

## Amalgamated version

Black text is the final published version, while purple text reflects most significant content present in earlier versions. Due to major textual rearrangements, this amalgamated version does not always flow sensibly. In many cases, this additional text is unrefined but presented for completeness.

## A Genome Commons

### The shadowy genome

#### 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 pioneering geneticists James Watson and J. Craig Venter elicited headlines in recent months, but most press reports struggled to offer meaningful interpretations. Astonishingly, most press reports offered no meaningful interpretation of these genomes other than that Watson's genome had many regions not previously seen, and that it masked information about an Alzheimer's gene—because Watson did not want to know about it. 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 the *New York Times* 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 detail and 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 upon 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. The manuscript also reveals that Venter has some gene variants associated with reduced cardiac risk, providing a more nuanced picture than the early news stories conveyed.

The very few definitive alleles reported include the observation that Venter has wet ear-wax, which presumably could have been determined by means less elaborate than genome sequencing.

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 of **definitive genes**. For example, variants of cytochrome P450 isozymes determine how rapidly individuals metabolize various drugs, and the US Food and Drug Administration has approved the **AmpliChip** microarray test for genotyping these enzymes to **determine their metabolic efficiency**. Venter's cytochrome P450 gene variants were not reported, but these variations can inform drug dosages. **One could painstakingly apply the rules in the AmpliChip package insert to Venter's genome sequence so as to determine whether Venter might need special dosing of beta-blockers in the event his statins don't do the job. Were Venter an infant, his genome could have predicted that he would be unafflicted by many diseases that time has now shown him to be free of. Indeed there are already more than a thousand genetic disease tests in clinical use.**

We are still waiting to learn if the analysis of Watson's genome will reveal more or less than Venter's, **beyond well-reported suppression of Watson's ApoE gene**. 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. **The most prominent of these is an allele associated with age-related macular degeneration, which Watson seems unlikely to suffer at this point.** 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 automatically linked directly to a unique single-nucleotide substitution[4].

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. **Many of these databases could use software like LOVD, which offers a consistent architecture, and facilitates their integration.** But merging all 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, 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 **in a systematic manner** by anyone **with an Internet connection**, 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 **integration**, 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, **just as the authors of Venter's genome paper noted that "the release of Celera Assembler as an open-source project has allowed us and others to continue to improve the assembly algorithms."**[1]

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 multi-layered that it presents a technical challenge. At one extreme, for sickle cell anemia, 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 analyzing these will require predictive approaches. Moreover, variations appear on different scales in the genome, ranging from small substitutions, insertions and deletions, **to copy number variation to chromosomal inversion** 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 medical 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 clearinghouse 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, **by reducing expenses associated with negotiating each individual license**, more assays become accessible and affordable to patients. **This approach also offers new revenue models for promulgating genetic discoveries.**

**The challenges of building a Genome Commons navigator are not trivial.** 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 support **on the order of** millions of dollars each year, **though lesser sums could be useful.** Most of this would be spent on salaries for curators and staff overseeing the informatics.

**While intellectual property management may benefit from a commercial aspect**, 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. **Because it appears to be a winner-take-all market**, 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.

Companies would also have incentives to support the Genome Commons and could aid its launch and development. Interest in personal genome sequencing is limited today by the lack of perceived value in the sequence generated. Thus, sequencing and genotyping device vendors will be able to grow the market for their products if information to interpret genomes is more readily available. Likewise, corporations that intend to use genetic assays for diagnostic testing whether in the traditional medical realm or in various more recent intriguing guises (i.e., 23andMe, etc), will benefit from the Genome Commons and navigator. Therapeutics companies could strategically subsidize tests that would help lure new consumers for their products, and thus they too would have incentives to support a Genome Commons.

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 & Microbial Biology, 111 Koshland Hall, University of California, Berkeley, California 94720, USA.**

[1] Levy S, et al. *PLoS Biology* **5**, e254 (2007). Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol* **5**: e254 doi:[10.1371/journal.pbio.0050254](https://doi.org/10.1371/journal.pbio.0050254)

[2] [www.jcvi.org/research/huref/](http://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). Maurer SM. 2006. Inside the anticommmons: academic scientists' struggle to build a commercially self-supporting human mutations database, 1999-2001. *Research Policy* **25**:839-853. doi:[10.1016/j.respol.2006.04.008](https://doi.org/10.1016/j.respol.2006.04.008)

