RESEARCH AGENDA

Genetic and molecular epidemiology

John P A Ioannidis

J Epidemiol Community Health 2007;**61**:757–758. doi: 10.1136/jech.2006.059055

Genetic and molecular epidemiology covers a vast area of research. Given the rapid changes in this field, discussing a research agenda is a precarious and ambitious task. A representative set of high-priority concepts will be presented here, each of which alone could be the topic of a long series of essays. The wish list includes issues of full transparency and integration of information, dealing efficiently with complex multidimensional biology, juxtaposing the genome and environmental exposures, and using robust randomised trials to advance our knowledge and its application in this field.

Genetic and molecular epidemiology, the investigation of genetic and molecular determinants of health and disease, has rapidly evolved into a highly prolific field. Its growth has been kindled by the decoding of the human genome and major advances in molecular biology and measurement platforms. Technological progress has continued to spark enthusiasm about future prospects.[1–3] However, real advances have been held back by a poor replication record, errors and biases, and inefficient translation to date of postulated discoveries for the improvement of individual and population health. Here, I will highlight four areas that may deserve more attention in the research agenda (table 1).

FULL TRANSPARENCY AND INTEGRATION OF INFORMATION

A major challenge for genetic and molecular epidemiology is the transparent and comprehensive availability of information. Current databases are increasing exponentially in volume. Selective availability of information and fragmented, selective publication of statistically significant results may be responsible for the poor replication of many research findings to date.[4] There are already major initiatives for comprehensive public availability of data for microarrays research (eg, Gene Expression Omnibus, Stanford Microarray Database) and for genome-wide association studies (eg, dbGaP and Wellcome Trust).[5–9] These efforts should be expanded and become more standardised, so as to allow the information to be used by all teams working on specific molecular epidemiology topics.[10] A careful balance must be struck between public sharing, investigator proprietary rights, and patient/participant confidentiality issues.

More consortia of investigators working on diverse diseases need to be created to facilitate the standardisation and harmonisation of phenotype definitions and laboratory, genotyping and statistical/bioinformatics methods.[11] Full standardisation across many teams may often be unfeasible. It would be difficult to revisit and modify the large amount of information that has already been collected, while for prospectively collected information consensus is lacking concerning the main definitions of disease outcomes, risk factors, and other measurements. However, it should always be possible to reach consensus on some minimum harmonisation of definitions using some common denominators. Some consensus on gold standards is also needed, even for statistical analysis methods.

Moreover, keeping track of the big picture is difficult given the very rapid pace of research in this field. The integration of evidence into regularly updated field synopses[12] should experiment with different flexible formats that would enhance inclusiveness, quality control and protection from bias.

DEALING WITH COMPLEX MULTIDIMENSIONAL BIOLOGY

The phenomena studied by genetic and molecular epidemiology are likely to be highly multifactorial and involve complex effects of many biological factors. However, the pursuit of complex models and interactions has been hampered by small sample sizes and inadequate study design in many molecular applications. Few claims for interaction effects have been rigorously validated so far.[13] In some fields focusing on multidimensional biology, for example gene expression profiling or proteomics, replication and validation of complex molecular signatures and multivariate molecular patterns has often been incomplete and has led to the propagation of potentially spurious claims.[14–16]

Dealing with one biological risk factor at a time is unlikely to get us very far. Epidemiology, bioinformatics and systems biology can learn from each other, and contact and collaboration between investigators in these disciplines should be facilitated.[17] Information on genes, gene variants, gene expression and modification, proteins, and signalling and metabolic pathways can be integrated across many levels. One also needs to improve the robustness of approaches that reduce the dimensionality of the data for problems where very many variables are available for testing. We need to explore approaches based on gene ontology, function and other classifications derived from biological information. Pathways and complex genic approaches will create further challenges for the replication process. Efficient designs need to be developed and mastered for the replication of such complex patterns. Success is not to be taken for granted. These designs should also accommodate and test the exchangeability of biological variables,

Correspondence to: Professor John P A Ioannidis, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece; jioannid@cc.uoi.gr

Accepted 7 March 2007

Table 1 Research areas which require strengthening

Full transparency and integration of information
Support for public databases
Standardisation and harmonisation of information
Generation of systematically updated field synopses
Dealing with complex multidimensional biology
Collaboration between epidemiology, bioinformatics, systems biology
Research on pathways, genic and other approaches that reduce dimensionality
Development of replication designs for complex models
Evaluation of exchangeability of biological risk factors
Juxtaposing the genome and the ''exposurome''
Incorporation of ''exposurome'' measurements in large-scale genomic projects
Understanding intermediate phenotypes
Use of mendelian randomisation to identify modifiable factors
Randomise!
Trials to prove the utility of new molecular technologies
Randomised trials in pharmacogenomics
Nested trials of lifestyle and other interventions in biobanks

that is, whether different sets of biological variables may achieve the same effects in different settings.[18]

JUXTAPOSING THE GENOME AND THE ''EXPOSUROME''

Most genetic and molecular epidemiology studies collect little or no information on environmental exposures and few manage to address even simple analyses where both genetic variables and environmental exposures are considered.[19] It makes sense (but we have few well validated examples) that disease risk is shaped by the interplay of both genetic and environmental factors. Genetic and molecular studies should aim to incorporate more measurements on non-genetic exposures. The misclassification error of these measurements should be carefully reduced to levels that are comparable with those of current genomic measurements. Large-scale studies should foster the development of a body of information on the ''exposurome'' of human experience, the totality of non-genetic exposures of individuals and populations. It will be difficult to reach the level of sophistication already achieved for measuring genomic variability. However, uneven focus between genetics and environment may also hinder our understanding of genetic factors and will likely perpetuate misconceptions and spurious research claims about environmental exposures. Enhancement of information on the exposurome and intermediate phenotypes may also facilitate use of the principle of mendelian randomisation[20] to disentangle true effects from confounding and identify important modifiable risk factors.

