Genome Medicine

OPINION

The Human Genome Project: big science transforms biology and medicine

Leroy Hood* and Lee Rowen*

Abstract

The Human Genome Project has transformed biology through its integrated big science approach to deciphering a reference human genome sequence along with the complete sequences of key model organisms. The project exemplifies the power, necessity and success of large, integrated, cross-disciplinary efforts - so-called 'big science' - directed towards complex major objectives. In this article, we discuss the ways in which this ambitious endeavor led to the development of novel technologies and analytical tools, and how it brought the expertise of engineers, computer scientists and mathematicians together with biologists. It established an open approach to data sharing and open-source software, thereby making the data resulting from the project accessible to all. The genome sequences of microbes, plants and animals have revolutionized many fields of science, including microbiology, virology, infectious disease and plant biology. Moreover, deeper knowledge of human sequence variation has begun to alter the practice of medicine. The Human Genome Project has inspired subsequent large-scale data acquisition initiatives such as the International HapMap Project, 1000 Genomes, and The Cancer Genome Atlas, as well as the recently announced Human Brain Project and the emerging Human Proteome Project.

Origins of the human genome project

The Human Genome Project (HGP) has profoundly changed biology and is rapidly catalyzing a transformation of medicine [1-3]. The idea of the HGP was first publicly advocated by Renato Dulbecco in an article published in 1984, in which he argued that knowing the human genome sequence would facilitate an understanding of cancer [4]. In May 1985 a meeting focused entirely on the HGP was held, with Robert Sinsheimer, the Chancellor of the University of California, Santa Cruz (UCSC), assembling 12 experts to debate the merits of this potential project [5]. The meeting concluded that the project was technically possible, although very challenging. However, there was controversy as to whether it was a good idea, with six of those assembled declaring themselves for the project, six against (and those against felt very strongly). The naysayers argued that big science is bad science because it diverts resources from the 'real' small science (such as single investigator science); that the genome is mostly junk that would not be worth sequencing; that we were not ready to undertake such a complex project and should wait until the technology was adequate for the task; and that mapping and sequencing the genome was a routine and monotonous task that would not attract appropriate scientific talent. Throughout the early years of advocacy for the HGP (mid- to late 1980s) perhaps 80% of biologists were against it, as was the National Institutes of Health (NIH) [6]. The US Department of Energy (DOE) initially pushed for the HGP, partly using the argument that knowing the genome sequence would help us understand the radiation effects on the human genome resulting from exposure to atom bombs and other aspects of energy transmission [7]. This DOE advocacy was critical to stimulating the debate and ultimately the acceptance of the HGP. Curiously, there was more support from the US Congress than from most biologists. Those in Congress understood the appeal of international competitiveness in biology and medicine, the potential for industrial spin-offs and economic benefits, and the potential for more effective approaches to dealing with disease. A National Academy of Science committee report endorsed the project in 1988 [8] and the tide of opinion turned: in 1990, the program was initiated, with the finished sequence published in 2004 ahead of schedule and under budget [9].

* Correspondence: Leroy.Hood@systemsbiology.org; Lee.Rowen@systemsbiology.org
Institute for Systems Biology, 401 Terry Ave N., Seattle, WA 98109, USA

What did the human genome project entail?

This 3-billion-dollar, 15-year program evolved considerably as genomics technologies improved. Initially, the HGP set out to determine a human genetic map, then a physical map of the human genome [10], and finally the sequence map. Throughout, the HGP was instrumental in pushing the development of high-throughput technologies for preparing, mapping and sequencing DNA [11]. At the inception of the HGP in the early 1990s, there was optimism that the then-prevailing sequencing technology would be replaced. This technology, now called 'first-generation sequencing', relied on gel electrophoresis to create sequencing ladders, and radioactive- or fluorescent-based labeling strategies to perform base calling [12]. It was considered to be too cumbersome and low throughput for efficient genomic sequencing. As it turned out, the initial human genome reference sequence was deciphered using a 96-capillary (highly parallelized) version of first-generation technology. Alternative approaches such as multiplexing [13] and sequencing by hybridization [14] were attempted but not effectively scaled up. Meanwhile, thanks to the efforts of biotech companies, successive incremental improvements in the cost, throughput, speed and accuracy of first-generation automated fluorescent-based sequencing strategies were made throughout the duration of the HGP. Because biologists were clamoring for sequence data, the goal of obtaining a full-fledged physical map of the human genome was abandoned in the later stages of the HGP in favor of generating the sequence earlier than originally planned. This push was accelerated by Craig Venter's bold plan to create a company (Celera) for the purpose of using a whole-genome shotgun approach [15] to decipher the sequence instead of the piecemeal clone-by-clone approach using bacterial artificial chromosome (BAC) vectors that was being employed by the International Consortium. Venter's initiative prompted government funding agencies to endorse production of a clone-based draft sequence for each chromosome, with the finishing to come in a subsequent phase. These parallel efforts accelerated the timetable for producing a genome sequence of immense value to biologists [16,17].

