Summary Report of the NIST-NCI

Biospecimen Quality Assessment & Standards Development Workshop

October 21-22, 2010

Office of Biorepositories and Biospecimen Research
National Cancer Institute
National Institutes of Health
U.S. Department of Health and Human Services

Contents

Background 4

Workshop Planning 4

Workshop Summary 6

Day 1: Thursday, October 21 6

Welcome/Overview/Context 6

Current State of Clinical Laboratory Medicine: Need for Standards in the Context of Next-Generation Technologies 9

How Standards are Developed and Used and the Role of NIST 11

Introduction to Case Studies for Afternoon Session 19

Importance of Experimental Design, Bias, and Rigorous Statistical Methods 19

Measurement Concerns from a Regulatory Perspective 23

Report Out from Each Table 26

Adjournment 29

Day 2: Friday, October 22 29

Report Out from Working Groups 29

Adjournment 35

Summary of workshop products and recommendations 36

RNA Molecular Quality 36

DNA Molecular Quality 37

Formalin-Fixed Tissue Quality 38

Serum-Level Protein Molecular Quality 39

Appendix A: Workshop Case Studies 42

RNA Integrity Case Study 42

DNA Case Study 43

Formalin-Fixed Tissue Case Study 44

Serum Protein Integrity Case Study 45

Appendix B: Day 1 Case Study Working Groups 46
RNA Integrity Case Study Discussion 46
DNA Integrity Case Study Discussion 49
Formalin-Fixed Tissue Integrity Discussion 53
Appendix C: RNA Integrity Working Group 57
Appendix D: DNA Integrity Working Group 65
Appendix E: Fomalin-Fixed Tissue Integrity Working Group 72
Appendix F: Serum Proteins Integrity Working Group 83
Appendix G: Final Workshop Agenda 90

Background

The US Congress has directed the National Institutes of Health (NIH), through recent health care reform legislation, to formulate a more aggressive strategy towards the translation of scientific discovery directly into the development of novel therapeutic and diagnostic capabilities. The National Cancer Institute (NCI) seeks to facilitate these efforts by targeting bottlenecks along the relevant development pipelines as it relates to treatment of cancer. Many of the bottlenecks that impede the pipelines for drug, biologic and device development are known, and include a lack of understanding and standardization surrounding the research biospecimens required to advance this research.

The National Cancer Institute Office of Biorepositories and Biospecimen Research (NCI/OBBR) hosted a workshop in coordination with the National Institute of Standards and Technology Materials Measurement Laboratory (NIST/MML) and the Food and Drug Administration (FDA) Office of Personalized Medicine to develop fit-for-purpose quality assessment tools for the human biospecimen samples used by the research community. Broad consensus by the scientific community has been repeatedly demonstrated that suggests the lack of sufficient quality human biospecimens is a critical limiting factor for translational research. The first step of this collaboration was to convene a workshop to gather recommendations from leaders in the scientific community.

Workshop Planning

The early vision for this collaboration would involve specimen procurement, financial support, and scientific guidance from NCI, regulatory guidance from FDA, and the bulk of the investigation performed by NIST (with any available assistance that could be provided by hiring temporary research staff using funds from the NCI Cancer Human Biobank, caHUB, appropriation). Four specific laboratories were available to explore projects for the initial pilot period to pursue biospecimen quality assessment metrics for research biospecimens. The four participating labs can roughly attributed as those investigating DNA quality, RNA quality, serum-level protein quality, and formalin-fixed tissue quality. Given that NIST would perform the bulk of the research, it was thought that workshop participants would require some guidance in understanding the constraints of the four participating NIST laboratories. In addition, it was thought that the caliber of participants being invited to participate would appreciate the opportunity to start interacting and contributing at the earliest opportunity, rather than receive a full day of lectures regarding the various constraints and considerations necessary to frame the vision. To accomplish these things, it was felt that the following agenda structure would achieve these goals:

1. Introductory presentations to describe the agencies involved and the primary issues for which we sought recommendations from participants.
2. Case studies to frame some of the constraints of the NIST research labs involved and allow for early participation of workshop attendees.

3. Intermittent presentations on key issues that were thought to be often overlooked for objective assessment of fitness-for-purpose
4. Extended breakout discussions by key individuals to explore potential projects.

