS” era and the advent of next-generation sequencing. The “$1000 Genome” [Mardis 2006] is nearly upon us, with the potential for truly “whole genome” association studies to generate the same excitement as GWAS did not long ago, while raising a host of new methodological challenges [Mardis 2010]. One interesting shift I am seeing is a resurgence of interest in family-based designs for dealing with multiple rare variants [Shi and Rao 2011; Zhu et al. 2010]. Although case-control designs using unrelated individuals have become the method of choice for GWAS, we can expect them to yield enormous numbers of novel variants when applied to whole genome sequence data. Family-based methods may offer greater potential for sifting the causal from noncausal variants by exploiting cosegregation information, eliminating genotype calling errors, detecting *de novo* variants and parent or origin effects, and increasing the yield of causal variants through oversampling multiple-case families (although this may be less helpful for heterogeneous causal variants with weak effects and their relatedness means smaller effective sample sizes).

There is also a parallel explosion coming on the environmental side, with novel –omics technologies being developed to measure exposures (e.g., metabolomics) [Gieger et al. 2008; Thomas and Ganji 2006], Environment-Wide Association Studies (EWAS) [Patel et al. 2010] (not to be confused with another kind of "EWAS," the epigenome-wide association study [Rakyan et al. 2011]), and Gene-Environment Wide Interaction Studies (GEWIS) [Khoury and Wacholder 2009]. The Exposome [Wild 2005] is only a vaguely conceived concept as the entirely of a person's history of exposures, with some arguments about whether the concept applies more to the external or internal environment, and still to be implemented in any meaningful way. But ultimately it has the potential to revolutionize the study of gene-environment interactions and will pose a host of novel methodological challenges: unlike genes, exposures are time-dependent, complex in the sense of each agent having multiple aspects, often highly correlated across agents, space, and time, fraught with measurement error, and subject to other forms of selection bias and confounding.

Another feature of GWAS having become routine is the emergence of consortia for discovering smaller and smaller risks because of the need for enormous sample sizes [Hunter et al. 2007]. This pressure towards Big Science will doubtless become even stronger as we move into sequence data and looking for rarer variants. There are some issues in the sociology of science that are worth attention: How are new investigators to find their niche in such an environment without becoming lost in a list of hundreds of authors? What is the role of investigator-initiated studies and novel or paradigm-shifting ideas? Is the dominance of a single journal, by virtue of its impact factor, in setting the agenda for the entire field a good thing? Is the huge burden of time and effort required to establish these consortia really worth the yield of smaller and smaller effect sizes? Certainly these consortia can be expected to yield more and more—and finer and finer— gold dust, but what about the nuggets? How are we to deal with the requirement of replication when a consortium has essentially the corner on all the available data or in unique situations (e.g., a gene-environment interaction with an unusual or unusually well characterized exposure)—perhaps by some form of internal cross-validation?

One aspect of GWAS becoming routine is the emergence of various guidelines for reporting their findings [Chanock et al. 2007; Hudson and Cooper 2009; Ioannidis et al. 2008; Khoury et al. 2009; Little et al. 2009], as discussed at some length by Khoury et al. These are undoubtedly useful for systematizing a chaotic literature and simplifying the task of synthesizing knowledge (through formal meta-analyses or carefully curated ontologies). But this needs to be done in a manner that does not stifle the creativity of investigators to address novel challenges with novel methodologies. Requiring consortia to make their raw data publicly available has certainly been a good step in this direction, as are forums for comparing novel methods like the Genetic Analysis Workshop.

An aspect of GWAS that is definitely not routine is the proliferation of novel methods of analysis for mining the "lower Manhattan" — the mass of highly significant findings that individually may fail to attain genome-wide significance but in the aggregate may point towards the involvement of certain pathways. Of these techniques, some variant of gene set enrichment analysis has been most widely used [Wang et al. 2010], but a wide range of other techniques are also being developed [Cantor et al. 2010]. Some of these are aimed at testing the involvement of already known pathways, others at discovering novel pathways by various high-dimensional exploratory analyses. These methods often rely heavily on our sister field of bioinformatics to guide the analysis or interpret the findings using the wealth of genomic, pathway, or other biological knowledge available in various on-line ontologies. Perhaps one of the most important developments is the emergence of "integrative genomics" [Schadt et al. 2005] as a tool for bringing together disparate data types — genomic, transcriptomic, metabolomic, proteomic, epigenomic, tumor mutations, etc. — and their respective external knowledge repositories, for a more comprehensive analysis of disease etiology. Filling the vacuum of ontologies that bring together information on genetic and environmental factors from epidemiology, toxicology, and molecular biology would greatly facilitate this goal [De Roos et al. 2001].

Translational value

Perhaps the most striking aspect of Khoury et al.'s vision is their transmutation of the letter "E" from Environment to Evaluation. Whereas they correctly note that my original piece highlighted three core principles of what I called genetic epidemiology—population-based, joint effects of genes and environment, and incorporation of biology—they expand the second of these into what has actually become a distinct topic, the use of epidemiologic methods to evaluate the clinical utility of emerging genetic knowledge — in short, its "translational" value: from bench to bedside, and from bedside to improved public health. Although not the focus of my own work, I do believe this is a reasonable goal for our field. So far, there have been few success stories in translating GWAS findings in the general population into clinically useful risk prediction models [Chatterjee et al. 2011; Jostins and Barrett 2011]. But in the field of pharmacogenomics, effect sizes may be much larger owing to the lack of time for evolutionary selection to weed out alleles conferring susceptibility to novel treatments [Altshuler et al. 2008]. Thus, the potential translational benefits to alter the risks (or benefits) posed to carriers of specific genetic variants are potentially major and immediate.

Certainly those in our field are well aware of the dangers posed by indiscriminant proliferation of genetic testing, including the growing industry of direct-to-consumer kits, some based on flimsy science and providing ill-founded advice on lifestyle changes to reduce risks of discovered variants [Kaye 2008], but these dangers are not well appreciated by the general public. Worse, they invite the "scare of the month" criticism that has plagued risk factor epidemiology even before Gary Taubes' "Epidemiology faces its limits" article was published in *Science* [Taubes 1995]. Our field must focus on the evaluation of the risks associated with genetic variants, the utility of interventions that can ameliorate them, and the accuracy of information about them that is provided to clinicians and the public. Nevertheless, there seems to be growing tendency for granting agencies to require translational impact, sometimes to the detriment of research that could yield fundamental biological insights without immediate clinical utility.

Population-based research

I was pleased to see Khoury et al.'s emphasis on population-based research as central to their vision of "evaluation." Too often have I felt like a voice crying in a wilderness of geneticists who don't seem to have a clue about the principles of epidemiologic study design and blithely proceed to use convenience samples, all too often taking cases and controls from distinct populations and hoping that statistical analyses using genomic control methods will correct for any of the subtle selection, information, or confounding biases that could arise [Chen et al. 2010; Sinnott and Kraft 2011]. This way of thinking is particularly endemic in consortia, of necessity as they are often cobbling sets of cases and controls obtained under many different designs. Perhaps this is OK for the purposes of discovery, knowing that subsequent replication and characterization studies will weed out the false positives, but at a cost of increased effort by the community pursuing dead ends. It is essential that we use the principles of population-based research to move from simple discovery of novel genetic variants to characterizing their effects in populations (e.g., estimating penetrance, attributable fractions, percent of heritability explained, modification by environmental or other host risk factors) and to evaluate the potential for interventions

Biological insights

Personally, I think the third of my principles—the incorporation of the underlying biology into our conceptual models—is going to become even more central to our field [Freedman et al. 2011]. For example, the growing recognition that common variants are unlikely to account for much of the "missing heritability" [Eichler et al. 2010; Manolio et al. 2009; Park et al. 2010] and the uncertainty over whether multiple rare variants will account for much more [Wray et al. 2011] has led to an interest in epigenetic mechanisms. But unless epigenetic variation is itself strongly heritable, this may explain more of the "missing causality" than "missing heritability" [Slatkin 2009]. So far, this line of research has focused primarily on the role of epigenetics in disease, with comparatively little attention to developmental and environmental influences on epigenetic changes. The emerging field of environmental epigenetics [Bollati and Baccarelli 2010; Jirtle and Skinner 2007] will instead focus on the mediation of exposure-response relationships through epigenetic mechanisms, both within the individual and transgenerationally. The growing interest in the developmental origins of adult disease [de Boo and Harding 2006; Symonds 2010] is one example of this.