As a key component of the HGP, it was wisely decided to sequence the smaller genomes of significant experimental model organisms such as yeast, a small flowering plant (*Arabidopsis thaliana*), worm and fruit fly before taking on the far more challenging human genome. The efforts of multiple centers were integrated to produce these reference genome sequences, fostering a culture of cooperation. There were originally 20 centers mapping and sequencing the human genome as part of an international consortium [18]; in the end five large centers (the Wellcome Trust Sanger Institute, the Broad Institute of MIT and Harvard, The Genome Institute of Washington University in St Louis, the Joint Genome Institute, and the Whole Genome Laboratory at Baylor College of Medicine) emerged from this effort, with these five centers continuing to provide genome sequence and technology development. The HGP also fostered the development of mathematical, computational and statistical tools for handling all the data it generated.

The HGP produced a curated and accurate reference sequence for each human chromosome, with only a small number of gaps, and excluding large heterochromatic regions [9]. In addition to providing a foundation for subsequent studies in human genomic variation, the reference sequence has proven essential for the development and subsequent widespread use of second-generation sequencing technologies, which began in the mid-2000s. Second-generation cyclic array sequencing platforms produce, in a single run, up to hundreds of millions of short reads (originally approximately 30 to 70 bases, now up to several hundred bases), which are typically mapped to a reference genome at highly redundant coverage [19]. A variety of cyclic array sequencing strategies (such as RNA-Seq, ChIP-Seq, bisulfite sequencing) have significantly advanced biological studies of transcription and gene regulation as well as genomics, progress for which the HGP paved the way.

Impact of the human genome project on biology and technology

First, the human genome sequence initiated the comprehensive discovery and cataloguing of a 'parts list' of most human genes [16,17], and by inference most human proteins, along with other important elements such as non-coding regulatory RNAs. Understanding a complex biological system requires knowing the parts, how they are connected, their dynamics and how all of these relate to function [20]. The parts list has been essential for the emergence of 'systems biology', which has transformed our approaches to biology and medicine [21,22].

As an example, the ENCODE (Encyclopedia Of DNA Elements) Project, launched by the NIH in 2003, aims to discover and understand the functional parts of the genome [23]. Using multiple approaches, many based on second-generation sequencing, the ENCODE Project Consortium has produced voluminous and valuable data related to the regulatory networks that govern the expression of genes [24]. Large datasets such as those produced by ENCODE raise challenging questions regarding genome functionality. How can a true biological signal be distinguished from the inevitable biological noise produced by large datasets [25,26]? To what extent is the functionality of individual genomic elements only observable (used) in specific contexts (for example, regulatory networks and mRNAs that are operative only during embryogenesis)? It is clear that much work remains to be done before the

functions of poorly annotated protein-coding genes will be deciphered, let alone those of the large regions of the non-coding portions of the genome that are transcribed. What is signal and what is noise is a critical question.

Second, the HGP also led to the emergence of proteomics, a discipline focused on identifying and quantifying the proteins present in discrete biological compartments, such as a cellular organelle, an organ or the blood. Proteins - whether they act as signaling devices, molecular machines or structural components - constitute the cell-specific functionality of the parts list of an organism's genome. The HGP has facilitated the use of a key analytical tool, mass spectrometry, by providing the reference sequences and therefore the predicted masses of all the tryptic peptides in the human proteome - an essential requirement for the analysis of mass-spectrometry-based proteomics [27]. This mass-spectrometry-based accessibility to proteomes has driven striking new applications such as targeted proteomics [28]. Proteomics requires extremely sophisticated computational techniques, examples of which are PeptideAtlas [29] and the Trans-Proteomic Pipeline [30].