It was felt that two days would be sufficient to cover this agenda, but likely the most important component for achieving the intended goals would be inviting the right individuals to participate. A copy of the final agenda can be found in Appendix G. It was decided that the following perspectives were required to gather the best recommendations for going forward:

- biomedical and biospecimen scientists
- both investigative and practicing pathologists and oncologists
- standards/metrology scientists
- commercial assay developers
- experimental design and research methods experts
- statisticians
- regulatory scientists

Specific individuals invited were those thought to be leaders in their respective fields based on personal recommendations and their publication records. Participants were for the most part restricted to domestic US individuals as it was thought that incorporating international regulatory issues would be beyond the scope of this pilot initiative.

Case studies were developed in close coordination with each of the relevant NIST investigators with guidance from clinical researchers. The point of these case studies beyond providing a mechanism for early interaction and activity of workshop participants was also to introduce the capacities and process considerations relevant to the specific labs through a hypothetical construct.

Finally, it was thought that the conversations themselves, beyond the final recommendations, would offer excellent insight into relevant quality issues surrounding research biospecimens. Four science writers were recruited to participate in the discussion groups for case studies as well as the breakout discussions of the second day to capture as much of the conversation as possible.

Please see Appendix G for the final agenda for the workshop.

Workshop Summary

Day 1: Thursday, October 21

Welcome/Overview/Context

Carolyn Compton, M.D., Ph.D.
Director, Office of Biorepositories and Biospecimen Research (OBBR)
Executive Director, The Cancer Human Biobank (caHUB)
National Cancer Institute, National Institutes of Health

Dr. Compton opened the workshop at 9:37 a.m. EDT and welcomed participants. She spoke about what brought the National Cancer Institute (NCI), National Institute of Standards and Technology (NIST), and Food and Drug Administration (FDA) together to plan this workshop and commented on the resources available for this pilot collaboration program. She said that this initiative could be the most important thing these agencies do, and she seeks for input. She hopes to leave the workshop with a roadmap for moving forward.

The questions behind this workshop are very complex, Dr. Compton said. As executive director of caHUB, NCI's newest enterprise, Dr. Compton had her own reasons for coming. About 6 years ago, she said, she experienced an epiphany. She was one of a group of 35 thought leaders in the field who were asked the question: Why is there no cure for cancer? It has been more than 35 years since the war on cancer was declared, with billions of dollars spent. But the progress has not been what it should be. The main bottleneck has been the lack of human specimens.

The war on cancer still needs victories, and more progress is needed to reduce death rates. Compared to other conditions, cancer rates have changed little over the past 60 years. While death rates from heart disease have more than halved since 1950, with even larger decreases for cerebrovascular diseases and pneumonia/influenza, the death rates from cancer have remained about the same. Nearly 600,000 Americans will die of cancer this year. One of three females and one of two males will develop cancer in their lifetimes. In 2001, only one of three patients benefited from cancer drug treatment. With only a hammer, Dr. Compton said, the tendency is to see everything as nail. Cancers are lumped together because they look alike. The best chance for a cure is surgery—to find the disease and cut it out. Diseases such as colorectal cancer are viewed as surgical diseases.

Cancer remains a formidable foe, but many new tools are available. Development is rapid. Molecular biology, advanced technologies, and bioinformatics offer promise that is not yet realized. The convergence of advanced technology has led to new defining moment in the war on cancer, with unprecedented potential for exponential progress toward molecular oncology. Treatments are orders of magnitude more powerful, sensitive, and specific than in the past. The field looks ahead to defining all cancers by molecular diagnosis, with assays to establish individual molecular signatures.

This sets an extremely high bar for discovery research: to develop assays that are reproducible in a single lab and between labs. These assays must be validated to allow data strong enough to convince regulators to release the test and disseminate it around the world. Then the special molecular features of tumors will allow individual treatment of everyone with cancer—the right

treatment for the right person at the right time and the right dose. Treatments will be more effective, less toxic, and less costly than those used in the past.

Dr. Compton described a schematic of the pathway forward for personalized medicine. It begins with and is centered on the patient, and the representative of the patient is the biospecimen. The biospecimen is the center of the personalized medicine universe. It provides information to classify the disease and determine its gene sequence and expression and protein profile. This determines diagnosis and targeted treatment and comes back to the patient in the form of new products developed by the bioresearch system. The biospecimen also provides data for prognosis and prevention of disease.

