Exploring Cancer through Genomic Sequence Comparisons

A National Cancer Institute–National Human Genome Research Institute Workshop

April 14–15, 2004

Bethesda Marriott Hotel
Bethesda, MD

Index

Executive Summary

Exploring Cancer through Genomic Sequence Comparisons

A National Cancer Institute–National Human Genome Research Institute Workshop

April 14–15, 2004

Bethesda Marriott Hotel
Bethesda, MD

Co-chairpersons

Anna D. Barker, Ph.D., Deputy Director for Advanced Technologies and Strategic Partnerships
National Cancer Institute, National Institutes of Health

Francis S. Collins, M.D., Ph.D., Director
National Human Genome Research Institute, National Institutes of Health

Executive Summary

On April 14–15, 2004, the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) convened a workshop, "Exploring Cancer through Genomic Sequence Comparisons." Participants included leaders from the Nation's cancer centers, genome centers, the NIH, biotechnology and pharmaceutical sectors, and also international scientists. The purpose of the workshop was to assess the value of a project to catalogue all of the DNA sequence changes that occur in tumorigenesis by comparing the sequences of multiple tumor genomes to reference sequences from normal tissue taken from the same individuals. Such a data set would provide new insights about the molecular events underlying the development of cancer and could lead to new strategies for studying cancer and eventually to the development of new interventions.

Background

The confluence of two major scientific advances made this workshop particularly timely. First, years of productive research have identified a large number of mutations involved with or implicated in tumorigenesis. In-depth study of the genes and pathways identified by these mutations has led to an understanding of many of the molecular details underlying tumorigenesis. The success of newly introduced cancer drugs, such as Gleevec®, Herceptin®, and Iressa®, which were designed on the basis of such understanding, proves that somatic genetic alterations are legitimate targets for therapy. Low resolution, genome-wide studies are beginning to catalog additional changes, such as small deletions or amplifications, and directed studies of individual genes implicated in tumor biology are steadily increasing our awareness of the variety of mutations underlying cancer. These data make it evident that only a modest fraction of the molecular targets involved in tumorigenesis has been identified and that cancer is a very heterogeneous disease due to many different mutations, environmental factors, and the interaction between the two. To develop targeted interventions, it will be important to identify all or most of these events.

The second major advance that precipitated this workshop is the availability of a reference human genome sequence, and rapidly advancing sequencing technologies that were stimulated to occur in concert with this effort, and which continue beyond it. These technological improvements make obtaining sequence information increasingly rapid and relatively inexpensive. It is now possible to contemplate something that was previously incomprehensible: obtaining comprehensive sequence information from multiple tumor types, at different stages—in a process that would be unbiased by our currently selective knowledge of the biology of cancer—to catalog all the genomic changes associated with cancer, and to render them accessible to study and intervention.

The NCI and the NHGRI convened this workshop to explore the unprecedented opportunity that the cancer community and the genomics community have to align their resources and foster a partnership for making significant advances toward achieving the goal of the NCI to reduce the burden of cancer on the American and world populations.

Charge to the Workshop

A project of this scope requires delineation of the complexities. These include technical and design challenges, design of a pilot project for sequencing cancer genomes to best inform decisions about optimal sequencing approaches and strategies for a full project, selection of tumor

types, and translation of results into clinically meaningful outcomes. To address these issues, workshop participants were asked to address the following questions:

- Which sequencing technologies would be the most appropriate for this application?
- Which tumor(s) should receive initial focus?
- What other data should be collected on selected tumor types?
- How could such an effort be piloted?

Workshop Outcome

The workshop explored state-of-the-art uses of genomic information to identify mutations that could be associated with tumorigenesis. The discussion began with descriptions of activities that are already under way including resequencing of regions already implicated in cancer, genome-wide analysis of gene expression, and low-resolution genome-wide assessments of chromosome instability such as rearrangements or loss of heterozygosity. Participants discussed the ability of current and anticipated sequencing technologies and strategies to address whole-genome sequencing of tumor genomes. Experts on individual cancer types discussed current knowledge that could be effectively used in developing and assessing the whole-genome approach to tumor analysis. By the end of the workshop, the participants had identified the elements that should be included in a publicly available database. This included nucleotide changes, insertions, deletions, and translocations.

Because a project to sequence tumor genomes would be one of significant scope, this workshop must be considered a first step toward a collaborative effort between the NCI, NHGRI, and the public and private research communities to employ sequencing technologies to address key scientific questions with respect to cancer.

Choosing the appropriate sequencing technology and strategy

Most, if not all, tumor cells will accumulate many mutations. To give meaningful results, a sequencing strategy must eventually allow identification of the mutations directly implicated in cancer origin and propagation, out of the relatively large set of background mutations that accumulate as a consequence of the neoplastic state. In addition, because the spectrum of genetic alterations (e.g., point mutations or small nucleotide changes, deletions, translocation, amplifications) associated with cancer processes is broad, the most appropriate tools for analysis of each must be considered. Workshop participants discussed the relative merits of several approaches, including bacterial artificial chromosome (BAC)-end sequence profiling, PCR-directed resequencing using array technology, massively parallel signature sequencing, and sequencing of coding regions.

Whole-genome analysis starting with BAC-based end sequence profiling is capable of revealing larger genomic changes, such as rearrangements or larger insertions or deletions, but it is more limited in discovering smaller events. It has the advantage that once a tumor is arrayed as a BAC library, its genome is effectively "immortalized," allowing it to be made available to the research community for additional studies involving, for example, resequencing or functional assessment.

Directed, or candidate, approaches—involving the sequencing of one or a few genes from a large number of tumors—have been effective, and the number of reasonable candidates is not so high that it challenges present technology. However, these approaches do not sample the genome in an unbiased way. Sequencing all coding and some control regions may be possible in a

comprehensive way relatively soon, but would still miss mutations in noncoding sequences that might be involved in gene regulation.

Comprehensive genomic sequencing would provide an unbiased view of the cancer genome. This approach would entail sequencing a sufficient number of genomes (perhaps 200) from the same tumor type in order have the statistical power to distinguish clinically significant changes from background mutations. The cost of using current sequencing technology in existing high-throughput production settings is still prohibitive. However, several resequencing technologies are in development that promise to drive costs down far enough within the next several years for such a project to become feasible.

Selecting the tumor type(s) for analysis

Workshop participants discussed the current state of the science with regard to the sequencing and molecular characterization of many common tumor types, including prostate, colon, breast, brain, lung, pancreas, and lymphomas. DNA can be obtained from primary samples or from cell lines. Advantages of cell lines as a source include the "unlimited" quantities of DNA available and the low level of heterogeneity. However, cell lines may distort the biology if they represent subclones of the sample that have a selective advantage for *in vitro* growth or contain additional mutations that permit immortalization. By contrast, tissue samples may exhibit dramatic heterogeneity, a factor that must be considered when selecting a tumor type for a pilot project.

It was agreed that it would be preferable to attempt to do this project with tissue samples. The samples collected should represent the following:

- Different anatomic locations (e.g., breast, colon, lung, prostate, and ovary)
- Specific genetic defects (e.g., mutations in mismatch repair genes, BRCA1 mutations)
- Several levels of genetic abnormality (i.e. tumors carrying low, intermediate, and high numbers of numerical and structural aberrations) and low and high levels of
- gene amplification
- Progression series within patients (e.g., normal, hyperplasia, *in situ* invasive and metastatic cancer, associated stroma)
- Mutagen induction (e.g., ionizing radiation)

Value of mouse models of human cancer

The workshop participants discussed the merits of including complementary mouse models of human cancer, as the ability to manipulate the mouse genome offers a controlled platform for investigating the roles of various genes in processes that are relevant to cancer. The project could be linked with the NCI-sponsored Mouse Models of Human Cancer Consortium (MMHCC), which builds and characterizes mouse models of common human cancers. Mouse models may add value to the pilot projects inasmuch as multiple models are available at each site based on common human signatures and abundant, well-characterized samples are available. Many of the models already have substantial data sets that include immunohistochemical data, comparative genome hybridization, gene expression profiling, serum and tissue protein, and karyotyping. Gene sequencing of appropriate mouse tumors in parallel with corresponding human tumors is underway at a low level and resequencing of large regions of loss of heterozygosity in mouse that are shared with human tumors can be highly informative. Preclinical trials are ongoing on validated mouse models and samples of responsive or recurrent tumors can be readily collected. Proof-of-principle data illustrate the feasibility of analyzing epistasis and of identifying interacting genes, though this requires a large sample size.

Designing a pilot project

The scope of the project contemplated would be significant and, because not all the required technologies are mature, it will be important to conduct a pilot project. The pilot project has to be designed to provide data on feasibility, to identify the critical technological and biological issues such as which tumor types are most appropriate, to drive the development of technology, to gather data about alternative approaches, and to assess the potential benefit against the cost of a full effort. Because it is currently impractical to identify each mutation within a specific tumor, a pilot project may instead target functional classes known to be implicated in cancer, such as kinases (potential targets for intervention), or it may pursue genes in genomic regions that are either deleted or amplified in the tumors. A great deal of additional planning will have to occur to formulate an appropriate pilot project that will effectively inform decisions about whether, and how, to proceed.

In combination with the opening emphasis on a few functional categories of genes, it was suggested that a pilot project initially be limited to a small "discovery" tumor set to reveal areas of focus for subsequent analysis. It will be important, however, for the long-range goals of the enterprise to extend well beyond the coding regions of a short list of candidate genes, and for participants to contemplate a general strategy for exploring many tumor types.

As the pilot evolves, the effort will likely shift from hypothesis-driven individual research to a coordinated strategy with the potential for large-scale discovery.

Summary and Next Steps

This workshop concluded that a pilot project to sequence cancer genomes is both feasible and timely. Advances in sequencing technology, bioinformatics, and understanding of cancer processes suggest that the results from such a project could have widespread ramifications. Participants offered a series of recommendations and next steps:

Organize only around the ultimate goal, using a pilot project as a catalyst

The ultimate goal defining all of the important mutations found in cancers through comprehensive sequencing of tumor genomes will lead towards cancer diagnosis, treatment, and prevention. The Human Genome Project (HGP) demonstrated the advantage of maintaining a diversity of approaches, all of which supported one well-defined goal. Maintaining multiple approaches during a pilot phase will promote the greatest number of innovations and breakthroughs. Seeding a few projects will be a key step in developing an early portfolio of approaches. Three or four main paths for progress must first be determined, with the understanding that they may converge at a later point. During this pilot phase, we will streamline the effort through economies of scale and guide the development of technologies towards use for detection of mutations in tumors at the scale needed to distinguish significant mutations from background.

Arrange a portfolio based on four major issues

It was suggested that a pilot project portfolio be arranged around four major issues related to different critical aspects of preparing for a high-throughput effort:

1. What technology will be used to accumulate genomic data (e.g., sequencing techniques)?
2. What will be identified or measured (e.g., whole candidate genes, exons only, a short list of likely candidates, the entire genome)?

3. What biological materials (e.g., cell lines, primary tissues (early, late lesions, metastasis)) will be used?
4. What other kinds of data are desired from these materials (e.g., proteomic analysis, information about pathways)?

Many parallels exist between this pilot project and the HGP. Thus, an agenda similar to that of the HGP, complete with milestones that are beyond current capabilities, can be created with the confidence that the technology will improve to meet the goals.

Devise a strategy that will promote synergy, partnerships, and profitable spin-offs

One of the most important lessons learned from the HGP is that great ideas , which capture the imagination, are more likely to attract support. Engaging the private sector will promote the development of new technologies and diagnostics as well as the development and delivery of viable therapies.

Determine deliverables and ways to engage the imagination of patients and the public

Many clinicians are unsure how to translate basic science results into clinical practice. Moreover, the general public must be convinced that these results will contribute to the armamentarium against cancer. Deliverables that can be understood by clinicians and patients will be necessary. The clear endpoint of the HGP helped translate the importance of the work to the public.

Build upon past successes

The cancer community has made a difference in the lives of many cancer patients, despite an incomplete understanding of all of the relevant genetics of the condition. By building upon current successes, this project will allow major impacts on the understanding of cancer in the next three to five years, opening the potential for an impact on treatment to follow close behind.

Create a working group to align biological research and clinical outcomes

Aligning mutations with disease progression will be critical to the success of this project. The NCI and NHGRI will convene a working group to outline long-range milestones and timetables for this enterprise, and to generate a portfolio of research components addressing the four issues identified at this workshop and listed above.. Proposals related to a pilot project would be solicited via competitive review.

Conclusion

This workshop reinforced the observation that current cancer research and genomic sequencing technology have reached a point of confluence. As the cost and speed of DNA sequencing continue to decrease, large-scale sequencing is becoming a viable tool for discovery. An exciting opportunity exists to create an ambitious and well-defined strategy that will provide information critical for reducing the Nation’s cancer burden. The cancer and genomics communities, together, have the means and motivation to embark on a project to describe the universe of genetic changes associated with cancer, which will in turn identify numerous molecular targets for diagnosis, prevention, and therapeutic intervention.

