Recommendation for a Human Cancer Genome Project

Report of Working Group on Biomedical Technology

February 2005

The Working Group on Biomedical Technology strongly recommends the creation of a Human Cancer Genome Project (HCGP). The project's goal would be to **obtain a comprehensive description of the genetic basis of human cancer**. Specifically, the project would aim to identify and characterize all the sites of genomic alteration associated at significant frequency with all major types of cancers.

Comprehensive knowledge of the genetic basis of cancer would provide a permanent foundation for all future cancer research and have far-reaching implications for basic, clinical and commercial efforts to understand, prevent and treat cancer. It has the capacity to reveal the subtypes of cancers and would systematically identify the cellular pathways that are deranged in each subtype. This would increase the effectiveness of research to understand tumor initiation and progression, susceptibility to carcinogenesis, development of cancer therapeutics, approaches for early detection of tumors and the design of clinical trials.

In this report, we describe the scientific rationale and feasibility of a Human Cancer Genome Project and the key scientific considerations in the design of such a project. We then offer a preliminary outline for such a project, including estimates of timeline and cost. We note that the outline is intended as a starting point that will need to be refined in the course of launching a Human Cancer Genome Project.

1. Biomedical Rationale

1.1 Cancer is a heterogeneous collection of heterogeneous diseases. The successful treatment of medical illness largely depends on the elucidation of disease pathogenesis. Achieving a deep understanding of the pathogenesis of cancer has been exceedingly difficult as there are distinct cancers arising from unique tissues (e.g., cancers of the breast, prostate, colon, lung, bladder, pancreas, brain and others) and within each tissue type there are distinct subgroups of cancers that manifest radically different clinical behavior. For example, prostate cancer can be an indolent disease remaining dormant throughout life or an aggressive disease leading to death. However, we have no clear understanding of why such tumors differ. Similar issues arise with respect to the wide diversity of clinical behaviors observed in most cancer types.

Cancer heterogeneity poses many problems for the clinical study and treatment of cancer. Patients with inherently different tumors may be lumped together as having a single disease (e.g., prostate cancer), when they should ideally be treated quite differently. Clinical trials to assess the therapeutic efficacy of pharmacologic agents may fail to recognize efficacious drugs because the trial fails to distinguish among cancers with distinct molecular mechanisms. Drugs developed for one purpose (related or unrelated to cancer) may never be tested against specific cancer types because it is not recognized that the cellular

target of the drug is deranged in that cancer type. Patients may receive therapy that is ineffective against their tumor-type rather than a targeted therapy that would be more effective, resulting in unnecessary morbidity.

More generally, drug development is currently impeded because it is difficult to identify the full range of molecular targets for potential cancer treatment. Without the ability to select molecular targets and patient populations in a rational manner, the cancer community has witnessed little progress in the development of novel drug candidates.

In addition, the heterogeneity of cancer limits the power of current epidemiological studies that aim to associate population-based factors (environmental and genetic) with cancer risk. Such studies will be enhanced if patients are correctly classified into more homogeneous groups according to their clinical characteristics and underlying molecular pathophysiology.

Systematic understanding of the genetic basis of all cancers would have a transforming effect on the study and treatment of cancer. It would resolve cancer into sets of more homogeneous sub-classes and identify the full range of molecular targets for intervention within the cancer cell.

Achieving this ambitious goal is now within reach as a result of several important developments. First, the Human Genome Project (HGP) provided the basic foundation by providing a roadmap of the normal human genome. Second, various genome-scale technologies have been developed that allow analysis of cancer genomes at high-throughput and increasingly affordable cost. Third, various pilot projects have demonstrated the power of such information by unveiling specific genetic alterations that have led to immediate clinical application. These developments lead us to conclude that there is no fundamental conceptual barrier to achieving a comprehensive understanding of the genetic basis of cancer. There are, however, numerous practical issues to overcome, including the need for focused technology development and improvement, for organized sample collection, and for efficient sample characterization. With appropriate focus and a unified approach, the overall goal could be achieved within a decade.

1.2 Cancer is fundamentally a disease of genomic alteration. Cancer cells typically carry genomic alterations that confer on tumors their distinctive abilities (such as the capacity to proliferate and metastasize, ignoring the normal signals that block cellular growth and migration) and liabilities (such as unique dependence on certain cellular pathways, which potentially render them sensitive to certain treatments that spare normal cells).

By the 1960s, the genetic basis of cancer was clear from cytogenetic studies that showed consistent translocations associated with specific cancers (notably the so-called Philadelphia chromosome in chronic myelogenous leukemia). But, the ability to recognize specific cancer-causing mutations awaited the recombinant DNA revolution of the 1970s. Following the identification of the first vertebrate and human oncogenes and the first tumor suppressor genes, there has been increasing work leading to identification of a number of such genes selectively mutated in human cancers. These discoveries have elucidated the cellular pathways governing processes such as cell-cycle progression, cell-death control, signal transduction, cell migration, protein translation, protein degradation and transcription.

At the same time, the scientific progress has underscored that we still understand only a fraction of the genes that play a crucial role in causing cancer. Cancer is a multi-step disease process, with experimental and epidemiological data in humans and model systems suggesting that the development of a cancer involves perhaps a dozen critical steps. Some of the genetic alterations may be inherited, but the vast majority are somatically acquired during progression from a normal cell to a cancer cell. For no human cancer do we have a comprehensive understanding of the events required.

1.3 Understanding of the genetic basis of cancer has the capacity to transform clinical treatment. Although scientific progress has been rapid, translation into clinical benefit has not been immediate. It will take time to develop effective strategies to exploit knowledge of the cancer genome. After a slow start, there has been an explosion in sophisticated efforts to exploit genomic alterations in the understanding, accurate diagnosis and safe and effective treatment of cancer. We briefly review some notable recent successes in which genomic understanding has led to important progress in clinical treatment through the development of therapeutic interventions or the identification of high-risk patients to receive intensive screening.

Chronic myelogenous leukemia (CML): The BCR-ABL translocation and imatinib (Gleevec™). Molecular and biochemical studies elucidated the function of the fusion protein product BCR-ABL encoded by the Philadelphia chromosome. Specifically, these studies convincingly demonstrated that the transforming capacity of this unique oncogene was directly linked to its biochemical activity as a kinase and provided the rationale and impetus for the development of selective kinase inhibitors against this protein. The molecule imatinib (Gleevec™) is a selective inhibitor of Abl, PDGFR and Kit kinases and has proven remarkably efficacious in the treatment of chronic phase CML. More generally, these findings demonstrated that kinase inhibitors were well-tolerated, contrary to initial concerns that such drugs would have dire toxicities. It is worth noting that the efficacy of imatinib was already clear in Phase I trials and FDA approval came after Phase II trials that demonstrated overwhelming clinical benefit in this genetically homogenous disease.

Gastrointestinal stromal tumors: c-Kit and PDGFR mutation and imatinib (Gleevec™). Gastrointestinal stromal tumor is an uncommon sarcoma poorly defined by histopathology. Recently, such tumors have been found to have mutations leading to activation of c-Kit or PDGFR. These findings led to the discovery that this type of tumor, when defined based on genetic alterations, is more common than previously recognized. It also suggested that imatinib might be efficacious against gastrointestinal stromal tumors, a prediction that has been dramatically confirmed by clinical trials.

