National Cancer Institute and National Human Genome Research Institute

National Institutes of Health U.S. Department of Health and Human Services

Executive Summary

Toward a Comprehensive Genomic Analysis of Cancer

Convened by:

Andrew von Eschenbach, M.D., *Director, National Cancer Institute*
Francis Collins, M.D., Ph.D., *Director, National Human Genome Research Institute*

July 20-22, 2005

The Ritz-Carlton Washington, DC

Purpose of the Workshop

"Toward a Comprehensive Genomic Analysis of Cancer" was co-organized by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) as part of their collaboration to implement the recommendation of the National Cancer Advisory Board's Working Group on Biomedical Technology to initiate a pilot phase of the Human Cancer Genome Project (HCGP). The long-term goal of the HCGP is to develop a comprehensive catalog of the genomic changes that occur in tumors and obtain an in-depth understanding about the relation of these changes to the biological processes in cancer. Such a complete information set will potentially revolutionize the cancer research community's strategies to develop and implement major new approaches to prevent, detect, diagnose, and treat cancer, through much more efficient clinical trials and accelerated introduction of more effective therapies.

The workshop explored critical issues in cancer biology, genomic technologies, bioinformatics, and the bioethical, consent, and legal issues that must be factored into the design and implementation of a two-phase approach to an HCGP, starting with a focused pilot program. The aim of the pilot phase is to demonstrate that comprehensive characterization of a tumor type[1], when analyzed in conjunction with carefully obtained, deep biological information about the tumor type, will facilitate the development of clinically meaningful applications in a way that has previously been unavailable to cancer researchers. Successful completion of the pilot project will provide a strong empirical basis for an expanded "production phase" that will rapidly and efficiently generate a genomic "atlas" for all major cancer types, ultimately benefiting cancer patients in many ways. The workshop attendees comprised a broad range of researchers and cancer survivors and explored a rich array of ideas related to the design and implementation of the pilot program. The discussion and recommendations of the workshop are described in the subsequent sections of this Executive Summary.

The Complexity of Cancer Requires a Comprehensive Atlas of Genomic Alterations to Derive Medical Solutions

Since cancer is a genetic disorder, in principle it should be possible to derive a complete catalog of inherited and acquired mutations, to understand fully the functional consequences of these alterations, and to use that information to develop and implement preventative or interventional strategies to eliminate or control the pain, debilitation, and death caused by cancer. However, the challenge of cancer lies in its complexity. The many different tumor types, each with distinct subgroups, present radically different clinical behaviors and treatment challenges. This heterogeneity arises, in part, from the fact that tumor genomes are dynamic and tumors are complex organ systems that are shaped by gene aberrations, cellular biological context, characteristics specific to the person, and environmental influences. While certain similarities exist across tumor types, any effort to characterize the genomes of tumors in a comprehensive, systematic manner must address the many questions related to heterogeneity, including:

- How many tumor cells can be distinguished from normal cells?
- What is the basis of tissue specificity for some oncogenes?

[1] Tumor type refers to the histologic type (*e.g.,* adenocarcinoma) of cancer within a given organ-specific cancer site (*e.g.,* lung).

- What is the nature of sexual dimorphism in cancer?
- Why do some people develop tumors at a younger age; *i.e.*, what are the genetic and environmental factors that influence the age of onset?
- Why are some tumors virulent and others indolent?
- Have the experimental approaches to gene discovery biased the types of genes that are studied?
- What roles do epigenomic factors and the tumor microenvironment play in cancer development?

Genomic profiles of cancers, to the level of DNA sequence, are needed to bring to bear the full complement of emerging molecular-based approaches on the diagnosis and treatment of tumors. The application of multiple optimized sequencing technologies and of characterization platforms will enable comparative oncogenomic analysis, enhance our understanding of cancer biology, establish a shared data foundation for mining and discovery, and provide a productive entry point for cancer gene discovery.

Several lessons learned from the Human Genome Project apply to the HCGP, including the value of large-scale biology, the importance of technology investment and of defining goals as well as costs upfront, and the empowerment of science by rapid data release. However, a large-scale project that seeks to obtain a comprehensive description of the genetic basis of human cancer will also have many unique features and requirements, necessitating an informed and stepwise design.

Reports from Breakout Discussions

Following plenary presentations from Drs. Zerhouni, von Eschenbach, and Collins and keynote presentations and panels, workshop participants convened in breakout groups to discuss key biological, technical, and legal challenges and issues. The conclusions from the breakout groups, which incorporated the earlier discussion from the plenary sessions, were then discussed and formed the basis of the overall conclusions of the Workshop.

A. Selection and collection of samples, heterogeneity, and other biological issues: Topics of breakout discussions in these areas included consideration of processes for prioritization of tumor sample candidates, collection of tumors, sample preparation and distribution, and accounting for the effects of cancer heterogeneity. Among the conclusions and recommendations for the pilot project were:

- Candidate sample selection for the pilot project should be based on the objective of generating meaningful data to evaluate the feasibility and utility of a future full-scale cancer genomics project and of creating compelling and novel data.
- The pilot phase of the HCGP is an opportunity to examine the range of abnormalities in a carefully selected, common set of biospecimens.
- The pilot should address the problems of tissue validation for use with multiple technologies and technology validation on a common set of quality-controlled specimens.
- For sample identification, an inventory of the current literature and available biospecimen sets, followed by an inventory of current relevant research using the multiple proposed methodologies and "specimen collection science" efforts (*e.g.*, such as those conducted by the NCI Biospecimen Coordination Committee) would be useful.
- Peer review should play a role in the choice of specific tumor types, cell lines, and xenografts.