Third, our understanding of evolution has been transformed. Since the completion of the HGP, over 4,000 finished or quality draft genome sequences have been produced, mostly from bacterial species but including 183 eukaryotes [31]. These genomes provide insights into how diverse organisms from microbes to human are connected on the genealogical tree of life - clearly demonstrating that all of the species that exist today descended from a single ancestor [32]. Questions of longstanding interest with implications for biology and medicine have become approachable. Where do new genes come from? What might be the role of stretches of sequence highly conserved across all metazoa? How much large-scale gene organization is conserved across species and what drives local and global genome reorganization? Which regions of the genome appear to be resistant (or particularly susceptible) to mutation or highly susceptible to recombination? How do regulatory networks evolve and alter patterns of gene expression [33]? The latter question is of particular interest now that the genomes of several primates and hominids have been or are being sequenced [34,35] in hopes of shedding light on the evolution of distinctively human characteristics. The sequence of the Neanderthal genome [36] has had fascinating implications for human evolution; namely, that a few percent of Neanderthal DNA and hence the encoded genes are intermixed in the human genome, suggesting that there was some interbreeding while the two species were diverging [36,37].

Fourth, the HGP drove the development of sophisticated computational and mathematical approaches to data and brought computer scientists, mathematicians, engineers and theoretical physicists together with biologists, fostering a more cross-disciplinary culture [1,21,38]. It is important to note that the HGP popularized the idea of making data available to the public immediately in user-friendly databases such as GenBank [39] and the UCSC Genome Browser [40]. Moreover, the HGP also promoted the idea of open-source software, in which the source code of programs is made available to and can be edited by those interested in extending their reach and improving them [41,42]. The open-source operating system of Linux and the community it has spawned have shown the power of this approach. Data accessibility is a critical concept for the culture and success of biology in the future because the 'democratization of data' is critical for attracting available talent to focus on the challenging problems of biological systems with their inherent complexity [43]. This will be even more critical in medicine, as scientists need access to the data cloud available from each individual human to mine for the predictive medicine of the future - an effort that could transform the health of our children and grandchildren [44].

Fifth, the HGP, as conceived and implemented, was the first example of 'big science' in biology, and it clearly demonstrated both the power and the necessity of this approach for dealing with its integrated biological and technological aims. The HGP was characterized by a clear set of ambitious goals and plans for achieving them; a limited number of funded investigators typically organized around centers or consortia; a commitment to public data/resource release; and a need for significant funding to support project infrastructure and new technology development. Big science and smaller-scope individual-investigator-oriented science are powerfully complementary, in that the former generates resources that are foundational for all researchers while the latter adds detailed experimental clarification of specific questions, and analytical depth and detail to the data produced by big science. There are many levels of complexity in biology and medicine; big science projects are essential to tackle this complexity in a comprehensive and integrative manner [45].

The HGP benefited biology and medicine by creating a sequence of the human genome; sequencing model organisms; developing high-throughput sequencing technologies; and examining the ethical and social issues implicit in such technologies. It was able to take advantage of economies of scale and the coordinated effort of an international consortium with a limited number of players, which rendered the endeavor vastly more efficient than would have been possible if the genome were sequenced on a gene-by-gene basis in small labs. It is also worth noting that one aspect that attracted governmental support to the HGP was its potential for economic benefits. The Battelle Institute published a report on the economic impact of the HGP [46]. For an initial investment of

approximately $3.5 billion, the return, according to the report, has been about $800 billion - a staggering return on investment.

Even today, as budgets tighten, there is a cry to withdraw support from big science and focus our resources on small science. This would be a drastic mistake. In the wake of the HGP there are further valuable biological resource-generating projects and analyses of biological complexity that require a big science approach, including the HapMap Project to catalogue human genetic variation [47,48], the ENCODE project, the Human Proteome Project (described below) and the European Commission's Human Brain Project, as well as another brain-mapping project recently announced by President Obama [49]. Similarly to the HGP, significant returns on investment will be possible for other big science projects that are now under consideration if they are done properly. It should be stressed that discretion must be employed in choosing big science projects that are fundamentally important. Clearly funding agencies should maintain a mixed portfolio of big and small science - and the two are synergistic [1,45].

Last, the HGP ignited the imaginations of unusually talented scientists - Jim Watson, Eric Lander, John Sulston, Bob Waterston and Sydney Brenner to mention only a few. So virtually every argument initially posed by the opponents of the HGP turned out to be wrong. The HGP is a wonderful example of a fundamental paradigm change in biology: initially fiercely resisted, it was ultimately far more transformational than expected by even the most optimistic of its proponents.