Several other cancers also respond to imatinib (Gleevec™). Three other cancer types have recently been shown to harbor genetic alterations leading to the constitutive activation of PDGFR signaling that is the target of imatinib. Hypereosinophilic syndrome and chronic eosinophilic leukemia are characterized by a FIP1L1-PDGFR-alpha translocation leading to constitutive kinase activation. Dermatofibrosarcoma protuberans is characterized by a translocation between the *COL1A1* gene and the gene encoding the ligand PDFGB, leading to an autocrine activation of PDFGR. In addition, some acute lymphoblastic leukemias (ALL) harbor the BCR-ABL translocation. For each of these diseases, imatinib has shown clinical activity that is often dramatic. In each case, there is a strong correlation between the genetic lesion and the effective therapeutic.

Breast cancer: *Her2* amplification and trastuzumab (Herceptin™). In 1987 it became clear that the *ErbB2* oncogene is amplified in 25–30% of breast carcinomas and that such amplification correlates with a poorer outcome. This finding, coupled with the discovery that *ErbB2* encodes a transmembrane receptor tyrosine kinase, gave rise to the development of therapeutic strategies based on neutralizing antibodies directed against the extra-cellular domain of the *ErbB2*. In patients whose tumors harbor amplification of the gene, these antibodies (trastuzumab) exert significant clinical activity. Notably, there is emerging data that the best predictor of therapeutic benefit is the presence of a genetic alterations of the gene rather than increased RNA or protein expression.

Lung adenocarcinoma: *EGFR* mutation and gefitinib (Iressa™)/erlotinib (Tarceva™). Systematic studies of tumors in lung cancer patients revealed that ~5–10% of lung adenocarcinomas from patients of European ancestry and 25–30% of lung cancers in Japanese patients harbor activating

mutations in the *EGFR* gene. Strikingly, patients whose tumors carry *EGFR* mutations tend to have dramatic response to the anti-cancer drug gefitinib (Irressa™), an *EGRF* inhibitor that was developed without knowledge of the presence of activating mutations in *EGFR* in lung cancer. By contrast, other patients show little or no response. Importantly, Iressa nearly failed to win FDA approval because randomized clinical trials in hundreds of unselected patients showed no statistically significant survival benefit. Recent clinical trials in unselected patients have shown similar results. It is now clear that these results reflect the drug's high efficacy in 5–15% of patients being masked by its relative ineffectiveness in the remainder. If the Iressa clinical trials had targeted *only* patients with *EGFR* mutations, it would likely have been possible to demonstrate dramatic efficacy in small Phase II trials with perhaps 50 patients (much as occurred for imatinib with CML). Thus, genetic selection strategies provide a mechanism for greatly decreasing the cost and accelerating the speed of such trials.

Genetic insights are driving additional clinical drug development. Clinical trials are currently underway with many additional novel therapeutic agents directed against targets identified through their genomic lesions in cancer. Examples include inhibitors of the serine-threonine kinase encoded by the *B-RAF* gene, found to be mutated in 80% of melanomas; inhibitors of the receptor tyrosine kinase encoded by the *FLT3* gene, found to be mutated in 25% of patients with acute myelocytic leukemia; and inhibitors of downstream components suppressed by the product of the *VHL* gene, which was found to be mutated in renal cell carcinoma. The early results of these trials are encouraging. Additional efforts are at earlier stages. In addition, striking new discoveries of mutations in cancer suggest further opportunities for drug development, such as the recent finding of mutations or amplification of phosphatidyl-inositial kinases in many epithelial and glial tumors and frequent mutations in the *Notch* gene in acute T cell leukemia

The examples above demonstrate the importance of understanding the genomic lesions underlying cancers. Specifically, these include:

- **Identification of the cellular pathways that underlie cancer**. From a fundamental standpoint, identification of genomic lesions underlying cancer has proven to be one of the most powerful ways to understand signaling and other pathways that are present in normal cells and that go awry in cancer. The understanding of these pathways has consistently proven essential for the development of strategies to treat cancer. The identification of genes mutated in cancer is thus important, even for genes that are not themselves good therapeutic targets.
- **Improved selection of therapeutic targets.** In an increasing number of cases, the identification of the genomic lesions in cancer has revealed cellular pathways on which a tumor has become critically dependent. Because these represent unique properties of cancer cells, they provide excellent potential targets for therapy. (It should be noted that some of the genes mutated in cancer may represent functions that were important to tumorigenesis but are not essential to continued tumor maintenance. Biological knowledge of pathways and functional studies with such tools as RNAi can help identify the best point of therapeutic intervention.)
- **Faster and more efficient clinical trials.** By selecting patients who are most likely to respond to a drug's mechanism of action (based on the mutations in the tumor), it should be possible to obtain indications of efficacy with a smaller number of patients (potentially even in Phase II rather than Phase III trials). This is important because it would make best use of the precious resource of patients participating in clinical trials. In addition, it would decrease the costs of clinical trials and correspondingly increase the number of trials that can be undertaken with fixed financial resources. Once a drug has been approved, the ability to identify patients most likely to respond will clearly be of great benefit to patients, physicians and payors.

- **Improved applications of drugs**. Identification of the molecular mechanisms in specific cancers may allow the extension of existing therapies to additional cancers, the understanding of drug interactions, and the rational development of combination therapies.
- **Resolution of cancer into more homogeneous groups.** Based on knowledge of underlying molecular mechanisms, patients can be differentiated into more appropriate subgroups for study and treatment. This will aid in understanding epidemiological risk factors and treatment responses.
- **Identification of markers for early detection**. Knowledge of the somatic genetic alterations that are commonly found in cancers provides targets for early detection strategies, based on analysis of DNA and protein in serum, urine and fecal samples.

1.4 Understanding inherited genomic variation has led to identification of patients at high risk. In addition to propelling therapeutic development, knowledge of the genetic basis of cancer has affected clinical practice by identifying patients at high risk and directing them to more intensive screening programs or other interventions. Some examples include the following:

Breast and ovarian cancer: *BRCA1* and *BRCA2*. At least 10% of breast cancer is thought to be familial and the analysis of large families with highly penetrant hereditary breast and ovarian cancers has led to the identification of two susceptibility genes: *BRCA1* and *BRCA2*. Carriers of germline mutations in these genes are at very high risk for early onset breast cancer. The discovery of these genes has allowed the reliable determination of carrier status, the implementation of aggressive screening strategies in affected pre-symptomatic individuals, and the clinical testing of prevention strategies based on anti-estrogen and other therapies.

Colorectal cancer: *HNPCC* and *FAP*. Familial predisposition to colorectal cancer is either associated or unassociated with polyposis. Familial adenomatous polyposis is a disorder related to mutations in the *APC* gene, a critical regulator of ß-catenin. Hereditary non-polyposis colon cancer arises as a result of mutations in the genes encoding mismatch repair proteins. Carriers of both disorders can now be detected early, offered aggressive screening, and, if needed, prophylatic colectomy (in the case of *APC*). In addition, considerable effort is now being made to identify dietary, life-style and environmental risks that predispose to cancer. For example, epidemiologic data strongly suggests that dietary or supplemental folate suppresses the risk of colon cancer in individuals with a family history of the disease. This example also highlights an example in which a lifestyle or dietary factor may influence a subset of cancers rather than cancers as a whole.

Genetic variants with lower penetrance. Familial correlation in cancer is much higher than can be explained by known high-penetrance cancer syndromes. For example, there is evidence that most of the breast cancer risk is carried by only a subset of the population. Studies are underway to evaluate common genetic variants that modulate risk, such as those involved in hormone metabolism in breast, ovarian, endometrial and prostate cancer, folate synthesis and metabolism genes in colon cancer, and so on. Systematic knowledge of the somatic changes in tumors may reveal specific genes and general pathways that should be evaluated in subsequent studies of individual risk.