Meeting Summary

Exploring Cancer through Genomic Sequence Comparisons

A National Cancer Institute–National Human Genome Research Institute Workshop

April 14–15, 2004

Bethesda Marriott Hotel
Bethesda, MD

Co-chairpersons

Anna D. Barker, Ph.D., Deputy Director for Advanced Technologies and Strategic Partnerships
National Cancer Institute, National Institutes of Health

Francis S. Collins, M.D., Ph.D., Director
National Human Genome Research Institute, National Institutes of Health

Table of Contents

Background

The purpose of the meeting was to explore the potential value of a pilot project to compare the genome sequence of a tumor to a reference sequence for obtaining information that will provide useful insight into human cancer biology and which will, ultimately, drive the development of new interventions. Is there broad research value in developing a publicly available data set representing all the sequence-based variations found (e.g., nucleotide changes, interstitial deletions, translocations)?

Why convene this meeting now? The sequence of the human genome has been essentially completed, making available a baseline against which nucleotide and structural changes identified in the genomes of tumors can be compared. Other resources that could aid in analysis, such as databases of variation and function, are increasingly available. Computational tools are advancing rapidly and may already be at point where comparative analyses could yield useful data.

Examples of questions that should be addressed by the meeting and its outcomes are as follows:

1. How much sequence information would be desirable? The whole genomic sequence of a tumor? 50 percent? What is the rationale for each approach and the cost/benefit ratio?
2. Is there evidence that this approach would yield useful data? Are there other approaches that would yield the same information?
3. Are the computational tools and scientific groups in place to analyze this data?
4. In addition to the human genome sequence, what other existing genomic resources might be complementary in yielding useful analyses?
5. What technologies/strategies could be applied to determine the genomic sequence of a tumor and what are the strengths and weaknesses of each? Consider, for example, current large-scale sequencing technology versus other sequence scanning methods.
6. What tumor type(s) should be included in the pilot? How could they be prioritized (e.g., morbidity, mortality, prevalence, tractability)? Which stages should be considered? What sample sizes will be needed to obtain meaningful data?
7. What will be the effect of cellular heterogeneity within the tumor on the ability to interpret the results?
8. Is it expected that the outcome of the approach will facilitate a better understanding of cancer? Is the potential for acceleration of progress sufficient to merit making this program (or the pilot) a high priority?
9. What is the study design that assesses the feasibility of the project? Should the pilot be confined to human tumors or should mouse models also be included?
10. Should each tumor have an experimentally sequenced "normal" control? What would be the benefit vs. the cost?

Yet another question to ask is, if a feasibility study is desirable, what would be the expected outcomes? A feasibility study would be expected to determine whether this approach will result in detection of genomic changes that might be candidate regions or markers implicated in the initiation and progression of cancer. There are many considerations, for example:

- Adequacy of sampling design (number of samples, size of region sequenced) needed to find candidate markers
- Feasibility and cost of technologies, strategies, and analysis of data
- Pilot data to inform the decision of whether this study should be scaled up to a larger sample size and larger number of cancer types

Introduction and Objectives of the Workshop

Anna Barker, Ph.D.
Deputy Director for Advanced Technologies and Strategic Partnerships
National Cancer Institute, National Institutes of Health

Dr. Barker opened the workshop by thanking attendees for their time and expertise. She noted that many NCI investigators wrestle with genomic issues and the translation of information about the genome into effective therapies, and she thanked attendees for their willingness to assist the NCI as it plans strategically for the upcoming decade.

The NCI has recently issued a "Challenge Goal," to eliminate the suffering and death due to cancer by 2015. Current paradigms for cancer treatment often involve intervention after tumors have been detected and diagnosed. To meet the Challenge Goal, however, the cancer community must intervene much earlier in the cancer process. Cancers must be detected and diagnosed much earlier in their progression toward metastatic disease. Progression toward tumor formation can then be modulated by targeted treatments. To reach this ambitious goal, the NCI must think differently about short-term goals and ways to build a community to reach these goals. The success of currently available cancer drugs such as Gleevec® and Herceptin® proves that molecular defects are legitimate targets for directed therapy and suggests that cancer may one day be managed in the same manner as other chronic diseases.

Many facets of the current environment in cancer research suggest that a pilot project of sequence comparisons has never been more timely:

- The current pace and climate of change are unprecedented, as evidenced by the completion of the Human Genome Project.
- Major advances have been made in the understanding of genomic changes in select cancers.
- Genomics and proteomics have the potential to dwarf prior advances in cancer research by identifying thousands of molecular targets, compared to a total number of targets to date of approximately 500.
- Cancer genomic resources (e.g., Cancer Gene Anatomy Project (CGAP), Mouse Genome Centre (MGC), transgenic animal models), a new bioinformatics infrastructure (Cancer Bioinformatics Grid (caBIG)), and novel analysis tools dramatically expand current capabilities.

The cancer community must leverage its resources to accelerate the pipeline for the discovery, development, and delivery of cancer therapies. Current statistics indicate that development of a successful drug requires approximately 15 years and upwards of $1 billion dollars. Marginal improvements to the current system will be insufficient to reach the NCI Challenge Goal and will have little impact on eliminating the suffering and death due to cancer.

Several hurdles must be overcome to optimize technology for the early detection of cancer. They include the following:

- Identification of at-risk populations
- Lack of validated biomarkers
- Limitations in the deployment of key technologies in the clinic
- Limited availability of annotated and quality-assured tissues
- Lack of a new business model that engages the private sector
- Issues related to regulatory science for early detection

To address these hurdles, the NCI has implemented a series of strategic initiatives in the following areas:

- Systems/integrative biology
- A national resource in bioinformatics, known as caBIG
- Partnerships with the FDA to enable drug development
- Initiatives in biomarkers and targeted interventions
- Support for key advanced biomedical technologies (e.g., proteomics, nanotechnology)
- An advanced technology initiative that will provide leadership for the convergence of science and technology
- Alignment of resources through new partnerships at the center, state, and regional levels

Given the current advances in technology, a resource that contains all of the sequence-based variations in cancer genomes has never been timelier. The rationale to explore cancer through sequence determination is supported by the following:

- The draft of the human genome is a great starting point. Sequence information can be generated from tumors and compared to reference standards.
- Technologies and public resources are becoming increasingly available.
- Tumor sequence information has the potential to inform human cancer biology through the development of new interventions that can lead to individualized strategies for disease management.

The NCI brings substantial expertise to this initiative. A large number of NCI investigators have moved the field of cancer genomics forward, and the timing is right to ask substantive questions. It is anticipated that this meeting will be the first step in a future collaborative effort between the NCI, NHGRI, and the cancer community to use enabling sequencing technologies to address key scientific questions with respect to cancer. Genomic sequence analysis can illuminate several key factors in the cancer process, including:

- Germ-line mutations that can inform early intervention strategies
- Somatic mutations that will provide clues to molecular signatures and downstream genetic changes
- Germ-line variations to identify predisposing variants and pharmacogenomics
- The mapping of chromosomal aberrations to inform the knowledge base of translocations

However, many complexities remain when designing a pilot project for sequencing cancer genomes. Among the issues that must be considered are:

Complexities of cancer choice

- Cancer type(s): common (breast, prostate, colon) versus molecularly well-characterized (e.g., acute lymphoblastic leukemia (ALL), B-cell lymphoma)
- Practical issues: stage of disease, tissue heterogeneity and availability, morbidity, mortality, and prevalence
- Sample size
- Cost benefit

Complexities of sequencing approaches

- Bacterial artificial chromosome (BAC)-end sequencing of tumor genomic DNA
- Candidate gene sequencing of tumor genomic DNA
- Hybridization to SNP chips

- Whole genome shotgun sequencing
- Sequencing of single cells

Technical and design challenges

- How much of the genome should be sequenced and why?
- When should the normal matching sample be sequenced?
- What is the most appropriate normal matching sample (e.g., lymphocytes, neighboring normal tissue)?
- Is the microenvironment a critical consideration?

Other complexities

- Transcription data: sequence-based (e.g., serial analysis of gene expression (SAGE) and hybridization-based (e.g., cDNA chips)
- Protein data: antibody structures, mass spectrometric analysis

These complexities suggest that simplicity and rigor in approach and organization will be critical to the success of any genomic sequencing pilot project. Although the NCI and the NHGRI have no preconceived notions regarding the outcome of this meeting, the goals of this workshop are five-fold:

1. Explore the value of comparing tumor sequence to reference standards.
2. Brainstorm ideas about preferred technologies and best study approaches.
3. Discuss the value of potential information gained.
4. Potentially consider the value of a pilot project.
5. Discuss the issue of the appropriate timing for such a project.

Dr. Barker also charged participants to ponder the following relevant questions:

- How much sequence information do we need? Is the whole genome necessary?
- What approaches offer the greatest yields?
- Are there optimal tissues/specimens for a pilot project?
- Do we have the tools and expertise to analyze the data?
- Are the existing resources that may inform genomic analysis (*i.e.*, animal models)?

Finally, Dr. Barker noted that the cancer community currently knows much about the genetics of many cancers. However, the NCI must consider whether the current time is appropriate for a pilot program in cancer gene sequencing that can establish proof-of-concept to empower a larger-scale effort. If a pilot project is deemed feasible, the major scientific questions and costs must be assessed.

She closed her introductory remarks by noting that the meeting participants represent diverse aspects of the cancer community. As such, each person has a contribution to offer, whether entertaining new ideas challenging conventional thought. She encouraged participants to explore the strengths and weaknesses of various technologies, cancers, and strategies. By doing so, new leverage points and partnership opportunities will emerge, thus eliminating the need to "reinvent the wheel."

Francis Collins, M.D., Ph.D.
Director
National Human Genome Research Institute, National Institutes of Health

Dr. Collins welcomed participants and noted that this workshop is the result of conversations held over several months. He noted that the field of genomics is progressing rapidly. Researchers are now approaching the point where large-scale sequencing can be considered as a tool for discovery. As a consequence, constraints that limit thinking can now be replaced by a global perspective.

He noted that the Human Genome Project (HGP) was conceived as a way to provide information that would lead to a medical benefit. Its completion in 2003 has provided a portal for many future research endeavors for curing diseases. Historically, the cancer community has embraced the sequencing of the human genome with great vigor, and the time has arrived to begin thinking about the sequencing of cancer genomes. Through enabling technologies, it is expected that the cost of sequencing will continue to decline, offering many opportunities for large-scale sequencing endeavors that can systematically identify somatic mutations.

In 2003, Dr. Collins authored a perspectives paper (Collins FS, et. al., *Nature* 2003;422:835–47) that positioned the HGP as the foundation of a multi-tiered building. The ground floor of such a structure is the application of *genomics to biology*, which encompasses the following goals:

- Define the structure of human variation, the human haplotype map.
- Sequence many additional genomes.
- Develop new technologies for sequencing, genotyping, expression analysis, and proteomics.
- Identify all functional elements of the genome.
- Identify all the proteins of the cell and their interactions.
- Develop a computational model of the cell.

The current capacity to sequence cancer genomes is aided greatly by the strengths of five NHGRI-funded centers for large-scale sequencing (Broad Institute, Washington University in St. Louis, Baylor College of Medicine, The Institute for Genomic Research (TIGR)/Joint Technology Center, and Agencourt). Collectively, these centers have the annual capacity to sequence 100 billion base pairs. At 4X coverage, this capacity translates into approximately eight mammalian genome drafts per year. Moreover, due to a high level of technology investment, both by the public and private sectors, sequencing costs have dropped in the past decade by a factor of two every 22 months.

Currently, techniques are needed to annotate the human genome with the most power. Ultimately, it is desirable to know the evolutionary changes of genomes from many species, and branches of the evolutionary tree that have heretofore been lightly sampled will provide insight into development. Today, nearly eight million single nucleotide polymorphisms (SNPs) are posted in public databases. One year ago, this number approximated two million (an estimated ten million are common in human genome). Thus, a sizeable fraction of the total number of human variants is currently available and accessible through public databases. How can the cancer community capitalize on such resources and utilize this capacity for the study of cancer?

Although cancer is a genetic disease, much remains unknown about the genetic changes that promote disease progression. In some respects, the cancer community has been "looking under the lamppost" when identifying genes implicated in cancer. Current technology provides the capability to examine chromosomal rearrangements, yet there is no strategy that casts a net broad

enough to capture all of the genes associated with cancer. However, the genome is a bounded set. If all of the targets for cancer can be identified, many new approaches for drug therapies will emerge. What is needed is a way to examine cancer cells exhaustively within the context of resource limitations. For example, if the cancer community were to characterize three genomes (germline, cancer, and metastasis) for each of two tissue types, the resultant six genomes would utilize a significant fraction of the current national resources available in a given year.

