A road map for efficient and reliable human genome epidemiology

John P A Ioannidis[1,2], Marta Gwinn[3], Julian Little[4], Julian P T Higgins[5,6], Jonine L Bernstein[7], Paolo Boffetta[8], Melissa Bondy[9], Molly S Bray[10], Paul E Brenchley[11], Patricia A Buffler[12], Juan Pablo Casas[13], Anand Chokkalingam[12], John Danesh[14], George Davey Smith[15], Siobhan Dolan[16], Ross Duncan[17], Nelleke A Gruis[18], Patricia Hartge[19], Mia Hashibe[8], David J Hunter[20], Marjo-Riitta Jarvelin[21,22], Beatrice Malmer[23], Demetrius M Maraganore[24], Julia A Newton-Bishop[25], Thomas R O'Brien[19], Gloria Petersen[26], Elio Riboli[8], Georgia Salanti[1,5], Daniela Seminara[27], Liam Smeeth[13], Emanuela Taioli[28], Nic Timpson[15], Andre G Uitterlinden[29], Paolo Vineis[21,30], Nick Wareham[31], Deborah M Winn[27], Ron Zimmern[6], Muin J Khoury[3] & the Human Genome Epidemiology Network and the Network of Investigator Networks

Networks of investigators have begun sharing best practices, tools and methods for analysis of associations between genetic variation and common diseases. A Network of Investigator Networks has been set up to drive the process, sponsored by the Human Genome Epidemiology Network. A workshop is planned to develop consensus guidelines for reporting results of genetic association studies. Published literature databases will be integrated, and unpublished data, including 'negative' studies, will be captured by online journals and through investigator networks. Systematic reviews will be expanded to include more meta-analyses of individual-level data and prospective meta-analyses. Field synopses will offer regularly updated overviews.

[1]Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, and Biomedical Research Institute, Foundation for Research and Technology-Hellas, Ioannina 45110, Greece. [2]Department of Medicine, Tufts University School of Medicine, Boston, Massachusetts 02111, USA. [3]Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia 30333, USA. [4]Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada. [5]Medical Research Council Biostatistics Unit, University of Cambridge, Cambridge CB2 2SR, UK. [6]Public Health Genetics Unit, Strangeways Research Laboratory, Cambridge CB1 8RN, UK. [7]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA. [8]International Agency for Research on Cancer, 69008 Lyons, France. [9]Department of Epidemiology, University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, USA. [10]United States Department of Agriculture/Agricultural Research Service Children's Nutrition Research Center, Baylor College of Medicine, Houston, Texas 77030, USA. [11]Renal Research Laboratories, Manchester Institute of Nephrology and Transplantation, Royal Infirmary, Manchester M13 9WL, UK. [12]University of California, Berkeley, California 94720-7360, USA. [13]Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK. [14]Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK. [15]Department of Social Medicine, University of Bristol, Canynge Hall, Whiteladies Road, Bristol, BS8 2PR, UK. [16]March of Dimes, White Plains, New York 10605, USA. [17]World Health Organization, Geneva CH-1211, Switzerland. [18]Department of Dermatology, Leiden University Medical Center, 2333 AL Leiden, The Netherlands. [19]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland 20852, USA. [20]Harvard School of Public Health, Boston, Massachusetts 02115, USA. [21]Department of Epidemiology and Public Health, Imperial College, London W2 1PG, UK. [22]Department of Public Health Science and General Practice, University of Oulu, Oulu, Finland. [23]Department of Radiation Sciences, Oncology, Umea University Hospital, Sweden. [24]Department of Neurology, Mayo Clinic, Rochester, Minnesota 55905, USA. [25]Genetic Epidemiology Division, CR-UK Clinical Centre, Leeds LS8 7FT, UK. [26]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota 55905, USA. [27]Division of Cancer Control and Population Sciences, National Cancer Institute, Rockville, Maryland 20892, USA.[28]University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania 15232, USA. [29]Departments of Internal Medicine and Epidemiology & Biostatistics, Erasmus MC, Rotterdam 3000DR, The Netherlands. [30]Institute for Scientific Interchange Foundation, Torino, Italy. [31]Medical Research Council Epidemiology Unit, Elsie Widdowson Laboratories, Cambridge CB1 9NL, UK.
e-mail: jioannid@cc.uoi.gr and muk1@cdc.gov