Whole labs can be experts in one molecule or mutation, but the power of new tools allows study of groups of molecules. New technologies raise the quality bar for human biospecimens. The evolving fields of genomics, proteomics, and metabolomics provide powerful tools to analyze specimens. This is seen in a range of endeavors from discovery to diagnosis and treatment. The new tools allow elucidation of the molecular mechanisms of neoplasia and identification of targets for drug development, treatment, and prevention. They have the potential to define markers for susceptibility, screening, and recurrence and will lead to the development of a molecular-based taxonomy of cancer. The new technologies will identify biologic variations that determine drug efficacy and toxicity and help with the development and validation of new therapeutics and diagnostics. But the bottom line is that they all depend on high-quality, annotated human biospecimens.

Powerful tools, however, carry with them powerful risks. The technological capacity exists to produce low quality data from low quality analytes with unprecedented efficiency. This means that wrong answers are now available with unprecedented speed. Unraveling the massive matrix of misleading data is compromising forward progress in unprecedented ways. It can be difficult to differentiate artifact from biology. No matter how powerful the tools become, good science and the progress the results from it requires adherence to the primary principles of scientific discovery. Quality data begins with quality analytes. Garbage in means garbage out, and quality in means quality out. This can be underappreciated in the world of biomedical research. Publishers and investigators rarely think about the intrinsic quality of human samples. They have little knowledge about interfering factors and do not understand how artifacts can be introduced. Research is limited to the samples investigators can get, not the quality of the samples.

The challenge is to define quality, and that is the purpose of this workshop. A sample must be “fit for purpose.” Purpose is the type of molecular analysis that is proposed. Fit means that the biomolecules are of sufficient quality for the specific analysis. Reproducibility of assay data, time, money, progress in molecular medicine, and possible loss of life all depend on this.

Defining what is meant by molecular quality is a huge challenge, Dr. Compton continued. NCI does not want to hear about a garbage in/garbage out model. Scientists do not understand human biology enough to know what in the genome controls expressions of molecules. They don’t know what to measure and how to measure it in this complex, reactive system. Variability occurs according to biologic stresses. NIST scientists measure many different things all the time and are undeterred by the physical sciences, but they are challenged by this. RNA in a cell, for example, is very difficult to gauge. Is there a grading scale of acceptability? Are there cutoffs or an absolute value that delineate quality as good or bad or on a sliding scale of acceptability? These

questions must be addressed. This workshop will try to answer a number of questions. What is measured? How is it measured? What constitutes quality in a sample and fitness for purpose? Timing and handling of procedures for measurement present other challenges.

Variables involved with biospecimen samples begin with taking the specimen from the patient through medical or surgical procedures. They continue from there. Many individuals are involved in the process, from acquisition through handling and processing, storage, distribution, scientific analysis, and restocking unused samples. A succession of powerful, iatrogenic variables can affect the sample. They can include antibiotics or other drugs the patient was taking and the type and duration of anesthesia. Multiple potent drugs are used during surgery; in an average colectomy, the patient receives 124 different pharmaceutical agents. Anesthesia has powerful physical effects in the body but is not factored into specimen quality. The time an arterial clamp is on the sample also varies. Other variables include time the sample remained at room temperature and what the room temperature was type of fixative, time the sample was in fixative, rate of freezing, and size of aliquots. Everything done in the pathology suite affects the sample, and it is not surprising that results are not easily reproducible.

OBBR has taken up this challenge. For the first time, NCI has funded initiatives in which the biospecimen itself is the target of research with the intent to document how variables change the biospecimen. There is a need to document metrics and a cutoff at which the change is so great that it renders a specimen inadequate for research. This must be addressed in every specimen. The metrics are needed now.

Dr. Compton offered participants a tool to use in their deliberations as they think about how to move forward to collect biospecimens in a way that will allow NIST to define the quality of samples. OBBR initiated caHUB, the first national biobank, to permit the acquisition of high quality specimens for research in a highly orderly way. The vision of caHUB is to fill the biobanking gap with a unique, centralized, nonprofit public resource. It is not intended to replace existing biobanks. It will serve as a source for human biospecimens and associated data of absolutely known quality and lineage, with controlled variables when possible and recorded variables when they are impossible to control (as in the operating room). All specimens will flow to a central point, and caHUB will function within an ethical framework.