2. Scientific Foundation for a Human Cancer Genome Project

Notwithstanding the encouraging progress described above, our knowledge of the genomic alterations underlying cancer remains only fragmentary. It has been the result of a largely piecemeal approach involving many individual studies, typically focused on individual genes or cancers.

Recently, pilot projects have begun to explore systematic approaches to discovering the genomic alterations underlying cancer. Such studies have only just become possible with the recent availability of an essentially complete sequence of the human genome (in rough draft form in mid-2000 and in "finished" form in mid-2003).

The results of these recent pilot projects on systematic analysis of cancer genomes make clear that there is still a great deal of important information that remains to be discovered.

- **Genomic loss and amplification**. High-resolution genome-wide studies of genomic loss and amplification have begun to be undertaken in the last 5 years using a variety of technologies. These studies show that specific cancer types typically show consistent association with genomic loss or amplification in many specific regions, indicating that these regions harbor key cancer-associated genes. **Importantly, the vast majority of cancer-associated genes underlying these consistent genomic losses and amplifications remain unknown.**
- **Gene resequencing**. Knowledge of the human genome has enabled resequencing of candidate genes through PCR-based approaches. Several groups have begun to systematically study specific gene classes (such as kinases and phosphatases) in particular cancer types. Already, these systematic efforts have led to discoveries with major clinical implications, including the presence of mutations in *B-RAF* in melanoma, *PI3K* in colorectal cancer, and *EGFR* in lung adenocarcinoma. Although these efforts are still small compared with the magnitude of the need, they clearly indicate that it is likely that **the vast majority of cancer-associated genes that are consistently mutated in specific cancer types remain unknown**.
- **Chromosome rearrangements**. Chromosome rearrangements frequently underlie crucial events in tumorigenesis, sometimes by activating kinase pathways through fusion proteins or inactivating differentiation programs through gene disruption. They have been extensively studied in hematological malignancies, where there can be a single stereotypical translocation in some diseases (such as CML) and as many as 20 important translocations in others (such as AML). Adult solid tumors have not been as well characterized, in part owing to technical hurdles. It is clear that **many of the key chromosomal rearrangements have yet to be identified and most of those that have been identified have yet to be characterized at the molecular level**.
- **Epigenetic changes**. It is becoming clear that loss of function of tumor suppressor genes can occur by epigenetic modification of the genome, such as DNA methylation and histone modification. Although the effect has been demonstrated in a number of cases, technology to monitor epigenetic changes is still quite new. **The range of important epigenetic changes in cancer thus remains largely unexplored.**

In summary, deep and broad discovery efforts are still needed to systematically identify the genomic lesions underlying cancer and thereby to dissect much of the heterogeneity of the disease.

It is now time to launch a systematic program to gain a comprehensive description of the genomic alterations underlying cancer. Such an effort would provide the most important and most general foundation for basic, clinical and commercial work in cancer in the future.

A Human Cancer Genome Project would be a natural successor to the Human Genome Project, which ran from 1990–2003. Indeed, one of the central reasons for sequencing the human genome was the goal of understanding cancer. In 1986, Renato Dulbecco published an influential article in *Science* entitled "A Turning Point in Cancer Research: Sequencing the Human Genome". Writing at a time when only a small proportion of all human genes were known, he stated: "We have two options: either try to discover the genes important in malignancy by a piecemeal approach, or to sequence the whole genome of selected

animal species [including the human]". He strongly advocated the latter systematic approach. Based on this and other calls, the HGP was eventually formulated and launched.

The program described in Dulbecco's article has now been completed in the sense that the sequencing of the human genome by the HGP has led to the discovery of essentially all human genes. It is no longer necessary for cancer researchers to devote huge amounts of time and effort to the discovery of human genes per se. This basic knowledge has greatly accelerated cancer research.

But, the program remains incomplete in the sense that we still do not know which genomic alterations play key roles in cancer. This will require studying the human genome in many tumor samples in order to identify those alterations that are significantly associated with each major type of cancer.

Dulbecco's question remains pertinent: Should the biomedical community accomplish this program through a piecemeal approach or through a systematic approach? We believe that a systematic approach will dramatically accelerate cancer research and treatment.

A systematic approach is appropriate for two reasons:

- The problem is reasonably well defined. It is possible to define a concrete goal that would provide a powerful and permanent foundation for future cancer research.
- The problem involves scalable work. It would be accomplished more cost effectively and more rapidly by mounting an organized project than through piecemeal efforts.

In both respects, the problem shares some similarities with the HGP.

3. Goals of a Human Cancer Genome Project

The general goal for a Human Cancer Genome Project could be stated as follows:

Identify all genomic alterations significantly associated with all major cancer types.

Achieving this goal will require:

i) **creating a large collection of appropriate, clinically annotated samples from all major types of cancer; and**

ii) **completely characterizing each sample in terms of:**
   - **all regions of genomic loss or amplification,**
   - **all mutations in the coding regions of all human genes,**
   - **all chromosomal rearrangements,**
   - **all regions of aberrant methylation, and**
   - **complete gene expression profile, as well as other appropriate technologies.**

Such knowledge will propel work by thousands of investigators in cancer biology, epidemiology, diagnostics and therapeutics.

Various issues need to be considered to convert this general goal into a feasible project. They include definition of the genomic alterations to be considered (e.g., somatic vs. inherited); specification of the threshold for significant association with cancer (e.g., occurring at a frequency of 5%); identification of the types of cancer to be studied; and assessment of currently available and expected technology. We consider such issues in the next section.

4. Scientific Issues in Designing a Human Cancer Genome Project

The Working Group on Biomedical Technology explored the scientific issues in designing a Human Cancer Genome Project. The analysis drew upon input from a focus group on "Characterization of Cancer in the Cell" and a workshop on "Exploring Cancer through Genomic Sequence Comparison" sponsored

jointly by the NCI and the National Human Genome Research Institute (NHGRI) that brought together ~50 scientists for a two-day meeting in April 2004. The working group also benefited from discussions with various knowledgeable individuals.

The discussion below is intended to serve as a starting point. We recognize that the issues should be carefully re-examined in the course of planning and throughout a HCGP.

The following major issues are addressed: identification of genes that are frequent sites of somatic genomic alteration in tumors; identification of inherited genomic variants often found in cancer; the types of cancers to be studied; technologies for genome analysis; and the process of sample acquisition.

4.1 Identifying the sites of somatic genomic alteration. Somatic genomic alterations underlie the initiation and progression of cancer. These somatic changes in the tumor genome can alter the underlying DNA sequence in various ways, including point mutations (nucleotide substitution and small deletions/insertions); larger-scale loss and amplification (affecting regions in the range of 1 kb to 1 Mb); and chromosomal translocations and other rearrangement. Tumor genomes may also harbor aberrant methylation, which may silence or activate genes. Such abnormal methylation is formally an "epigenetic" change, but will be included here as a genomic alteration.

In principle, it is straightforward to identify all of the somatic alterations present in a tumor genome. One need only compare it with normal genome from non-tumor tissue from the same individual. Genomic differences between tumor and matched normal tissue necessarily represent somatic mutations. Rare genomic variants found in *both* the tumor and matched normal tissue represent novel polymorphisms, which may be informative for epidemiological studies. (It should be noted that epigenetic alterations may be harder to detect, because normal tissue may not serve as an adequate control.)