**Join the discussion at [www.GenomeCommons.org](http://www.GenomeCommons.org)**

## SOME EXISTING SOURCES FOR INTERPRETING HUMAN GENOMES

Name	Website	Brief description	Restrictions on use
dbSNP	<a href="http://www.ncbi.nlm.nih.gov/SNP/">www.ncbi.nlm.nih.gov/SNP/</a>	Repository for short nucleotide polymorphisms	None
OMIM, Online Mendelian Inheritance in Man	<a href="http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM">www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM</a>	Catalogue of 18,000 essays on human genes and genetic disorders	Licence for commercial use or redistribution
dbGaP	<a href="http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap">www.ncbi.nlm.nih.gov/sites/entrez?db=gap</a>	Mainly a database from genome-wide association studies	None on open data, some on personal
SNPedia	<a href="http://www.snpedia.com">www.snpedia.com</a>	Wikipedia-style site for single-nucleotide polymorphisms	None
HGMD, Human Gene Mutation Database	<a href="http://www.hgmd.cf.ac.uk/ac/index.php">www.hgmd.cf.ac.uk/ac/index.php</a>	Catalogue of gene mutations responsible for human inherited disease	Fee-based for full access; no redistribution
GeneTests	<a href="http://www.genetests.org/">www.genetests.org/</a>	Summarizes more than 1,000 diagnostic genetic tests	None with proper attribution
PharmGKB	<a href="http://www.pharmgkb.org/">www.pharmgkb.org/</a>	Pharmacogenetics and pharmacogenomics knowledgebase	Some privacy restrictions
Locus Specific Mutation Databases	<a href="http://www.hgvs.org/dblist/qlsdb.html">www.hgvs.org/dblist/qlsdb.html</a>	Lists over 600 locus specific databases	Some copyright restrictions
SIFT	<a href="http://blocks.fhcrc.org/sift/SIFT.html">http://blocks.fhcrc.org/sift/SIFT.html</a>	Software predicting sequence effects on protein function	None with proper attribution
SNPs3D	<a href="http://www.snps3d.org/">www.snps3d.org/</a>	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>

### Some resources for interpretation of human genome variation

*Table: each entry here has a database name, URL, a brief description, information I could elicit regarding restrictions on use (in italics), and the preferred citation, if one is given.*

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>

#### dbSNP

<http://www.ncbi.nlm.nih.gov/SNP/>

Maintained by the National Center for Biotechnology Information (NCBI), this is a central repository for short nucleotide substitution, deletion, and insertion polymorphisms. The database has very little phenotypic information, but it has some links to OMIM, and NCBI is planning to add the ability to permit users to submit annotations.

*There are no restrictions on use of dbSNP.*

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001 dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29:308-11. doi:[10.1093/nar/29.1.308](https://doi.org/10.1093/nar/29.1.308)

#### OMIM, Online Mendelian Inheritance in Man

<http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM>

This catalog of human genes and genetic disorders is authored and edited by Victor McKusick and colleagues. Its 18,000 entries are short essays summarizing the literature related to the disease or gene. Mutations and phenotypes are often textually described.

*OMIM cannot be used commercially or redistributed without a license.*

Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). World Wide Web URL: <http://www.ncbi.nlm.nih.gov/omim/>

## dbGaP

<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap>

The NCBI database of genotype and phenotype provides results from genome-wide association studies and other analyses. dbGaP provides two levels of access, open and controlled. Often this allows everyone to see the overall associations, while authorization is needed for access to personal health information. The European Bioinformatics Institute (EBI) is constructing a similar resource, known as the European Genotype Archive.

*The open data in dbGaP are freely available and no data are protected by patents, while controlled data are restricted in use for privacy reasons.*

## SNPedia

<http://www.snpedia.com>

This is a fledgling wikipedia-style uncurated effort to describe the functional consequences of SNPs. Open to anyone to provide information, it now has 1,713 described SNPs. Pages have Google ads, though these are not expected to cover costs.

*SNPedia will use a Creative Commons Attribution-Share Alike 3.0 Unported License.*

## HGMD, Human Gene Mutation Database

<http://www.hgmd.cf.ac.uk/ac/index.php>

This database collates published gene mutations and variations thought to be responsible for human inherited disease; this database includes 73411 total mutation entries in over 2000 genes.

*To provide support for its maintenance, the complete database is available commercially, with limited usage of content that is more than 2 years old available without charge to academics.*

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeyasinghe S, Krawczak M, Cooper DN. 2003. The Human Gene Mutation Database (HGMD®): 2003 Update. *Hum Mutat* 21:577-581. doi:[10.1002/humu.10212](https://doi.org/10.1002/humu.10212)

## Mitomap

<http://www.mitomap.org/>

Mitomap provides annotations of the human mitochondrial sequence, including information about polymorphisms and mutations. Most information is taken from the literature, but it is also possible to contribute new information online. For many variations, it indicates associated disease and the level of confidence in the association.

*Several relevant tables are copyrighted by Elsevier and presumably not redistributable.*

Ruiz-Pesini E, Lott MT, Procaccio V, Poole J, Brandon MC, Mishmar D, Yi C, Kreuziger J, Baldi P, Wallace DC. 2007. An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Research* 35 (Database issue):D823-D828. doi:[10.1093/nar/gkl927](https://doi.org/10.1093/nar/gkl927)

## GeneTests

<http://www.genetests.org/>

GeneTests provides summary information on 600 laboratories offering more than a thousand genetic tests for clinical and research use. It also includes 400 expert-authored, peer-reviewed “GeneReviews” that discuss application of genetic tests for diagnosis, management, and counseling of patients.

