Molecular Systems Biology (2005) doi:10.1038/msb4100040

www.molecularsystemsbiology.com
Article number: 2005.0030

molecular systems biology

EDITORIAL

The Personal Genome Project

Molecular Systems Biology 13 December 2005; doi:10.1038/msb4100040

Large potential benefits for systems biology reside in applications to human health and identity. To develop our community's skills in these directions, ready access to *highly integrated and comprehensive* human genome and phenome data sets is extremely important and increasingly feasible technically. The few human 'functional genomics' data sets available today tend to be isolated from one another. Some of the tools needed to break through this impasse are addressed below in the context of a Personal Genome Project (PGP) as a natural successor to the Human Genome Project (HGP)—two recent buds in the ancient field of genetics.

From my first interaction with Wally Gilbert in 1976, it seemed that a large (but appealing) leap would be to go from his new method for sequencing 30 bp segments to a method to get everyone's full genome sequenced. Six billion base pairs for six billion people had a nice ring to it. This was still merely a fantasy when we published a paper called 'Genomic Sequencing' in 1984 (Church and Gilbert, 1984) and conspired to create a 3 billion dollar HGP later that year (Cook-Deegan, 1989). For the subsequent 16 years, radical technology development (while kept alive in a few 'back-rooms') was clearly a minor funding priority relative to 'production' sequencing. However, by 2001, the criticisms of the old technology grew and the call for affordable personal genomes became irresistible (Jonietz, 2001). In early 2004, the NIH-NHGRI posted a request for applications, and in October 2004 and August 2005, announced grant awards totaling $70 million for technology leading to human genome sequences for $100 000 in 5 years and $1000 in 10 years (http://www.nih.gov/news/pr/aug2005/nhgri-08.htm). As if the motivation were not already high enough, at the recent Genome Sequencing & Analysis Conference in Hilton Head (October 18), the prospect of a new X-prize arose to encourage this new Personal Genomics field. (The first X-prize, $10 million for re-usable spacecraft, was awarded in October 4, 2005 and is followed by a $50 million prize for orbiting.) Amid all of this positive reinforcement, some key points were left fuzzy—What exactly is meant by sequencing a human genome? What is the utility of personal genomes? What are the ethical, legal, and social implications (ELSI)? The time has come to sharpen these points up. As we begin to purchase personal genomes, we want to know what we are paying for.

What is meant by sequencing a human genome? To quantitatively assess progress in this field, we need engineering specifications for the quality and cost of genomes and practical ways for standardizing and validating progress. The quality of genomes can be stated by the fraction of the genome sequenced and the accuracy of those covered regions. Both of these measures have nuances. The fraction of the genome can be of the whole genome or just the protein-coding or just the euchromatic regions. In our current rate of discovering new genetic elements and phenomena that can affect human health, how close we can get to a whole genome sequence is what we want to measure. The quality can count an insertion or deletion of 4 bp as one error or four errors. The former seems like an appropriate measure for tandem repeats. An erroneous 'crossover' from one haplotype to another would count as a single error rather than counting up all downstream errors. An individual genome can be estimated to within 1E−3 simply by guessing that the individual is the same as the hodgepodge HGP genome currently in GenBank. It can be guessed to within 1E−4 by using occasional SNPs to drag in 10 kb or so of linked common SNPs. However, an error rate of 1E−4 means that there will be 600 000 errors in a diploid human genome sequence each of which would require additional sequencing or functional tests (more base-pair errors than some entire genomes). The genomic differences between a cancer cell and a normal cell might be around 1E−6 per base pair (Wang, 2004) (only some of those changes being causative for the pathological proliferation). So a reasonable X-prize challenge would be to sequence 100 diploid (and some aneuploid) human genomes at the same coverage and accuracy as today's haploid mosaic, that is, 99% of the euchromatin and 93% of the whole genome at an error rate better than 1E−6. The results could be checked by the other teams using a shared set of DNAs (see discussion of cell lines below).

The utility of the first personal genome is analogous to the first fax machine, web page, or computer. Until communities of resources build up, these revolutionary new tools serve mainly the 'early adopters'. These initial participants are heroes and human guinea-pigs paving the way for potentially increasing utility for the general public. For personal genomics, we are already seeing market activity in genetic diseases, cancer, and pharmacogenomics, for example, tests for BRCA1/2, EGFR that impact diagnostics and drug choice and dose. This is expanding rapidly to include personalized nutrition and lifestyle decisions. As DNA is only a small part of destiny, personal genomics might fruitfully de-emphasize 'prediction' and focus on augmenting systems biology interpretations and prioritizations of actual day-to-day measurements of our physiological states (Hood *et al*, 2004). Even before the PGP is fully ready for medical practice, it will aid research in human functional genomics and systems biology. These topics need the perspective of inter-individual variation (Cheung, 2005). A system model that describes a canonical or consensus cell is not as informative as one that embraces natural genetic variation and epigenetic stochastics as well.

What would we do if we were given our full genome tomorrow? We would first rank our 6 million differences relative to a reference genome, looking for mutations which (a) affect known disease genes (or genes related by function or homology), (b) affect conserved genetic elements (Ramensky,

2002; Vitkup, 2003), and/or (c) create homozygous or (co)dominant alleles. Then, we generate hypotheses about potential consequences of the top ranked changes including customizing existing system models. Then, we would want to test these hypotheses with focused population association studies, animal models, and functional genomics on the cells from the PGP subjects. Results from those studies will allow us to better personally prioritize clinical diagnostics, therapeutics, lifestyle, and nutritional changes.

What are the ELSI of personal genomics? Although easy to get carried away by the excitement of the technology, we need to also ask what is needed to help it play out in society positively. Will acceptance be rapid, like the worldwide web in the year 1993, or go through a series of major setbacks as has occurred with genetically modified organisms? Two top issues that come up are privacy and insurance. First, privacy. An ideal systems biology resource would be a highly integrated data set for all aspects of human phenotype for a genetically diverse set of subjects. This would include full medical records, omics data, and potentially identifying data like craniofacial MRI and 3D photogrammetry. Facial features are among the most noticeable and socially significant of human phenotypes. Correlating them with genotype is of huge significance for forensics and security applications and will encourage early visualization of positive and negative uses. Research on human subjects, especially when de-identification is impractical (as above), requires Institutional Review Board (IRB) approval not only for collection of the data, but also for accessing it. Occasionally, researchers deviate from the filed IRB plans or evade the IRB entirely, risking loss of funding, loss of access to clinical resources, and potential harm to the subjects. It is clearly far more beneficial to communicate the full study plan to the IRB and the public in advance to catch broader ramifications than single researchers typically imagine (Kennedy, 2002; Church, 2005a).

The more popular and broadly distributed the PGP genome and phenome data, the more likely that accidental or deliberate release to a public part of the internet (and Google) or re-identification (Kohane and Altman, 2005) will occur. If the study subjects are consented with the promise of permanent confidentiality of their records, then the exposure of their data could result in psychological trauma to the participants and loss of public trust in the project. On the other hand, if subjects are recruited and consented based on expectation of full public data release, then the above risks to the subjects and the project can be avoided. How many volunteers are willing to participate initially and how many as the study expands (each noting the cumulative experiences of the earlier participants) is what we can call PGP-ELSI question #1. The fraction of people willing to volunteer may surprise us. Like pioneers, health-care workers, and astronauts, they will put themselves (and their families) at risk, but with rewards for society (and their families). The number of personal facts considered stigmatizing has been dropping since the 1960s when cancer, depression, sexual dysfunction, and sexually transmitted diseases were taboo topics, while today discussion of personal decisions on Iressa, Viagra, Prozac, and AZT are common. The key point is that an open-PGP might be able scale up to thousands of subjects without relying on perfect data security, and whether it will is not a theoretical question. As full exposure is a new concept, the PGP will stimulate ELSI research examining broader implications including insurance, workplace discrimination, and profiling.

How do we minimize risk to the participants? The initial participants should be diverse, yet very familiar with research on human subjects, genetics, information technology, and ELSI. They should have thoroughly considered a variety of worst-case scenarios. The subject ideally (and close relatives too) would not be hesitant about knowing and sharing ANY part of the subject's genome or Personal Health Records (Sands and Halamka, 2004; Kohane and Altman, 2005), as whatever makes that part scary to them could make some other part scary at a later date (after it is too late to remove the data from the public domain). The goal of the PGP should be to slowly ramp up the number and diversity of participants including outreach and broad-audience genetics education. At each stage, a Data Safety Monitoring Committee will assess the need for discontinuation or mid-course corrections. Initial trust may be higher if the PGP is a non-profit entity closely affiliated with medical schools and teaching hospitals rather than a commercial enterprise. To increase trust further, PGP question #2 will be whether it is feasible to protect participants (beyond their normal health, life, and employment insurance) specifically for genetic discrimination (Geller *et al*, 1996) consequential to the PGP study itself. Some initial participants may be healthy senior geneticists and hence intrinsically low risk of being surprised, but in the interests of generalizability, subjects should also be chosen to be closer to normal risk levels. Nevertheless, the PGP insurance fund could still be 'practical' as it does not have to be immediately 'profitable'. It could leverage donations, possibly including participation from insurance companies recognizing a potential sea change.

What happens if subjects learn about predispositions to diseases without cures? One classic approach to this problem has been to hide the data from the subject. An alternative is that the subjects can become expert advocates for research on the disease afflicting their families. Notable examples of this are Augusto Odone (adrenoleukodystrophy and Lorenzo's oil), Doug Melton (diabetes and the Harvard Stem Cell Institute), Nancy Wexler (Huntington's disease), and Mike Milken (prostate cancer). This might be PGP question #3.

Scaling up genetic epidemiology. Initially, the PGP will have a small number of participants and hence might focus on the hypothesis generation and the cell-based testing described above. Epidemiological and genetic association studies will benefit from thousands or even millions of participants, and an early successful PGP experience might open up new study design options for multidisease studies on large populations like the Nurses Health Study and Health Professionals' Follow-up Study (Forman, 2005). Radically new social and business models of data sharing might emerge, analogous to peer-to-peer, wiki, and blog phenomena. So, question #4 of the PGP—What new software will arrive to utilize the 'openness' of the PGP that would not (or did not) apply to previous (closed) studies? Once approved for publication by the primary IRB, the PGP data would be available for further research just like any

other public domain data without reapproval by other IRBs. Software developers and users could be contacted to assess the impact they felt from such a policy.

In order to standardize and validate the new technologies, we need standard genomic DNA (and other molecules and cells) from a reliable, renewable source, that is, cell lines. Ideally, these cells would correspond to individuals participating in an IRB-approved 'open' project like the PGP. The industry-standard for genomic resources is an EBV-transformed B-cell line. The PGP has initiated a collaboration with the Coriell NIGMS repository to provide community access to the cells and DNA corresponding to each PGP volunteer (Church, 2005a). To aid in collecting transcriptome and proteome data from a variety of cell types for each individual, a pluripotent cell line would also be extremely useful (Hwang, 2005). These cells would represent the fusion of analytic and synthetic (Church, 2005b) tools at the cutting edge of personalized medicine where histocompatible stem cells are established for each individual and vetted for full genomic and epigenomic quality in advance of use.

In summary, biological and medical research need not only new 'omics' technology and systems models, but also new ways to assess technology and to obtain low-risk, 'open', integrated data sets focused on inter-individual variation. The PGP and projects like it are works in progress and likely to change and diversify in response to a variety of inputs.

Acknowledgements

This work has been supported in part by the ELSI component of an NIH-NHGRI CEGS grant with many helpful discussions over the past year with the Harvard Medical School-IRB, members of the Personal Health Records and ELSI communities including Ting Wu, John Aach, Zak Kohane, Esther Dyson, John Halamka, Eric Juengst, Lynn Dressler, Mildred Cho, and Vivian Ota Wang. This note of thanks is not meant to imply that they agree with this paper.

References

Cheung VG *et al* (2005) Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437:** 1365–1369

Church GM (2005a) Personal Genome Project. HMS-IRB approval Aug 2005. http://pgen.us

Church GM (2005b) From systems biology to synthetic biology. *Mol Systems Biol* doi:10.1038/msb4100007

Church GM, Gilbert W (1984) Genomic sequencing. *Proc Natl Acad Sci USA* **81:** 1991–1995

Cook-Deegan RM (1989) The Alta summit, December 1984. *Genomics* **5:** 661–663. http://www.ornl.gov/sci/techresources/Human_Genome/project/alta.shtml

Forman JP *et al* (2005) Vitamin D intake and risk of incident hypertension: results from three large prospective cohort studies. *Hypertension* **46:** 676–682

Geller LN, Alper JS, Billings PR, Barash CI, Beckwith J, Natowicz MR (1996) Individual, family, and societal dimensions of genetic discrimination: a case study analysis. *Sci Eng Ethics* **2:** 71–88

Hood L, Heath JR, Phelps ME, Lin B (2004) Systems biology and new technologies enable predictive and preventative medicine. *Science* **306:** 640–643

Hwang WS *et al* (2005) Patient-specific embryonic stem cells derived from human SCNT blastocysts. *Science* **308:** 1777–1783

Jonietz E (2001) Personal genomes. *Technol Rev* 30

Kennedy D (2002) Not wicked, perhaps, but tacky. *Science* **297:** 1237

Kohane IS, Altman RB (2005) Health-information altruists—a potentially critical resource. *N Engl J Med* **353:** 2074–2077

Ramensky *et al* (2002) Human non-synonymous SNPs. *Nucleic Acids Res* **30:** 3894

Sands DZ, Halamka JD (2004) PatientSite: patient centered communication, services, and access to information. In *Consumer Informatics: Applications and Strategies in Cyber Health Care*, Nelson R, Ball MJ (eds). New York: Springer-Verlag

Vitkup *et al* (2003) Amino-acid mutational spectrum of human genetic disease. *Genome Biol* **4:** R72

Wang Z *et al* (2004) Mutational analysis of the tyrosine phosphatome in colorectal cancers. *Science* **304:** 1164–1166

GM Church
Department of Genetics, Harvard Medical School, Boston, MA, USA