Tumor specimens will be collected from Commission on Cancer (CoC)-accredited institutions, which can guarantee a certain level of care in patients. Also, CoC-approved institutions are required to have a tumor registry with standardized data elements that are entered in the National Cancer Database, which contains diagnosis, surgery, pathology, treatment, and follow up data for 80 percent of the cancer patients in country. caHUB will partner with the National Cancer Database, and samples will be tracked through preoperative, postoperative, and laboratory settings. The information will be placed in the public domain, with biospecimen access for research, consulting services, training and education, and pathology reference.

The data from caHUB should become most powerful tool for cancer research in the next 5 years, Dr. Compton said, with the ability to carry the work forward *in silico*. Approximately 30 to 50 types of tissue from about 500 individuals will be needed. Harvesting at autopsy will allow capture of tissues in high demand and short supply such as tissues from metastatic disease. Users of the data will be required to redeposit their data back into the database.

Data available through caHUB will include specimen type, amount, diagnosis, pathological characteristics, and macro- and microscopic appearance. Details about collection, processing, storage, and distribution will be specified, along with quality control metrics, clinical information about patients at multiple time points, and results of molecular analysis from different platforms. The comprehensive database could, with maturation over time, become more useful to the scientific community for *in silico* research than the specimens themselves. caHUB's policies and procedures will be publicly available and could serve as standards for the biobanking community.

caHUB represents an advance in biomolecular science and medicine and might not have been possible at an earlier date, Dr. Compton said. It would not be possible without the development and integration of advanced technologies, transdisciplinary approaches, new research tools and infrastructure, and standards. It marks a movement from qualitative to quantitative science and a convergence of the biological and physical sciences. caHUB will play a role in translational science, regulatory science, and molecular medicine.

This workshop marks the start of a bold new collaborative initiative to bring together stakeholders to define the needs for quality metrics in human biospecimens. Translational researchers and funders from NCI, metrologists from NIST, medical product regulators from the FDA, medical product and assay developers from industry, clinical investigators and practicing clinicians, and statisticians are in the room. The meeting objective is to define the roadmaps for projects to collect human biospecimens that will allow NIST to identify molecular quality metrics that have regulatory and clinical relevance and can be performed in a reasonable timeframe. Everyone will be asked for input. The endeavor represents an investment in the immediate professional future and a much larger investment in the future of medicine.

Following the morning session, a question-and-answer panel will respond to participants' questions.

Current State of Clinical Laboratory Medicine: Need for Standards in the Context of Next-Generation Technologies

James A. Robb, M.D.
Governor of the College of American Pathologists (CAP)
Consulting Pathologist, OBBR

Dr. Robb commented that it is challenging to put together a workshop with three agencies. His background is as a professor and an investigator and clinical researcher. When he relocated to Miami in the 1990s, he was driven by the community viewpoint, although he understands academic research. But cancer exists in the community, not in the research setting.

The stars are now aligned to drive best standards through research and into the community, Dr. Robb said. He and his colleagues firmly believe that research will rapidly translate to care in the community. This workshop is not just for research, it is for the patient. Scientists are gathered here for the patient and no other reason.

Dr. Robb described a case history that is relevant to this workshop. A 39-year old mother of three young children in Wisconsin, a friend of his daughter, was diagnosed with a 2.5-centimeter triple-negative breast cancer, with a poor prognosis. Her mRNA expression array was favorable, however. But her hope turned to confusion and anger when the clinical trial she was enrolled in was canceled because the mRNA profile could not be duplicated. Three trials were canceled by Duke University because the mRNA expression profiles could not be duplicated.

This type of replication problem certainly is not unique to Duke. In a study of breast cancer gene expression signatures, a major problem in validating studies that involve tissue samples is the lack of detailed description of how the biospecimens are handled—how they were collected, transported, preserved, processed, and stored. Some hospitals had no idea how specimens were handled and they could not be validated.

Dr. Robb asked why pathology is changing so much. Evolving the field into personalized molecular medicine will depend on the development and implementation of evidence-based, best practice biospecimen annotating, collecting, processing, storing, and privacy protocols. The tests are excellent. Already the technology is amazingly advanced, with sensitivity of tests such as ELISA [enzyme-linked immunosorbant assay] and polymerase chain reaction (PCR) that would have been unimaginable in years past. But pathology will descend rapidly into the abyss if its excellent tests are conducted on inadequately annotated, collected, processed, and stored biospecimens. Pathologists must know what they are measuring.