Impact of the human genome project on medicine

Since the conclusion of the HGP, several big science projects specifically geared towards a better understanding of human genetic variation and its connection to human health have been initiated. These include the HapMap Project aimed at identifying haplotype blocks of common single nucleotide polymorphisms (SNPs) in different human populations [47,48], and its successor, the 1000 Genomes project, an ongoing endeavor to catalogue common and rare single nucleotide and structural variation in multiple populations [50]. Data produced by both projects have supported smaller-scale clinical genome-wide association studies (GWAS), which correlate specific genetic variants with disease risk of varying statistical significance based on case–control comparisons. Since 2005, over 1,350 GWAS have been published [51]. Although GWAS analyses give hints as to where in the genome to look for disease-causing variants, the results can be difficult to interpret because the actual disease-causing variant might be rare, the sample size of the study might be too small, or the disease phenotype might not be well stratified. Moreover, most of the GWAS hits are outside of coding regions - and we do not have effective methods for easily determining whether these hits reflect the mis-functioning of regulatory elements. The question as to what fraction of the thousands of GWAS hits are signal and what fraction are noise is a concern. Pedigree-based whole-genome sequencing offers a powerful alternative approach to identifying potential disease-causing variants [52].

Five years ago, a mere handful of personal genomes had been fully sequenced (for example, [53,54]). Now there are thousands of exome and whole-genome sequences (soon to be tens of thousands, and eventually millions), which have been determined with the aim of identifying disease-causing variants and, more broadly, establishing well-founded correlations between sequence variation and specific phenotypes. For example, the International Cancer Genome Consortium [55] and The Cancer Genome Atlas [56] are undertaking large-scale genomic data collection and analyses for numerous cancer types (sequencing both the normal and cancer genome for each individual patient), with a commitment to making their resources available to the research community.

We predict that individual genome sequences will soon play a larger role in medical practice. In the ideal scenario, patients or consumers will use the information to improve their own healthcare by taking advantage of prevention or therapeutic strategies that are known to be appropriate for real or potential medical conditions suggested by their individual genome sequence. Physicians will need to educate themselves on how best to advise patients who bring consumer genetic data to their appointments, which may well be a common occurrence in a few years [57].

In fact, the application of systems approaches to disease has already begun to transform our understanding of human disease and the practice of healthcare and push us towards a medicine that is predictive, preventive, personalized and participatory: P4 medicine. A key assumption of P4 medicine is that in diseased tissues biological networks become perturbed - and change dynamically with the progression of the disease. Hence, knowing how the information encoded by disease-perturbed networks changes provides insights into disease mechanisms, new approaches to diagnosis and new strategies for therapeutics [58,59].

Let us provide some examples. First, pharmacogenomics has identified more than 70 genes for which specific variants cause humans to metabolize drugs ineffectively (too fast or too slow). Second, there are hundreds of 'actionable gene variants' - variants that cause disease but whose consequences can be avoided by available medical strategies with knowledge of their presence [60]. Third, in some cases, cancer-driving mutations in tumors, once identified, can be counteracted by treatments with currently available drugs [61]. And last, a systems approach to blood protein diagnostics has generated powerful new

diagnostic panels for human diseases such as hepatitis [62] and lung cancer [63].

These latter examples portend a revolution in blood diagnostics that will lead to early detection of disease, the ability to follow disease progression and responses to treatment, and the ability to stratify a disease type (for instance, breast cancer) into its different subtypes for proper impedance match against effective drugs [59]. We envision a time in the future when all patients will be surrounded by a virtual cloud of billions of data points, and when we will have the analytical tools to reduce this enormous data dimensionality to simple hypotheses to optimize wellness and minimize disease for each individual [58].

Impact of the human genome project on society

The HGP challenged biologists to consider the social implications of their research. Indeed, it devoted 5% of its budget to considering the social, ethical and legal aspects of acquiring and understanding the human genome sequence [64]. That process continues as different societal issues arise, such as genetic privacy, potential discrimination, justice in apportioning the benefits from genomic sequencing, human subject protections, genetic determinism (or not), identity politics, and the philosophical concept of what it means to be human beings who are intrinsically connected to the natural world.

Strikingly, we have learned from the HGP that there are no race-specific genes in humans [65-68]. Rather, an individual's genome reveals his or her ancestral lineage, which is a function of the migrations and interbreeding among population groups. We are one race and we honor our species' heritage when we treat each other accordingly, and address issues of concern to us all, such as human rights, education, job opportunities, climate change and global health.

What is to come?

There remain fundamental challenges for fully understanding the human genome. For example, as yet at least 5% of the human genome has not been successfully sequenced or assembled for technical reasons that relate to eukaryotic islands being embedded in heterochromatic repeats, copy number variations, and unusually high or low GC content [69]. The question of what information these regions contain is a fascinating one. In addition, there are highly conserved regions of the human genome whose functions have not yet been identified; presumably they are regulatory, but why they should be strongly conserved over a half a billion years of evolution remains a mystery.

There will continue to be advances in genome analysis. Developing improved analytical techniques to identify biological information in genomes and decipher what this information relates to functionally and evolutionarily will be important. Developing the ability to rapidly analyze complete human genomes with regard to actionable gene variants is essential. It is also essential to develop software that can accurately fold genome-predicted proteins into three dimensions, so that their functions can be predicted from structural homologies. Likewise, it will be fascinating to determine whether we can make predictions about the structures of biological networks directly from the information of their cognate genomes. Indeed, the idea that we can decipher the 'logic of life' of an organism solely from its genome sequence is intriguing. While we have become relatively proficient at determining static and stable genome sequences, we are still learning how to measure and interpret the dynamic effects of the genome: gene expression and regulation, as well as the dynamics and functioning of non-coding RNAs, metabolites, proteins and other products of genetically encoded information.

The HGP, with its focus on developing the technology to enumerate a parts list, was critical for launching systems biology, with its concomitant focus on high-throughput 'omics' data generation and the idea of 'big data' in biology [21,38]. The practice of systems biology begins with a complete parts list of the information elements of living organisms (for example, genes, RNAs, proteins and metabolites). The goals of systems biology are comprehensive yet open ended because, as seen with the HGP, the field is experiencing an infusion of talented scientists applying multidisciplinary approaches to a variety of problems. A core feature of systems biology, as we see it, is to integrate many different types of biological information to create the 'network of networks' - recognizing that networks operate at the genomic, the molecular, the cellular, the organ, and the social network levels, and that these are integrated in the individual organism in a seamless manner [58]. Integrating these data allows the creation of models that are predictive and actionable for particular types of organisms and individual patients. These goals require developing new types of high-throughput omic technologies and ever increasingly powerful analytical tools.

The HGP infused a technological capacity into biology that has resulted in enormous increases in the range of research, for both big and small science. Experiments that were inconceivable 20 years ago are now routine, thanks to the proliferation of academic and commercial wet lab and bioinformatics resources geared towards facilitating research. In particular, rapid increases in throughput and accuracy of the massively parallel second-generation sequencing platforms with their correlated decreases in cost of sequencing have resulted in a great wealth of accessible genomic and transcriptional sequence data for myriad microbial, plant and animal genomes. These data in turn have enabled large- and

small-scale functional studies that catalyze and enhance further research when the results are provided in publicly accessible databases [70].

One descendant of the HGP is the Human Proteome Project, which is beginning to gather momentum, although it is still poorly funded. This exciting endeavor has the potential to be enormously beneficial to biology [71-73]. The Human Proteome Project aims to create assays for all human and model organism proteins, including the myriad protein isoforms produced from the RNA splicing and editing of protein-coding genes, chemical modifications of mature proteins, and protein processing. The project also aims to pioneer technologies that will achieve several goals: enable single-cell proteomics; create microfluidic platforms for thousands of protein enzyme-linked immunosorbent assays (ELISAs) for rapid and quantitative analyses of, for example, a fraction of a droplet of blood; develop protein-capture agents that are small, stable, easy to produce and can be targeted to specific protein epitopes and hence avoid extensive cross-reactivity; and develop the software that will enable the ordinary biologist to analyze the massive amounts of proteomics data that are beginning to emerge from human and other organisms.

Newer generations of DNA sequencing platforms will be introduced that will transform how we gather genome information. Third-generation sequencing [74] will employ nanopores or nanochannels, utilize electronic signals, and sequence single DNA molecules for read lengths of 10,000 to 100,000 bases. Third-generation sequencing will solve many current problems with human genome sequences. First, contemporary short-read sequencing approaches make it impossible to assemble human genome sequences *de novo*; hence, they are usually compared against a prototype reference sequence that is itself not fully accurate, especially with respect to variations other than SNPs. This makes it extremely difficult to precisely identify the insertion-deletion and structural variations in the human genome, both for our species as a whole and for any single individual. The long reads of third-generation sequencing will allow for the *de novo* assembly of human (and other) genomes, and hence delineate all of the individually unique variability: nucleotide substitutions, indels, and structural variations. Second, we do not have global techniques for identifying the 16 different chemical modifications of human DNA (epigenetic marks, reviewed in [75]). It is increasingly clear that these epigenetic modifications play important roles in gene expression [76]. Thus, single-molecule analyses should be able to identify all the epigenetic marks on DNA. Third, single-molecule sequencing will facilitate the full-length sequencing of RNAs; thus, for example, enhancing interpretation of the transcriptome by enabling the identification of RNA editing, alternative splice forms with a given transcript, and different start and termination sites. Last, it is exciting to contemplate that the ability to parallelize this process (for example, by generating millions of nanopores that can be used simultaneously) could enable the sequencing of a human genome in 15 minutes or less [77]. The high-throughput nature of this sequencing may eventually lead to human genome costs of $100 or under. The interesting question is how long it will take to make third-generation sequencing a mature technology.