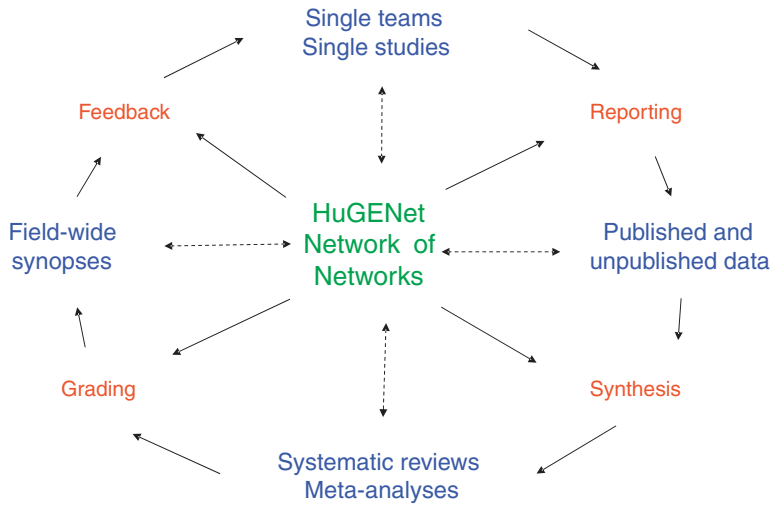

Figure 1 Framework for risk evaluation in genetic association studies.

Genetic epidemiologists and geneticists face the challenge of creating an efficient and reliable compilation of the evidence for genetic risk contributions to common human diseases (a 'risk engine'[1]). Although data relating DNA sequence variation to disease states and/or intermediate traits are accumulating exponentially, the current situation is plagued with problems[2,3]. These include the prevalence of small, underpowered studies, often with flawed designs, suboptimal conduct and biased analyses; selective reporting of 'positive' results; lack of standardization among studies; poor reporting of results even from well-conducted studies; and difficulties in assessing interactions with environmental risk factors[4]. Consequently, the research evidence is fragmented, and the interface between epidemiological and other biological evidence is poorly developed. It remains unclear how to keep track of the rapidly evolving evidence across fields that can be defined by disease, genes or exposures, and how to rate the credibility of this evidence.

The Human Genome Epidemiology Network (HuGENet), a global initiative committed to the development and integration of the knowledge base on human genetic variants and health (http://www.cdc.gov/genomics/hugenet), proposes a plan for developing this knowledge base and making it efficient and reliable. Several interrelated and synergistic actions are currently underway within this framework (**Table 1** and **Fig. 1**). Whatever advances are to be achieved in human genome epidemiology, they should be developed, adopted and promoted by the investigators themselves. The first step was the creation of a Network of Investigator Networks[3] in 2005 (**Table 1**, Step 1). Investigator networks comprise teams of researchers working on a common theme (for example, on a specific disease, a set of genes or on modulating exposures). These networks will promote large-scale evidence with rigorous methods and will be instrumental in driving Steps 2–5. At an October 2005 meeting in Cambridge, UK, experiences were shared by representatives of 27 such networks; the networks comprise several hundred research teams and several thousand investigators working in human genetics (see http://www.hugenet.org.uk). As it may be impractical to create a single consortium for all investigators in some themes, we encourage both plurality and large-scale evidence through communication of consortia with complementary research agendas.

Collaborating investigators can reach common agreement on overall study design, definitions of phenotypes, exposures and endpoints, as well as analyses of gene-disease associations and gene-gene and gene-environment interactions (**Table 1**, Step 2). Agreement on gene variants to be genotyped is critical at the design stage, if the subsequent information is to be combined. In the absence of common, completely standardized methods or genotyping platforms, only prospectively planned consortia can achieve this.

Moreover, some issues are specific to a field or research question, but many issues that arise in conducting and reporting genetic association studies are common across diverse fields. To help provide guidance for reporting study results, HuGENet is planning the development in 2006 of an extension of the STROBE (for 'strengthening the reporting of observational studies in epidemiology') statement for genetic epidemiology (http://www.strobe-statement.org). A HuGE-STROBE guideline statement will offer an objective checklist that can be widely adopted by journals, analogous to similar initiatives in clinical trials (CONSORT)[5] or microarray studies (MIAME)[6]. Such guidelines will also be useful for future studies of large population cohorts and biobanks; many such efforts are underway or being planned in different parts of the world, harmonized for international collaboration under the Public Population Project in Genomics initiative (http://www.p3gconsortium.org). We do not expect that the HuGE-STROBE guideline will replace peer review, but it will offer guidance for investigators, peer reviewers and editors on improving the quality of the studies and will aid their assessment before publication. Meeting the reporting requirements may result in longer manuscripts, but all statistical steps, assumptions and processing can be documented in supplementary files.