Biospecimens are saved in biobanks, and caHUB is the first national public biobank. But the field has a long way to go. Lack of significant standardization of biospecimen annotating, collecting, qualifying, processing, storing, and transferring are obstacles to progress on the road to molecular personalized health care. No quality standardized biospecimen markers have been identified. High quality biospecimens in high quality biorepositories are necessary for improved patient care and safety.

The pillars of caHUB are evidence-based protocols and quality at every step every time for every specimen. As a molecular virologist and professor of pathology in California, Dr. Robb said, he was subject to research standards. As a clinical pathologist and medical director of a lab that conducted 8 million tests per year in Florida, he was subject to standards. But the field of anatomic pathology has no quality standards.

Reference standards for clinical lab analytes improve patient outcomes and represent good medicine. Three conditions for which application of NIST standards has had a major impact are diabetes mellitus, cardiovascular disease, and kidney disease. Among early measures, glucose levels and hemoglobin A1C have allowed uniform diagnosis of diabetes with technology that has developed since the 1940s. Measurement of the different types of cholesterol is more recent, and standardization has refined diagnosis of coronary artery disease. All vendors can now measure their assays in a consistent, validated way. And standards for measurement of creatinine for kidney disease varied greatly until a standard was issued.

Many tests have been cleared by the FDA, but without reference standards. They can be used for monitoring but not for diagnosis. These include tests for carcinoembryonic antigen (CEA) for colon cancer, prostate-specific antigen (PSA) for prostate cancer, and human epidermal growth factor receptor 2 (HER2) for breast cancer.

HER2 marks the first time pathology has directed diagnosis. Hundreds of similar applications are now in the pipeline, including a pharmacogenomics approach that will attempt to determine the level of drug needed by each individual patient.

Dr. Robb emphasized the need for contributions to the field from the participants in this workshop. Initiative, drive, and resources exist, but DNA quantity is a poor quality measure for the quality of DNA. Good information to qualify specimens is needed. The RNA integrity number (RIN) is a poor quality measure for mRNA. While miRNA is stable in formalin-fixed, paraffin-embedded (FFPE) tissue, there is a need to measure phosphoproteins, polypeptides, and lipoproteins, and little is known about this. Research results are rapidly translated to patient care, but without validation. Standardized reference biospecimen quality markers are needed for both research and patient care.

How Standards are Developed and Used and the Role of NIST

Laurie Locascio, Ph.D.
Chief, Biomedical Science Division
National Institute of Standards and Technology

Dr. Locascio spoke about the challenges and importance of establishing quality measures from the NIST point of view. The goal of NIST is to develop standards, and she said she hoped she could excite workshop participants about standards.

NIST is a non-regulatory agency established in 1901 in the U.S. Department of Commerce. Its mission is to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology. NIST has established standards in the U.S. in a wide variety of categories including physical standards, test methods, and calibrations. Much of NIST's work involves biological standards.

NIST was established by the Organic Act of 1901, updated in 2008. The Act emphasized the importance of establishing standards for the work of manufacturing, commerce, scientific apparatus, the scientific work of government, and academic institutions. Functions and activities of the Act consist of custody and dissemination of national standards including calibrations, certified reference materials, reference data, and reference methods; determination of physical constants and properties of materials; comparison of U.S. national standards with those of the rest of the world; and solution of standards and measurement problems for industry and other agencies.

An early driver for the establishment of U.S. standards was the 1904 Baltimore fire, when out-of-town fire companies responding to the emergency found that their hoses could not couple to the Baltimore hydrants. More than 1,500 buildings were destroyed in the fire. The following year, the National Fire Protection Association adopted a national hose coupling standard developed by the National Bureau of Standards (the name for NIST until 1988).

NIST evolved through the century, establishing standards for train tracks, couplings, and steel manufacturing in the early 1900s, through its current work setting standards for clinical analytes, medical imaging, and cybersecurity. NIST employs about 2,900 workers at its two primary

campuses in Gaithersburg, Maryland and Boulder, Colorado. The agency has about 2,600 associates and facilities users and approximately 400 NIST staffers serve on about 1,000 national and international standards committees. Three NIST scientists have received Nobel Prizes.

Major programs at NIST include the NIST laboratories, the Baldrige Performance Excellence Program, the Hollings Manufacturing Extension Partnership, and the Technology Innovation Program. Joint institutes between NIST and academia are JILA (with the University of Colorado at Boulder), the Joint Quantum Institute (University of Maryland at College Park), the Institute for Bioscience and Biotechnology Research (University of Maryland), and the Hollings Marine Laboratory (College of Charleston and Medical University of South Carolina).

