Lessons from the Human Genome Project that may be applicable to a Human Toxicology Project

Christopher P. Austin. M.D.
Director, NIH Chemical Genomics Center
Senior Advisor to the NHGRI Director for Translational Research
National Institutes of Health

Human Toxicology Project Consortium
Washington, D.C.
November 9, 2010
The Human Genome Project: Lessons from Large-Scale Biology

Francis S. Collins,¹* Michael Morgan,² Aristides Patrinos³

The Human Genome Project has been the first major foray of the biological and medical research communities into “big science.” In this Viewpoint, we present some of our experiences in organizing and managing such a complicated, publicly funded, international effort. We believe that many of the lessons we learned will be applicable to future large-scale projects in biology.

“It is essentially immoral not to get it [the human genome sequence] done as fast as possible.”

James D. Watson (I)

We each joined the ranks of HGP management during the challenging period of the early to mid-1990s, with Francis Collins assuming the lead role at the NIH in 1993, Michael Morgan at The Wellcome Trust in 1992, and Aristides Patrinos at the DOE in 1995. The next several years were turbulent, as we learned “on the job,” made lots of mistakes, and experienced more than a few moments of great anxiety that the whole enterprise might fail; but ultimately, we watched the creativity, talent, and dedication of
Points from 2003 HGP Lessons Paper

• Build the best teams
• Process must be science-driven
• Meet managerial challenges
• International participation important
• Explicit milestones and quality assessment are valuable
• Technology matters
• Rapid prepublication data release demonstrates value to community
• Social consequences should be included as part of project (ELSI program)

From Collins et al., Science 300:286, 11 April 2003
HGP Stats

- HGP officially launched 1990
- Completion projected 2005, finished 2 years early
- Projected cost $3B; actual cost $2.7B
- 20 sequencing centers from 6 countries: China, France, Germany, Great Britain, Japan, U.S. Five institutions generated majority of sequence:
  - Whitehead Institute/MIT Center for Genome Research
  - Washington University School of Medicine
  - Wellcome Trust Sanger Institute
  - DOE's Joint Genome Institute
  - Baylor College of Medicine
# Importance of staging

<table>
<thead>
<tr>
<th>Area</th>
<th>Goal</th>
<th>Achieved</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic map</td>
<td>2- to 5-cM resolution map (600 to 1,500 markers)</td>
<td>1-cM resolution map (3,000 markers)</td>
<td>September 1994</td>
</tr>
<tr>
<td>Physical map</td>
<td>30,000 sequence-tagged sites (STTs)</td>
<td>52,000 STTs</td>
<td>October 1998</td>
</tr>
<tr>
<td>DNA sequence</td>
<td>95% of gene-containing part of human sequence finished to 99.99% accuracy</td>
<td>&gt;98% of gene-containing part of human sequence finished to 99.99% accuracy</td>
<td>April 2003</td>
</tr>
<tr>
<td>Capacity and cost of finished sequence</td>
<td>Sequence 500 Mb/year at &lt;$0.25 per finished base</td>
<td>Sequence &gt;1,400 Mb/year at &lt;$0.09 per finished base</td>
<td>November 2002</td>
</tr>
<tr>
<td>Human sequence variation</td>
<td>100,000 mapped human SNPs</td>
<td>3.7 million mapped human SNPs</td>
<td>February 2003</td>
</tr>
<tr>
<td>Gene identification</td>
<td>Full-length human cDNAs</td>
<td>15,000 full-length human cDNAs</td>
<td>March 2003</td>
</tr>
<tr>
<td>Model organisms</td>
<td>Complete sequences of <em>E. coli</em>, <em>S. cerevisiae</em>, <em>C. elegans</em>, <em>D. melanogaster</em></td>
<td>Finished sequences of <em>E. coli</em>, <em>S. cerevisiae</em>, <em>C. elegans</em>, <em>D. melanogaster</em>, plus whole-genome drafts of several others, including <em>C. briggsae</em>, <em>D. pseudoobscura</em>, mouse, and rat</td>
<td>April 2003</td>
</tr>
<tr>
<td>Functional analysis</td>
<td>Develop genomic-scale technologies</td>
<td>High-throughput oligonucleotide synthesis</td>
<td>1994</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DNA microarrays</td>
<td>1996</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Normalized and subtracted cDNA libraries</td>
<td>1996</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eukaryotic, whole-genome knockouts (yeast)</td>
<td>1999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scale-up of two-hybrid mapping</td>
<td>2002</td>
</tr>
</tbody>
</table>

*From Collins et al., Science 300:286, 11 April 2003*
Stages in Deciphering the Genome

2000: First Draft

2001: Working Draft

April 15, 2003: “Finished” reference sequence

2002-2007: Defining sequence variation in populations

2010 - : Defining individual genomes for medical purposes

2007 - 2010: Defining sequence variation in a few individuals
Importance of technology

Fig. 1. (A) Decrease in sequencing costs, 1990–2005. (B) Increase in DNA sequence in GenBank, 1990–2005.