The numerous types of genetic alterations (e.g., point mutations, mismatch repair/slippage, loss of copy number, translocation, amplifications) require different tools for analysis. Gross mutations will be detected more easily than subtle ones. Genetic instabilities in cancer can be considered in three broad categories: chromosome instability (CIN), microsatellite instability (MIN), and nucleotide excision repair (NER)-related instability (NIN) (Lengauer C, *et. al. Nature* 1998;396:643–49), each of which offers a unique set of challenges when characterizing its phenotype.

Characterizing the CIN phenotype involves tools that have not all been widely applied, although the potential to do so exists, including:

- Karyotyping
- Comparative genome hybridization (CGH) (against metaphase chromosomes or BAC-, cDNA-, or oligonucleotide arrays)
- Identifying specific genes activated or inactivated at the breakpoints, sequencing BACs to show evidence of internal rearrangement

Characterizing the MIN phenotype requires identification of the target genes in the tumor where the instability affects the tumor progression. Approaches to characterize the MIN phenotype include:

- Testing for microsatellite instability using standard markers
- Identifying cancer-relevant genes that are potential targets for activation or inactivation

Characterizing the NIN phenotype may be approached, too:

- Testing a short-list of candidate genes
- Testing all exons from a longer list of candidate genes
- Complete tumor genome sequencing

Dr. Collins concluded his remarks by noting that the cost of sequencing continues to decline, perhaps leading to the sequencing of an entire genome for $1,000 in the next ten years. He noted that, with the advent of increasingly large volumes of genomic data, it will become critical to isolate mutations implicated in cancer from the “haystack” of irrelevant mutations. He asked attendees to consider the following questions with respect to this meeting:

- How well have we done so far in identifying the genetic basis of cancer?
- What more could be done using large-scale genomic sequencing?
- Which sequencing technologies would be most appropriate for this application?
- Which tumor(s) should receive initial focus?
- What other data should be collected on selected tumor types?
- How could such an effort be piloted?

The Case for a Tumor Genome Sequencing Project

Joe W. Gray, Ph.D.
Director
Life Sciences Division, Lawrence Berkeley National Laboratory

Dr. Gray thanked meeting organizers for the opportunity to speak. He noted that the time is right for a tumor genome sequencing project and outlined his proposal for a phased way to move into such a project.

Several factors argue for a tumor genome sequencing project:

- Genome sequence information has transformed our understanding of the structural organization, evolution, and function of normal genomes.
- Genome-wide analysis of gene expression, copy number, loss of heterozygosity (LOH), structural aberrations, methylation and mutations in human and model cancers have revealed deregulation at many levels.
- Evolutionary mechanisms vary dramatically between tumors.

Current technology enables the testing of hypotheses about DNA sequence evolution within tumors. However, gene composition varies dramatically in chromosomal instability, even within clinically similar cancers. Moreover, a substantial fraction of the genome is influenced by recurrent aberrations that are implicated in the evolution of cancers; thus, most of the genome is present in a given abnormality. Also, expression and copy number have been qualitatively correlated in cell lines derived from tumors. Thus, it is possible that the level of amplification does not predict the level of overexpression. A recent study by Volik and colleagues (*PNAS* 2003;100:7696–7701) using BAC-end sequencing on the breast cancer cell line, MCF-7, demonstrates multiple levels of genomic rearrangement in addition to amplification of a target amplicon.

Two important issues arise when considering a tumor genome sequencing project. First, researchers currently have very limited genome structure or sequence information about events that alter genome function. Second, information that is generated is typically not comprehensive or on well-characterized, widely available tumor material. However, current technical capabilities can address these limitations.

Two general approaches—gene-based resequencing and clone-based sequencing—can be applied to sequence the genome.

The gene-based methods are well supported and have advantages and limitations. Advantages are that they are straightforward and supported by commercial primers for most exons and junctions. They are also suited to analysis of limited regions of large numbers of tumors. Disadvantages of the gene-based methods is that they are insensitive to structural changes that alter regulation, they miss mutations in unsuspected regulatory regions, and they are not well suited for large-scale, community-based studies.

Clone-based sequencing approaches, such as transformation-associated recombination (TAR) and end-sequence-directed BAC libraries, offer a different set of options.

TAR cloning is well suited to targeted resequencing. Although the technique can capture a substantial part of a genome, it does not immortalize the tumor genome for communal access.

Furthermore, the approach is DNA-intensive and requires relatively large amounts of sample tissue.

By contrast, whole genome analysis starting with BAC-based end sequence profiling immortalizes tumors in arrayed BAC libraries. Clones are end-sequenced to identify structural rearrangements and to create a tumor BAC map. Mapped library clones can then be made available to the research community for sequencing or functional assessment. This approach is becoming increasingly straightforward and requires modest amounts of DNA. The procedure generates a crude contig map of a tumor, which can be sequenced efficiently. The sequence for each BAC can be made available, thus allowing easy resequencing from a BAC repository.

However, the full spectrum of benefits from this approach is not yet known, suggesting that a BAC-based effort to sequence a cancer genome should begin with a phased approach, as follows.

Phase I: BAC library production for selected, well-characterized primary tumors, precursor lesions, associated stroma, and metastases

This phase should consider the analysis of a series of samples that will immortalize the complete tumor progression in several patients. Using the current resources, the creation and management of 6X libraries for 100 tumors can be realized for less than $10M. However, it would be necessary to consider the depth of the libraries when calculating project's scope and cost. There would be specific tissue requirements for the sample series:

- Different anatomic locations (e.g., breast, colon, lung, prostate, and ovary)
- Specific genetic defects (e.g., mutations in mismatch repair genes, BRCA1 mutations)
- Several levels of genetic abnormality (e.g., tumors carrying low, intermediate, and high numbers of numerical and structural aberrations and low and high level of gene amplification)
- Progression series within patients (e.g., normal, hyperplasia, *in situ* invasive and metastatic cancer, associated stroma)
- Mutagen induction (e.g., ionizing radiation)

It is also critical to recognize the tremendous heterogeneity of tumor tissues and the limitations that instability may impose on the information obtained. For example, a 1X analysis of copy number abnormalities will fail to reveal chromosomal instability.

Phase II: BAC End sequencing fingerprinting

- Approximately 60,000 BACs (~2X coverage) end sequenced for each sample*
- BAC clones containing inversions, translocations, and amplicons associated with genome breakpoints by analysis of paired end sequences
- End sequences used to assemble rough sample contigs

*End sequencing of 60,000 BACs currently costs approximately $500,000.

*Phase IIIa. Targeted sequencing (BACs on demand)**

- Guided by positional cloning efforts (e.g., from recurring numerical and structural aberrations)
- Matched to chimeric cDNAs
- Selected to reveal information about mechanisms of aberration formation (amplification, structural aberrations)

- Correlated with changes in gene expression
- From regions being accessed by the Encyclopedia of DNA Elements (ENCODE) project

*Current BAC sequencing costs approximately $5,000/BAC.

Phase IIIb. Full sequencing of a few tumors selected with community input

- Consider normal tissue, *in situ* cancer, invasive cancer and metastatic lesion from the same patient
- Coordinated with spectral karyotyping (SKY), LOH, CGH, exon resequencing, methylation analyses
- Sequencing of a minimal tiling path (depth to be decided)

Although tumor heterogeneity will complicate this process, BAC sequencing offers a way to immortalize cancer genomes to approach them rationally. A clone-based strategy provides the community with access to regions of the genome without having to sequence it themselves, making it a useful technique for a cancer genome sequencing pilot project.

Discussion

One participant inquired about the amount of DNA required to create a BAC library. Libraries have been constructed using as little as 50 mg of tumor tissue, although extensive microdissection is required. If the whole genome is amplified, however, it is difficult to identify structural rearrangements.

Another attendee asked about the level of coverage necessary for BAC sequencing of a triploid genome. This level is not known.

The level of expression is not correlated with CGH, which suggests that the genome is deregulated in unknown ways. By cataloguing all of the unusual rearrangements, how can a meaningful rearrangement be identified over background "noise"? As time goes on, the regulatory regions of the genome will be better understood. It was suggested that the best strategy at present is to target attention to select regions and see whether gene expression matches copy numbers. One attendee commented that it is possible to overlook the key mutation when it is grouped among a series of mutations. However, rearrangement occurs during amplification, so the amplification process itself offers some selective benefit.

One participant asked about the extent to which analysis should focus on regular intercellular somatic cell variations outside of malignancy. It was observed that the rate for mutations in the HPRT gene in normal calls is 10^{-13} to 10^{-15} mutations per generation, so there will necessarily be a level of background noise.

One participant noted that BAC has not successfully provided much information on recurrent abnormalities or to identify the changes that are most biologically relevant.

Another attendee asked if there is another method that rapidly provides the high-density structure of a genome. For example, recircularized jumping libraries are useful for locating breaks in the genome, although the approach is not as amenable for development as a community resource. Although noise is a problem with this approach, the technique may be worth considering as an adjunct issue.

Another comment was made concerning the consideration of metaplastic cells as a sample source. These cells appear phenotypically normal, yet they contain indications that they will progress toward carcinoma *in situ* or dysplasia.

One attendee also suggested adding a cDNA library to the immortalized genome from BAC as part of the public repository.

A Vision for Finding Genomic Changes in Cancer

Eric S. Lander, Ph.D.
Director
Broad Institute

The time has arrived to transition from an era marked by the characterization of individual oncogenes to one that focuses on the overall characterization of cancers. The cancer community now has the technological basis to consider the comprehensive characterization of all genomic mutations associated with cancer (e.g., germline, somatic (point mutations, deletions, amplifications, and translocations) and epigenetic changes associated with cancer) as well as the functional pathways associated with cancer. This "wholesale" approach to cancer will require the community to devise a strategy that capitalizes on new technologies and optimizes the analysis and application of the data thus generated.

In the previous decade, several papers have called for the sequencing of a cancer genome. The NCI has responded to community interest in the comprehensive characterization of cancer by forming a technology working group with the following goals:

- A **structural goal** to identify all genomic mutations associated with cancer (e.g., germline, somatic, epigenetic)
- A **functional goal** to identify all functional pathways associated with cancer (e.g., tumor expression, synthetic lethality, perturbation maps)

What can be learned by resequencing tumor genomes in their entireties? The background rate of mutations is approximately 3×10^{-6}/base (range: 10^{-5} to 10^{-6}), and the frequency of mutations is 5×10^{-3}/gene, or 0.5 % (coding: 1500 bp). Using these parameters, approximately 200 tumors per cancer type will be required to locate all genes with a mutation rate less than or equal to five percent. (If one mutation per gene is expected on average and ten are observed, the probability of a gene with 10 mutations is 0.0003. This is a rough estimate of what to look for in order to draw meaningful conclusions given the mutation frequency and background rate).

Several currently available sequencing options can be used as reference points for assessing the scope of such a proposal. Conventional shotgun resequencing currently costs approximately $\$10^{-3}$/base, or $3M per 1X coverage. For a target coverage of 2X (stochastic coverage of two alleles), the cost of sequencing a tumor genome is approximately $6 million. For targeted resequencing, the current cost is $\$10^{-2}$ to 10^{-3}/base. The future goal of sequencing a genome for $1,000 remains beyond the current capacities.

The feasibility of studying 200 tumors, as a function of sequencing costs, is projected below:

Cost/Tumor	Cost/Cancer Type	Cost/30 Cancer Types
$6M	$1.2B	$36B
$100,000	$20M	$600M
$10,000	$2M	$60M
$1,000	$200,000	$6M

Using these projections, the cost of such a project becomes justified if the technology permits a tumor genome to be sequenced for $10,000 or less. To reach this point, the process must begin now with bold goals, and we must ride the learning curve.

Dr. Lander then proposed a two-part "Strawman Plan" to reach the goal*:

Step 1: Identify altered regions in the genome

Amplification/deletion (SNP, representational oligonucleotide microarray analysis (ROMA), etc.); $100–$1000
Re-sequence 1 Mb regions; $1000–$10,000
Translocation breakpoints; cost unknown

Step 2: Target functional classes (e.g., kinases)

Re-sequence favorite 700 genes; $1000–$10,000

*Note: 100 tumors may suffice for targeted regions and classes.

Discussion

It was noted that it is difficult to write a compelling study that argues for total resequencing of just a few tumors. Instead, an effective strategy focuses on point mutations based on regional or functional class mutations. Although point mutations are difficult to identify, they must be studied to understand the changes affecting the genome. Such a strategy allows a "big picture" sense of the changes, although it will overlook certain aspects in the process.

One participant noted that, as cancer becomes more dissected by gene expression profiling, the concept of "tumor type" becomes more fluid. The sub-classification of cancer will lead to more types. Dr. Lander agreed but responded that the Human Genome Project began as a process toward a goal. Sequencing of tumor genomes can be viewed within a similar lens, with the immediate goal of spurring technology development.