In its new structure, six NIST laboratory programs—material measurement, physical measurement, engineering, information technology, nanoscale science and technology, and neutron research—are organized under the associate director for laboratory programs. NIST has traditionally focused research and measurement service activities on physical science and engineering disciplines, but biology has been a continuing thread at NIST with medical services since the 1940s. Bioscience and health have been identified as a new area for significant emphasis for NIST labs.

Standards and metrology are needed to supply good data that can be used to confidently support decision-making. Good decisions are informed by good data, which are the results of excellent measurements. Metrology and standards comprise the formal systems that establish trust in data. They are necessary for comparisons—results are only useful when compared to other results (such as to observe a trend), to limits (such as a threshold for action), or when compared over time and space to different results in different places or measured at different times. Comparability over space and time is critical. Only when results can be linked to a common reference is it possible to compare data, and scope of reference defines the scope of comparability.

Another problem can be measurement uncertainty. How well does an investigator know a result? That is an essential part of being able to compare results from one study to another. Are the results good enough? It might be necessary to think about the umpteenth decimal point for launching satellites but clinical data might not need that level of detail. Fit-for-purpose does not mean beating data to get to a reference point that is not needed.

Standards allow comparison to the work of others. Documentary standards and certified or standard reference material are examples of two types of standards. Documentary standards are guideline documentation that reflects agreements in practices by governmental bodies or nationally or internationally recognized industrial, professional, or trade associations. Certified or standard reference material is a physical entity that serves as a reference in measuring quantities or qualities, establishing practices or procedures, or evaluating results.

Part of the NIST mission is demonstrated through standards development and dissemination. The agency establishes and maintains national measurement standards for chemical and biochemical measurements and chemical and physical properties. It disseminates these national standards through provision of standard reference materials, standard reference data, and instrument calibrations to facilitate trade; regulatory compliance; design, control, and optimization of new technologies; and confidence in data. NIST sells about 25,000 units per year of standard

reference materials. It calibrates and tests approximately 1,000 items per year and sells some 6,000 units per year of standard reference data.

Recently, standards have been established for gene expression. NIST hosted the External RNA Control Consortium (ERCC), an industry-initiated ad-hoc consortium to address inconsistencies in microarray data. The agency is developing spike-in controls for microarray validation, protocols, and novel data analysis approaches. Approximately 90 organizations from private, public, and academic sectors, including microarray and reagent developers, are involved. Probe content and controls will be widely available. This endeavor is expected to help benchmark understandings of microarray data.

Another focus is on standards for cell line identification. Surveys of scientific literature suggest that up to 20 percent of published work uses human cell lines that are known to be misidentified. NIST is working with a group at ATCC, an international private, nonprofit biological resource center, to develop documentary standards for cell lines. This is a robust program that involves in-house authentication of cell lines using short tandem repeats (STR) used for human identification, Dr. Locascio said. NIST recently developed STR 26plex for measuring 26 STRs simultaneously, and the technique is readily adaptable to human cell line authentication. NIST develops databases to accompany its data, and the data will be published in a database on cell lines that are typed.

NIST's structural database to facilitate HIV research can improve access to biological data with the largest existing collection of three-dimensional structures of interest to AIDS researchers and the largest collection of two-dimensional structures of proven inhibitors. It includes preclinical data (e.g., structures, biological data, references) on all classes of inhibitors as well as drug resistance information. The database allows visualization and animation of drug–target interactions for education purposes. Integrated access is available through Chem-BLAST, a chemical taxonomy-based search engine; Semantic Concepts; and visualization tools.

Dr. Locascio emphasized that all standards—whether they apply to materials, methods, or data—must be supported by the best science. Therefore, fundamental world-class research programs are critical for NIST to do its job in standards development, which allows investigators to know their standards are comparable to someone else's.

An example of an area that is moving from basic research to standards is DNA damage and repair. Basic measurement science is pushing capabilities in quantitative biology and can now measure 1 mutated base in 100,000,000 bases. This capability has been used to build a program to study oxidative DNA damage, which has a role in many diseases including carcinogenesis and aging. This has led to discovery and isolation of a human DNA repair enzyme that is linked to metabolic disease and certain types of cancers and development of reference material to support mass spectrometry measurements of mutated bases.

