The case for a global human genome epidemiology initiative

To the editor:

Collins argues for a large population-based prospective cohort study in the US to assess the role of genes and environment in common diseases[1]. Without such a study, he maintains that the promise of genomic research for improving population health will remain out of reach. This study is worthy of serious consideration but will be expensive, take years to implement and not guarantee the desired benefit of translating human genome discoveries into population health benefits. Here, I contend that what is urgently needed is a coordinated global initiative to carry out and synthesize human genome epidemiologic research worldwide. I discuss three needs driving this initiative and argue that this effort could accelerate translation of human genome discoveries into population health benefits.

First, we need global collaboration in population genomic cohort studies. Advances in genomics have inspired the development of large longitudinal studies, of entire populations, to establish repositories of biological materials (*e.g.*, UK Biobank and Iceland)[2]. Collaboration across these cohort studies is crucial to allow validation of initial findings by minimizing false alarms and to increase statistical power to detect gene-environment interactions, especially for rarer health outcomes. Because of the expected large number of false positive associations (type I errors) between health outcomes and genetic variants, hypothesis testing across sites will have to be accomplished as part of validation of results from hypothesis-generating studies.

The problem of type II errors or poor statistical power is even more challenging. Consider for a moment the staggering implication of the interactions of numerous gene variants and their products. Let us assume that for a common disease only ten genes contribute a substantial population attributable fraction. Even if variation at each locus can be classified dichotomously (susceptible versus nonsusceptible genotype), this will create 2^{10} (>1,000) possible strata. Classification based on just 20 genes will produce more than one million strata. This is methodologically challenging, especially considering interactions of these genes with other genes and environmental factors[2]. No single cohort study, no matter how large, will have adequate power to detect gene-environment interaction for numerous gene variants, especially for rarer health outcomes. Appropriate pooled analyses will increase the chance of finding true associations of relevance to public health. The full potential of cohort studies to shed light on the occurrence of complex diseases will probably be realized only by pooling and synthesis across multiple populations with different genetic, environmental and sociocultural factors. Integrating data across studies will require developing approaches for facilitating pooled analyses and synthesis. We are seeing the beginning of such a global movement across international boundaries with the establishment of P3G by Bartha Knoppers and her colleagues (Public Population Project in Genomics; http://www.p3gconsortium.org/index.cfm).

Second, we need systematic integration of all human genome epidemiology studies. To build our knowledge base on human genes and health, we need to carry out different types of epidemiologic studies and synthesize their results. Epidemiologic studies can be cohort, case-control or cross-sectional in nature. The strengths and limitations of each design are well-known[3]. Cohort studies are often erroneously perceived as inherently superior to case-control studies. Given the large variation in funds and time needed to conduct cohort studies, every effort should be made to conduct case-control studies that are based on a valid population sampling scheme of newly diagnosed cases in well-defined communities and appropriately selected controls. Well-designed population-based incident case-control studies can even be nested in a larger population cohort or population under surveillance[4].

To develop a systematic approach to the integration of epidemiologic data on human genes, the Human Genome Epidemiology Network (HuGENet; http://www.cdc.gov/genomics/hugenet/default.htm) was launched in 1998. This network of individuals and organizations continuously assesses the impact of human genome variation on population health. HuGENet develops and applies systematic approaches to build the global knowledge base on genes and diseases. In May 2004, the network has more than 700 collaborators from 40 different countries. Its website featured 26 reviews of specific gene-disease associations. In addition, HuGENet has continuously abstracted epidemiologic articles on human genes in an online searchable database, by gene, outcome and risk factor. Because of the tendency for publication bias, an ongoing serious systematic evaluation is now needed for published and unpublished data.

Third, we need evidence-based processes that use epidemiologic information. The synthesis of epidemiologic and biologic data should lead to an evidence-based process that assesses the value of genomic information in health care and disease prevention. For example, an interaction between factor V Leiden (FVL) and use of oral contraceptives has been documented (joint relative risk of 30; ref. 5). But the absolute risk is relatively low (28 per 10,000 person-years) among women with FVL who use oral contraceptives. Whether it is beneficial to screen women for FVL before prescribing oral contraceptives is unclear. Venous thrombosis is relatively rare, and mortality from venous thrombosis is low in young women[6]. For healthy women contemplating using oral contraceptives, the risk-benefit equation would not currently favor screening[7]. This

example illustrates that epidemiologic data need to be collected to inform clinical trials and decision-making for health practice. As single cohort studies are carried out around the world, we can begin to synthesize the incomplete epidemiologic knowledge base for use in policy and practice. These reviews will also uncover gaps in our knowledge base that can be filled by new research from ongoing studies.

It is time that we develop a global public health genomics initiative that builds on the currently fragmented efforts of genetic-epidemiologic research around the world. This initiative can be developed through public-private-academic collaborations. In particular, we need to build a robust process that allows data from many biobanks to be integrated through standardized platforms for joint analyses. Also, we need to integrate data obtained from all valid epidemiologic study designs, notably population-based incident case-control studies. Systematic synthesis of epidemiologic data takes time and skills and should be allocated sufficient resources. This proposed initiative can take us a long way towards translating human genome discoveries into population health benefits for citizens of the twenty-first century.

Muin J Khoury

Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, 1600 Clifton Road, Mailstop E82, Atlanta, Georgia 30333, USA. Correspondence should be addressed to M.J.K. (mkhoury@cdc.gov).