From Collins et al., Science 300:286, 11 April 2003
Importance of actively managing scientific and public expectations and perceptions

From Collins et al., Science 300:286, 11 April 2003

Table 2. Comparison of prices of large government projects circa 1990 with their projected useful life-span.

<table>
<thead>
<tr>
<th>Proposed project</th>
<th>Projected cost ($ billion)</th>
<th>Target completion date</th>
<th>Estimated life-span (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space Station Freedom</td>
<td>30.0</td>
<td>1999</td>
<td>30</td>
</tr>
<tr>
<td>Earth Observing System</td>
<td>17.0</td>
<td>2000</td>
<td>15</td>
</tr>
<tr>
<td>Superconducting Super Collider</td>
<td>11.0</td>
<td>1999</td>
<td>30</td>
</tr>
<tr>
<td>Human Genome Project</td>
<td>3.0</td>
<td>2005</td>
<td>Perpetual</td>
</tr>
<tr>
<td>Hubble Space Telescope</td>
<td>1.5</td>
<td>1990</td>
<td>15 to 20</td>
</tr>
</tbody>
</table>
A ‘community resource project’ is a research project specifically devised and implemented to create a set of data, reagents or other material whose primary utility will be as a resource for the broad scientific community. Recent examples of community resource projects include the International Human Genome Sequencing Consortium, the Mouse Genome Sequencing Consortium, the Mammalian Gene Collection, the SNP Consortium, and the International HapMap Project.

The products of community resource projects have, over the past several years, become increasingly important as drivers of progress in biomedical research. The scientific community will best be served if the results of community resource projects are made immediately available for free and unrestricted use by the scientific community to engage in the full range of opportunities for creative science. At the same time, it is crucial that the scientific community recognizes and respects the important contribution made by the scientists who carry out community resource projects.
A more recent lesson: Publish a project paper

The Knockout Mouse Project

Mouse knockout technology provides a powerful means of elucidating gene function in vivo, and a publicly available genome-wide collection of mouse knockouts would be significantly enabling for biomedical discovery. To date, published knockout lines exist for only about 1% of mouse genes. Furthermore, many of these are limited in utility because they have not been made or phenotyped in standardized ways, and many are not freely available to researchers. It is time to harness new technologies and efficiencies of production to mount a high-throughput international effort to produce and phenotype knockouts for all mouse genes, and place these resources into the public domain.

The ENCODE (ENCyclopedia Of DNA Elements) Project

The ENCODE (ENCyclopedia Of DNA Elements) Project aims to identify all functional elements in the human genome sequence. The pilot phase of the Project is focused on a specified 30 megabases (~1%) of the human genome sequence and is organized as an international consortium of computational and laboratory-based scientists working to develop and apply high-throughput approaches for detecting all sequence elements that confer biological function. The results of this pilot phase will guide future efforts to analyze the entire human genome.

The International HapMap Project

The goal of the International HapMap Project is to determine the common patterns of DNA sequence variation in the human genome and to make this information freely available in the public domain. An international consortium is developing a map of these patterns across the genome by determining the genotypes of one million or more sequence variants, their frequencies and the degree of association between them, in DNA samples from populations with ancestry from parts of Africa, Asia and Europe. The HapMap will allow the discovery of sequence variants that affect common disease, will facilitate development of diagnostic tools, and will enhance our ability to choose targets for therapeutic intervention.
Advances in molecular genetics made over the past two decades are already having a major impact on medical research and clinical care. The ability to clone and analyze individual genes and to deduce the amino acid sequences of encoded proteins has greatly increased our understanding of genetic disorders, the immune system, endocrine abnormalities, coronary artery disease, infectious diseases, and cancer. A few proteins produced on a commercial scale by recombinant DNA methods are available for therapeutic use or in clinical trials, and many more are in earlier developmental stages. Recent progress in determining the genetic basis for such neurological and behavioral disorders as Huntington’s disease (Gusella et al., 1983), Alzheimer’s disease (St George-Hyslop et al., 1987), and manic-depressive illness (Egeland et al., 1987) promises new insights into these common and serious conditions. Higher resolution maps of the human genome will accelerate progress in understanding disease pathogenesis and in developing new approaches to diagnosis, treatment, and prevention in many areas of medicine. In Chapter 3 the potential medical impact of a detailed human genomic map is discussed further.
TOXICITY TESTING IN THE 21ST CENTURY: A VISION AND STRATEGY