Identifying the subset of genomic alterations that are *functionally important* to the cancer is somewhat more complex. Some of the genomic alterations will be responsible for the initiation and progression of the cancer through the loss or gain of important cellular functions. However, most of the genomic alterations will simply reflect the high background rate of random mutation that occurs in tumors.

Recent studies suggest that tumor genomes may have a typical nucleotide substitution rate in the range of ~1–2 nucleotide substitution per Mb[1] relative to the normal somatic genome. This corresponds to a total of ~10,000 mutations in a tumor genome. Of these, one would expect more than 100 that alter an amino acid in protein-coding regions or affect regulatory sites of genes. A typical gene might thus be mutated in perhaps 0.5% of tumors, but only a minority of these changes will be functionally relevant in the cancer. Recognizing the *functionally important* genomic alterations thus requires overcoming this background noise. This can be accomplished by statistical analysis — that is, by examining a sufficient number of tumors to reliably detect changes that occur at a frequency significantly above the background noise.[2]

Important functional changes will not be expected to occur in 100% of tumors of a given type, but rather only in a subset. For example, mutations in the *EGFR* gene occur in ~5–10% of lung cancers but play a crucial role in determining patient response to gefitinib. Similarly, amplification of the *Her2* gene is present in 25% of ductal breast adenocarcinomas but is a key determinant of response to Herceptin.

Comprehensive identification of genomic alterations underlying cancer will thus require examining a substantial collection of tumors of each type to identify genomic alterations that occur at a *significantly higher rate than the background mutation rate.*

[1] The mutation rate is likely to vary among tumor types, especially those with different types of genomic instability.

[2] Identification of functionally important genomic alterations may also be greatly aided by improved bioinformatics approaches to interpret the likely consequences of specific changes in nucleotide and protein sequences.

We propose the following threshold at least for planning purposes: **For each important cancer type, identify all genes in which the total frequency of genomic alterations exceeds 5%**.

How large a sample collection of tumors of a given type is required to detect all genes having genomic alterations in at least 5% of cases? A simple statistical analysis indicates that **a collection of ~250 tumors of any given cancer type should suffice**.[3] Such a collection should be feasible for all important cancer types, making it possible to analyze each type individually. Moreover, additional power can be gained by searching for genomic alterations seen in multiple cancer types.

It should be emphasized that the evaluation of the genes that emerge from such analysis will need to draw on extensive biological insights gained from ongoing and future research. These will be needed to infer the likely mechanism of action of the genes, which will be crucial for clinical application. In addition, it will be valuable in assessing the importance of the less frequently occurring alterations.

4.2 Issues in identification of inherited risk factors. In addition to somatic mutations, inherited variation also plays an important role in cancer. The discovery of germline variants that influence cancer risk is currently the subject of intense research, building on i) family-based and population-based clinical collections of patients with cancer, ii) catalogs of the human genome sequence and its common variation in the population (SNP maps and haplotype maps), iii) rapidly improving technologies for detecting genotyping of SNPs in large patient samples, and iv) an emerging suite of analytic methods drawn from epidemiology, population and statistical genetics. These methods should make it practical in the coming years to comprehensively test common genetic variation in large patient samples, and they hold substantial promise to elucidate inherited risk factors that can be used to predict individual risk of cancer, direct screening paradigms, and most importantly, discover causal pathways that can then be targeted for prevention and therapy.

A Human Cancer Genome Project, as currently conceived, would not be designed to comprehensively identify all inherited risk factors for cancer. Rather, large epidemiological studies are required to detect variants that confer increased risk (for example, an allele present in 10% of people that increases risk by 10%). Such epidemiological studies exist in some cases and others are being planned. Nonetheless, a HCGP will contribute to such ongoing activity in a number of powerful and important ways.

First and most fundamentally, by illuminating causal pathways in cancer a HCGP will provide a framework for designing and interpreting future studies that aim to understand the inheritance of cancer. Pathways found to be causal in somatic studies would be targets for more intense study as candidate genes in population-based association studies. Even when technology allows association studies to expand beyond candidate genes to a whole-genome search, knowledge of cancer pathways will continue to be of great value for interpreting association data. It will inform both the statistical analysis (in terms of prior probabilities of association) and the biological analysis of potential risk factors.

Second, the current efforts in genetic epidemiology are largely focused on *common* genetic variants, with catalogs of such variants becoming increasingly complete and technologies for large-scale genotyping being developed. However, it is also important to consider the role of germline variants that are rare in the general population. By identifying the genetic variations (both somatic and germline) in

[3] Suppose that we wish to find the genes with genomic alterations in at least 5% of tumors of a given type in the presence of random genomic alterations occurring at a background rate of 0.5%. By studying N=250 tumors, one can statistically expect to identify ~94% of all true-positive genes while having <1 false-positive signal out of 20,000 genes tested.

This is intended only as a simple analysis to indicate the approximate range needed to achieve high sensitivity and specificity. More sophisticated analyses should be performed; for example, to consider potential variation in mutation rate across tumor types, variation in mutation rates across genes, ways to use information about mutation types (such non-synonymous vs. synonymous changes), and so on.

tumors, we will obtain information about the aggregate frequency of rare inherited variants in each gene in cancer patients. With such information, it will be possible to test the frequency of such rare changes in cancer cases and controls by genotyping or resequencing in appropriately designed family and population-based studies.

Third, epidemiological studies are hampered by the heterogeneity of cancer. When multiple types of cancer are lumped together, the power to detect risk factors affecting one type is greatly diminished. As a HCGP allows scientists to classify tumors into more homogeneous groups based on underlying molecular pathophysiology, this knowledge can be applied to epidemiological studies.

4.3 Cancers to be studied. Defining the cancer types to be included in a HCGP is a complex question. Ideally, "cancer types" would correspond to biologically homogeneous groups. Unfortunately, no such taxonomy is available at present. Ultimately, the HCGP will greatly aid in clarifying the types of cancer. In the meanwhile, though, we must rely on available classifications.

4.3.1 Human cancers. Table 1 lists 34 major cancer types having combined incidence of ~1.4 millon in the United States (with the individual types ranging from 230,000 to 1,500). Most of these clinical types can be further divided on the basis of histopathological or molecular properties into important subtypes; a few examples of subtypes are listed in Table 2. Although the precise definition of cancer types to be collected and analyzed by a Human Genome Cancer Project will need to be carefully considered, Table 1 provides a reasonable starting point. This suggests that the **number of relevant cancer types may be in the range of ~50**. Assuming ~250 samples of each type, this would imply an overall collection of ~12,500 samples.

For at least some of the cancers, it may be possible to make collections of both primary tumors and metastases. This would enable study of the characteristic genomic lesions associated with metastasis.

4.3.2 Cell lines. In addition to tumor samples from patients, it would also be important to include samples from cancer cell lines. Cancer cell lines are a key resource because they allow reproducible studies on the properties of cancer cells, including the derangement of cellular pathways, response to chemical and biological modulators and behavior in xenograft models. Although cancer cell lines are not perfectly representative of patient tumors (because they have undergone additional mutations ex vivo), studies have shown that most of the clonal alterations observed in such cell lines are also present in uncultured cells. Understanding the genomic alterations in widely available cancer cell lines would be an invaluable asset in the study of cancer.

The precise number of human cell lines to be included within a HCGP needs to be carefully considered. For planning purposes, we have estimated that it might be appropriate to include ~1,000 such cell lines.