Another attendee asked about deeper coverage of select genomes. Is it better to cover one tumor at 6X or many tumors at 2X coverage? Dr. Lander noted that, given the current level of available funding, it is impractical to identify each mutation (at a rate of one in 300,000), since the majority of individual changes are meaningless in terms of disease progression. The value arises from seeing the effects across several tumors.

One participant noted that the mutation rate/base will vary between tumors. Does this imply that a way to sample the background of a tumor is needed? Dr. Lander responded that doing so would be relatively straightforward, and that price calculations may vary by a factor of two due to somatic tissue and polymorphisms.

Dr. Lander concluded by observing that the challenge is to choose the areas in which to delve more deeply. Regarding concerns of overhead costs, the cancer community currently has enough experience to recognize that this strategy is feasible. However, it should begin with pilot projects, followed by choice of specific regions of the selected tumors.

Session 1: Approaches for Determining Cancer Gene Sequences

Moderator: Joe Gray, Ph.D.

A Sequenced-Based Approach to Analysis of Tumor Genome Structure and Function

Colin Collins, Ph.D.
Assistant Professor
Cancer Center, University of California, San Francisco

There are several challenges of this undertaking, including creating a methodology capable of:

- creating a single sequence-based methodology capable of mapping and cloning all structural rearrangements *en masse*, mapping copy number, and integrating transcriptome and proteome data.
- providing a rational framework for sequencing tumor genomes while making high throughput functional studies straightforward.
- immortalizing the tumor genomes being studied.

One approach that addresses these challenges is end-sequence profiling (ESP). This approach begins with the construction of BAC and cDNA libraries from a selected cancer cell line or tumor (cost: ~$100K per 10X BAC library; $10K per cDNA library). End-sequencing the generated clones can then be performed at an approximate cost of $7 per clone ($1.4M for 10X clonal coverage and $100K for cDNA sequencing) at an efficiency of 90%. End sequences are then aligned to the normal genome to identify aberrations (efficiency: ~90% for BAC ends; ~80% for cDNAs). Genome breakpoints identified in this manner can then be identified and cloned, and their copy numbers mapped at high resolution (cost: 100 Kb resolution, ~$200K; 10–15 Kb, ~$1.6M). This procedure has recently been demonstrated using the breast cancer cell line, MCF-7 (Volik, *et. al. PNAS* 2003;100:7696–701) with a whole genome resolution of approximately 200 Kb. Using ESP on this cell line, resolution was increased up to 10 Kb for high copy-number regions.

The ability to make BAC libraries is not tumor-specific, and ESP has been extended to the sequencing of primary (breast, brain, ovary) and metastatic tumors (prostate). These applications generally require 50–200 mg of tumor tissue. The libraries thus created contain approximately 20,000 clones with estimated yields of 200,000–400,000 clones.

The translation of back-end sequence data to structural information can be obtained through spectral karyotyping (SKY) to identify break points in tumor genomes. In addition to specific genome location, the approach offers the possibility of mechanistic information. When ESP is combined with digital SKY (Raphael, *et al. Bioinformatics* 2003;19 Suppl 2:II162–71), insight may be gained into the molecular archeology of tumor genomes and the sequences of episome substructures.

ESP is also useful for "functional oncogenomics" studies by addressing the significance of the genome breakpoints thus identified. The availability of BAC clones allows direct isolation of loci from tumor genomes. Accordingly, it becomes possible to reconstruct segments of human tumor genomes in model systems, thus preserving the nuances of transcriptional regulation, alternative splicing, and novel genes to model more accurately human tumor biology.

ESP is emerging as a powerful technology for structural genomics and may revolutionize the understanding of tumor genome evolution and biology. With this technique, approximately 100–200 tumors can be analyzed and immortalized for a cost less than that required to sequence a single Gb genome. Strengths of ESP are several:

- Unimpeded by heterogeneity
- Can detect rare events
- Reveals tumor genome structures at high resolution
- Directly clones genome breakpoints en masse
- Identifies fusion transcripts
- Can integrate transcriptome and proteome data
- May enable high throughput identification of tumor-specific proteins

Sequencing of Coding Regions

Richard Wooster, Ph.D.
Cancer Genome Project, Wellcome Trust Sanger Institute

At least 291 genes are known to be causally implicated in human oncogenesis. Descriptions of the structure of cancer genomes are required to identify these mutated genes and to detect patterns of genomic abnormality that reveal information about processes of mutagenesis (e.g., past exposures and alterations in DNA maintenance).

The Cancer Genome Project (CGP) was begun in 2000 with a $60M donation from The Wellcome Trust and the Institute of Cancer Research. In 2003, the CGP completed two million sequence reads. Current experiments at the CGP include genotyping (fingerprinting, LOH prediction), copy number analysis (homozygous deletion mapping, amplification mapping), BAC end sequencing analysis of rearrangements and copy number, sequencing coding regions for the discovery of small intragenic mutations, and a catalog of somatic mutations in cancer by curation of data from the literature. Resequencing projects at the CGP include gene-based efforts (ultimate goal: the entire coding genome; initial efforts: the protein kinase gene family and DNA synthesis and repair) and samples from tumors and cell lines. Currently, the CGP is resequencing approximately 100 samples from each of several primary tumor types, including breast, lung, colon, testis, gastric, and sarcomas. Tumors are selected for characterization by pathologists who evaluate the tissues for tumor content and appearance. Cell lines currently being resequenced include all cancer cell lines from the European Collection of Cell Cultures (ECACC) and the American Type Culture Collection (ATCC), as well as from other public repositories.

Researchers at the CGP have recently developed a kinase gene mutation screen, based on their discovery of somatic mutations in the BRAF oncogene in many human neoplasms. This screen has identified a point mutator phenotype in human beast cancers for which there is no clear precedent in the literature. It has also identified other small intragenic somatic mutations of unknown consequence.

The CGP has developed a database, the Catalogue of Somatic Mutations in Cancer (COSMIC; www.sanger.ac.uk/perl/CGP/cosmic) as a public resource to catalog known cancer genes. The database features negative and positive results and may be extended to include SNP data.

An attendee inquired about the amount of tissue required for a typical resequencing experiment. Dr. Wooster replied that 200 μg genomic DNA is necessary to get 500 genes. No microdissection is required.

Approaching Cancer Genomes: Current Practice and Future Prospects

Elaine Mardis, Ph.D.
Co-Director
Genome Sequencing Center, Washington University School of Medicine

Dr. Mardis used acute myeloid leukemia (AML) as a model to illustrate her institution's approach to sequencing cancer genomes. AML is a group of diseases caused by a variety of inherited and acquired genetic and epigenetic changes. Although genetic therapeutic approaches exist, most patients still die from these diseases. Numerous genes are implicated in all stages of the pathogenesis of AML, including susceptibility, initiation, progression, and relapse/resistance.

The AML Program Project Grant (PPG) at the Washington University School of Medicine (WUSOM) centers around creating diagnostics and therapeutics for AML patients. Input toward this goal is provided by programs focusing on stem cell transplants and leukemia, mouse models of AML, the Siteman Cancer Center, and the genome sequencing center.

Numerous genomic discovery tools are required to detect the full complement of genetic alterations associated with AML. These tools include lower-resolution techniques, such as cytogenetics and spectral karyotyping, medium-resolution approaches such as SNP analysis, BAC array CGH, and RNA profiling, and higher-resolution tools such as DNA sequencing. DNA may be obtained from either primary samples or from cell lines. Advantages of cell lines include the "unlimited" quantities of DNA and the lack of heterogeneity. However, cells and/or passaged cells may represent subclones of a sample that are selected for growth *in vitro* or that contain additional mutations that permit immortalization.

At the WUSOM, AML patients provide skin and bone marrow biopsies for RNA preps, profiling, and RNA databasing as well as DNA preps and sequencing/array CGH. Patients also enable the collection of clinical data on treatment and outcomes. All of these data are captured in a clinical database maintained by bioinformatics and biostatistics core facilities.

The current approach to sequence genomes from AML begins by arraying CGH with BAC tile path clones. Preliminary experiments have been carried out using a mouse chromosome 2 BAC tile path queried with mouse AML models to develop techniques and benchmark sensitivity and specificity. This experience can then be used to examine patient tumor samples via DNA analysis with a human chromosome 7/11 composite BAC tile path array. Detection intervals can be elucidated by analysis of differences in signal strengths.

The study design involves a pilot set of 46 samples used to establish technical feasibility (Ley, *et. al. PNAS* 2003;100:14275–80). A discovery set (n = 46), designed principally to survey 450 genes, is currently in process. This prospective study examines samples from 11–12 patients from the hospital's leukemia program who represent each stage of disease classification

(e.g., M0/1, M2, M3, M4) with matched germline samples. The final step will be a validation set of patients (n = 94) in which AML target gene categories are specified.

The current GSC mutational profiling pipeline utilizes PCR with tailed primers to sequence products. Data is analyzed through customized informatics integrated with standard production process software. Using this approach, polyphred tags are verified manually, making data analysis for mutation and indel detection labor intensive.

Pitfalls of the current approach include the following:

- A PCR-based approach requires a separate production pipeline, including separate primer and template validation/QC infrastructures.
- PCR-based approaches tend to focus almost exclusively on exonic and putative regulatory regions, so the sequence that is potentially informative is not assayed.
- Certain genes (e.g., the Hox family) are not easily accessible by PCR.
- Manual review of assembled sequences for mutation/indel detection is time-consuming and demanding of personnel.

There are also challenges to cancer genome analysis:

- Alternatives to PCR-based approaches are needed, such as a simple, low-coverage genome sequencing approach that is time- and cost-efficient and more likely to target relevant regions.
- Sample abundance is a limiting factor.
- A cancer discovery database is needed to integrate information on each patient from disparate sources and to enable intelligent data interpretation and hypothesis generation.

Session 2: Description of Basic Sequencing Technologies

Moderator: Joe Gray, Ph.D.

Analyzing a Tumor Genome by Traditional Sequencing and Other Methods

Richard Gibbs, Ph.D.
Director
Human Genome Sequencing Center, Baylor College of Medicine

The genome is incredibly plastic, and researchers are currently positioned at the tip of the iceberg with respect to unlocking its intricacies. Shotgun sequencing has been used in the sequencing of three mammalian genomes. PCR/directed resequencing has been in use since 1988 and has proven successful when evaluating small changes in most of the genome or in targeted areas. However, the technique does not allow whole genome sequencing or the resolution of "difficult" sequences, such as duplications. Cost per base is the guiding factor in choosing this approach; a genome with 1X coverage currently costs approximately $3M.

Current 454 sequencing methods can be applied to analyze human BACs. However, BACs may be difficult to work with and may be harder to sequence than plasmids. BACs can also be pooled for sequencing using clone-array pooled shotgun sequencing (Cai *et.al*, *Gen Res* 2001;11:1619-23).

Pooled genomic indexing (PGI) can be used for pooled sequencing for BAC mapping and breakpoint detection. One PGI "short tag" read can identify eight locations on the genome. Moreover, short tags and full-length reads can identify the same loci.

One proposal for sequencing the cancer genome is to use short-tag PGI to locate breakpoints in tumor cell lines. The technique can be combined with "padlocks"[1] and molecular inversion probes (MIPs; ParAllele) to investigate allele copy number and to genotype triploid DNA. MIP can be used to detect LOH and other local changes in copy number.

Such a strategy will provide a shotgun sequence for coverage of the whole genome, a pooled, clone-based approach to help find and resolve "difficult" regions, and a quantitative dosage scan to locate LOH.

Microarray Chip Technology

Janet Warrington, Ph.D.
Vice President
Clinical and Applied Genomics Research and Development, Affymetrix

Current breakthroughs in technology have required us to ask, "What is the answer desired at the end of a study?" Multiple formats (e.g., expression analysis, gene copy, LOH, transcription regulation, RNA binding sites) are now readily available, and throughput and scalability continue to expand with the advent of high throughput arrays and CustomSeq™ Resequencing Arrays.

The high throughput array (HTA) platform has scaled up current technology by:

- Industrializing downstream applications
- Packaging GeneChip® technology in microtiter plates
- Provides benefits such as high-throughput automation, increased reproducibility and ease of use, and significant cost reduction
- Will be expanded to DNA analysis offering

With the HTA (6mm x 6 mm microtiter format), two technicians (@ 5 plates/day/technician) can generate 960 U133+ expression profiles, 10 entire transcritpome scans at 15 bp resolution, and 35 million SNPs.

CustomSeq® Resequencing Arrays also allow scale-up of sequencing capacity. As the feature size on arrays is reduced, the price point continues to decrease. Current capacity of the high-density wafer is approximately 65 million probes; one technician who uses three wafers can generate the same amount of information as 100+ DNA sequencers in a 24-hour period. These high-density arrays have been used to identify a fraction of all common human chromosome 21 SNPs and to observe directly the haplotype structure of these SNPs (Patil *et. al, Science* 2001;294:1719–23). Other genome-wide analyses made possible using these techniques include gene copy, LOH, transcriptome, transcription factor binding sites (Cawley, *et al. Cell* 2004;116:499–509), RNA binding protein sites, sites of chromatin modification and DNA methylation, and the chromosomal origins of replication.