POLICYFORUM

TOXICOLGY

Transforming Environmental Health Protection

Francis S. Collins, M.D., George M. Gray, Ph.D., John R. Bucher

In 2005, the U.S. Environmental Protection Agency (EPA), with support from the U.S. National Toxicology Program (NTP), funded a project at the National Research Council (NRC) to develop a long-range vision for toxicity testing and a strategic plan for implementing that vision. Both agencies wanted future toxicity testing and assessment paradigms to meet ever-evolving regulatory needs. Challenges include the large numbers of substances that need to be tested and how to incorporate recent advances in molecular toxicology, computational sciences, and information technology to rely increasingly on human as opposed to animal data, and to offer increased efficiency in design and costs (2-5). In response, the NRC Committee on Toxicity Testing and Assessment of Environmental Agents produced two reports that reviewed current toxicity testing, identified key issues, and developed a vision and implementation strategy to create a major shift in the assessment of chemical hazard and risk (6, 7). Although the NRC reports have laid out a solid theoretical rationale, comprehensive and rigorously gathered data (and comparisons with historical animal data) will determine whether the hypothesized improvements will be realized in practice. For this project, NTP, EPA, and the National Institutes of Health Chemical Genomics Center (NCGC) (organizations with expertise in experimental toxicology, computational toxicology, and high-throughput technologies, respectively) have established a collaborative research program.

EPA, NCDC, and NTP Joint Activities

In 2004, the NTP released its vision and roadmap for the 21st century (7), which established initiatives to integrate high-throughput screening (HTS) and other automated screening assays into its testing program. In 2006, the EPA established the National Center for Computational Toxicology (NCCT). Through these initiatives, NTP, EPA, and the NCGC are promoting the evolution of toxicology from a predominantly observational science at the level of disease-specific models in vivo to a predominantly predictive science focused on broad inclusion of target-specific mechanisms-based, biological observations in vivo (1, 4) (see figure, below).

Toxicity pathways. In vivo and in vitro tools are being used to identify cellular responses after chemical exposure expected to result in adverse health effects (1). HTS methods are a primary means of discovery for drug development, and screening of >100,000 compounds per day is routine (8). However, drug discovery HTS methods traditionally test compounds at one concentration, usually between 2 and 10 μM, and tolerate high false-negative rates. In contrast, in the EPA, NCDC, and NTP combined effort, all compounds are tested at as many as 15 concentrations, generally ranging from ~5 nM to ~100 μM, to generate a concentration-response curve (9). This approach is highly reproducible, produces significantly lower false-positive and false-negative rates than the traditional HTS methods (9), and facilitates multiassay comparisons. Finally, in silico platforms have been built to compare results among HTS screens; this is being expanded to allow comparisons with historical toxicologic NTP and EPA data (http://ncgc.nih.gov/pubs/roehl.pdf). HTS data collected by EPA and NTP, as well as by the NCGC and other Molecular Libraries Initiative centers (http://mlni.nih.gov), are being made publicly available through Web-based databases [e.g., PubChem (http://pubchem.ncbi.nlm.nih.gov)]. In addition,
Prospective community engagement is important

Integrating ethics and science in the International HapMap Project

The International HapMap Consortium*

Community Engagement/Public Consultation and Sample Collection Groups

This risk of group stigmatization is inherent in any study of samples from identified populations. Nevertheless, the limitations and ambiguities of population identifiers must continually be emphasized. For example, the individuals sampled from the residential community at Beijing Normal University do not represent all people in China, where there are 56 officially recognized ethnicities. Nor do the people sampled in Ibadan, Nigeria, represent all Africans or even all Yoruba people. Such limitations will be noted explicitly in Project publications that report the study’s findings, and researchers who do future studies with these samples or with Project data will also need to be aware of these complexities when designing and reporting their studies. Although there are differences among populations in the frequencies of some genetic variants, it is important that the findings of the HapMap Project not be over-simplified to perpetuate social and historical stereotypes.
Consider using a Time-Risk Matrix for planning

<table>
<thead>
<tr>
<th>Time-Risk (Difficulty) Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increasing Level of Difficulty</td>
</tr>
<tr>
<td>1-3 years</td>
</tr>
</tbody>
</table>

Time
The three important parts of a large project:
Example of the Molecular Libraries Initiative

- **Technology Development**
  - Chemical Diversity
  - Assay Development
  - Instrumentation
  - Predictive ADMET

- **Data Production**
  - Molecular Libraries Screening Centers Network (MLSCN)

- **Data Analysis/Dissemination**
  - Cheminformatics Research Centers
  - Compound Repository (MLSMR)