4.3.3 Mouse and other model systems. We also believe that genomic analysis should be carried out on key animal models of cancer. The most important are mouse models, for which there are an increasing number of carefully constructed models with precisely defined genetic etiology. Understanding of the cancer genome in these mouse models will greatly assist in relating these powerful models to human cancer and thereby in developing their full potential to assist in the development of human therapies. In addition, there may be value in characterizing samples from larger models that develop spontaneous tumors (for example, dogs) in which therapeutics can be tested.

The precise number of tumor samples from mouse and other models to be included in a HCGP should be carefully considered. Because tumors from such model systems tend to be more homogeneous than human samples, many fewer samples (perhaps 20–50) should likely need to be analyzed. For planning purposes, we have estimated that it might be appropriate to include a total of ~1,000 samples.

4.3.4 Total sample collection. Based on the considerations above, we estimate that the full sample collection might contain in the range of ~15,000 samples. Each would be represented by a high-quality well-annotated DNA sample in sufficient quantity to allow ongoing analyses of its genome by multiple groups.

4.4 Technologies for sample preparation and genomic characterization of tumors. Cancer samples would need to be analyzed with multiple technologies over the course of a HCGP, with the ultimate goal of identifying all significant genomic alterations.

Technologies for genome analysis have improved dramatically in terms of power and cost over the past decade, and progress is expected to continue the coming years. We discuss below the current state of technologies that would be required for a HCGP. In many cases, current technology is already adequate to make the project feasible — although further efficiencies would be desirable and appear to be in prospect. In some cases, current technology is not yet adequate to the task and new methods will need to be developed. In both cases, the HCGP itself would serve to propel technology improvement and development. The timing is thus right for the launch of a project.

The discussion below includes estimates of current costs and projection of likely future costs, to provide a baseline for project planning[4]. These are intended to be only approximate.

We consider six technology areas. The first pertains to sample preparation and the remaining five concern genomic analysis.

4.4.1 Whole-genome amplification (WGA). It would be desirable to have effectively unlimited quantities of genomic DNA from each tumor sample. This would allow each sample to undergo many analyses by many groups over the course of a project. With the development of "whole-genome amplification" (WGA) methods, this increasingly appears to be feasible. Recent analyses indicate that WGA appears to yield an amplification of at least 5,000-fold, while maintaining a high degree of fidelity in terms of sequence accuracy and relative allellic representation. (Such results will, of course, need to be broadly confirmed.) With such techniques, it should be feasible to collect typical quantities of high-quality tumor samples and produce enough material to permit many analyses. (N.B. Current versions of WGA are not applicable to the study of epigenetic changes.)

4.4.2 Whole-genome loss and amplification analysis (WG-LAA). Various high-throughput and high-resolution techniques have been recently developed that make it possible to survey the entire genome to identify all regions of loss-of-heterozygosity and of amplification. One of the first techniques was array-CGH (in which tumor and normal DNA are hybridized to an array of large-insert clones). The most cost-efficient current techniques are array-based methods, known as Representational Oligonucleotide Analysis (ROMA) and Single Nucleotide Polymorphism genotyping (SNP-Chips). These can simultaneously assay loss and amplification at genomic locations at a density sufficient to detect deletions and loss-of-heterozygosity within the typical size range of a single gene. Future generations of these technologies should provide even greater resolution.

Such techniques could be used to identify those regions that are consistently lost or amplified in each type of cancer. These regions would be expected to harbor a gene (or genes) that is mutated in and plays an important role in cancer, which could subsequently be identified by gene resequencing (see below). Moreover, the pattern of loss and amplification would immediately help classify subtypes of the cancer.

[4] The costs are "fully loaded", in that they reflect labor, reagents, equipment amortization and indirect costs. They assume a relatively high sample throughput, to allow efficient operation and full amortization of fixed costs. The costs, however, do not include sample acquisition or data analysis. The cost estimates were based on information from various laboratories, centers and manufacturers.

The current estimated cost of WG-LAA is ~$3,000–5,000 per tumor sample. Based on ongoing technology development, it is likely that these costs will fall by at least 5-fold in the coming 5 years.

4.4.3 Chromosome rearrangement analysis. Larger chromosomal rearrangements, such as translocations, can play an important role in the initiation and progression of cancer. A well-known example is the t(9;22) translocation that results in the BCR-ABL fusion gene in CML.

Technologies are available for systematic detection and precise genomic characterization of all chromosome rearrangements, but they are not well suited to the throughput and scale of a HCGP. Chromosomal rearrangements can be detected by multi-color fluorescent hybridization, but this technique requires whole-cell preparations (rather than just genomic DNA) and has relatively low resolution. In principle, chromosomal rearrangements can be detected by shotgun sequencing of paired ends from large DNA fragments prepared from a tumor, but this procedure is currently too costly for large-scale application (although this may well change with new sequencing technologies). It may be possible to develep effective techniques that do not require whole-cell preparation can be developed, but this will require focused efforts. It is thus not yet possible to give meaningful cost estimates.

4.4.4 Large-scale resequencing of genes. High-throughput resequencing of genes is becoming increasingly efficient. Exons can be readily amplified with flanking PCR primers, subjected to fluorescent dideoxy-sequencing and analyzed to identify mutations.

In the initial phases of a HCGP, we envisage two types of systematic resequencing efforts. 1) Genes whose function implicates them as potential targets for cancer-related mutations, including kinases, phosphatases, G-protein coupled receptors, transcription factors, non-tyrosine-kinase receptors, proteases, and others. A list of 1,000–2,000 high-priority candidates could be readily generated based on the known biology of cancer, to be resequenced in all cancers. 2) Genes across regions identified as lost or amplified in specific cancer types. Experience to date suggests that genes that are deleted or amplified in some human tumor samples are often a target for point mutation or epigenetic changes in other tumors. These genes would be resequenced in the cancer types that show consistent loss or amplification.

As sequencing technology continues to improve and costs fall, it should become possible to resequence all human genes in all samples. This would eliminate the bias of selecting specific sets of genes. (In the further future, it may someday become practical to resequence the *entire genome* of each sample. However, the launch of a HCGP should not wait for such advances in technology.)

The cost of resequencing is rapidly changing. At present, we estimate that the cost[5] to resequence ~2,000 genes in a tumor sample would be less than $75,000. Various technologies are currently in development (involving single-molecule sequencing or related techniques) that seem likely to afford at least a 10-fold reduction within the next 5 years. This would make it possible to include all ~22,000 human genes at roughly the same cost.

4.4.5 Genomic methylation analysis. Efficient, high-throughput methods for the analysis of epigenetic changes are not yet broadly available, although there are growing efforts toward this goal. It is likely that the ROMA technique (applied with methylation-sensitive restriction enzymes) and new sequencing technologies (applied to bisulfite-treated DNA) will aid in this goal. Focused efforts will be required and it is not yet meaningful to estimate costs.

[5] This estimate assumes an average of ~10 exons per gene, double-stranded sequencing, and realistic failure rates. It does not include the cost of sequencing all of the genes in paired normal tissue; candidate mutations would thus need to be followed-up by targeted resequencing testing in normal tissue.

4.4.6 Genome-wide RNA expression analysis. The tumors collected and analyzed by a HCGP will be a crucial resource for understanding human cancer. Ideally, they should thus be characterized with additional genomic tools that would provide important information to elucidate the tumors' biology.

RNA expression analysis is a particularly powerful approach that provides a near-complete picture of the genes active in the cell. This can provide valuable complementary information about the pathways activated and inactivated in tumors. RNA expression analysis has already been widely applied to cancer samples and has provided important insights, including revealing subtypes of cancer. Progress has been limited, however, by the lack of standardized methods and controls that would allow complete integration of RNA expression information produced in diverse laboratories (as, for example, can be readily done for DNA sequence information). Within a HCGP, close attention must be given to such standardization.