Other work moving from basic research to standards is being done with microfluidic cell culture. Basic measurement science also is pushing capabilities in cellular measurements. A microfluidic cell culture system is developed in a controlled environment with a multiplexed culture. Investigators are working to link that directly to single cell transcriptome, protein content, and microRNA. NIST used microfluidic cell culture to build a program to perform highly controlled and multiplexed studies on cells for toxicology and cancer biology. It proposes to study the

relationship between measurement variability and biological variability, which is critical to understand the value of biological data. NIST has a new partnership with the Early Detection Research Network at NCI to develop a reference laboratory to validate measurements by the cancer biology community on circulating tumor cells and a partnership with the FDA to validate measurements of stem cells.

Dr. Locascio concluded with questions that she asked participants to provide feedback on. How can standards help to understand the quality of biological specimens? Is it possible to prescribe methods for obtaining the "best sample" using existing or new tools and technologies? Is it possible to develop methods or standards to help define the quality of a human specimen, either through use of single or multiple molecules, molecular signatures, spike in molecular standards to determine stability, or retrospective analysis of samples for specific analytes? The overarching goal is to develop methods of standards to define the quality of human specimens.

Question and Answer Panel: Goals of Workshop and Available Resources

Dr. Compton; Dr. Locascio; Dr. Robb; Jim Vaught, Ph.D., Deputy Director, OBBR, NCI; Stephen A. Wise, Ph.D., Chief, Analytical Chemistry Division, NIST

Robert W. Veltri, Ph.D. associate professor of urology and oncology, Johns Hopkins University School of Medicine, said that he works with the physical sciences on oncology studies in areas of microfabrication and nanotechnology. He asked about plans to work in this area. He has set up a microfabrication manufacturing laboratory to design and manufacture devices. Dr. Locascio said that Robert Austin, Ph.D., of Princeton University, is working on this. At NIST, the question is how microtechnology can be used to facilitate standards. If it were possible to control the micro environment around cell culture, it would be possible to preserve a sample in a unique way.

Dr. Compton said that these issues are of great interest to OBBR. NCI's Center for Strategic Scientific Initiatives (CSSI), which reports to the director, works on broad initiatives that are on the cutting edge of technology development and application. Investigators who work with bacterial systems might not be familiar with human systems, but CSSI can help create a Venn diagram of overlapping interests so that scientists who have never been focused on cancer research and representatives from other fields such as engineering can be brought together in an uber-team. The NCI Alliance for Nanotechnology in Cancer has connections to both OBBR and CSSI. OBBR also houses an office devoted to pure technology development, the Innovative Molecular Analysis Technologies Program, a high risk, high impact program. Many groups work together on this effort. She did not mean to imply otherwise. OBBR works with technology developers, standardizers, and validators to move forward to clinical application as soon as possible.

Lynne Rainen, Ph.D., scientific director of Preanalytix, Becton, Dickinson and Company, asked Dr. Robb about the role of CAP in driving changes in standards by which tissue specimens are collected, preserved, and processed. Dr. Robb said that is being done with evidence-based standards and accreditation power. A letter of intent is in place between NCI and CAP allowing accreditation of tissue samples; accreditation will be incorporated into CAP cancer protocols within the next 3 or 4 years. Protocols will include instructions in how to report out.

Dr. Compton added that NCI's strategy in partnering with CAP and NIST is that NCI is not a regulatory agency, and it cannot make anyone do anything (except indirectly through requirements of grants and contracts). Achieving personalized medicine will require a behavior change in the way that personnel who handle specimens do their work. It will require collaboration between CAP and FDA. One way to change behavior is to pay people to do something, an effective method but unlikely solution in this era of cost constraint. Using accreditation or licensure is another approach. Dr. Compton said that she believes that scientists and doctors want to be good doctors or scientists, and if they are shown data about how handling of biospecimens affects outcomes, they will change their behavior. The ultimate outcome is the death of a patient. On all levels, it is necessary to create partnerships to move the agenda forward.

Dr. Robb noted that OBBR is leading in a biospecimens program to take specimens to the community. OBBR created a resource for 30 community cancer centers, which do what OBBR says because it's best practice.

