Lessons from the Human Genome Project that may be applicable to the Exposome Endeavor

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

Exposome Workshop
NAS Emerging Science for Environmental Health Decisions
February 26, 2010
The Exaposome
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)

when a handful of visionaries dared to break ranks with the prevailing view that biological research must always be conducted as a hypothesis-driven enterprise. The first serious discussion of the possibility of sequencing

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 1. HGP goals and dates of achievement.

<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 (STSs)</td>
<td>52,000 STSs</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

2007 - 2010: Defining sequence variation in a few individuals

2010 - : Defining individual genomes for medical purposes?
Importance of technology

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

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>
NAS Reports are very helpful but must be specific in recommendations

National Research Council, 1988
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.

From: Sharing Data from Large-scale Biological Research Projects
Fort Lauderdale, January 2003
A more recent lesson: Publish a project paper

The Encode (Encyclopedia Of DNA Elements) Project

The Encode Project Consortium

The Encode Project Consortium aims to identify all functional elements in the human genome sequence. The project's focus is on a specified set of nucleotides (~1% of the human genome sequence) and is organized as an international consortium of computational and laboratory-based scientists working together to develop and apply high-throughput approaches for identifying all sequence elements that confer biological functions. The results of this initial phase will guide future efforts to analyze the entire human genome.

The International HapMap Project

The International HapMap Consortium

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.

Policy Forum

Molecular Biology

NIH Molecular Libraries Initiative

Christopher P. Austin, Linda S. Brady, Thomas R. Insel, and Francis S. Collins

12 November 2004

VOL 306

SCIENCE

www.sciencemag.org

Interface of the MLII and drug development.
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>
<tr>
<td>Time</td>
</tr>
</tbody>
</table>
Importance of Exposure is appreciated at NIH

The Genes, Environment and Health Initiative (GEI)

On February 8, 2006 Health and Human Services Secretary Michael O. Leavitt announced that the President’s 2007 budget proposal included $40 million for the National Institutes of Health to plan and implement a Genes and Environment Initiative (GEI). Approved by Congress, federal funding began in fiscal year 2007 and will continue until 2010, with $26 million annually going to genetic analysis and $14 million annually designated for the development of new tools to measure environmental exposures that affect health.

The GEI has two main components:

- The Genetics Program is a pipeline for analyzing genetic variation in groups of patients with specific illnesses.
- The Exposure Biology Program is an environmental technology development program to produce and validate new methods for monitoring environmental exposures that interact with a genetic variation to result in human diseases.

The Genetics Program is spearheaded by NIH’s National Human Genome Research Institute (NHGRI). The genetic analysis focuses on the alternative spellings – called single nucleotide polymorphisms or SNPs – that normally occur in the order of the 3 billion DNA letters that make up a person’s genome. SNPs are like single-letter variant spellings in a word. Most of these variations occur at single letters in the genetic code and are biologically meaningless. But a small fraction of these changes alter the function of a gene – often only slightly. The sum of many slightly altered genes may significantly increase the risk of a specific disease, but identifying such a complex set of genetics changes is challenging. Finding these disease-causing variants is one of the highest priorities of current biomedical research.

The Exposure Biology Program is led by NIH’s National Institute of Environmental Health Sciences (NIEHS). Genes alone do not tell the whole story. Recent increases in chronic diseases like diabetes, childhood asthma, obesity, or autism cannot be due to major shifts in the human genome. They must be due to changes in our environments, diets, and activity levels, which may produce disease in genetically predisposed persons. Therefore, GEI also invests in innovative new technologies to measure environmental toxins, dietary intake, and physical activity, and to determine an individual’s biological response to those influences, using new tools of genomics, proteomics, and metabolomics.
“This 2007 National Academy of Science report envisions a not-so-distant future in which virtually all routine toxicity testing would be conducted in vitro in human cells or cell lines by evaluating perturbations of cellular responses in a suite of toxicity pathway assays using high throughput robotic assisted methodologies.”
The Tox21 Community

Tox21

is becoming....
The Tox21 Community
# The Tox21 Community

<table>
<thead>
<tr>
<th>Activities</th>
<th>NTP</th>
<th>NCGC</th>
<th>EPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Historical Toxicology Data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental Toxicology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ultra High-Throughput Testing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mid- to High Throughput Systems</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower Organism Model System</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Zebrafish</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In Vitro 3-D Model Systems</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect of Human/Rodent Genetic Background on Toxic Effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computational Toxicology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validation Experience (NICEATM-ICCVAM)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Tox21: Organization

- **Leadership**: meets every 2 wks
  - B. Kavlock (EPA), R. Tice (NTP), C. Austin (NCGC)
- **Working Groups**: chairs meet together every 4 wks
  - Compounds, Assays, Informatics, Targeted Testing
  - Co-leads from each agency
- **Community**: meets every 3 months
  - Larger group of interested parties from 3 agencies
- **Oversight**: component Scientific Advisory Boards
  - Reports at least once/yr
## Tox21 Candidate Chemicals

<table>
<thead>
<tr>
<th></th>
<th>Current</th>
<th>Additional</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Universe</strong></td>
<td>13,247</td>
<td></td>
</tr>
<tr>
<td><strong>With structures</strong></td>
<td>8,277</td>
<td></td>
</tr>
<tr>
<td><strong>Plausible P-chem (logP)</strong></td>
<td>7,116</td>
<td></td>
</tr>
<tr>
<td><strong>NTP</strong></td>
<td>1353</td>
<td>~1400</td>
</tr>
<tr>
<td><strong>EPA</strong></td>
<td>1330</td>
<td>~2800</td>
</tr>
<tr>
<td><strong>NCGC</strong></td>
<td>~3000 drugs</td>
<td>-</td>
</tr>
<tr>
<td><strong>Target library, Summer 2010</strong></td>
<td>~10,000</td>
<td></td>
</tr>
</tbody>
</table>

Sources include NTP, EPA HPV, CCL, OPPIN, OW, Inerts, ToxCast, DSSTox, EU Carcinogenomics, Pharmaceuticals, others
Lessons perhaps applicable to Exposome - 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

• Have “elevator speech” – 2 sentences max – what project is, why it’s critical, what bad will happen if it’s not done

• Team of people who WANT to collaborate is critical

• “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)

• Expect pushback, plan for objections
  – 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
  – Data spawn entire fields of investigation, analyzing data (“expo-informatics”?) 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 and ultimately individualized - first genome $3B, next year will be $1000
Lessons perhaps applicable to Exposome - 2

• Have 3 parts: data generation, technology development, informatics/data release
  – Is inherent tension between production and.techdev; ideal techdev 30%
• 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 – Exposome as a “community resource” project
  – Helps maintain support of community by demonstrating value
  – Manage IP so that have diagnostic tests broadly available so don’t segment measurements since they only make sense in the context of many others
• Include ELSI component
• When dealing with communities, have explicit community consultation to avoid perception of stigma
• Project plan must be science based, not politically based
  – Can be difficult with big projects that require substantial ongoing political support