The HGP has thus opened many avenues in biology, medicine, technology and computation that we are just beginning to explore.

Abbreviations

BAC: Bacterial artificial chromosome; DOE: Department of Energy; ELISA: Enzyme-linked immunosorbent assay; GWAS: Genome-wide association studies; HGP: Human Genome Project; NIH: National Institutes of Health; SNP: Single nucleotide polymorphism; UCSC: University of California, Santa Cruz.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors gratefully acknowledge support from the Luxembourg Centre for Systems Biomedicine and the University of Luxembourg; from the NIH, through award 2P50GM076547-06A; and the US Department of Defense (DOD), through award W911SR-09-C-0062. LH receives support from NIH P01 NS041997; 1U54CA151819-01; and DOD awards W911NF-10-2-0111 and W81XWH-09-1-0107.

Published: 13 September 2013

References

1. Hood L: **Acceptance remarks for Fritz J. and Delores H. Russ Prize.** *The Bridge* 2011, **41:**46–49.
2. Collins FS, McKusick VA: **Implications of the Human Genome Project for medical science.** *JAMA* 2001, **285:**540–544.
3. Green ED, Guyer MS, National Human Genome Research Institute: **Charting a course for genomic medicine from base to bedside.** *Nature* 2011, **470:**204–213.
4. Dulbecco R: **A turning point in cancer research: sequencing the human genome.** *Science* 1984, **231:**1055–1056.
5. Sinsheimer RL: **The Santa Cruz workshop - May 1985.** *Genomics* 1989, **5:**954–956.
6. Cooke-Degan RM: *The Gene Wars: Science, Politics and the Human Genome.* New York: WW Norton; 1994.
7. *Report on the Human Genome Initiative for the Office of Health and Environmental Research.* http://www.ornl.gov/sci/techresources/Human_Genome/project/herac2.shtml.
8. National Academy of Science: *Report of the Committee on Mapping and Sequencing the Human Genome.* Washington DC: National Academy Press; 1988.
9. Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431:**931–945.
10. *Understanding Our Genetic Inheritance. The United States Human Genome Project, The First Five Years: Fiscal Years. 1991–1995.* http://www.genome.gov/10001477.
11. Collins FS, Galas D: **A new five-year plan for the U.S. Human Genome Program.** *Science* 1993, **262:**43–46.
12. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH, Hood LE: **Fluorescence detection in automated DNA sequence analysis.** *Nature* 1986, **321:**674–679.
13. Church G, Kieffer-Higgins S: **Multiplex DNA sequencing.** *Science* 1988, **240:**185–188.

14. Strezoska Z, Paunesku T, Radosavljević D, Labat I, Drmanac R, Crkvenjakov R: **DNA sequencing by hybridization: 100 bases read by a non-gel-based method.** *Proc Natl Acad Sci USA* 1991, **88:**10089–10093.
15. Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M: **Shotgun sequencing of the human genome.** *Science* 1998, **280:**1540–1542.
16. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860–921.
17. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, *et al*: **The sequence of the human genome.** *Science* 2001, **291:**1304–1351.
18. *International Human Genome Sequencing Consortium.* http://www.genome.gov/11006939.
19. Shendure J, Aiden ER: **The expanding scope of DNA sequencing.** *Nat Biotechnol* 2012, **30:**1084–1094.
20. Hood L: **A personal journey of discovery: developing technology and changing biology.** *Annu Rev Anal Chem* 2008, **1:**1–43.
21. Committee on a New Biology for the 21st Century: *A New Biology for the 21st Century.* Washington DC: The National Academies Press; 2009.
22. Ideker T, Galitski T, Hood L: **A new approach to decoding life: systems biology.** *Annu Rev Genomics Hum Genet* 2001, **2:**343–372.
23. *Encyclopedia of DNA Elements.* http://encodeproject.org/ENCODE/.
24. ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57–74.
25. Editorial: **Form and function.** *Nature* 2013, **495:**141–142.
26. ENCODE Project Consortium: **A user's guide to the Encyclopedia of DNA Elements (ENCODE).** *PLoS Biol* 2011, **9:**e1001046.
27. Aebersold R, Mann M: **Mass spectrometry-based proteomics.** *Nature* 2003, **422:**198–207.
28. Picotti P, Aebersold R: **Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions.** *Nat Methods* 2012, **9:**555–566.
29. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R: **The PeptideAtlas Project.** *Nucleic Acids Res* 2006, **34:**D655–D658.
30. Deutsch ED, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, Sun Z, Nilsson E, Pratt B, Prazen B, Eng JK, Martin DB, Nesvizhskii A, Aebersold R: **A guided tour of the Trans-Proteomic Pipeline.** *Proteomics* 2010, **10:**1150–1159.
31. *Genomes Online Database: complete genome projects.* http://www.genomesonline.org/cgi-bin/GOLD/index.cgi?page_requested=Complete+Genome+Projects.
32. Theobald DL: **A formal test of the theory of universal common ancestry.** *Nature* 2010, **465:**219–222.
33. Wolfe KE, Li W-H: **Molecular evolution meets the genomics evolution.** *Nat Genet* 2003, **Suppl 33:**255–265.
34. Marques-Bonet T, Ryder OA, Eichler EE: **Sequencing primate genomes: what have we learned?** *Annu Rev Genomics Hum Genet* 2009, **10:**355–386.
35. Noonan JP: **Neanderthal genomics and the evolution of modern human.** *Genome Res* 2010, **20:**547–553.
36. Stoneking M, Krause J: **Learning about human population history from ancient and modern genomes.** *Nat Rev Genet* 2011, **12:**603–614.
37. Sankararaman S, Patterson N, Li H, Paabo S, Reich D: **The date of interbreeding between Neanderthals and Modern Humans.** *PLoS Genet* 2012, **8:**e1002947.
38. Schatz MC: **Computational thinking in the era of big data biology.** *Genome Biol* 2012, **13:**177.
39. Mizrachi I: **GenBank: the Nucleotide Sequence Database.** In *The NCBI Handbook.* Edited by McEntyre J, Ostell J. Bethesda: National Center for Biotechnology Information; 2002.
40. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996–1006.
41. *SourceForge.* http://sourceforge.net/.
42. *Bioconductor: open source software for bioinformatics.* http://www.bioconductor.org/.
43. Field D, Sansone S-A, Collina A, Booth T, Dukes P, Gregurick SK, Kennedy K, Kolar P, Kolker E, Maxon M, Millard S, Mugabushaka M, Perrin N, Remacle JE, Remington K, Rocca-Serra P, Taylor CF, Thorley M, Tiwari B, Wilbanks J: **Omics data sharing.** *Science* 2009, **326:**234–236.
44. Knoppers BM, Harris JR, Tasse AM, Budin-Ljosne I, Kaye J, Deschenes M, Zawati M: **Towards a data-sharing Code of Conduct for international genomic research.** *Genome Med* 2011, **3:**46.
45. Hood L: **Biological complexity under attack: a personal view of systems biology and the coming of "big science".** *Genet Eng Biotechnol News* 2011, **31:**17.
46. Tripp S, Grueber M: *Economic Impact of the Human Genome Project.* Columbus: Battelle Memorial Institute; 2011.
47. International HapMap Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437:**1299–1320.
48. The International HapMap3 Consortium: **Integrating common and rare genetic variation in diverse human populations.** *Nature* 2010, **467:**52–58.
49. Abbott A: **Neuroscience: solving the brain.** *Nature* 2013, **499:**272–274.
50. The 1000 Genomes Project Consortium: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491:**56–65.
51. *A Catalog of Published Genome-wide Association Studies.* http://www.genome.gov/gwastudies/.
52. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ: **Analysis of genetic inheritance in a family quartet by whole-genome sequencing.** *Science* 2010, **328:**636–639.
53. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, *et al*: **The diploid genome sequence of an individual human.** *PLoS Biol* 2007, **5:**e254.
54. Wheeler DA, Srinivasian M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song X, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452:**872–876.
55. *International Cancer Genome Consortium.* http://icgc.org/.
56. *The Cancer Genome Atlas.* http://cancergenome.nih.gov/.
57. Pandey A: **Preparing for the 21st century patient.** *JAMA* 2013, **309:**1471–1472.
58. Hood L, Flores M: **A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory.** *Nat Biotechnol* 2012, **29:**613–624.
59. Price ND, Edelman LB, Lee I, Yoo H, Hwang D, Carlson G, Galas DJ, Heath JR, Hood L: **Systems biology and the emergence of systems medicine**. In *Genomic and Personalized Medicine: From Principles to Practice. Volume 1.* Edited by Ginsburg G, Willard H. Philadelphia: Elsevier; 2009:131–141.
60. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, McGuire A, Nussbaum RL, O'Daniel JM, Ormond KE, Rehm HL, Watson MS, Williams MS, Biesecker LG: *ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing.* Bethesda: American College of Medical Genetics and Genomics; 2013.
61. Meyerson M, Gabriel S, Getz G: **Advances in understanding cancer genomes through second-generation sequencing.** *Nat Rev Genet* 2010, **11:**685–696.
62. Qin S, Zhou Y, Lok AS, Tsodikov A, Yan X, Gray L, Yuan M, Moritz RL, Galas D, Omenn GS, Hood L: **SRM targeted proteomics in search for biomarkers of HCV-induced progression of fibrosis to cirrhosis in HALT-C patients.** *Proteomics* 2012, **12:**1244–1252.
63. Li X-J, Hayward C, Fong P-Y, Dominguez M, Hunsucker SW, Lee LW, McClean M, Law S, Butler H, Schirm M, Gingras O, Lamontague J, Allard R, Chelsky D, Price ND, Lam S, Massion PP, Pass H, Rom WN, Vachani A, Fang KC, Hood L, Kearney P: **A blood-based proteomic classifier for the molecular characterization of pulmonary nodules.** *Sci Transl Med*, in press.
64. Knoppers BM, Thorogood A, Chadwick R: **The Human Genome Organisation: towards next-generation ethics.** *Genome Med* 2013, **5:**38.
65. Hood L: **Who we are: the book of life. Commencement Address.** In *Whitman College Magazine* 2002:4–7.
66. Foster MW, Sharp RR: **Beyond race: towards a whole-genome perspective on human populations and genetic variation.** *Nat Rev Genet* 2004, **5:**790–796.
67. Royal CDM, Dunston GM: **Changing the paradigm from 'race' to human genetic variation.** *Nat Genet* 2004, **36:**S5–S7.
68. Witherspoon DJ, Wooding S, Rogers AR, Marchani EE, Watkins WS, Batzer MA, Jorde LB: **Genetic similarities within and between populations.** *Genetics* 2007, **176:**351–359.

