New Technologies for Integrating Genomic, Environmental and Trait Data

George M. Church, Ph.D.

Department of Genetics, Harvard Medical School, Boston, MA

Rare diseases, which (by definition) occur at a frequency less than 1/2000 per allele – are individually rare, yet common collectively (10% affected and 50% carrier rates). There are 1800 genes which have tests considered highly predictive and actionable. Human genes with known variants causing insomnia, narcolepsy, and circadian variation include Prion Protein Fatal Familial Insomnia (PRNP), hypocretin (HCRT), DQ beta 1 (DQB1), and period circadian protein homolog (PER2). We have developed human genome sequencing technology that lowered costs a million-fold over

the past 6 yr. This has increasingly enabled the use of the causative alleles above, which are far more valuable than merely correlated or common variants. To expand this further we have established community resources for open access collection, integration and interpretation of diverse personal genomic, environmental and trait data (evidence. personalgenomes.org)

Citation: Church GM. New technologies for integrating genomic, environmental and trait data. *J Clin Sleep Med* 2011;7(5):Supplement S43-S44.

There are a number of new and different technologies that are now available for studying genetic traits including a potential biomarker for the predisposition of individuals to be sleepy. In addition, the cost of sequencing the entire genome has declined enormously, by about 40,000 fold in the last four years and almost by a million-fold in the last six or seven years. This has been made possible by a number of “next-generation” sequencing technologies that allow for rapid and relatively inexpensive sequencing of the entire genome.

With respect to genomes, there are approximately three million differences between 2 individuals. Some of them are single-base differences, some of them are copy-number repeats and some are other differences. The genome can be analyzed by looking at any cell in the body. With the exception of cancer and immune cells, as far as we know, all cells essentially have the same genome.

The following is an example of how sequencing the genome might identify a gene or genes responsible for a disease. Consider examining the genome of an individual with a particular trait. There may be a large number of potential candidate genes, but probably only one or two are really causing disease. If additional genomes are examined from individuals with the same trait, especially if it is a very distinctive trait and if criteria for important changes in the DNA are made more strict, one can narrow the abnormality down to one change. This was done for Freeman-Sheldon Syndrome, an inherited craniofacial dysmorphism, where 2 different dominant alleles were recently identified as producing the disease.¹

However, it is not just the genome. Expression of a trait also is markedly affected by the environment such as various therapies, immune responses through the microbial environment, allergens, chemicals, and nutrition. These are all part of our epigenome. In contrast to the genome, almost every cell is dif-

ferent. Therefore, to analyze the epigenome, every cell would need to be examined, or at least the right cells for a particular trait or purpose.

One approach to identify relevant genomic and epigenomic variations responsible for particular phenotypes is the establishment of databases containing genetic and environmental information from large numbers of individuals. One such database is the Personal Genome Project.² It is the first and only open access database of its kind, and has a branch in Boston and several other countries. It is approved for 100,000 individuals in Boston and there are different goals for the other centers. There have been 16,000 persons who so far have volunteered. It is intended to study multiple traits, not just one trait like sleepiness, but as many traits as possible. Because, as additional traits are added, the set of genomic data becomes increasingly richer. Thus, the cohort becomes more and more valuable both as cases and controls for a whole variety of medical and non-medical traits. One can then apply to test for a trait in this cohort. For example, in one project, variation in fMRI performance is being correlated with variations in the genome.

The Personal Genome Project will allow exploration of associations among the genome, microbiome and the immunome. In one such study, the dynamic antibody response over 28 d to vaccination against Hepatitis A and B, and several flu viruses is being studied in relationship to variations in immunoglobulin genomic sequences.

As previously noted, ideally to assess the impact of genes and environment in determining the complex traits that define individual humans, it would be necessary to examine every cell. Obviously, this is not possible. However, one approach is to use pluripotent stem cells derived from the skin which are then reprogrammed to become a variety of other cells such as respiratory epithelium, bone and neuroectoderm.³ One can then

analyze differences in protein production and function, and differences and ratios of RNA in among these different cell types.

While it may seem that genomics has little current clinical value, there are over 2000 genes which have tests considered to be highly predictive and actionable. These are tests that medical geneticists routinely use (some since 1991). Forty of these are required by every state to be done on every infant born. Testing for these genes costs ~\$1000 each, and thus all are not routinely done unless there is clinical suspicion such as a positive family history. However, this could change because although rare diseases, which (by definition) occur at a frequency less than 1/2000 per allele and are individually rare, they are common collectively (10% affected and 50% carrier rates).

There are several human genes with known variants causing insomnia, narcolepsy, and circadian variation including PRNP, HCRT and PER2. With respect to PRNP, there is some evidence suggesting that a PRNP polymorphism may affect sleep in a healthy population in addition to being the cause of fatal familial insomnia.⁴ Deficiency in hypocretin (orexin) is now known to be a common cause of cases of narcolepsy,⁵ and the HLA-DQB1*0602 allele is another putative causative factor (possibly acting through an autoimmune response to cells producing hypocretin). Familial advanced sleep phase syndrome is caused by a mutation in PER2.⁶

In summary, the ability to measure all gene variants especially those that are highly predictable and medically actionable is improving. Although variants in this latter category are rare, they are collectively common (*vide supra*). There are now a variety of technologies that are available for sequencing one's individual genome, a once in a lifetime event. However, these same technologies can be used in the context of large databases

such as the Personal Genome Project to study the impact of such factors as the microbiome and immunome on determining complex human traits.

REFERENCES

1. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009;461:272–6.
2. Personal Genome Project. www.personalgenomes.org
3. Lee JH, Park IH, Gao Y, et al. A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *PLoS Genet* 2009;5:e1000718.
4. Plazzi G, Montagna P, Beelke M, et al. Does the prion protein gene 129 codon polymorphism influence sleep? Evidence from a fatal familial insomnia kindred. *Clin Neurophysiol* 2002;113:1948–53.
5. Peyron, C., Faraco, J., Rogers, et al. A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains. *Nat Med* 2000;6:991–7.
6. Toh KL, Jones CR, He Y, et al. An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome. *Science* 2001;291:1040–3.

ACKNOWLEDGMENTS

Editing of the conference proceedings supported by HL104874.

CORRESPONDENCE

Address correspondence to: George M. Church, Ph.D., Department of Genetics, Harvard Medical School, Genetics NRB Room 238, 77 Avenue Louis Pasteur, Boston, MA 02115; Tel: (617) 432-7562; Fax: (617) 432-6513; E-mail: gmc@harvard.edu

DISCLOSURE STATEMENT

Dr. Church has indicated no financial conflicts of interest.