[1] Hardenbol P, Baner J, Jain M, Nilsson M, Namsarnev EA, Kurlin-Neumann GA, Fakhrui-Rud H, Ronaghi M, Willis TD, Landegren U, et al. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol* 2003; 21:673-678.

This technology can be used for the following outcomes:

- Adequately powered studies providing useful information for understanding risk, cause, progression, or the likelihood of relapse, survival, or response to treatment
- Correlation of the many levels of molecular information from DNA, RNA, protein, regulatory and structural elements, and their impacts on stage, type, and outcome
- Model development for integrating molecular and clinical information into meaningful, clinically useful output

In summary, Dr. Warrington offered the following suggestions for optimizing the potential of this technology:

- Leverage economies of scale.
- Keep in mind that the real cost is in well-characterized patient samples.
- Work backwards from desired outcomes.

Single Molecule Sequencing: Novel Methods for Whole Genome Sequencing

Shaun Lonergan, M.S.
Vice President of Business Development, Sales and Marketing
454 Life Sciences, Inc.

There is currently a need for a robust, cost-effective, and high throughput sequencing "system" that must be fully integrated and complete and exceed current capacities.

Four areas of concern exist with respect to nano/micro technologies.

- **Sample preparation**: reduction in sample size, automation of sample preparation to reduce effort and costs and increase throughput
- **Physics of nanoscale and smaller assays**: single molecule detection and handling, DNA structure, fluidics at smaller scale, overcoming signal-to-noise limitations at the single molecule level
- **Biological systems integration**: Enzyme selection and modifications for small-scale reactions, signal molecule development
- **Data management and bioinformatics**: innovating data management speed and volume to handle sheer mass of data points collected in a short time span, innovating bioinformatics to process signal data and complete assembly with equal speed, *de novo* assembly of small and large genomes from shotgun reads

The majority of commercialization of technology development is occurring in academic centers, although some companies are becoming involved. Single molecule sequencing can be divided into two classes of commercial technology. Solexa, 454 Life Sciences, and VisiGen are involved in sequencing by synthesis. Helicos, Eagle Research and Development, and Advanced Research Corporation are involved in nanopore technology.

A "whole system," from sample prep to the bioinformatic assembly of sequence fragments into a genome sequence, is required for a whole-gene approach to sequencing. 454 Life Sciences has recently developed a proprietary system for gene sequencing that begins with one sample per genome. Individual DNA fragments are then applied to picotiter plates (up to 1.6M wells/plate) and subject to massively parallel sequencing. Instrumentation and software support integrated microfluidics, field-programmable gate array (FPGA)-based signal processing, and real-time

basecalling. Bioinformatics capabilities support high-speed genome assembly and sequence analysis, either through resequencing with single-ended sequencing using a proprietary genome assembler, or through *de novo* sequencing with double-ended sequencing using a proprietary genome assembler. The system enables simultaneous sequencing of the entire genome, sequencing-by-synthesis, and pyrophosphate signal generation.

This approach creates the first scalable sequencing method and has contributed the first whole genome sequence (adenovirus) to GenBank by a new method in 25 years. In this proof-of-principle example, massively parallel sequencing was conducted using one 300,000 well picotiter plate, giving 100% coverage with greater than 99.97% accuracy. Instruments will be available for sale to collaborators in 2004.

(A question was asked about the sensitivity and costs associated with this method. The answer is that sensitivity decreases with longer homopolymers and the actual cost per base pair is 0.1 cents/base. In the future, decreases in the pitch size of the wells in the picotiter plate should enable a single-pass human genome sequence for approximately $3M. Although 10X coverage is achievable with this method, physical constraints preclude using this approach to sequence an entire genome for $1,000.

Another participant asked about the success rate of the methodology at first time. The technology governing the bead emulsions used in the process continues to evolve. The current amount of starting material required is approximately 15 μg, which is then nebulized to produce 200 ng of assayed material.

Massively Parallel Signature Sequencing

Thomas Vasicek, Ph.D.
Vice President, Business Development
Lynx Therapeutics, Inc.

Massively parallel signature sequencing (MPSS) allows the capture of a large number of DNA molecules that have been tagged with a unique oligonucleotide. Tagged fragments are amplified and hybridized to a population of one million beads (each bead has 10^5 identical molecules). This approach enables the sequencing of one million DNA clones in parallel, and signatures can be counted and assigned to the genome or to transcripts.

Features of chromosome breakpoint analysis using this approach include:

- 180K BamHI BAC library with insert size of 130–170 K at 9X coverage of haploid genome (3X for triploid genome)
- Paired BAC-end sequences uniquely identify breakpoints (BAC-insert ends can be joined to form a 30-bp ditag; 97% of these uniquely to genome)
- Arraying individual BACs is unnecessary; individual breakpoints may be amplified by PCR for sequencing
- Cost: less than $0.03 per ditag sequenced

Considerations for sample selection (e.g, cell lines versus tumor samples for proof-of-principle and process optimization) using this approach include:

- Many cancer cell lines are extensively characterized
- Many cancer cell lines are relatively less heterogeneous than primary tumors

- Cell lines are immediately available
- Cell lines as evolutionary systems
- Known cell line complexity offers the depth required for coverage

The technique can also be used to sequence restriction fragment ends for prospective SNPs. Also, resequencing is possible through genomic DNA capture and tagging, whereby genomic DNA is subjected to hydro-shearing, followed by ligation of blunt ends and MPSS analysis. Furthermore, recent work has suggested that MPSS can be used to map candidate DNaseI hypersensitive sites.

MPSS applications include:

- comparative genomics (e.g., BAC end sequencing (chromosome breakpoint mapping and whole genome sequence assembly), re-sequencing, SNP discovery and genotyping, and insertion element mapping)
- epigenomics (e.g., DNaseI hypersensitive site mapping, chromatin immunoprecipitation, and methylation mapping)
- aptamer and phage display library characterization
- gene expression (e.g., mRNA, micro-RNA (siRNA identification and profiling))

Discussion

One participant commented that there seems to be a consensus that identifying drug targets in cancer is a primary mission. The capacity to identify activating and inactivating mutations in important druggable genes (those coding for any enzymes that are implicated in tumor pathogenesis) is crucial.

Another participant noted that druggable targets that will be identified in the future can not be known at present. Therefore, it may be useful to begin with enzyme families, such as the kinases, that have known roles in the pathogenesis of cancer. From this point, studies could be extended to look at all enzymes and, ultimately, all genes. Moreover, a tradeoff must be made with respect to the number of tumors/genes sequenced; investigation of a larger number of tumors will provide a solid assessment of the major activating mutations. Using this strategy, druggable subtypes will be identified. This strategy is amenable to a tiered system with pilot projects that include rapid, real time data release mechanisms.

Participants suggested several key requirements and outcomes for a tumor genome sequencing project, including:

- proper annotation of tissue and an understanding of related tumor biology.
- uniformity in classification of tumor progression in humans.
- gene analysis available for each patient's unique diagnosis (a long-term goal from the cancer diagnostic viewpoint).
- technology development toward an inexpensive genome for analysis as part of individual diagnosis.
- enhanced SNP identification capabilities.

One attendee noted that superimposed perspectives currently challenge the cancer community in assembling data into meaningful analyses. Clinical data is often annotated using relatively crude classifications (e.g., life, death, development of cancer). Unfortunately, much of the excitement of today's discussion will be lost in translation because there is no clear outcome defined in the clinic. However, electronic patient information provides the ability to examine all information

about a particular patient as part of a clinical data set. It is essential to look at forms of data that can remove biases. The subsets of clinical data that can be queried as "genetic data" will be crucial. A hierarchy must be included for the available structural data.

Another participant commented that recent progress in sequencing technologies holds great promise. Techniques of the past decade have focused on shotgun sequencing, but the ability to resequence targeted areas suggests new directions. This pilot project could be a driver for resequencing. Some may suggest that resequencing is a distraction, since the whole genome will eventually be sequenced. However, the challenge is to sequence regions of interest. This pilot project is a natural partnership between the cancer and genome-sequencing communities, and it must be viewed as the initial commitment to a decade-long project.

One participant observed that many tumors, such as gliomas, exhibit dramatic heterogeneity, complicating the analysis of even a single tumor sample. There may be tumor types that will be refractory to this type of analysis, and it will be important to be selective when choosing targets for study in a pilot project.

However, different technologies may be melded nicely. By starting with model systems, an array of candidates may be identified that can then be tested using specific systems. Cell lines can complement information gleaned from primary tumors, as these lines can be screened for identified pathways or specific genes.

One attendee suggested that the effort should begin with what is known about the genes, as this offers the greatest opportunity for therapy. Next, the costs must be defined clearly in terms of scientific knowledge lost in exchange for focusing. "Noise" in the data is no different than that generated from other techniques, and recognizable criteria must be established for this project as they would be for other approaches.

Another attendee suggested that colon cancer, in which chromosomally unstable variants do not appear to have a mutator phenotype, may be a useful model to evaluate background mutation frequency in a given tumor type. A suggestion was made to consider prioritizing entire coding regions within the genome, as it is not known how pathways are formed and how rates of mutation vary for elements within a pathway. It was also recommended that additional variations (e.g., amplification without modification) be considered when addressing large-scale genome content.

One attendee commented that heterogeneity represents a potential source of resistance to therapy. As such, the term "tumor genome" is a misnomer. More information can be gleaned by using "genomic instability" as a verb, through considering the very process of how genetic mutations arise as a druggable target. A word of caution was also offered regarding the use of cell lines, namely that a cell line can be manipulated to behave in a desired fashion. A cell line tells what is possible, but not necessarily what selection pressures take place *in vivo*.

One commenter noted thatit is essential during the planning phase to think more prospectively by identifying sets of patients in whom cell lines and primary tissue are available for different points of disease progression. By starting with well-characterized tumors, specific characteristics in genetic instability can be dissected.

Session 3: State-of-the-Science in Sequencing Various Tumors

Moderator: Gregory Riggins, M.D., Ph.D.

Presenters in this session highlighted the latest technologies for sequencing various tumors that may serve as pilot projects for this initiative. Prior to the meeting, presenters were asked to consider the following four questions in the context of the tumor types featured in their presentations:

1. What is the current state of sequence determination of human tumors, and are these studies yielding insights that are informative for diagnostic or therapeutic development (i.e., molecular targets)?
2. What methods are being used to determine sequence aberrations among human tumors in your field?
3. Are phenotypic/genotyipc subtypes of cancer a significant issue for sequence determinations for this cancer?
4. If a sequencing project for human tumors were to be conducted, are their currently ongoing clinical studies that have specimens and clinical data that would be informative? Is laser capture/microdissection needed, and are the tissues suitable?

Prostate

William Sellers, M.D.
Assistant Professor of Medicine
Dana-Farber Cancer Institute, Harvard Medical School

The clinical problem with the treatment of prostate cancer speaks to the heterogeneity of the underlying tumor and the heterogeneity of its clinical outcome. The difference between annual incidence and annual mortality is relatively large, although not due to success of treatments. The genetic classification of prostate cancer will allow stratification of existing therapeutics and define the dependence of prostate cancer on genetic alteration, which could lead to target identification and drug discovery and development.

Current efforts to understand the genetics of prostate cancer include sequencing of nucleotide changes, investigation of copy number changes (through array CGH, SNP arrays, representational display analysis (RDA), and digital karyotyping), and investigation of structural changes.

Project and proposal:

- Use 100 tumors (discovery set) representative of lymph node metastases
- Provide reproduced DNA by whole-gene amplification, SNP arrays, and exon sequencing directed at gene classes and candidate regions

Reasonable success has been achieved using whole genome amplification, followed by copy number analysis using 120 K Affymetrix SNP arrays. Many SNPs can be analyzed for the cost of a single CT scan, and the depth to which clinicians could investigate tumors is remarkable. It must be noted, however, that methods to examine copy number changes should have sub-gene resolution.

Breast

Thea Tlsty, Ph.D.
Professor
Department of Pathology, University of California, San Francisco

Sequencing efforts related to breast cancer began at the stage of predispositionary lesions (e.g., BRCA1). Currently, points further along the progression to disease are being examined through sequencing of specific genes associated with predisposition (genotyping), re-sequencing of identified genes that are functionally implicated, target sequencing of regions identified in tumor process, and pilot projects for assessing widespread regions that are identified as amplified or deleted.

Clinical considerations and biology:

In order to impact outcome, annotated data must be linked to various endpoints through the integration of clinical and biological data. Lesions must be sequenced within several complementary contexts, including occurrence, process of mutation, and reaction to therapy.