The current cost of comprehensive RNA expression analysis is in the range of ~$2,000 per tumor sample. Based on ongoing technology development, it is likely that these costs will fall by at least 5-fold in the coming 5 years.

4.4.7 Other technologies. It may be desirable to include additional methods for characterization of tumors (for example, proteomics), as they become available and affordable. These options should be considered on an ongoing basis by the project.

4.4.8 Technology summary. The discussion above indicates that it is already possible to perform extensive analysis of tumor genomes (including whole-genome loss and amplification analysis, systematic resequencing of 2,000 genes and expression analysis) at a fully loaded cost of less than $100,000 per sample. With reasonable assumptions about technology improvement and development over the next 5 years, it seems likely that comprehensive tumor analysis (including systematic analysis of all human genes, chromosomal rearrangement analysis and epigenetic analysis) will become feasible at roughly the same cost.

Assuming a comprehensive collection of ~15,000 samples, this would correspond to a total of ~$1.5B over perhaps a 10-year period to obtain comprehensive knowledge of the genetic basis of cancer. To put this in perspective, this would represent ~0.5% of the overall NIH budget.

4.5 Sample acquisition. The second critical component of a HCGP is the acquisition of suitable tumor samples to be analyzed.

4.5.1 Considerations in selecting samples. There are a number of important considerations in selecting samples:

Patient consent. It is essential that appropriate patient consent be obtained to allow multiple laboratories to perform comprehensive tests for genomic information from the tumor, to place the results (with usual identifiers removed) in a shared database for further analysis, and to report the results publicly. To the extent possible under HIPAA rules, NCI should develop uniform model consent forms that can be applied across institutions to meet these requirements for optimally effective use of the data.

Clinical annotation. The tumors to be analyzed should ideally be accompanied by detailed clinical information (including family history, medical history, onset and course of illness, nature and time of medical, surgical, and radiological treatments, responses to therapy, and outcome of disease) while being stripped of conventional identifying information (name, address, social security numbers and other unambiguous identifiers). Conditions for updating of information about living patients should be established.

Sample quality. The tumor samples to be analyzed should be of high quality. Multiple sections should be carefully evaluated by at least two certified pathologists, ideally including one who serves as coordinator for all samples of each tumor type. The pathological evaluation should include, at a

minimum, an estimate of the fraction of each sample composed of tumor cells, stromal cells, inflammatory cells, and vasculature. Especially in the cases of tissues or organs that give rise to multiple histological types of cancers, tests for appropriate developmental and differentiation markers should be conducted. In general, samples stored as paraffin blocks are unlikely to provide material of sufficient quality and quantity to permit a full range of analyses. Preference should be given to samples that have been quickly frozen after removal from the patient during surgery.

Sample quantity. Samples of adequate size will be required in order to permit enough material for the required genomic tests. Standards for the minimum amount of tissue to be collected will depend on the type of tumor, purity of the cell population in the tumor, and incidence of the tumor type (to meet the required number of tumors of each type to be analyzed in this project).

Availability of matched normal DNA. Samples of normal DNA (from blood samples, buccal smears, or normal tissue obtained at surgery) should be available from all patients whose tumors will be analyzed, so that it is possible to distinguish whether genomic variants are somatic mutations or newly discovered germline polymorphisms. This aspect of the study needs to be mentioned in the consent forms for the study, recognizing that germ-line mutations predisposing to cancer may occasionally be found and might be clinically important to cancer risk for the patients or their relatives.

Ethnic diversity. The tumors should ideally be selected to include representative samples from ethnically diverse populations. This may require focused sampling efforts to boost the number of cases from particular ethnic groups.

4.5.2 Logistics of sample collection. Sample collection should be coordinated by appropriate Cancer Sample Acquisition Centers.

In most cases, it will be necessary to collect new tissue samples that meet the requirements above. It will be necessary to establish guidelines for i) the protocols for collection, storage and DNA preparation to ensure uniformly high quality, ii) the clinical information to be obtained, and iii) the patient consent to be obtained. Such guidelines will need to be established in collaboration with knowledgeable physicians and scientists. The guidelines should be enforced by careful oversight of centers, with continued funding dependent on proper adherence. Such coordination will be particularly important because sample acquisition for many cancers may require multi-institutional efforts. Given the importance of patient consents, NCI will likely want to ensure that proposed Cancer Sample Acquisition Centers resolve key IRB issues before the commencement of funding.

In some cases, it may be possible to use existing clinical collections of tumor samples that have been systematically collected with both contemporaneous clinical annotation and subsequent patient outcome data. The availability of such patient outcome data would be advantageous, because it may immediately shed light on the prognostic power of genomic lesions identified. Such existing tumor collections will need to be carefully evaluated, however, to be sure that appropriate patient consent has been granted, that the tumors samples are of a suitable uniform standard and that the data derived can be made available without restriction to the entire scientific community.

DNA extracted from tumors will need to be distributed to project researchers under uniform access and distribution policies. This will likely require amplifying the genomic DNA by appropriate methods that ensure faithful representation of the genome. (It may also be possible to establish cell lines in some cases, but research would be required to determine whether such lines faithfully represent the genome.) It will need to be decided whether DNA distribution should be centralized. In any case, an ultimate repository of the DNA samples will be required. Because the Cancer Sample Acquisition Centers should not exist in perpetuity, long-term storage centers to house the primary DNA samples and the representations of the DNA will also need to be established.

4.6 Informatics. Medical informatics and bioinformatics will play a crucial role in a HCGP. There will be needs for:

- systems to manage collection, integration, storage and dissemination of samples and associated clinical data;
- systems to manage collection, integration, storage and dissemination of genomic information from the samples;
- new analytical tools to integrate and interpret experimental data about genomic alterations (including processing of raw data for statistical analysis and association of mutations with cancer types and clinical outcomes); and
- national database(s) to make the primary information and the scientific results broadly available to the biomedical community.

Informatics support will likely be needed in at least three forms: individual investigator grants to develop new analytical methods; components of center grants to develop and maintain production informatics systems; and national databases. In all case the support should be allocated based on competitive peer review.

5. Experience from the Human Genome Project

In planning a Human Cancer Genome Project, it may be useful to briefly consider the recent experience of the Human Genome Project.

One of the most important lessons is: **it is important to have a clear goal at the outset, but the operational definition of the goal, costs and timeline will likely need to be continually refined over the course of the project**. In the HGP, the degree of completeness and accuracy that could be achieved was not known in advance and only a rough estimate of the projected costs and timeline could be made. Some degree of ambiguity and uncertainty were tolerated, in order to take advantage of rapid technological change.

Other key lessons are:

- a program to achieve a focused goal should nonetheless include diverse scientific activities of different types and scales;
- funding should be awarded on the basis of rigorous peer review and should be subject to re-competition on a regular basis;
- rapid release of data before publication can greatly accelerate scientific progress;
- peer review is important for all components of the program;
- public policy issues raised by the science (as done by the ethical, legal and social issues component of the HGP) must be addressed; and
- involvement of multiple U.S. funding agencies and international funding agencies can improve the quality and effectiveness of a project;

Although the HGP is by no means a perfect analogy for the proposed project, these particular observations may well be applicable and are worth considering.

6. Initial Outline for a Human Cancer Genome Project

The working group recognizes that the ultimate design for a Human Genome Cancer Project will require further careful study and should continue to evolve over the course of the project. Here, we offer an initial outline as a starting point for further refinement.

6.1 Project goal. The project goal would be to **obtain a comprehensive description of the genetic basis of human cancer** by identifying all the sites of genomic alterations present at significant frequency in all major types of cancers.

6.2 Operational definition. The operational plan will need to be carefully considered. As discussed above, an initial plan would be to:

- Perform comprehensive genomic analysis of ~15,000 samples (including ~250 tumor samples from each of ~50 major cancer types, together with samples from cancer cell lines and appropriate animal models).
- Identify the comprehensive list of all somatic genomic alterations that occur with a frequency of at least 5% in any of the major cancer types (for which the samples should provide sufficient power).
- Capture information about the inherited genomic variations seen in patient samples, to provide information to be used in subsequent studies of cancer risk.

6.3 Project organization. The Human Genome Cancer Project would be largely carried out through an extramural network consisting of two kinds of centers:

- Cancer Sample Acquisition Centers, with the ability to collect samples from various tumor types with high quality and at appropriate scale.
- Cancer Genome Analysis Centers, with the ability to characterize tumor samples with high quality and at appropriate scale.

We envision a network consisting of multiple centers of each type. The appropriate support mechanism (e.g., grant vs. contract) will need to be considered. We note that the grant mechanisms may offer greater opportunity for innovation, while contract mechanisms may offer tighter accountability (which may be particularly important for sample collection). In either case, centers should be selected through rigorous peer review and funding should be contingent on the accomplishment of milestones. Centers should have sufficient size and stability to be able to efficiently accomplish their missions, but should have no expectation of permanent existence. Although existing NCI cancer centers or NHGRI genome centers may well prove excellent sites for some centers, the competition for centers should be open to any groups without preference.

The Human Genome Cancer Project should also include other mechanisms beyond production centers. These should include:

i) investigator-initiated research grants for the improvement and development of technologies (e.g., methods for characterization of chromosomal rearrangements or genomic methylation) and computational tools for cancer genome analysis, and
ii) databases for maintaining information produced by the project.

Finally, a Human Genome Cancer Project should include a component to address ethical, educational, medical and regulatory issues (EEMRI). It is important to ensure that the health care system will be adequately prepared to deal with the changes that will be catalyzed by the project. An EEMRI program would study and (in coordination with other stakeholders, such as FDA) explore mechanisms that prepare patients, physicians, regulators, medical educators, health care agencies, biopharmaceutical companies, insurers and others for the consequences of systematic information about the genetic basis of cancer.

6.4. Data release. The Human Genome Cancer Project should adopt a policy that information about cancer genomes is rapidly released into the public domain without restriction on scientific use. The specific details of the data release policy will need to be worked out and may depend on the precise nature of the data.

The goal of the data release policy should be to protect the public interest in at least two important respects. The data release policy must ensure the protection of patient confidentiality. Also, it should

ensure that companies have freedom-to-operate with respect to the information to maximize commercial progress in clinical diagnostics and therapeutics. (This might involve such steps as having grantees file of statutory invention reports with the Patent and Trademark Office to put information into the public domain, as was done for the SNP Consortium.) Clear data release policies should be incorporated as a condition of each grant award.

6.5 Project management. The Human Cancer Genome Project should be jointly managed as an equal partnership by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). Such an arrangement would take advantage of the extraordinary depth of scientific and management expertise in these two sister institutes. Effective joint management will be best accomplished by ensuring that funding flows through both institutes.

In addition, a U.S. HCGP should encourage and cooperate with similarly directed efforts in other countries. Notably, the Wellcome Trust Sanger Institute in Cambridge, England, already has experience in analyzing cancer genomes.

6.6 Project timeline and costs. We believe that most of the ambitious goals of a HCGP could be accomplished within ~10 years of launch, based on current and projected technology.

Although the required project cost cannot be stated with certainty. our best judgment is that a HCGP could be accomplished with an average annual budget of ~$150M. (As discussed above, comprehensive characterization of tumor genomes would cost considerably more at current costs. However, we believe that continued improvements in technology will make it possible to accomplish the goal with this budget.)

The working group strongly believes that the long-term budget for a HCGP must come from increased funds appropriated by Congress, rather than from existing funds in the current NIH budget. Current projections show cuts or sub-inflationary increases in the NIH budget for at least the next several years. Without dedicated funding, a HCGP would require major cuts in existing programs. The working group, however, does support the immediate launch of pilot projects and recognizes that these will need to be funded from existing budgets. The recommended scope of such pilot projects is discussed below.

In relative terms, the projected cost of a HCGP would be modest given its broad impact across cancer research. The proposed average annual cost of ~$150M would correspond to an increase of ~3% in the combined annual budget of NCI and NHGRI. (As discussed below, the cost would begin at a lower level during a pilot phase and then increase to a higher level during full-scale implementation.)

In our judgment, such supplemental funding would seize the opportunity to create a permanent foundation of knowledge to transform the understanding and treatment of cancer. A compelling case can and should be made to Congress and the American people to support this project through the appropriation of new funds.

6.7 Related research. The information generated by a HCGP will propel progress in many areas. It will need to be followed up by studies in cancer biology (e.g., to analyze pathways biochemically or to construct animal models), in cancer epidemiology (e.g., to assess the attributable population risk of particular inherited genetic variants) and, of course, in cancer diagnostics and therapeutics. Although such follow-on studies fall outside the scope of the HCGP itself, they will surely require significant support from NIH sources, industry and public-private cooperation. This need should be incorporated into programmatic planning at NCI.

7. Assessing Success

How should the success of a Human Cancer Genome Project be assessed? In the short term, it will be important to drive the project with clear goals and milestones (such as samples collected, samples analyzed, technologies developed). In the long term, however, the ultimate measure of success must be

the impact on the lives of patients. It is important that the project design be continually optimized to maximize the likelihood and speed of such impact.

For diagnostics and prognostics, the project should aim to propel the development of routine "DNA biopsy" of tumors to supplement the current pathological practice of visual grading of tumors. DNA biopsies should require only a small sample from a patient's tumor and should add useful prognostic information — including information about which cancers are likely to metastasize and should therefore be treated aggressively and which cancers are likely to respond to particular treatments.

For therapeutics, the project should aim to propel the development of new, targeted treatments that are more effective and less toxic than conventional chemotherapy. When coupled with long-term research on the biochemical pathways in which the discovered cancer genes participate, additional targets for cancer therapy will emerge. A clear picture of the underlying defects of cancer cells will provide a coherent framework for drug development and testing.

These benefits are already beginning to accrue and should accelerate long before the project itself is completed. Partial information about specific subsets of genes and specific cancer types has immediate value in propelling work in diagnostics and therapeutics across industry and academia.

The project should thus be designed to ensure that it bears early fruit. Moreover, it should be periodically reviewed to evaluate its impact on both basic science and clinical medicine.

8. Next Steps

The working group recommends that NCI and NHGRI take the next steps toward launching a Human Cancer Genome Project, even in advance of appropriation of new federal funds. Toward this end, we recommend concrete next steps and sketch an approximate timeline.

8.1 Initiation of management and oversight (2005).

8.1.1 Joint Working Group. A Joint Working Group (JWG) of experienced scientific staff from NCI and NHGRI should be established now to guide and coordinate the role of the funding agencies.

8.1.2 External Scientific Committee. An External Scientific Committee (ESC) should be established as soon as possible, consisting of ~5–8 senior scientific advisors to the HCGP.

The members of the ESC should have experience in implementing large-scale biomedical research projects and should span a broad range of expertise. They should primarily be scientists who use, in their own research, the kind of information that would come from this project.