- Critical issues that need to be addressed in sample collection include the availability of same-patient control samples, standard operating procedures, and standardization guidelines (quality control criteria).
- Possible models for biospecimen collection for the HCGP include those based on clinical annotation, disease progression, signal-to-noise ratio, incidence, metastasis, indolence/aggression, treatment response, mutagen induction, and hereditary predisposition.
- The two major types of heterogeneity that must be addressed: 1) different types and subtypes of cancer and 2) different cellular and genetic components of a single cancer. An important focus of the pilot should be the development of methods to evaluate type/subtype heterogeneity.
- Desired biospecimen characteristics include full morphologic characterization, fresh frozen tissue, the capability to support longitudinal studies, flexible informed consent, suitability for laser capture microdissection analysis, matched normal and blood samples, and ample clinical data.
- All tissues should be deposited in a central repository.

B. Sequencing and other technologies for genome characterization: Breakout discussions in this topic area included sequencing technologies, other genomic technologies for the measurement and interpretation of cancer-related genomic changes, and informatics.

- Clearly defined data standards, quality control, and quality assessment criteria, with well-defined deliverables, must be developed for every molecular technology employed in the HCGP and are one of the important goals of the pilot project.
- The pilot project should include a data production effort designed to sequence a substantial set of genes from a statistically determined number of samples of each tumor type.
- While production sequencing of PCR products will undoubtedly be initiated in the pilot phase with the well-proven Sanger capillary electrophoresis methods, new sequencing technologies should be evaluated at the production level as soon as possible for the HCGP. The large-scale sequencing of human cancer genes and the need for developing a more extensive knowledge base about the genetic basis of cancer can be a driver of effective new sequencing technologies.
- Sequencing challenges in the HCGP will include:
  - highly heterogeneous samples resulting in low signal-to-noise ratios,
  - limited amount of tissue per sample requiring the development of faithful amplification technologies,
  - DNA from unusual sources, such as paraffin-embedded samples and laser-microdissected samples, and
  - the need to continue to reduce sequencing costs to the point where whole genome analysis by sequencing is feasible.
- Other technologies used to support the HCGP must include a global view of cancer genomes, high resolution, and high-throughput capability.
- Technology platforms for determination of copy number variation, loss of heterozygosity, and structural aberrations were considered to be relatively mature, although increased capabilities in resolution, throughput, and data analysis need to be supported in the HCGP.
- The HCGP should place sufficient emphasis on epigenomic analysis (*e.g.*, DNA methylation status, histone modification), as epigenomic information is necessary to enable a full understanding of the gene inactivation events in cancer and to distinguish driver and passenger mutations; it is not clear whether the technologies for epigenomic analysis are mature enough

for large-scale application or whether the HCGP should support further technology development in this critical area.

- Meeting the informatics needs of the HCGP will require flexibility in responses (in terms of sample choices, measurements, and study design) and, at the same time, rigorous attention to the question of adequately powered study designs.
- The HCGP will require a data archive for de-identified sample annotation as well as raw and derived data from many data acquisition modalities. Such an archive should be accessible to a wide range of users without restriction and should allow the ability to browse, query, and download data; juxtapose investigator-generated data within the context of the cancer genome data; and track data provenance. The NCI Cancer Biomedical Informatics Grid (caBIG) may support this activity.
- The HCGP should create a data standards group in which all grantees must participate but which should have dedicated personnel for the maintenance of ontologies, vocabularies, and unique identifiers; these individuals should have expertise with data standards and dictionaries from relevant large-scale research initiatives and bioinformatics organizations.
- Each experimental application used in the pilot project must describe a process for regular delivery of data to the data archive in a way that complies with the data standards.
- The HCGP's data archive does not mitigate the need for appropriate funding of internal bioinformatics and other resources in the production laboratories that are dedicated specifically for data delivery.
- The pilot project needs a component for the development of new statistical and computational biology approaches that are directly relevant to the biological goals of the HCGP.

C. Legal and Ethical Questions and Issues: The breakout groups discussed several topics, including intellectual property (IP), legal issues, and guidelines for data release and informed consent policies.

- There is a "New World Order" for scientific research, embodied in projects such as the HCGP, that includes the identification of a project as a "community resource project," the release of prepublication data, and a focus on the patient.
- The HCGP should adopt data release guidelines that maximize the amount of information that can be released rapidly pre-publication and that provide access to all of the data for research purposes. There are some existing data release models (*e.g.*, HGP, HapMap, ENCODE) that can provide useful models for developing an HCGP data release policy.
- To address the patient-oriented issues that will inevitably arise, a tiered access model should be considered that identifies as much data as possible for open, unrestricted release but that also allows other data to be obtained only under certain conditions, such as only for IRB-approved research projects.