Sample choice (e.g., tissue versus cell lines) is crucial to inform about the actual process, instead of merely identifying a set of changes. Investigating early stages in the tumor process will provide added benefits. The number of genomic alterations is low in early stages of the disease process, followed by an exponential increase late in the process. This contributes to a difficulty, namely that significant genomic instability is present when the cancer is diagnosed. At this point, identification of a clear target is obscured and the tumor becomes more resistant to therapy.

Breast tumor tissue must, therefore, be considered in terms of its overall context. The epithelial response to the environment provides possible targets, and future sequencing efforts should investigate hereditary susceptibility (beyond BRCA 1 and 2), target resequencing, and genome-wide methylation changes.

Brain

Howard Fine, M.D.
Chief
Neuro-Oncology Branch, National Cancer Institute, National Institute of Neurological Disorders and Stroke, National Institutes of Health

Primary brain tumors are the leading cause of cancer-related mortality in children under age 18 and the third leading cause of cancer-related mortality in males 18–54 years old and the fourth leading cause in females between the ages of 18 and 34. Gliomas are the most common primary brain tumors, with few therapeutic advances during the previous three decades. Medulloblastomas are curable in 50% of affected children, although significant lifelong neurocognitive, neuroendocrine, and physical deficits occur.

Because the brain is composed of numerous cell types, different progenitors can give rise to a variety of tumor types. Genomic analysis of glial tumors has been limited:

- Historically limited to karyotyping
- Individual genes (e.g., EGFR, PDGF, p16, MDM2) assessed for expression and sequence

- Significant work with CGH over the last decade
- Limited amount of recent gene expression data (e.g., SAGE, microarray)

Two molecular pathways that lead to glioblastoma have been identified. As the cells of origin for these pathways are not known, the pathways may be completely independent. Expression data have identified at least two types of glioblastoma based on the quantitative expression levels of EGFR. These tumors also differ clinically and genetically from anaplastic oligodendrogliomas.

Current genomic efforts in brain tumors can be summarized as follows:

- Continued work in CGH
- Continued efforts in gene expression profiling
- Traditional chromosomal walking for deleted genes (e.g., 1p, 19q, 10q)
- Individual gene sequencing
- The Glioma Molecular Diagnostic Initiative (see below)

The Glioma Molecular Diagnostic Initiative (GMDI) is a joint effort between the Neuro-Oncology Branch of the National Institute of Neurological Disorders and Stroke (NINDS) and the NCI. The goal of the initiative is to create, within three years, a publicly accessible Web-based glioma database and informatics platform consisting of in-depth pathologic, molecular, and genetic data with detailed clinical corollary data for hundreds of individual brain tumors.

The GMDI currently supports the following initiatives:

- A national study (designed to include more than 1000 patients) through two NCI-funded brain tumor consortia comprised of 16 institutions
- Extensive prospective clinical data to be correlated with molecular data, including:
  - gene expression profiles (microarray: "Glioma Chip"; Affymetrix U133 2.0 chip)
  - chromosomal abnormalities (CGH and Affymetrix 100K SNP array)
  - germ line genetic variations (SNPs)
  - genetic mutations (high-throughput sequencing)
  - exploratory proteomics (with Lance Liotta)
  - tissue arrays to be generated for IHC

Colon

Victor Velculescu, M.D., Ph.D.
Associate Professor
Oncology Center, Johns Hopkins University School of Medicine

Digital karyotyping can be used to analyze short sequence tags from genomic DNA and can be linked for rapid sequence-based analysis for matching with chromosomes. The technique can identify regions of amplification/deletion on chromosomes.

In colon cancer, there is a genetic amplification of thymidylate synthase (TS) in metastases that are resistant to 5-fluorouracil. Metastatic patients who exhibit TS amplification have a lower rate of survival. Early identification of an amplified level of thymidylate synthase may have therapeutic benefit.

Somatic mutational analysis of gene families (e.g., tyrosine kinases, PI3 kinases, chromosomal instability genes) will be valuable for understanding the progression of colorectal cancer.

Background somatic mutations are present in colorectal cancers at the rate of one to two somatic mutations per Mb of tumor DNA, implying that approximately 10,000 alterations are present per genome (with more than 99% being nonfunctional). The prevalence of these mutations is consistent with that observed in normal cells, indicating that there is no mutator phenotype in these tumors.

Future directions include:

- extending digital karyotyping analyses to identify recurring genetic content alterations.
- conducting mutational analysis of other gene families.
- determining roles of mutated genes in tumorigenesis.

Lung

Matthew Meyerson, M.D., Ph.D.
Assistant Professor of Pathology
Dana-Farber Cancer Institute, Harvard Medical School

Lung cancer is the leading cause of cancer death in men and women, with a five-year survival rate of 15–20%. Current chemotherapy is largely inadequate, and new targets are needed. Dramatic responses to Iressa® (gefitinib; AstraZeneca Pharmaceuticals) have been observed recently in some patients with certain types of lung cancer, although these effects are not understood. It can be inferred that lung cancer is highly heterogeneous; different, unidentified subtypes may harbor different causative mutations.

Systematic kinase sequencing has identified two cancer-specific mutations in the EGFR gene, both in non-smoking women with adenocarcinoma (a subpopulation of patients who responds to treatment with Iressa). Additional EGFR mutations have been identified that suggest that EGFR mutation may predict response to Iressa®. The implication for the cancer community and the pharmaceutical industry is that "rare" mutations may represent clinically significant therapeutic targets.

Recent work using exon resequencing in lung cancer tumor tissues indicates that the combination of genomic technology and targeted therapy provides a unique opportunity to transform the understanding of cancer pathogenesis as well as cancer treatment.

Pancreas

Michael Hollingsworth, Ph.D.
Professor
Eppley Institute, University of Nebraska Medical Center

Pancreatic cancer is extremely lethal, with a two percnt survival rate in two years following diagnosis. Ninety percent of cases demonstrate mutations in kRAS; thus, some information about the biology has been elucidated. Pancreatic cancer is a polygenic disease, and we must look beyond tyrosine kinases to the entire expressed gene profile.

It is not known why the lethality rate is so high in pancreatic cancer and what factors are being generated that alter the body's function and response to this disease. Sequencing the tumor genome for pancreatic cancer will require a successful program for the harvesting of organs.

Tissue acquisition will be key and an organ donation program for research, such as the Rapid Autopsy Program, will be essential.

Lymphoma

Louis Staudt, M.D., Ph.D.
Chief
Lymphoid Malignancies Section, Metabolism Branch, National Cancer Institute, National Institutes of Health

Diffuse large B cell lymphoma accounts for 40% of non-Hodgkin lymphomas. With approximately 23,000 new diagnoses and 10,000 deaths per year, this cancer has a 40% cure rate. The disease is relatively heterogeneous and may actually represent three independent diseases (not subtypes), each with a unique cell of origin. These lymphomas exhibit a mutator phenotype, somatic hypermutation of immunoglobulin genes, with the following characteristics:

- Mutation is initiated by a germinal-center, B cell-specific protein, AID.
- Mutation rate of immunoglobulin genes is 10-3/bp/generation.
- Mutation machinery is also directed to some non-immunoglobulin genes.
- Mutations can accumulate in proto-oncogenes in germinal center B cell-derived lymphomas.

Efforts to understand this complex cancer must be carried out in conjunction with large-scale clinical trials. Clinical data are crucial for understanding the molecular alterations in lymphoma. Also, the Lymphoma/Leukemia Molecular Profiling Project is a useful resource.

Suggestions for sequencing the diffuse large B cell lymphoma genome include the following:

- Initially tackle a smaller number of druggable targets in a larger number of cases.
  - Pathogenetically distinct subgroups may yield different mutational targets.
  - Further heterogeneity in survival suggests further molecular diversity.
  - Specialized high-frequency mutational mechanism known to hit oncogenes and potential drug targets.
- Leverage the availability of tumor specimens from clinical trials.

Mouse Models of Human Cancer

Tyler Jacks, Ph.D.
Director
Center for Cancer Research, Massachusetts Institute of Technology

The ability to manipulate the mouse genome offers a controlled platform to investigate the roles of various genes in cancer-relevant processes. The NCI-sponsored Mouse Models of Human Cancer Consortium (MMHCC) builds and characterizes mouse models of major human cancers (e.g., gastrointestinal, prostate, lung, ovarian, mammary) and promotes mouse model development more generally by:

- developing and/or disseminating technology and research tools.
- establishing a cancer repository of mouse models.
- facilitating interactions between modelers and industry for preclinical testing.

- organizing modeling/characterization workshops.
- establishing an MMHCC web site (eMICE; http://emice.nci.nih.gov) and Mouse Models Database.

The eMICE web site provides a portal into information stores about cancer models and their applications, with links to the NCI Mouse Repository and the Cancer Models Database (caMOD), the Cancer Images Database (caIMAGE), and the Cancer Array Database (caARRAY). It also offers disease-specific tutorial web pages.

The MMHCC collaborates extensively with the NCI CB Models Infrastructure to provide the vocabulary, pathology, and other descriptive data needed to integrate human and model system cancer research. The Consortium is also a core element of a new NCI CB pilot project, the caBIG Consortium, which includes NCI Cancer Centers, SPOREs, the NCI Intramural Program, specific initiatives such as the Cancer Genome Anatomy Project (CGAP) and Clinical Cooperative Groups, and other biomedical groups and consortia. All of the MMHCC groups are located in NCI Cancer Centers that are included in the first phase of caBIG.

Mouse models have been used recently to analyze tumor progression in K-ras initiated lung tumors. A gene set enrichment analysis (GESA) procedure revealed an increased expression in K-ras signature genes in K-ras mutant tumors. Furthermore, mouse expression profiles have shown a high correlation with human adenocarcinomas that have K-ras mutations, suggesting that mouse models are useful for comparative analysis of comparable human tumors. In addition, human profiling experiments can be used to inform the mouse model, and the mouse can serve as a reference point to clarify confusing human data (e.g., copy number changes).

Advantages of human cancer/mouse model cross-validation:

- Multiple models are available at each site (e.g., prostate, pancreas, lung, ovary, breast, colon, brain, leukemia) based on common human signatures.
- Samples are available in abundance.
- All are well-characterized histopathologically.
- Many models have substantial datasets that include IHC, CGH, SAGE, gene expression profiling, serum and tissue protein, and karyotyping.
- Gene sequencing of appropriate mouse tumors in parallel with corresponding human tumors is underway at a low level.
- Resequencing of large regions of LOH/CAN in mouse that are shared with human tumors can be highly informative.

Susceptibility and resistance:

- Exceptional proof-of-principle data illustrate the feasibility of analyzing epistasis and identifying interacting genes.
- Acceleration of human/mouse integrative susceptibility research is feasible if haplotyping were broadly available.

Interventions-related discovery:

- Ongoing preclinical trials on validated mouse models
- Samples of responsive/recurrent tumors can be readily collected
- Allows for investigation of genetic basis of sensitivity and resistance

Discussion

One attendee noted that a large number of samples are required to correlate data on tissue samples with clinical data. When expression analysis became a viable technique, many investigators used the approach to examine readily available samples in a haphazard fashion. Useful sequencing of a cancer tumor genome requires a planned process in which samples are frozen, with consent, and bundled with other useful information. The NCI has collected serum and plasma for several decades using an organized system to capture tumors. A similar strategy will be applied in this pilot project. One attendee suggested that the NIH help to establish central databases (e.g., CGAP) for cell lines and tissue samples.

Another participant noted the vast contrast between the scientific rigor of sequencing techniques and the variation in clinical medicine classifications. Currently, the analysis and tracking of patients is imprecise, leading to an intra-observer variation of as much as 15%. Different objective criteria, such as mathematical image analysis of spectra, are needed to classify tumors in the clinic. Tumor collection must be uniform and characterization of the collected tumors must be rigorous. Numerous variables must be addressed before sample analysis can commence. One attendee commented that the technology must also be improved, although the Glioma Molecular Diagnostic Initiative demonstrates proof-of-concept that pilot projects can be carried out today using imperfectly characterized and complex samples.

One commenter noted the crucial importance of early strategic planning. Project organizers must decide whether the effort to understand the cancer genome will focus primarily on research or on discovery. It was suggested that the approach taken by the Sanger Center could be emulated, by analyzing a discovery tumor set that will reveal areas of focus when analyzing annotated data sets. Using this strategy, a brief discovery effort will usefully inform future efforts.

It was also noted that clinical parameters may be helpful when ascertaining the most useful properties of a given data set. Even retrospectively-annotated tumor samples that have not been rigorously prepared have generated insights into clinical behavior.

Dr. Barker summarized by noting that several projects are evolving from this discussion. The NCI is interested in using genomic information to change lives in the clinic. However, there is also an opportunity to create a taxonomy for the entire genome, although the required technology may develop more slowly and possibly change over time. The NCI currently has a wealth of resources, including ongoing Phase III trials, SPOREs, and databases, that can be utilized in this effort. On any given day, thousands of cancer researchers carry out the types of experiments that have been discussed in this workshop. Change will be slow unless some level of uniformity is applied to the process, and a shift from hypothesis-driven individual research to a coordinated, "Big Science" strategy will be required. She noted that she has been greatly encouraged by the discussions, which may lead to several future projects.