HuGENet has already created a database (HuGEPubLit) that aims to capture published genetic association articles as they are indexed in Medline. As of November 2005, the database has over 18,500 entries. This effort will be extended to other databases such as EMBASE and will seek synergy with ongoing, similar initiatives such as the US National Institutes of Health (NIH)-sponsored Genetic Associations Database (GAD, http://geneticassociationdb.nih.gov). Similar efforts are ongoing in pharmacogenetics and pharmacogenomics (http://www.pharmgkb.org)[7].

These databases need search tools, and they should encompass both published and deposited data (**Table 1**, Step 3). Retrieving unpublished data is currently very problematic, and unpublished reports often present 'negative' results from well-conducted studies. A research environment that promotes and rewards only results that reach formal statistical significance[8,9] is likely to foster data dredging and will create a distorted literature with very low credibility[10–12]. Comparisons of primary outcomes defined in trial protocols with those defined in published articles have provided empirical evidence of selective reporting even for randomized controlled trials[13]. Selective reporting of extensive exploratory analyses would be almost impossible to detect in studies of gene-disease associations and related interactions, even by the most sophisticated methodologists and expert peer reviewers. The protocols of these studies are rarely available for scrutiny, and there is currently no formal way for registering analyses in advance of publication.

HuGENet is working with journals and collaborating in efforts to create online journals to encourage publication of 'negative' results after appropriate methodological appraisal and with due credit to the investigators. The citable, peer-reviewed Molecule Pages of the Alliance for Cell Signalling

(http://www.signaling-gateway.org/molecule), which compiles information about proteins involved in cell signaling, offers an analogous example where online publications with DOI numbers can have high visibility. These online deposited association data should be searchable in the integrated databases. Peer review will be promoted and facilitated, not replaced, by guidance as in STROBE. In addition, investigators participating in the Network of Networks have also agreed that network-specific, distributed databases can be developed to capture the data produced by network members. 'Negative' results from members of a consortium would get authorship credit when their results are incorporated in large-scale data syntheses, including prospective meta-analyses in which single teams work on studies with the explicit purpose of meta-analysis. Finally, journal editors can have an important role in preventing bias by basing editorial decisions primarily on study quality, relevance and methodological rigor, rather than formal statistical significance alone[14].

The epidemiologic evidence for gene-disease association requires replication, validation and synthesis (**Table 1**, Step 4). Forty systematic reviews have been published according to specific HuGENet guidelines, and the HuGE Pub Lit database includes more than 250 meta-analyses and other systematic reviews that have been published elsewhere (http://www.cdc.gov/genomics/hugenet). However, the effort needs to be intensified as data increase exponentially. Many journals have already shown interest in publishing these research products. Meta-analyses are widely accepted as the highest level of evidence in medicine and they are currently the most cited study design in the health sciences[15]. HuGENet recently created an updated, detailed guidance document, which will be available online in 2006, for the conduct of systematic reviews and meta-analyses, and we are actively developing new data synthesis methodologies. Increasingly, we would like to promote meta-analyses conducted by consortia of investigators working in the same field, including meta-analyses of individual-level data and prospective designs[16].

A key aspect of the road map is the development of widely accepted rules for assessing the evidence for causal inference in genetic association studies. Several investigators and journals have already suggested criteria[1,17–23], including the transparency of the data processing, magnitude and significance of the proposed genetic effect, extent of replication, protection from bias and concomitant supporting biological evidence. Adequate sample size is essential but not necessarily sufficient. We should distinguish between tools such as STROBE, which try to enhance the transparency and quality of reporting of single studies, from the tools that would grade the cumulative evidence in all the studies (both published and unpublished data) on a research question. To this end, HuGENet is planning a meeting in 2006 to develop a consensus on grading the evidence.

Finally, there is a need for up-to-date summaries of the genetic association knowledge base in order to identify gaps, avoid wasteful duplication and promote the translation of this knowledge to public health and medical applications (**Table 1**, Step 5). Systematic reviews and meta-analyses are tailored towards addressing specific gene-disease associations—usually one or a few at a time. However, for many diseases or fields, the number of tested associations is currently large and growing[24]. Faced with many 'negative' and few credible positive results, one should consider the development of 'field synopses', succinct summaries of the evidence from genetic epidemiology in a particular field. Annual synopses by consortia of authors of unpublished studies might be submitted to journals annually for peer review. Synopses will become even more important as whole-genome association studies become increasingly common. Pilot projects have already begun for osteoporosis, Parkinson disease, acute leukemia and preterm birth. Our goal is to establish a dynamic, online encyclopedia of the association between genetic variation and human health that is updated regularly and linked to studies that meet quality criteria. Synopses will also identify knowledge gaps, helping to guide future research efforts.