The ESC would provide advice to program staff about the implementation of the HCGP — including initiation, specific areas of focus, assessment of progress and timing of scale-ups.

It would be advantageous if the ESC membership included at least one member on the National Cancer Advisory Board and at least one member was on the NHGRI Advisory Council.

8.1.3 ESC subgroups. The ESC would promptly establish subgroups to assist in developing the details of the project and provide appropriate input to program staff concerning solicitations for competitive awarding of funds for initial pilot projects. (NIH staff should do the actual preparation and writing of the solicitation.) We envisage four subgroups:

- **Sample selection subgroup**. This subgroup would consider: i) cancer types to be used in pilot phase and source of samples; ii) process for establishment of sample collection centers; and iii) sample types to be used for full project, including human cancers, cell lines and animal models (including decisions about metastases and sources of germline DNA).

- **Genomics subgroup.** This group would consider: i) initial genes to be prioritized in resequencing; ii) use of LOH information to guide selection of regions to be analyzed; and iii) priorities for technology development**.**
- **Bioinformatics subgroup**. This group would consider: i) storage and public access for project data and ii) data analyses that will be needed and mechanisms for catalyzing such work.
- **Ethical, educational, medical and regulatory issues (EEMRI) subgroup**. This group would consider: i) consent issues involved in carrying out the project and ii) unique intellectual property issues raised by the project.

8.2 Pilot phase (2006–2008). We envisage a pilot phase of ~3 years in duration, to be followed by full-scale production work. This phase would have three important components.

8.2.1 Sample collection. The primary goals of this work would be to provide ~1250 samples from five specific cancer types for the genomic analysis pilots (see below); provide a range of ~500 additional samples from a wider range of cancers for technology development projects (see below); create and validate general procedures for collection, characterization and maintenance of samples; and begin collection of larger sample collections for the full project. Decisions will need to be made about the appropriate mechanism for this solicitation, but it should be flexible enough that sample gathering can change to accommodate the needs of the technological applications, as those are discovered and refined through the life of the project.

Timeline. We suggest a solicitation for pilot sample collection centers, to be issued by summer 2005 and funded by February 2006. Initial samples would be collected and ready for distribution by mid-to-late 2006. Some tissue resources already exist and may be suitable for use in the pilot phase.

Approximate cost. Total of $3–4M/yr (including direct and indirect costs), across ~3–4 pilot sample collection centers.

8.2.2 Genomic analysis. The primary goals of this work would be to undertake initial genomic analysis by applying available technologies to an initial sample collection. The initial collection would consist of ~1,250 samples, with ~250 tumors from each of five cancer types. The available technologies consist of genome-wide LOH analysis, resequencing of ~2,000 genes and RNA expression analysis. This pilot work will allow an evaluation of the results and will also help drive significant gains in efficiency.

This work effort should be undertaken through cooperative agreements rather than contracts, because there are still many open scientific and technical questions. A contract mechanism is unlikely to have the flexibility to rapidly respond to new information or technical advances or to appreciably stimulate cost reductions during the time of the pilot project. Several awards (3–5) should be made to ensure competition, but should not be so many that coordination becomes unwieldy. Ideally, these efforts should be scalable over time.

Timeline. We suggest a solicitation for pilot genome analysis centers, to be issued by summer 2005 and funded by February 2006. Because pilot sample collection centers would not deliver projects samples until mid-to-late 2006, initial work would begin on a preliminary set of sample (which the JWG/ESC would need to identify and procure).

Approximate cost. Total of $40M/yr (including direct and indirect costs), across ~3–4 pilot genome analysis centers. (Assuming some gains in efficiency, this should allow resequencing of ~2000 genes and genome-wide LOH analysis in ~250 samples from each of five tumor types.)

8.2.3 Technology development. Truly comprehensive analysis of cancer genomes will require further technology development. The needs include optimizing existing technologies for the setting of

tumor samples; creating effective techniques for studying genomic rearrangements and epigenetic modifications; and improving the efficiency of methods for comprehensive resequencing.

We would propose that applications for technology development projects be solicited as soon as practical (once samples are available). Projects to optimize existing technologies would apply them to a variety of tumor types, to demonstrate feasibility and establish costs. If successful, such efforts could be scaled up. Projects to develop new technologies might involve substantial research components.

Timeline. We suggest a solicitation for applications (R01, R21, R21/R33) to be issued by summer 2005 and funded by February 2006.

Approximate cost. $5M in Year 1, growing to $15M in Years 2 and 3).

8.3 Full-scale implementation (2009–2014). The pilot phase is intended to resolve open scientific and technological questions, including providing a clear picture of feasibility and cost. The ESC would be asked to evaluate these questions on a regular basis.

The HCGP should be ready for full-scale ramp-up by 2008. The project would involve a network of integrated or specialized centers, managed through cooperative agreements and/or contracts with a coordinating center.

Timeline. We project that the full-scale implementation could be completed over a period of ~5 years, with a budget of ~$200M/year. The details, including the number and structure of centers and other activities, should be defined by the JWG with advice from the ESC.

Total cost. The total cost of the project would thus be ~$1.35B over 9 years, corresponding to an average annual budget of ~$150M. This figure is an estimate, which should be revisited in the course of further project planning.

9. Conclusion

A major barrier in the fight against cancer has been the extraordinary complexity and heterogeneity of the disease. Scientific advances over the past decade have finally provided, in principle, the ability to gain a comprehensive picture of the genetic basis of cancer. Such systematic knowledge would lead to dramatic progress in the understanding, classification, detection, diagnosis and therapy of cancer, by accelerating research in thousands of laboratories throughout academia and industry. We owe it to generations of cancer patients to come to seize this opportunity.

Table 1. Cancer Types and Subtypes

Incidence in United States >100,000 cases/yr	Incidence
Prostate cancer	230,110
Breast cancer	217,440
Lung cancer	173,770
Colon and rectal cancer	146,940

Incidence in United States >10,000 cases/yr	Incidence
Bladder cancer	60,240
Melanoma	55,100
Non-Hodgkin's lymphoma	54,370
Cancer of the uterus	40,320
Cancer of the head and neck	38,530
Kidney and renal pelvis	35,710
Cancer of the pancreas	31,860
Cancer of the ovary	25,580
Thyroid cancer	23,600
Cancer of the stomach	22,710
Liver and intrahepatic ductal cancer	18,920
CNS tumors	18,400
Multiple myeloma	15,270
Cancer of the esophagus	14,250
AML	11,920
Cancer of the cervix	10,520
Gallbladder and other biliary	6,950

Incidence in United States >1000 cases/yr	Incidence
Testicular cancer	8,980
Soft-tissue sarcomas	8,680
CLL	8,190
Hodgkin's disease	7,880
Cancer of the small intestine	5,260
CML	4,600
Anal cancer	4.010
Cancer of the vulva	3,970
ALL	3,830
Ureteral cancer	2,450
Sarcomas of the bone	2,440
Cancers of the eye and orbit	2,090
Cancer of the urethra and penis	1,570

Table 2. Examples of Cancer Subtypes

Lung cancer:	Non-small cell, small cell
Breast cancer	Ductal, invasive, baseloid, estrogen receptor +/–, Her2 +/–
Cancer of the head and neck	Squamous cell, nasopharyngeal, adenoid cystic
Cancer of the uterus	Endometrioid, adenosquamous, papillary serous, clear cell
Cancer of the ovary	Serous, mucinous, endometrioid, clear cell
CNS tumors	Glioblastoma multiforme, anaplastic astocytoma, etc.