Is there still a meaningful distinction between genetic and molecular epidemiology? In the concluding pages of my textbook [Thomas 2004], I discussed this question and argued that the distinction was mainly historical and terminological, although there are real differences in emphasis (e.g., explaining heritability vs. understanding biological mechanisms). That said, it is clear that the genetic and epidemiology communities (and their subdivisions) remain quite distinct and we are the poorer for it. My own vision of the future would have the two talking to each other more, even merging.

The place of the journal

I find it ironic that the journal *Genetic Epidemiology*, which is publishing this piece and which carried my original article, ranks only 312th on Khoury et al's list of journals having published "genetic epidemiology articles." Obviously it must be said that this journal has become arguably *the* premier journal for publishing novel methodological papers on statistical genetics (most of which are probably not included in Khoury et al.'s tally because it is focused on substantive rather than methodological papers). In the most recent ISI Journal Citation Reports, the journal now ranks 11th out of 142 in the category of Public Environmental and Occupational Health, and 41st out of 156 in the category of Genetics and Heredity. It is not currently included in the category of biostatistics journals, but when that ranking appears it will likely be near the top. The new Editor in Chief has expressed priorities to publish more applied genetic epidemiology and software articles (Shete, 2012 in press). Hopefully this will reverse the tendency of many of my epidemiology colleagues not to read the journal or consider submitting their applied work to it because they perceive it as a purely statistical journal. While most applied research in genetic epidemiology may continue to appear in disease-specific journals, this journal has the opportunity to attract new readership by focusing on the Grand Challenges posed by the emergence of new technologies and the novel biological insights they are revealing.

References

Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008; 322(5903):881–8. [PubMed: 18988837]

Bollati V, Baccarelli A. Environmental epigenetics. Heredity. 2010; 105(1):105–12. [PubMed: 20179736]

Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet. 2010; 86(1):6–22. [PubMed: 20074509]

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE. Replicating genotype-phenotype associations. Nature. 2007; 447(7145):655–60. others. [PubMed: 17554299]

Chatterjee N, Park J-H, Caporaso N, Gail MH. Predicting the Future of Genetic Risk Prediction. Cancer Epidemiology Biomarkers & Prevention. 2011; 20(1):3–8.

Chen GK, Millikan RC, John EM, Ambrosone CB, Bernstein L, Zheng W, Hu JJ, Channock SJ, Ziegler RG, Bandera EV. The potential for enhancing the power of genetic association studies in African Americans through the reuse of existing genotype data. PLoS Genetics. 2010; 6(9):e1001096. others. [PubMed: 20824062]

de Boo HA, Harding JE. The developmental origins of adult disease (Barker) hypothesis. Aust N Z J Obstet Gynaecol. 2006; 46(1):4–14. [PubMed: 16441686]

De Roos, AJ.; Smith, M.; Channock, S.; Rothman, N. Toxicologic considerations in the application and interpretation of susceptibility biomarkers in epidemiologic studies. In: Bird, P.; Boffetta, P.; Buffler, P.; Rice, J., editors. Mechanistic Considerations in the Molecular Epidemiology of Cancer. IARC Scientific Publications; Lyon: 2001. p. 105-125.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010; 11(6):446–50. [PubMed: 20479774]

Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, De Biasi M, Carlson C, Duggan D. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet. 2011; 43(6):513–8. others. [PubMed: 21614091]

Gieger C, Geistlinger L, Altmaier E, Hrabe de Angelis M, Kronenberg F, Meitinger T, Mewes HW, Wichmann HE, Weinberger KM, Adamski J. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet. 2008; 4(11):e1000282. others. [PubMed: 19043545]

Hudson TJ, Cooper DN. STREGA: a ‘How-To’ guide for reporting genetic associations. Hum Genet. 2009; 125(2):117–8. [PubMed: 19184667]

Hunter DJ, Thomas G, Hoover RN, Chanock SJ. Scanning the horizon: What is the future of genome-wide association studies in accelerating discoveries in cancer etiology and prevention? Cancer Causes Control. 2007; 18(5):479–84. [PubMed: 17440825]