The success of the proposed initiatives hinges on investigators around the globe joining forces in these initiatives and on the collaboration of journal editors. We believe that such inclusiveness coupled with methodological rigor will be instrumental in developing an efficient and reliable human genome epidemiology risk engine.

Table 1 Steps and near-term action items in the human genome epidemiology road map

Steps		Action items
1.	Develop a Network of Investigator Networks	Create capacity to accomplish steps 2–5 below
2.	Improve study conduct, reporting and harmonization across studies	Share among networks best practices, tools and analytic methods; develop STROBE criteria for genetic association studies (consensus workshop planned for 2006); single studies performed with eventual meta-analysis in mind
3.	Capture published and unpublished data regardless of 'positive' or 'negative' results	Integrate published literature databases; capture unpublished data (online journals, networks); enhance transparency of methods and critical appraisal; develop comprehensive search engines
4.	Improve data synthesis methods and integrate the evidence on specific associations	Finalize HuGE handbook for conducting systematic reviews; promote methods for meta-analyses of individual-level data; expand database of HuGE reviews and other systematic reviews and meta-analyses; facilitate meta-analyses from consortia
5.	Capture and appraise the evidence on the evolving 'big picture' across whole fields	Develop widely accepted criteria for appraising evidence; initiate pilot phase for specific fields; hold consensus meeting for guidelines on grading the evidence in 2006; continue empirical research; publish regularly updated synopses of the knowledge base; identify knowledge gaps

1. Anonymous. *Nat. Genet.* **37**, 1153 (2005).
2. Little, J. *et al. Am. J. Epidemiol.* **157**, 667–673 (2003).
3. Ioannidis, J.P. *et al. Am. J. Epidemiol.* **162**, 302–304 (2005).
4. von Elm, E. & Egger, M. *Br. Med. J.* **329**, 868–869 (2004).
5. Altman, D.G. *et al. Ann. Intern. Med.* **134**, 663–694 (2001).
6. Brazma, A. *et al. Nat. Genet.* **29**, 365–371 (2001).
7. Thorn, C.F., Klein, T.E. & Altman, R.B. *Methods Mol. Biol.* **311**, 179–191 (2005).
8. Pan, Z., Trikalinos, T.A., Kavvoura, F.K., Lau, J. & Ioannidis, J.P. *PLoS Med.* **2,** e334 (2005).
9. Easterbrook, P.J., Berlin, J.A., Gopalan, R. & Matthews, D.R. *Lancet* **337**, 867–872 (1991).
10. Greenland, S. *J. R. Stat. Soc. [Ser. A]* **168**, 267–306 (2005).
11. Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. *J. Natl. Cancer Inst.* **96**, 434–442 (2004).
12. Ioannidis, J.P. *PLoS Med.* **2**, e124 (2005).
13. Chan, A.W., Hrobjartsson, A., Haahr, M.T., Gotzsche, P.C. & Altman, D.G. *J. Am. Med. Assoc.* **291**, 2457–2465 (2004).
14. Patterson, M. & Cardon, L. *PLoS Biol.* **3**, e327 (2005).
15. Patsopoulos, N.A., Analatos, A.A. & Ioannidis, J.P. *J. Am. Med. Assoc.* **293**, 2362–2366 (2005).
16. Ioannidis, J.P., Rosenberg, P.S., Goedert, J.J. & O'Brien, T.R. *Am. J. Epidemiol.* **156**, 204–210 (2002).
17. Rebbeck, T.R. *et al. Cancer Epidemiol. Biomarkers Prev.* **13**, 1985–1986 (2004).
18. Ioannidis, J.P. *Int. J. Epidemiol.* (in the press).
19. Huizinga, T.W., Pisetsky, D.S. & Kimberly, R.P. *Arthritis Rheum.* **50**, 2066–2071 (2004).
20. Freimer, N.B. & Sabatti, C. *Hum. Mol. Genet.* **14**, 2481–2483 (2005).
21. Wacholder, S. *Cancer Epidemiol. Biomarkers Prev.* **14**, 1361 (2005).
22. Weiss, S.T. *Am. J. Respir. Crit. Care Med.* **164**, 2014–2015 (2001).
23. Cooper, D.N., Nussbaum, R.L. & Krawczak, M. *Hum. Genet.* **110**, 207–208 (2002).
24. Marchini, J., Donnelly, P. & Cardon, L.R. *Nat. Genet.* **37**, 413–417 (2005).