General Group Discussion

Moderator: Maynard Olson, Ph.D.

Dr. Olson remarked that it is important to step back for a moment and consider why this discussion is occurring. Characteristics of this particular moment include the following points:

- An important problem has arisen.
- We appear to be at an historical inflection point.

- Established technologies are orders of magnitude too expensive.
- Many highly promising newer technologies are in various stages of development.
- An open-ended potential for technological improvements exists.
- Pilot-project/evaluation cycles are needed before a major scale-up occurs.
- Scientific logic favors lighter analysis of many samples rather than deep analysis of a few.
- We are faced with numerous biological and epidemiological uncertainties.

To seize this moment, the following points should be considered:

- There is a remarkable sense of excitement and opportunity.
- Creative science policy would capture this sense and do something new, ambitious, and well-defined.
- A successful initiative would need to achieve a delicate balance between focus and breadth.

The Alta Conference (1984), convened to discuss the long-term effects of radiation from World War II, offers several lessons that can be applied to this workshop. First, we will not likely converge on a winning idea this afternoon. Second, funding agencies should take the lead in building on this workshop with continuing input from the scientific community.

Participants continued with a discussion of possible strategies and starting points for a cancer genome sequencing pilot project. Related suggestions are grouped thematically and detailed below:

Begin by selecting events instead of anatomy

Current clinical understanding is empirical, and setting events in place to model disease is a step toward the treatment of malignant disease. These events must be linked to the most completely annotated data available. By approaching the problem in this manner, the project will capture the imagination of clinical biologists and have a transformative effect. However, the topic chosen must be important. For example, sufficient tissue representing a single metastatic site that covers all important clinical events and endpoints can be acquired within a relatively short time period. Prioritization of acquisition and analysis would be based on the quality of the sample and its potential to generate useful data, important thematic events, and acquisition parameters. This scenario must be conducted carried out knowing that it is a starting point for an emerging initiative.

Pattern the pilot project on the Sanger Center model

Examples of viable projects include the investigation of EGFR and its possible predictive impact on lung cancer. Also, reasonable targets for small molecules have been identified in colon cancers. By choosing an area such as these, less information is required up front to begin the process.

Use this project as a catalyst

Because this pilot project cannot provide an enormous scope of outcomes, it may be useful to decide whether it will yield primarily biological insight or clinical impact. Some funding could be directed into sequencing technologies and software that will identify diverse signatures in heterogeneous tumors. It will also be useful at this stage to set intermediate goals that will challenge the technology.

Focus on pharmacological selectivity of tumors

A richer holistic understanding is needed to refine current models of cancer. For a treatment perspective, the community is currently poor in effective drugs, not in targets. Focusing on the factors that govern drug efficacy will spur thinking about endpoints.

Select an area of focus (e.g., a protein family)

Many disperse activities can be consolidated to maximize efficiency and efficacy. One strategy is to select a thematic area, such as the kinase family, and tackle it in an ordered way. Doing so will force the community to develop standards and to articulate the biggest unknowns, which will ultimately streamline the effort through economies of scale.

Immortalize cancer genomes in a public resource

Re-creating the Sanger Center approach may preclude the understanding of certain aspects of cancer genomes, such as mechanisms of deregulation. By immortalizing cancer genomes so that changes in DNA are captured in an accessible form, a public resource can be created that will enable the research community in a variety of projects. Immortalizing and organizing cancer genomes, perhaps through centralized BAC clone libraries, should be considered as part of this effort.

Organize only around the ultimate goal

The HGP demonstrated the advantage of maintaining a diversity of approaches, all of which supported one well-defined goal. Maintaining multiple approaches will promote the greatest number of innovations and breakthroughs. Seeding a few BAC projects will be a key step in developing an early portfolio of approaches. However, the three or four main trajectories for progress must first be determined, with the understanding that they may converge at a later point.

Devise a strategy that will produce profitable spin-offs.

Arrange a portfolio based on several "menus"

Four menus should be considered when arranging a portfolio for a pilot project:

1. What technology will be used to accumulate genomic data (e.g., sequencing techniques)?
2. What will be identified or measured (e.g., whole candidate genes, exons only, a short list of likely candidates, the entire genome)?
3. What are the biological materials (e.g., cell lines, primary tissues (early, late lesions, metastasis))?
4. What other kinds of data are desired from these materials (e.g., proteomics, information about pathways)?

Many parallels exist between this pilot project and the HGP. Thus, an agenda similar to that of the HGP, complete with milestones that are slightly beyond reach, can be created with the confidence that the technology will improve to meet the goals.

Determine deliverables and ways to engage the imagination of the patients and public

Many clinicians are unsure of how to translate basic science results into the clinic. Moreover, the general public must be convinced that these tests will be a useful component of the armamentarium of diagnostic techniques. Deliverables that can be understood by both clinicians and patients will be necessary. Although technology and sample sets will vary, funding is likely to remain constant. Thus, a rational reason for undertaking this effort is needed now. The HGP had a clear endpoint, which helped to translate the importance of the work to the public. We must

define clearly what we are trying to achieve with this effort and design explicit yet achievable milestones and timelines.

Find or create opportunities for synergy

One of the most important lessons learned from the HGP was that great ideas, which capture the imagination, will produce funding. However, government representatives must be made to understand that this project cannot be shoehorned into the current NIH budget. Nonetheless, the ability to say that we know of all of the changes that cause cancer will be an incredible engaging principle for creating synergy and partnerships.

Build upon past successes

This is an unprecedented opportunity to redefine disease based on aligning clinical investigation with basic science. The cancer community has made a difference in the lives of many cancer patients, despite an incomplete understanding of all of the relevant genetic nuances of the condition. By building upon current successes, this project will allow major impacts on the understanding and treatment of cancer in the next three to five years. Several developments, including the discovery of Bcr-Abl in leukemia and the success of Iressa® in select lung cancers, can be used as "bullets of success" for the identification and development of drug targets.

Retain focus on long-term objectives

One participant noted that the long-term vision, to identify the full complement of genetic changes implicated in cancer, must not be disengaged from clinical outcomes.

Create a working group to align biological research and clinical outcomes

Aligning mutations with disease progression will be critical to the success of this project. Currently known and clinically meaningful insights can be used as the first phase of the project, bundled with the longer term goal of sequencing and fully interpreting the cancer genome. Starting with clinically meaningful points will leverage the power of the NCI and NIH. It is likely that more than one project will arise from the discussions of this workshop. The NCI should think more about how to unite the cancer community around shared concepts, such as kinases or proliferation, to align the biology research with a sequencing project. Also, proof-of-concept will be necessary to bring clinicians on board; it will be necessary to generate output that can be related to early disease. This meeting has provided a consensus for direction, but it remains to be seen how the biology impacts this direction. A working group is needed to address this alignment of biology with sequencing.

One attendee cautioned against one working group or agency making any central decision as to which genes to sequence or which methods to select. Competitive review (e.g., a Request for Applications) provides a time-tested route to solicit the best proposals, especially when the goal is to get the most important information available from tumors. Although the overall goal should be central, the organizational principles should result from competition. Using this structure, the community can unite and harness the energy of a generation.

Keep the long-term goal in sight

A grand vision is essential to keep efforts unified and organized. It will thus be useful to periodically assess the greater biological or clinical questions that the effort seeks to address (e.g., What is the taxonomy of cancer? How can we determine which patients should receive certain therapies?)

Consider the earliest changes in disease progression and induction of metastasis as possible organizing principles

Drug targets that focus on changes that take place during the earliest stages of disease progression may represent a "home run strategy" to prevent cancer altogether (e.g., Barrett's esophagus). At early stages of disease, fewer drug targets exist. Moreover, some tumors, such as brain tumors, never metastasize, and an exploration of the factors that prevent metastasis may have great value in all tumors.

Next Steps and Final Comments

Anna Barker, Ph.D., and *Francis Collins, M.D., Ph.D.*

This workshop has clearly demonstrated that the cancer community has the means and motivation to describe the genetic changes associated with cancer. However, an organized approach is necessary to obtain short, medium, and long-term clinical benefits. There is a clear need to understand cancer at the genomic level and to inform biology. Making a difference in cancer requires that mission-critical questions be integrated into this process in an organized fashion. Although the strengths of individual investigators are not currently harnessed, opportunities exist to change the ways that science is conducted. Colleagues will have to be convinced to use central bioinformatics platforms and share resources and knowledge.

Next steps include:

- The NCI and NHGRI will devise a working group within the near future to generate a portfolio of research enterprises.
- The group will be charged with arranging the portfolio using the "menu model" presented at this workshop.

On behalf of the NCI and the NHGRI, Drs. Barker and Collins thanked attendees for their expertise and enthusiasm. The meeting was then adjourned.

National Cancer Institute and National Human Genome Research Institute

Exploring Cancer Through Genomic Sequence Comparisons

April 14-15, 2004
Bethesda Marriott Hotel
Bethesda, MD

AGENDA

Wednesday, April 14

Salon D & E

6:30 p.m.	**Welcome and Dinner**
7:00 p.m.	**Introduction and Objectives of the Workshop** Anna Barker Deputy Director for Strategic Scientific Initiatives National Cancer Institute Francis Collins Director National Human Genome Research Institute
8:00 p.m.	**The Case for a Tumor Genome Sequencing Project** Joe Gray Lawrence Berkeley National Laboratory
8:30 p.m.	**Discussion**
9:00 p.m.	**Adjournment**

Thursday, April 15

Salon D & E

7:00 a.m.	**Breakfast**
8:00 a.m.	**A Vision for Finding Genomic Changes in Cancer** Eric Lander Broad Institute
8:30 a.m.	**Session 1: Approaches for Determining Cancer Gene Sequences** Moderator: Joe Gray ***BAC End Sequencing*** Colin Collins University of California, San Francisco

Sequencing of Coding Regions
Richard Wooster
Wellcome Trust Sanger Institute

Approaching Cancer Genomes: Current Practice and Future Prospects
Elaine Mardis
Washington University School of Medicine

9:30 a.m. **Session 2: Description of Basic Sequencing Technologies**
Moderator: Joe Gray

ABI 3730 "Traditional" Large-Scale Sequencing
Richard Gibbs
Baylor College of Medicine

Microarray Chip Technology
Janet Warrington
Affymetrix

Single Molecule Sequencing
Shaun Lonergan
454 Life Sciences

Massively Parallel Signature Sequencing
Thomas Vasicek
Lynx Therapeutics

10:30 a.m. **Discussion**

11:00 a.m. **Break**

11:15 a.m. **Session 3: State-of-the-Science of Sequencing Various Tumors**
Moderator: Gregory Riggins, Johns Hopkins University

Prostate
William Sellers
Dana-Farber Cancer Institute

Breast
Thea Tlsty
University of California, San Francisco

Brain
Howard Fine
National Cancer Institute/National Institute of Neurological Disorders and Stroke

Colon
Victor Velculescu
Johns Hopkins University School of Medicine

Lung
Matthew Meyerson
Dana-Farber Cancer Institute

Pancreas
Michael Hollingsworth
University of Nebraska

Lymphoma
Louis Staudt
National Cancer Institute

Mouse Models of Human Cancer
Tyler Jacks
Massachusetts Institute of Technology

12:15 p.m. **Working Lunch**

12:30 p.m. **Discussion of Session 3**

1:15 p.m. **General Group Discussion**
Moderator: Maynard Olson, University of Washington

3:00 p.m. **Discussion of Potential Pilot Projects**

3:30 p.m. **Adjournment**

National Cancer Institute and National Human Genome Research Institute

Exploring Cancer Through Genomic Sequence Comparisons

April 14-15, 2004
Bethesda Marriott Hotel
Bethesda, MD

PARTICIPANT LIST

Co-Chair
Anna D. Barker, Ph.D.
Deputy Director for Advanced Technologies
and Strategic Partnerships
National Cancer Institute
National Institutes of Health
Building 31, Room 10A-52
MSC 2580
31 Center Drive
Bethesda, MD 20892-2580
(301) 496-1550
(301) 496-7807 Fax
barkera@mail.nih.gov

Co-Chair
Francis S. Collins, M.D., Ph.D.
Director
National Human Genome Research Institute
National Institutes of Health
Building 31, Room 4B-09
MSC 2152
31 Center Drive
Bethesda, MD 20892-2152
(301) 496-0844
(301) 402-0837 Fax
fc23a@nih.gov

J. Carl Barrett, Ph.D.
Director
Center for Cancer Research
National Cancer Institute
National Institutes of Health
Building 31, Room 3A-11
MSC 2440
31 Center Drive
Bethesda, MD 20892-2440
(301) 496-4345
(301) 496-0775 Fax
barrett@mail.nih.gov