1. Collins, F.S. *Nature* **429**, 475–477 (2004).
2. Khoury, M.J., Millikan, R., Little, J. & Gwinn, M. *Int. J. Epidemiol.* (in the press).
3. Khoury, M.J., Beaty, T.H. & Cohen, B.H. *Fundamentals of Genetic Epidemiology* (Oxford University Press, New York, 1993).
4. Thomas, D.C. Statistical issues in the design and analysis of gene-disease association studies. in *Human Genome Epidemiology: A Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease* (eds. Khoury, M.J., Little, J. & Burke, W.) 92–110 (Oxford University Press, New York, 2004).
5. Vanddenbroucke, J.P. *et al. Lancet* **344**, 1453–1457 (1994).
6. Sass, A.E. & Neufeld, E.J. *Curr. Opin. Pediatr.* **14**, 370–378 (2002).
7. Khoury, M.J., McCabe, L.L. & McCabe, E.R. *N. Engl. J. Med.* **348**, 50–58 (2003).

Phylogenetic validation of horizontal gene transfer?

To the editor:

The study by Nakamura and coworkers[1] offers insight into the computational analysis of horizontal gene transfer. Their results seem to be convincingly supported by phylogenetic validation of the supplied examples of calculated horizontal gene transfer events. An outstanding example, validating their method, concerns the presence of the gene encoding an rRNA adenine N-6-methyltransferase, NMB0066, in the genome of *Neisseria meningitidis* MC58 (ref. 2). According to the authors' results, NMB0066 originates from plasmids naturally occurring in *Staphylococcus aureus*, such as pE5. In fact, this gene, being an erythromycin resistance cassette (*ermC*), was horizontally acquired, because it was deliberately introduced in the *N. meningitidis* MC58 genome by genetic modification using plasmid pIP10 (ref. 3) to reduce virulence. In the pIP10 construct, the gene encoding the polysialic acid capsule biosynthesis protein SiaD (NMB0067) is inactivated by insertion of cloning vector sequences and the *ermC* gene originally derived from plasmid pIM13, a naturally occurring plasmid found in *Bacillus subtilis*[4]. Remnants of cloning vector sequences flanking NMB0066 are noticeable in the genome sequence of *N. meningitidis* MC58. The sequences of NMB0066 and *ermC* of pIM13 are identical, whereas that of *ermC* of pE5 contains one nonsynonymous mutation and an insertion of 107 nucleotides upstream of the open reading frame. This means that, although NMB0066 is clearly horizontally acquired by *N. meningitidis* MC58, its origin remains at best obscure. In addition, it is implausible that the surrounding genes, NMB0065 through NMB0070, were acquired in one event from the same donor as *ermC*, opposing the authors' suggestion that they were transferred simultaneously with NMB0066. In conclusion, although the algorithm by Nakamura and coworkers correctly identified the acquisition of NMB0066 by *N. meningitidis*, their suggestion that *S. aureus* was the donor organism is improbable. Moreover, their interpretation concerning the simultaneous acquisition of NMB0066 and its surrounding genes is inappropriate.

Mark van Passel, Aldert Bart, Yvonne Pannekoek & Arie van der Ende

Academic Medical Center, Department of Medical Microbiology, University of Amsterdam, Amsterdam, the Netherlands. Correspondence should be addressed to A.v.d.E. (a.vanderende@amc.uva.nl).

1. Nakamura, Y., Itoh T., Matsuda, H. & Gojobori, T. *Nat. Genet.* **36**, 760–766 (2004).
2. Tettelin, H. *et al. Science* **287**, 1809–1815 (2000).
3. Virji, M. *et al. Mol. Microbiol.* **18**, 741–754 (1995).
4. Monod, M., Denoya, C. & Dubnau, D. *J. Bacteriol.* **167**, 138–147 (1986).

The use of genome annotation data and its impact on biological conclusions

To the editor:

We were interested to read the recent paper by Nakamura *et al.*[1] describing a new technique to identify horizontally acquired genes in bacterial genomes. But we were surprised to see that NMB0066, a gene from the *Neisseria meningitidis* MC58 genome, was used as an example of horizontal transfer. In fact, NMB0066 is part of an artificial erythromycin resistance cassette that was inserted into the capsule gene *siaD* (NMB0067) to disrupt it, rendering the MC58 strain less virulent and therefore less hazardous to manipulate in the laboratory. The annotation of NMB0066 submitted to the public databases clearly indicates that it is foreign: "NMB0066 rRNA adenine N-6-methyltransferase (ErmC); foreign cassette inserted to disrupt NMB0067

(SiaD) to reduce virulence." This gene may be a good positive control for the *in silico* approach used by Nakamura *et al.*, but it is biologically irrelevant in the context of this genome. Therefore, the further discussion on how this gene and the capsule locus in which it is inserted were horizontally transferred from *Staphylococcus aureus* is meaningless. The capsule locus probably was recently acquired by MC58 but probably not directly from *S. aureus*.

This example highlights the importance of not taking the output of any bioinformatics program at face value; the results should always be interpreted in the biological context of the organism or sequence under study, and the relevant literature should be thoroughly examined. In this case, however, reading the literature may not have helped; in the original sequence paper[2] the gene NMB0066 was mentioned as being within an island of horizontal transfer (the capsule locus), but, due to an oversight, the specific reason for its presence was not spelled out. Notwithstanding the fact that this was described in the annotation submitted to the public databases, this may have been misleading for the casual reader, for which we apologize. This serves to underscore the need to be rigorous in interpreting data, both one's own and those from other groups.

***Hervé Tettelin*[1] *& Julian Parkhill*[2]**

[1]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA. [2]The Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. Correspondence should be addressed to H.T. (Tettelin@tigr.org).

1. Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. *Nat. Genet.* **36**, 760–766 (2004).
2. Tettelin, H. *et al. Science* **287**, 1809–1815 (2000).