Ioannidis JP, Boffetta P, Little J, O’Brien TR, Uitterlinden AG, Vineis P, Balding DJ, Chokkalingam A, Dolan SM, Flanders WD. Assessment of cumulative evidence on genetic associations: interim guidelines. Int J Epidemiol. 2008; 37(1):120–32. others. [PubMed: 17898028]

Jirtle RL, Skinner MK. Environmental epigenomics and disease susceptibility. Nat Rev Genet. 2007; 8(4):253–62. [PubMed: 17363974]

Jostins L, Barrett JC. Genetic risk prediction in complex disease. Hum Mol Genet. 2011; 20(R2):R182–8. [PubMed: 21873261]

Kaye J. The regulation of direct-to-consumer genetic tests. Hum Mol Genet. 2008; 17(R2):R180–3. [PubMed: 18852208]

Khoury MJ, Bertram L, Boffetta P, Butterworth AS, Chanock SJ, Dolan SM, Fortier I, Garcia-Closas M, Gwinn M, Higgins JP. Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. Am J Epidemiol. 2009; 170(3): 269–79. others. [PubMed: 19498075]

Khoury MJ, Gwinn M, Clyne M, Yu W. Genetic epidemiology with a capital E: Ten years after. Genet Epidemiol. 2011 in press.

Khoury MJ, Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies--challenges and opportunities. Am J Epidemiol. 2009; 169(2):227–30. discussion 234-5. [PubMed: 19022826]

Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, Khoury MJ, Cohen B, Davey-Smith G, Grimshaw J. STrengthening the REporting of Genetic Association Studies (STREGA): an extension of the STROBE statement. PLoS Med. 2009; 6(2):e22. others. [PubMed: 19192942]

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747–53. others. [PubMed: 19812666]

Mardis ER. Anticipating the 1,000 dollar genome. Genome Biol. 2006; 7(7):112. [PubMed: 17224040]

Mardis ER. The $1,000 genome, the $100,000 analysis? Genome Med. 2010; 2(11):84. [PubMed: 21114804]

Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet. 2010; 42(7):570–5. [PubMed: 20562874]

Patel CJ, Bhattacharya J, Butte AJ. An environment-wide association study (EWAS) on type 2 diabetes mellitus. PLoS One. 2010; 5(5):e10746. [PubMed: 20505766]

Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet. 2005; 37(7):710–7. others. [PubMed: 15965475]

Shi G, Rao DC. Optimum designs for next-generation sequencing to discover rare variants for common complex disease. Genetic Epidemiology. 2011; 35(6):572–579. [PubMed: 21618604]

Sinnott JA, Kraft P. Artifact due to differential error when cases and controls are imputed from different platforms. Hum Genet. 2011

Slatkin M. Epigenetic inheritance and the missing heritability problem. Genetics. 2009; 182(3):845–50. [PubMed: 19416939]

Symonds ME. Epigenomics ? Grand Challenge: much more than the developmental origins of adult health and disease. Frontiers in Genetics. 2010; 1

Taubes G. Epidemiology faces its limits. Science. 1995; 269(5221):164–9. [PubMed: 7618077]

Thomas CE, Ganji G. Integration of genomic and metabonomic data in systems biology--are we ‘there’ yet? Curr Opin Drug Discov Devel. 2006; 9(1):92–100.

Thomas D. Genetic epidemiology with a capital “E”. Genet Epidemiol. 2000; 19:289–3000. [PubMed: 11108640]

Thomas, DC. Statistical methods in genetic epidemiology. Oxford University Press; Oxford: 2004.

Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet. 2010; 11(12):843–54. [PubMed: 21085203]

Wild CP. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev. 2005; 14(8):1847–50. [PubMed: 16103423]

Wray NR, Purcell SM, Visscher PM. Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol. 2011; 9(1):e1000579. [PubMed: 21267061]

Zhu X, Feng T, Li Y, Lu Q, Elston RC. Detecting rare variants for complex traits using family and unrelated data. Genet Epidemiol. 2010; 34(2):171–87. [PubMed: 19847924]