*GeneTests provides aggregate data reports and grant permission to redistribute materials with attribution.*

GeneTests: Medical Genetics Information Resource (database online). Copyright, University of Washington, Seattle. 1993-2007.

Available at <http://www.genetests.org>.

Pagon RA. 2006. GeneTests: an online genetic information resource for health care providers. *J Med Libr Assoc* 94:343-8.

## **PharmGKB**

<http://www.pharmgkb.org/>

The pharmacogenetics and pharmacogenomics knowledgebase relates genomic variation to phenotypes associated with pharmaceutical activity. It offers integrated knowledge in terms of gene summaries, pathways and annotated literature. Currently, more than 22,400 samples have been genotyped with 2,581,832 variants reported and 11,353 polymorphisms reported. There are approximately 250,000 phenotype measurements.

*The data in PharmGKB are available to all research scientists, with individual data access requiring registration and subject to limitations to protect subject privacy.*

Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, Hewett M, Lin Z, Liu Y, Liu S, Oliver DE, Rubin DL, Shafa F, Stuart JM, Altman RB. 2001. Integrating genotype and phenotype information: an overview of the PharmGKB project. *The Pharmacogenomics Journal* 1:167-170.

## **Chromosomal Variation in Man Online**

<http://www.wiley.com/borgaonkar>

This is a compendium of 24,000 citations regarding chromosomal alterations, phenotypes, and abnormalities, collected by Digamber S. Borgaonkar. Citations are typically to individual cases, with a brief description of pathology observed.

*Web access is free and results of searches can be redistributed for research purposes.*

## **Locus Specific Databases Database**

<http://www.hgvs.org/dblist/glsdb.html>

This is an enumeration of over 600 locus specific databases.

*The site is freely available and copyright restrictions on redistribution are enforced.*

Horaitis O, Talbot Jr CC, Phommavanh M, Phillips KM, Cotton RGH. 2007. A database of locus-specific databases. *Nature Genetics* 39:425. doi:[10.1038/ng0407-425](https://doi.org/10.1038/ng0407-425)

## **SIFT**

<http://blocks.fhcrc.org/sift/SIFT.html>

SIFT is a program for predicting whether an amino acid substitution affects protein function, based on sequence homology and the physical properties of amino acids. It is one of several programs using sequence information for this purpose.

*The software is freely available and can be modified and redistributed with attribution.*

Ng PC, Henikoff S. 2002. Accounting for human polymorphisms predicted to affect protein function. *Genome Research* 12:436-446. doi:[10.1101/gr.212802](https://doi.org/10.1101/gr.212802)

## **SNPs3D**

<http://www.snps3d.org>

SNPs3D has predictions of phenotypic impact of SNPs based on sequence, structure, and cellular networks. It is one of several resources using protein structure for this purpose.

*Predictions on website can be browsed freely, but the software is not available online.*

Yue P, Melamud E, Moulton J. 2006. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* 7:166. doi:[10.1186/1471-2105-7-166](https://doi.org/10.1186/1471-2105-7-166)

## **MutaGeneSys**

<http://www.cs.columbia.edu/~jds1/MutaGeneSys/>

This software accepts SNP data of individuals and maps OMIM records to them. Using natural language processing, there are only 133 parsed unique participating SNPs found associated with OMIM disease records, but the data are enriched with marker correlation data to yield a total of 1300 population-specific correlations.

*The software can be freely downloaded.*

Stoyanovich J, Pe'er I. 2007. MutaGeneSys: making diagnostic predictions based on genome-wide genotype data in association studies. *Columbia University Technical Report, February 16, 2007.*

<http://mice.cs.columbia.edu/getTechreport.php?techreportID=448&format=pdf>

## **Craig Venter's genome**

<http://www.jcvi.org/research/huref/>

This is a launch point for the Venter Institute's sequence of J. Craig Venter, including the sequence, variants, traces, and the open-access assembler.

*The data are freely available, the manuscripts is open access with the Creative Commons Attribution License, and the Celera Assembler is open source.*

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol* 5: e254. doi:[10.1371/journal.pbio.0050254](https://doi.org/10.1371/journal.pbio.0050254)

## **James Watson's genome**

<http://jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/jwsequence/>

This is a browser for Jim Watson's genome variants overlaid on a reference genome, with some OMIM entries automatically associated using MutaGeneSys.

*Anyone can browse and download the sequence data and variants. The GMOD browser is open source.*

## **Navigenics and 23andMe**

<http://www.navigenics.com>

<http://www.23andMe.com>

23andMe and Navigenics are two of the most prominent new companies intending to provide personal genome interpretation.

## **Genome Commons**

<http://www.GenomeCommons.org>

This new website supports the genome commons and is a portal to associated resources.

*Contents copyrighted, with most content available under the Creative Commons Attribution License.*

Brenner SE. 2007. Common sense for our genomes. *Nature* 449:783-784. doi:[10.1038/449783a](https://doi.org/10.1038/449783a)