RANDOMISE!

Even though we cannot randomise individuals to genomic patterns, randomised trials have an important place in enhancing genetic and molecular epidemiology efforts.

First, novel discoveries and technologies of the -omics era need to be tested in randomised trials before introduction into clinical and public health use. Discoveries pertaining to diagnosis and disease prediction should show convincing evidence that they can improve outcomes in target populations. These trials may have to wait until we have biomarkers offering incremental advantages over routine diagnostic and predictive tools.[21 22] Study design and selection of outcomes requires imaginative thinking, and may have to encompass hard outcomes, quality of life, utilisation of medical treatment and services and cost. Contrasts need to offer control groups the opportunity to obtain the full benefits of standard (non-molecular) diagnostic or predictive procedures and information.

Randomised trials may have a particularly important role in pharmacogenomic research, where appropriate randomised designs can maximise power and efficiency.[23] This may eventually facilitate individualised treatment choices.

We should also seriously consider the introduction of randomised trials nested into large biobanks. Biobanks represent the new generation of cohorts. Nested randomised trials would benefit from routinely using the data machinery of biobanks with linkage to existing registries for death and other hard outcomes (eg, cancer, coronary artery disease, end-stage renal failure) and practically ''unlimited'' follow-up. Tested interventions may pertain both to lifestyle changes and medical drugs or technology. Traditional long-term trials outside biobanks have become prohibitively expensive and difficult to conduct. With appropriate collection and storage of biological samples, markers can be measured on these samples. This could include markers currently unknown which may be identified and routinely measured in the future. This information may be used to identify treatment–effect modifications based on genetic variants and genetic effects that manifest under specific lifestyle exposures or other interventions.

Competing interests: None declared.

REFERENCES

1 **Cardon LR**. Genetics. Delivering new disease genes. *Science* 2006;**314**:1403–5.
2 **Todd JA**. Statistical false positive or true disease pathway? *Nat Genet* 2006;**38**:731–3.
3 **Gullans SR**. Connecting the dots using gene-expression profiles. *N Engl J Med* 2006;**355**:2042–4.
4 **Ioannidis JP**. Why most published research findings are false. *PLoS Med* 2005;**2**:e124.
5 **Ball CA**, Brazma A, Causton H, *et al*. Submission of microarray data to public repositories. *PLoS Biol* 2004;**2**:e317.
6 **Edgar R**, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization assay repository. *Nucleic Acids Res* 2002;**30**:207–10.
7 **Brazma A**, Parkinson H, Sarkans U, *et al*. Array Express – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 2003;**31**:68–71.
8 **Brazma A**, Krestyaninova M, Sarkans U. Standards for systems biology. *Nat Rev Genet* 2006;**7**:593–605.
9 **dbGaP**. Genotype and Phenotype. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gap (accessed 22 June 2007).
10 **Ioannidis JP**, Gwinn M, Little J, *et al*. A road map for efficient and reliable human genome epidemiology. *Nat Genet* 2006;**38**:3–5.
11 **Seminara D**, Khoury MJ, O'Brien TR, *et al*. The emergence of networks in human genome epidemiology: challenges and opportunities. *Epidemiology* 2007;**18**:1–8.
12 **Bertram L**, McQueen MB, Mullin K, *et al*. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet* 2007;**39**:17–23.
13 **Clayton D**, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 2001;**358**:1356–60.
14 **Ransohoff DF**. Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer* 2005;**5**:142–9.
15 **Simon R**, Radmacher MD, Dobbin K, *et al*. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 2003;**95**:14–18.
16 **Ioannidis JP**. Microarrays and molecular research: noise discovery? *Lancet* 2005;**365**:454–5.
17 **Hood L**, Heath JR, Phelps ME, *et al*. Systems biology and new technologies enable predictive and preventative medicine. *Science* 2004;**306**:640–3.
18 **Fan C**, Oh DS, Wessels L, *et al*. Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med* 2006;**355**:560–9.
19 **Hunter DJ**. Gene-environment interactions in human diseases. *Nat Rev Genet* 2005;**6**:287–98.
20 **Davey Smith G**, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;**32**:1–22.
21 **Weedon MN**, McCarthy MI, Hitman G, *et al*. Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLoS Med* 2006;**3**:e374.
22 **Wang TJ**, Gona P, Larson MG, *et al*. Multiple biomarkers for the prediction of first major cardiovascular events and death. *N Engl J Med* 2006;**355**:2631–9.
23 **Cardon LR**, Idury RM, Harris TJ, *et al*. Testing drug response in the presence of genetic information: sampling issues for clinical trials. *Pharmacogenetics* 2000;**10**:503–10.