69. Genovese G, Handsaker RE, Li H, Altemose N, Lindgren AM, Chambert K, Pasaniuk B, Price AL, Reich D, Morton CC, Pollak MR, Wilson JG, McCarroll SA: **Using population admixture to help complete maps of the human genome.** *Nat Genet* 2013, **45:**406–414.
70. Fernandez-Suarez XM, Galperin MY: **The, *Nucleic Acids Research* Database Issue and the online Molecular Biology Database Collection.** *Nucleic Acids Res* 2013, **2013:**D1–D7.
71. *Human Proteome Project.* http://www.hupo.org/research/hpp/.
72. Hood LE, Omenn GS, Moritz RL, Aebersold R, Yamamoto KR, Amos M, Hunter-Cevera J, Locascio L, Workshop Participants: **New and improved proteomics technologies for understanding complex biological systems: addressing a grand challenge in the life sciences.** *Proteomics* 2012, **12:**2773–2783.
73. Editorial: **The call of the human proteome.** *Nat Methods* 2010, **7:**661.
74. Schadt E, Turner S, Kasarskis A: **A window into third-generation sequencing.** *Hum Mol Genet* 2010, **19:**R227–R240.
75. Kim JK, Samaranayake M, Pradhan S: **Epigenetic mechanisms in mammals.** *Cell Mol Life Sci* 2009, **66:**596–612.
76. Hon G, Ren B, Wang W: **ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome.** *PLoS Comput Biol* 2008, **4:**e1000201.
77. Hayden EC: **Nanopore genome sequencer makes its debut.** *Nature News* 2012. doi:10.1038/nature.2012.10051.

doi:10.1186/gm483
Cite this article as: Hood and Rowen: **The Human Genome Project: big science transforms biology and medicine.** *Genome Medicine* 2013 **5**:79.