Lisa D. Brooks, Ph.D.
Program Officer
Genetic Variation Program
National Human Genome Research Institute
National Institutes of Health
Building 31, Room B2-B07
MSC 2033
31 Center Drive
Bethesda, MD 20892-2033
(301) 435-5544
(301) 480-2770 Fax
lisa.brooks@nih.hhs.gov

Kenneth H. Buetow, Ph.D.
Director
Center for Bioinformatics
National Cancer Institute
National Institutes of Health
Suite 403, Room 4001
MSC 8335
6116 Executive Boulevard
Bethesda, MD 20892-8335
(301) 435-1520
(301) 480-4222 Fax
buetowk@nih.gov

Stephen J. Chanock, M.D.
Senior Tenured Investigator
Director
Core Genotyping Facility
Center for Cancer Research
National Cancer Institute
National Institutes of Health
Advanced Technology Center, Room 134D
8717 Grovemont Circle
Gaithersburg, MD 20892-4605
(301) 435-7559
(301) 402-3134 Fax
chanocks@mail.nih.gov

Deanna M. Church, Ph.D.
Staff Scientist
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
MSC 6510
Building 45, Room 5AS-43
45 Center Drive
Bethesda, MD 20892-6510
(301) 594-5695
(301) 480-2484 Fax
church@ncbi.nlm.nih.gov

Colin Collins, Ph.D.
Assistant Professor
Cancer Center
University of California, San Francisco
Room S-151
2340 Sutter Street
San Francisco, CA 94143-0808
(415) 502-7065
(415) 476-8218 Fax
collins@cc.ucsf.edu

Carolyn Compton, M.D., Ph.D.
Professor and Chair
Department of Pathology
McGill University
3775 University Street
Montreal, Quebec H3A 2B4
Canada
(514) 398-7192, ext. 00515
(514) 398-7446 Fax
carolyn.compton@mcgill.ca

Gregory J. Downing, D.O., Ph.D.
Director
Office of Technology and Industrial Relations
National Cancer Institute
National Institutes of Health
Building 31, Room 10A-52
MSC 2580
31 Center Drive
Bethesda, MD 20892-2580
(301) 496-1550
(301) 496-7807 Fax
downingg@mail.nih.gov

Adam Felsenfeld, Ph.D.
Program Director
National Human Genome Research Institute
National Institutes of Health
Building 31, Room B2-B07
MSC 2033
31 Center Drive
Bethesda, MD 20892-2033
(301) 496-7531
(301) 480-2770 Fax
adam_felsenfeld@nih.gov

Howard A. Fine, M.D.
Chief
Neuro-Oncology Branch
National Cancer Institute
National Institute of Neurological Disorders
and Stroke
National Institutes of Health
Bloch Building, Room 235
MSC 8200
9030 Old Georgetown Road
Bethesda, MD 20892-8200
(301) 402-6383
(301) 480-2246 Fax
fineh@mail.nih.gov

Stephen P.A. Fodor, Ph.D.
Chairman and Chief Executive Officer
Affymetrix, Inc.
3380 Central Expressway
Santa Clara, CA 95051
(408) 731-5000
(408) 481-0920 Fax
steve_fodor@affymetrix.com

Daniela S. Gerhard, Ph.D.
Acting Director
Office of Cancer Genomics
National Cancer Institute
National Institutes of Health
Building 31, Room 10A-07
31 Center Drive
Bethesda, MD 20892
(301) 451-8027
(301) 480-4368 Fax
gerhardd@mail.nih.gov

Richard A. Gibbs, Ph.D.
Director
Human Genome Sequencing Center
Baylor College of Medicine
Alkek Building, Room N1519
1 Baylor Plaza
Houston, TX 77030
(713) 798-6539
(713) 798-5741 Fax
agibbs@bcm.tmc.edu

Joe W. Gray, Ph.D.
Director
Life Sciences Division
Lawrence Berkeley National Laboratory
Mail Stop 84-171
1 Cyclotron Road
Berkeley, CA 94720-8268
(510) 495-2438
(510) 495-2535 Fax
jwgray@lbl.gov

Mark Guyer, Ph.D.
Director
Division of Extramural Research
National Human Genome Research Institute
National Institutes of Health
Building 31, Room B2-B07
MSC 2033
31 Center Drive
Bethesda, MD 20892-2033
(301) 496-7531
(301) 480-2770 Fax
guyerm@exchange.nih.gov

Michael A. Hollingsworth, Ph.D.
Professor
Eppley Institute
University of Nebraska Medical Center
986805 Nebraska Medical Center
Omaha, NE 68198-6805
(402) 559-8343
(402) 559-3339 Fax
mahollin@unmc.edu

David Hunter, Sc.D., M.D.
Professor of Epidemiology and Nutrition
Harvard School of Public Health
Brigham and Women's Hospital
Channing Laboratory
181 Longwood Avenue
Boston, MA 02115
(617) 525-2755
(617) 525-2008 Fax
david.hunter@channing.harvard.edu

Tyler Jacks, Ph.D.
Director
Center for Cancer Research
Massachusetts Institute of Technology
Room E17-517
77 Massachusetts Avenue
Cambridge, MA 02139
(617) 253-0262
(617) 253-9863 Fax
tjacks@mit.edu

David B. Jaffe, Ph.D.
Manager
Whole Genome Assembly
Broad Institute
320 Charles Street
Cambridge, MA 02141-2023
(617) 258-0923
(617) 258-0903 Fax
jaffe@broad.mit.edu

Ilan R. Kirsch, M.D.
Chief
Genetics Branch
Center for Cancer Research
National Cancer Institute
National Institutes of Health
Naval Hospital Building 8, Room 5101
8901 Wisconsin Avenue
Bethesda, MD 20889-5105
(301) 402-6382
(301) 496-0047 Fax
kirschi@exchange.nih.gov

Eric S. Lander, Ph.D.
Director
Broad Institute
320 Charles Street
Cambridge, MA 02141-2023
(617) 252-1906
(617) 258-0903 Fax
lander@broad.mit.edu

Christopher J. Logothetis, M.D.
Professor and Chairman
Department of Genitourinary Medical Oncology
M. D. Anderson Cancer Center
Unit 427
1515 Holcombe Boulevard
Houston, TX 77030
(713) 794-1468
(713) 745-1625 Fax
clogothetis@mdanderson.org

Shaun C. Lonergan, M.S.
Vice President of Business Development,
Sales and Marketing
454 Life Sciences, Inc.
20 Commercial Street
Branford, CT 06405
(203) 871-2425
(203) 481-2074 Fax
slonergan@454.com

Elaine Mardis, Ph.D.
Co-Director
Genome Sequencing Center
Washington University School of Medicine
Campus Box 8501
4444 Forest Park Avenue
St. Louis, MO 63108
(314) 286-1807
(314) 286-1810 Fax
emardis@watson.wustl.edu

Paul S. Meltzer, M.D., Ph.D.
Senior Investigator
National Human Genome Research Institute
National Institutes of Health
Building 50
MSC 5000
50 South Drive
Bethesda, MD 20892-5000
(301) 594-5283
(301) 480-3281 Fax
pmeltzer@mail.nih.gov

Matthew Meyerson, M.D., Ph.D.
Assistant Professor of Pathology
Dana-Farber Cancer Institute
M446
44 Binney Street
Boston, MA 02115
(617) 632-4768
(617) 582-7880 Fax
matthew_meyerson@dfci.harvard.edu

Jeffrey D. Milbrandt, M.D., Ph.D.
Professor of Pathology, Immunology
and Medicine
Washington University School of Medicine
Box 8118
660 South Euclid Avenue
St. Louis, MO 63110-1093
(314) 362-4650
(314) 362-8756 Fax
jeff@pathbox.wustl.edu

Michael J. Morin, Ph.D.
Vice President, Discovery, Antibacterials, Immunology, and Cancer
Pfizer Global R&D
Groton Laboratories
MS 8118-B3
Eastern Point Road
Groton, CT 06340
(860) 441-5476
(860) 715-3409 Fax
morinmj@groton.pfizer.com

Jim Mullikin, Ph.D.
Associate Investigator
Genome Technology Branch
National Human Genome Research Institute
National Institutes of Health
Building 50, Room 5318
MSC 8004
50 South Drive
Bethesda, MD 20892-8004
(301) 496-2416
(301) 480-0634 Fax
mullikin@mail.nih.gov

Maynard V. Olson, Ph.D.
Professor of Medicine
University of Washington
Box 352145
Seattle, WA 98195
(206) 685-7346
(206) 616-5242 Fax
mvo@u.washington.edu

Jane L. Peterson, Ph.D.
Associate Director
Division of Extramural Research
National Human Genome Research Institute
National Institutes of Health
Building 31, Room B2-B07
MSC 2033
31 Center Drive
Bethesda, MD 20892-2033
(301) 496-7531
(301) 480-2770 Fax
jane_peterson@nih.gov

Brian J. Reid, M.D., Ph.D.
Member
Divisions of Human Biology and Public Health Sciences
Fred Hutchinson Cancer Research Center
Mail Stop C1-157
1100 Fairview Avenue, North
P.O. Box 19024
Seattle, WA 98109
(206) 667-6792
(206) 667-6132 Fax
bjr@fhcrc.org

Gregory J. Riggins, M.D., Ph.D.
Associate Professor of Neurosurgery, Pathology, Oncology, and Genetic Medicine
Johns Hopkins University School of Medicine
Mason F. Ford Center Tower, Suite 5200
5200 Eastern Avenue
Baltimore, MD 21224
(410) 550-9686
(410) 550-9689 Fax
griggin1@jhmi.edu

Jane Rogers, Ph.D.
Head of Sequencing
The Wellcome Trust Sanger Institute
Hinxton
Cambridgeshire CB10 1SA
United Kingdom
44 1223 834244
44 1223 494919 Fax
jrh@sanger.ac.uk

Julie A. Schneider, D.Phil.
Program Manager
Office of Technology and Industrial Relations
National Cancer Institute
National Institutes of Health
Building 31, Room 10-A52
31 Center Drive
Bethesda, MD 20892
(301) 496-1550
(301) 496-7807 Fax
schneidj@mail.nih.gov

William R. Sellers, M.D.
Assistant Professor of Medicine
Dana-Farber Cancer Institute
Harvard Medical School
D720C
44 Binney Street
Boston, MA 02115
(617) 632-5261
(617) 632-5417 Fax
william_sellers@dfci.harvard.edu

Dinah S. Singer, Ph.D.
Director
Division of Cancer Biology
National Cancer Institute
National Institutes of Health
Executive Plaza North, Suite 5000
6130 Executive Boulevard
Rockville, MD 20892
(301) 496-8636
(301) 496-8656 Fax
ds13j@nih.gov

Louis M. Staudt, M.D., Ph.D.
Chief
Lymphoid Malignancies Section
Metabolism Branch
National Cancer Institute
National Institutes of Health
Building 10, Room 4N-114
9000 Rockville Pike
Bethesda, MD 20892
(301) 402-1892
(301) 496-9956 Fax
lstaudt@mail.nih.gov

Robert L. Strausberg, Ph.D.
Vice President for Research
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
(301) 795-7890
(301) 838-0209 Fax
rls@tigr.org

Thea D. Tlsty, Ph.D.
Professor
Department of Pathology
University of California, San Francisco
513 Parnassus Avenue
San Francisco, CA 94143-0511
(415) 502-6116
(415) 502-6163 Fax
ttlsty@itsa.ucsf.edu

Thomas J. Vasicek, Ph.D.
Vice President, Business Development
Lynx Therapeutics, Inc.
25861 Industrial Boulevard
Hayward, CA 94545
(510) 670-9471
(510) 670-9303 Fax
tvasicek@lynxgen.com

Victor Velculescu, M.D., Ph.D.
Assistant Professor
Oncology Center
Johns Hopkins University School of Medicine
Cancer Research Building, Room 590
1650 Orleans Street
Baltimore, MD 21231
(410) 955-8878
(410) 955-0548 Fax
velculescu@jhmi.edu

Janet A. Warrington, Ph.D.
Vice President
Clinical and Applied Genomics Research
and Development
Affymetrix
3380 Central Expressway
Santa Clara, CA 95051
(408) 731-5000
(408) 481-0422 Fax
janet_warrington@affymetrix.com

Roberto Weinmann, Ph.D.
Director
Oncology Discovery
Bristol-Myers Squibb Pharmaceutical Research
Institute
Route 206 and Provinceline Road
Princeton, NJ 08543
(609) 252-3648
(609) 252-6171 Fax
roberto.weinmann@bms.com

Richard Wooster, Ph.D.
Wellcome Trust Sanger Institute
Wellcome Trust Genome Campus
Hinxton
Cambridgeshire CB10 1SA
United Kingdom
44 1223 494951
44 1223 494809 Fax
rw1@sanger.ac.uk