- **PubChem**

---

**Example of the Molecular Libraries Initiative**

**Molecular Libraries Screening Centers Network (MLSCN)**

**Compound Repository (MLSMR)**

**Cheminformatics Research Centers**

**PubChem**

---
Large-scale biological science projects since the HGP

- HapMap
- ENCODE
- Knockout Mouse Project
- Molecular Libraries Initiative
- Tox21
- Cancer Genome Anatomy Project
- Human Microbiome Project
- 1000 Genomes Project
- GTEx
- TRND
NIH Therapeutics for Rare and Neglected Diseases (TRND) Program

*Creating a Drug Development Pipeline for RND at NIH*

- Congressionally-mandated effort to speed development of new drugs for rare and neglected diseases
- Administration and governance at NIH
  - Governance/oversight by Office of Rare Diseases Research
  - Administered by NHGRI
- Operations: collaboration between intramural and extramural labs with appropriate expertise
- Projects will:
  - Enter TRND at a variety of stages of development
  - Be taken to phase needed for external organization to adopt for clinical development
The long pathway to drug development

Basic Research

NCGC, Molecular Libraries Initiative

Indefinite

Identify Target

3 yrs

Identify chemical starting point for drug

4 yrs

Make many chemical modifications to give drug beneficial effect without side effects

NIH RAID

IND

Ph I

Ph II

Ph III

1 yr

2 yrs

~3 yrs

1 yr

Pharmas, Biotechs
NIH Clinical Center, Academic Clinical Centers

FDA

“Valley of Death”

Patient Use

Review
The long pathway to drug development

- **Basic Research**
  - Indefinite
  - Identify Target

- **NCGC, Molecular Libraries Initiative**
  - 3 yrs
  - Identify chemical starting point for drug
  - Make many chemical modifications to give drug beneficial effect without side effects

- **NIH RAID**
  - 4 yrs
  - IND

- **Pharmas, Biotechs NIHC Clinical Center, Academic Clinical Centers**
  - ~3 yrs
  - Ph I
  - Ph II
  - Ph III
  - Review

- **FDA**
  - 1 yr
  - Patient Use
Toxicity is the most common reason for attrition in drug development.
TRND Science

- Distinguishing features
  - Disease agnostic, take advantage of cross-cutting mechanisms
    - “Diseaseome” approach
  - *Science* of preclinical drug development
    - Reasons for successes and failures will be investigated and published
  - Technology/paradigm development
    - 20% of effort, toward improving success rates
    - Focused on toxicology and efficacy
  - Large-scale systematic repurposing

- Project-specific activities
  - Medicinal chemistry
  - Efficacy, pharmacology, ADME, toxicology, PK/PD
  - Compound scale-up, formulation
  - First in human clinical trials as needed for project
Lessons perhaps applicable to HTP - 1

• Large scale biology projects are now accepted and generally appreciated
  – Some scientific questions are simply so complex, so large in scope, and/or so inherently multidisciplinary as to require a “big science” approach for success – use HGP example

• Large-scale projects spawn entire fields of investigation, analyzing data (toxicoinformatics) and testing hypotheses generated from data analysis – large scale data is hypothesis generating

• New technologies and companies are frequently created in course of projects and/or to address their needs
  – Technologies become smaller/cheaper/faster under pressure from larger project - first genome $3B, this year $3K

• ELSI component can be important and helpful
  – Note ABA interest in Tox21
  – When dealing with communities, have explicit community consultation to avoid perception of stigma
Lessons perhaps applicable to HTP - 2

• Have 3 parts: data generation, technology development, informatics/data release
  – Is inherent tension between production and techdev; ideal techdev 30%
  – HGP would have taken 130 years using technologies available when it was begun

• Staging is critical, with measurable outcomes, to assure all parties – the scientists involved, the larger scientific community, and the funders – that progress is being made and project is worthwhile

• Make data as freely available as possible, with as little patenting of data as possible – HTP as a “community resource” project
  – Helps maintain support of community by demonstrating value
Lessons perhaps applicable to HTP - 3

• Project plan must be science based, not politically based
  – Can be difficult with big projects that require ongoing political support

• Have “elevator speech” – 2 sentences max – what opportunity/need is, who’s doing it, what bad will happen if it’s not done

• Expect pushback, plan for objections
  – “Generating a phonebook no one will be able to interpret or use”
  – “Ome” approach requires fundamental change in thinking that is antithetical to how most science is done and most scientists are trained (and equally importantly, promoted)
  – Address perception of drawing funds from individual investigator grants by pointing out that large scale biology projects, when designed correctly, enable individual investigators to do things they otherwise could not
  – Prospective expectation management is a must

• Team of people who WANT to collaborate is critical
  – They will be together for a long time and must sing from same songbook