  The major differences between the HCGP, as proposed, and previous projects in the area of cancer genomics relate to data access and the possibility of detailed personal information being made available on the Internet. Therefore, it is critical to examine the adequacy of informed consent for existing samples, for example, in terms of whether the consent requirements for the pilot project differ from those of the projects for which consent was originally obtained, or whether there are additional risks from combining data sets (*e.g.*, mutations, epigenetics) that

would affect the use of existing consents. A number of suggestions regarding informed consent and ethical guidelines were made:

  - o Examine consent issues for using biospecimens from deceased persons.
  - o Examine consent agreements on existing biospecimens for re-contact provisions or for clauses that prohibit their use in future research projects.
  - o Explain policies regarding incidental results and results reporting in the consent form.
  - o Examine resources such as HapMap for best practices.
  - o Create a tiered-access policy with creative approaches to consent.
  - o Work with caBIG to establish an "early warning system" that indicates attempts to hack online project-related data.

- IP issues need to be analyzed further and the applicability of precedents set by other human genetics projects (*e.g.,* SNP Consortium) should be considered.
- The boundaries between protectable IP and free access must be defined.
- Commercial development must be encouraged, although there is no consensus on best practices for licensing mechanisms and royalties.
- Licensing issues (*e.g.,* fields of use, exclusivity, terms and conditions, royalties) must be considered.
- Best practices for license agreements should be encouraged.
- Materials transfer agreement (MTA) policies must be harmonized.

Next Steps and Future Vision

The Human Genome Project opened up new approaches to study human biology by providing a common reference sequence for the human genome and by supporting the development of high-throughput technologies to sequence DNA inexpensively and to characterize genomic features comprehensively. The promise of the HCGP is to leverage genomic and other large-scale methods to open new approaches to obtaining the critical information needed to control and conquer cancer. The overall goal of the HCGP should be to develop a public database of comprehensive genomic alterations and biological changes that lead to cancer. Such a data set will provide the basis for revolutionary new research that will lead to new and effective strategies for prevention, early detection, diagnosis, and treatment of cancer.

There was no one consensus reached among the Workshop attendees as to how the NCI and NHGRI should initiate the HCGP. However, there was general agreement that a pilot phase is neded to attack this enormous problem with multiple projects that use a range of genomic technologies coordinated around the same samples. Access to these different data views will unleash the full creativity of the scientific community to make significant advancements in fighting cancer.

Many challenges remain to be addressed for the full realization of the pilot phase. Among the different strategies discussed was a bipartite approach. The first component of this proposal would involve in-depth analysis of a small number (one to three) of carefully chosen tumors using several robust technologies that could be applied cost-effectively to a large enough number (to be determined) of

samples to obtain results that were statistically significant to a pre-defined level. Such technologies may include targeted DNA sequencing of a large number (to be determined) of candidate genes and interesting genomic regions, chromosomal copy number alteration detection, and gene expression analysis. The second component would involve preliminary analysis of up to a dozen tumors using a considerably smaller number of samples and fewer (cheaper) technologies. This two-pronged approach was proposed as a way to maximize the probability of success in deriving important biological

 information in the pilot, while also illuminating potential technical obstacles in an HCGP production phase. The pilot phase would also include a technology development component to improve additional methods to the point at which they could be applied in a high-throughput, cost-effective manner to identify genomic changes in tumor samples; this effort would necessarily include the identification of criteria to define when a technology is ready to scale for use in the HCGP. The technology development effort would also involve the identification of new approaches to find genome alterations in tumor samples while requiring a very small amount of biological material. Informatics is another serious challenge to the project. The pilot project will need to provide access to the data in a manner that both clinical and basic scientists can use productively. New data analysis and integration tools also will need to be developed as part of an emphasis on technology development. Finally, there is the challenge of releasing the data to the public as rapidly as possible in a manner that protects the patient, ensures future access to the data, and encourages the development of new approaches to the control of cancer. The best approach to address these challenges is through the careful design and management of the HCGP pilot project.

The HCGP would be a complex undertaking with enormous potential benefit to society. Solving the many challenges outlined in the pilot project will justify the undertaking of the full-scale HCGP. Moreover, it is likely that medically important discoveries will result from the pilot phase itself. The planning and implementation of the pilot project should leverage information and approaches from existing projects. Furthermore, as the science of cancer genomics continues to develop, emerging issues with biospecimens, new technologies, data and informatics, data release policies, and intellectual property will necessitate the need for new collaborative partnerships and a careful review of existing best practices.

In the near term, the NCI and NHGRI will use the input from this workshop to develop a series of solicitations to support the many components of the HCGP pilot. These solicitations will include sample collection, sequencing, genome analysis technologies, new technology development, and informatics. For additional input into data release and IP issues, an additional workshop will be planned to explore means to accommodate different ideas within the HCGP. To maximize the government investments in this project, new partnerships should be explored to include patient advocates, international partners, and the private sector. Finally, the goals and progress of this pilot project should be communicated to a wide range of groups, including patients, scientists, and the general public.