David Rimm, M.D., Ph.D., professor of pathology at Yale University, said that he understands the possibility of worldwide standards for measures such as glucose, but how can worldwide standards be developed for items that are harder to regulate such as proteins? Dr. Locascio said that NIST has been looking at the delivery of cell-based data and is collaborating with WHO and other global groups to establish standard reference materials of known stability. Many biospecimen materials are DNA and RNA. As materials become increasingly complex, the field is struggling with how to supply them and how to handle population variability. All related topics are under discussion and NIST is intimately involved with the process.

Dr. Wise said that most of the newer technologies are being handled by Dr. Locascio's division. His division (Analytical Chemistry) has traditionally worked with sodium, cholesterol, and glucose and is now moving into protein markers such as PSA. NIST is examining how to develop standards for the metabolomics and proteomics communities and is very aware of the needs in this area. Nothing is off the table, and NIST is looking for input from workshop participants and the wider biospecimen community about what their needs are.

Dr. Robb said that international standards are a significant issue. CAP shifts materials all over the world, and without measurement standards, if 80 percent of pathologists get the same answer and agree, that represents truth.

In response to a question about national service standards, Dr. Robb said that such standards are a major issue. He agreed that performance of a diagnostic procedure in a local or national facility is not the same. The same is true of biopsies. CAP has representatives from reference labs on their committees and is working toward national standards, but it is difficult. Dr. Compton added that this is part of the vacuum that caHUB wants to fill. Currently, no source of standardized benchmark samples exists. Those that are used are taken from clinics where no elements are required to report or control except in very new specific circumstances for handling breast tissue for HER2 analysis. The American Society of Clinical Oncology (ASCO) asked CAP for standardization. It is necessary to create reference samples. Dr. Robb commented that HER2 results have been shown to have an accuracy of plus or minus 20 percent, which is not good for women receiving treatment with Herceptin®. About 8 to 12 percent of women who take

Herceptin® are false positives and negatives for HER2 because of staining or other errors. This is a major issue. How does one inform a woman that her diagnosis was wrong?

Dr. Compton observed that problems should be regarded as opportunities. Once quality metrics are defined, they can be applied with molecular standards to any sample, and that will be an enormous benefit. The caHUB metrics can be used as a yardstick of truth. She has attended numerous workshops sponsored by the Institute of Medicine, the FDA, and other groups and agencies where a common focus was personalized medicine, but the subject of qualifying the specimen for the assay was ignored. Data-driven steps to qualify the specimens for assay are imperative. The perfect assay will provide the wrong answer if the quality of the biospecimen is not adequate.

David Eberhard, M.D., Ph.D., director of clinical trials pathology services for LabCorp, said that he is an anatomic pathologist and the comments of Drs. Compton and Robb resonated with him. A huge need and hunger exist for standardization and definition around histomorphological characterization. Standardized descriptions are critical. He has been involved in research and is now aware of what goes on behind the scenes. He was a coauthor of a recent paper in *Nature* about diverse mutation patterns in human cancers and criticized the uncertainty of material that is used for assays. It is necessary to histomorphologically characterize the tissues that are used.

Dr. Compton agreed on the need for metrics for histomorphological characterization. A morphology dataset will be available publicly and will be part of the inventory of data that surrounds each sample. She added that the pathology community at large can help put metrics around areas that for more than 100 years have been based on estimations by experienced pathologists who try to slot each new slide into their experience as a reference set. Studies of the human brain estimate that a human can interpolate five data points when making a decision. Molecular analysis yields terabytes of data, and it is far beyond the capacity of the human brain to absorb this and make consistent decisions. It is essential to involve information technology (IT) colleagues to develop tools that lead to consistent findings. This is absolutely not done now in anatomic pathology, yet anatomic pathology is considered the gold standard for cancer diagnosis worldwide. The procedure of eyeballing a biospecimen and loosely classifying it with no standard terms must be addressed.

Dr. Locascio said that NIST has a very robust collaboration with IT personnel. She looks forward to working with the IT community with real samples. Part of the challenge is not only developing analysis but also evaluating the results. It is not even clear what metrics for evaluation will look like. Dr. Robb added that CAP is working on correlating genomic data with pathology in order to modify targeted therapy in patients.

Elizabeth Mansfield, Ph.D., senior policy analyst for the Office of *In Vitro* Diagnostic Devices, Center for Devices and Radiological Health, FDA, said that Dr. Compton provided a somewhat frightening context for the discussion. Much of personalized medicine is related to genomics. Will caHUB address the variety of possible mutations, or will that be done by another group? Rare findings require validation. