COMMENTARY

Common sense for our genomes

A personal DNA sequence is not yet practically useful. But it could be, argues **Steven E. Brenner**, if we had the right resources available to interpret genomes.

Revelation of the complete DNA sequences of James Watson and J. Craig Venter elicited headlines in recent months, but most press reports struggled to offer meaningful interpretations. The most noted observation was that Venter has a particular gene variant predisposing him to cardiac disease, although his family history was enough to let him know about this general risk. If the genome is so revealing, why was so little revealed?

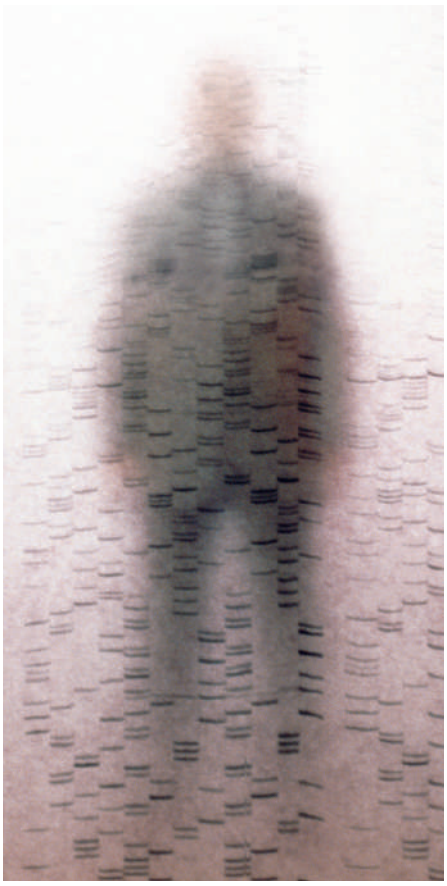
It is telling that Venter said he learned about the cardiac disease gene in a newspaper report. Put simply, even we in the scientific community can't easily come to grips with what we know. The effects of gene variations are scattered in hundreds of databases, across hundreds of interpretative reports in clinical laboratories, and among millions of manuscripts and patent applications. And although some papers discuss the precise effects of a single DNA base change, many analyses offer simple rules of thumb rather than specific guidance.

Moreover, even as we celebrate the advent of personal genome sequencing, we should maintain realistic expectations. Given that most common drug prescriptions don't even consider a patient's weight, it is unclear how many future therapies will depend on the minutiae of our genomic make-up. Indeed, it remains to be seen whether we will typically learn anything more important from our genomes than the need to use sunscreen, eat better and exercise more. However, I believe that if we don't seize the initiative and develop the necessary resources to interpret our genomes, the Venter and Watson genomes will be seen as missed opportunities.

Even the scientific paper reporting Venter's genome revealed less than it might[1]. The gene variants described in the initial analysis, intended to engage a wider audience, could have been selected to elicit guffaws, touching on associations with alcoholism, obesity, novelty-seeking and antisocial behaviour. However, these are all statistical likelihoods, and their relevances are hard to decipher.

Yet, after learning of the genetic variations that render him susceptible to cardiac disease, Craig Venter reportedly assumed a new level of personal responsibility by altering his diet and taking a cholesterol-lowering statin. So personal genomes may offer a way to translate genomic knowledge into better preventive medicine.

Even now, further analyses of the Venter genome[2] could reveal more useful gene variants.



R. WOODWARD/GETTY

For example, cytochrome P450 isozymes determine how rapidly individuals metabolize various drugs, and the US Food and Drug Administraton has approved a microarray test for genotyping these enzymes. Venter's cytochrome P450 gene variants were not reported, but these variations can inform drug dosages.

We are still waiting to learn if the analysis of Watson's genome will reveal more or less than Venter's. Watson's sequence is available online[3] and a small number of gene variants have been automatically annotated using the Online Mendelian Inheritance in Man (OMIM) database. OMIM has 18,000 entries summarizing the literature related to human genes and genetic disorders (see table overleaf). But because such mutations and their effects are described textually, only 133 of the 18,000 could be linked directly to a unique single-nucleotide substitution[4].

"It remains to be seen whether we will learn anything more important from our genomes than the need to use sunscreen, eat better and exercise more."

Visionary geneticists have long contemplated building a resource to consolidate our understanding of genome variation. However, academic squabbles and misunderstandings caused the most comprehensive effort — involving hundreds of scientists backed with millions of dollars — to founder[5]. Perhaps they were premature? Until recently, it was rarely productive to look beyond a single gene known to be of research or clinical interest. Today, the situation has changed radically. With the prospect of inexpensive personal genome sequences, there is profound impetus for integrating our knowledge of genetic variation and its effect on a genomic scale.

Covering the bases

Many of the foundations for describing human genome variation and integrating this knowledge are already in place. The Human Genome Variation Society has defined a standard nomenclature for precisely describing small variants, which makes it possible, for example, to consistently ascertain whether two polymorphisms are the same or different. Central publicly funded databases have repositories of genetic variation information and offer reference genes and genomes on which the variation can be mapped. Among these, dbGaP is an example of a database of genotype–phenotype relationships generated largely from genome-wide association studies. There are also more than 600 locus-specific databases that focus on narrow areas of the genome. But merging these databases with dbGaP, and other data sources, would be a complex task.

I propose establishing a Genome Commons, a public knowledgebase of human genetic variation and its effect, culled from databases, diagnostic laboratories, and the scientific literature. Ultimately, such a repository of our common human inheritance would be a vast resource for research, medicine and understanding ourselves.

There are many ways in which the Genome Commons could be constructed, but I offer some general guiding principles. It would certainly build on the curation of hundreds of small locus-specific and other databases today. This is an often used and successful model, employed for example at GeneTests,

SOME EXISTING SOURCES FOR INTERPRETING HUMAN GENOMES

Name	Website	Brief description	Restrictions on use
dbSNP	www.ncbi.nlm.nih.gov/SNP/	Repository for short nucleotide polymorphisms	None
OMIM, Online Mendelian Inheritance in Man	www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM	Catalogue of 18,000 essays on human genes and genetic disorders	Licence for commercial use or redistribution
dbGaP	www.ncbi.nlm.nih.gov/sites/entrez?db=gap	Mainly a database from genome-wide association studies	None on open data, some on personal
SNPedia	www.snpedia.com	Wikipedia-style site for single-nucleotide polymorphisms	None
HGMD, Human Gene Mutation Database	www.hgmd.cf.ac.uk/ac/index.php	Catalogue of gene mutations responsible for human inherited disease	Fee-based for full access; no redistribution
GeneTests	www.genetests.org/	Summarizes more than 1,000 diagnostic genetic tests	None with proper attribution
PharmGKB	www.pharmgkb.org/	Pharmacogenetics and pharmacogenomics knowledgebase	Some privacy restrictions
Locus Specific Mutation Databases	www.hgvs.org/dblist/glsdb.html	Lists over 600 locus specific databases	Some copyright restrictions
SIFT	http://blocks.fhcrc.org/sift/SIFT.html	Software predicting sequence effects on protein function	None with proper attribution
SNPs3D	www.snps3d.org	Website that predicts phenotypic impact of SNPs	Software not downloadable
A more comprehensive list is compiled by Rania Horaitis and available on the Human Genome Variation Society (HGVS) website at http://www.hgvs.org/dblist/dblist.html			

a reference database of thousands of gene and disease tests for diagnostic use. The editors of GeneTests benefit from contributions by hundreds of experts who volunteer their knowledge. Similarly, quality controls in the Genome Commons would be provided by experts overseeing entries in their domain of expertise, typically a set of genes or diseases. In addition to their own contributions, they would collate and review entries that could be submitted by anyone with access to academic journals and appropriate training.

Share and share alike

To work on a genomic scale, the Genome Commons would need to be carefully structured, incorporating statistical details about data quality and the strength of associations for researchers, as well as clinical references for eventual use by medical practitioners. It is essential that the Genome Commons be open for remixing, augmentation and redistribution of content. It is only in this way that researchers can fully share their knowledge and allow others to build on it.

An individual genome will typically have millions of differences when compared with a reference genome; most differences are of little consequence, but some single mutations can be fatal. The Genome Commons itself need not contain any individual's information and thus raises few ethical or privacy concerns. However, both for research purposes and for clinical interpretation, we will need a navigation tool to relate each individual's variations to the knowledge compiled in the Genome Commons.

But sequenced genomes do not come indexed for easy analysis and our knowledge is so multilayered, that it presents a technical challenge. At one extreme, for sickle-cell anaemia, we understand the molecular mechanism by which mutation leads to disease. In many more instances, however, there is a single-gene association, without any mechanistic understanding. In general, we are happy to find any significant association of phenotype with a genetic marker. Most variations have never been phenotypically characterized — Venter's genome had more than a million variants never seen before — and analysing these will require predictive approaches. Moreover, variations appear on different scales in the genome, ranging from small substitutions, insertions and deletions, to large-scale chromosomal restructuring.

Initially, I imagine that a Genome Commons navigator would amalgamate observed variation, and propose phenotypic interpretations. This first step would allow researchers to assess the challenge and promise of these data, and to design further research and analysis methods. Later versions of navigators will incorporate the best methods from many research groups. But to truly interpret a genome, we face the more daunting challenge of sifting through the millions of variations and ranking them so that we are not deluged with genomic marginalia. The navigator would eventually present a status report focusing on genetic differences of greatest medical or personal importance.

Private enterprise would play a vital part by providing an interface between the Genome Commons and the wider community. Researchers would access the Genome Commons directly, but companies would mediate its delivery to patients and physicians. Just as clinical laboratories are used by physicians to perform diagnostic testing today, I would expect clinical labs to perform large-scale genome sequencing in the future. I envisage these labs — and new companies such as 23andMe and Navigenics — using the Genome Commons navigator as a reference tool for producing diagnostic reports.

Much genomic variation information is not free, or is encumbered with intellectual-property protection. To be fully successful, companies must also contribute discoveries to the Genome Commons. As a central clearing house of intellectual property, the Genome Commons could reduce transaction costs. Companies could contribute information and accept a standard agreement for diagnostic use, making it easier for clinical laboratories to license large quantities of intellectual property with minimal overheads. In this way, more assays become accessible and affordable to patients.

The cost to create and maintain the Genome Commons will be considerable, even if many volunteers assist the effort. Extrapolating from the costs of other resources, such as OMIM, PharmGKB and GeneTests, the core knowledgebase may require millions of dollars in support each year. Most of this would be spent on salaries for curators and staff overseeing the informatics.

Ideally, the Genome Commons would be primarily funded as a government resource or by a major charity, although many companies will have strategic economic reasons to financially support an open resource. If a public Genome Commons fails to emerge, we may instead get a private resource with similar content, but whose licensing requirements stymie research and innovation. A single private resource would also lead to monopoly pricing for diagnostic information. After the huge investments made to ensure that a human genome sequence was public and free, additional outlays for the Genome Commons seem prudent so that genomes can be readily interpreted for medical practice and research.

The challenges of building a Genome Commons and navigator are not trivial, but this resource could affect us all personally. In a world where we all face limited time, resources and personal restraint, an open Genome Commons would eventually enable productive use of the wealth of information available, helping us to prioritize healthy activities and therapies to give us the most productive and enjoyable lifespans. ■

Steven E. Brenner is at the Department of Plant and Microbial Biology, 111 Koshland Hall, University of California, Berkeley, California 94720, USA.

1. Levy S, *et al. PLoS Biol.* **5,** e254 (2007).
2. www.jcvi.org/research/huref/
3. http://jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/jwsequence/
4. https://mice.cs.columbia.edu/getTechreport.php?techreportID=448&format=pdf
5. Maurer, S. M. *Res. Policy* **35,** 839–853 (2006).

Join the discussion at www.GenomeCommons.org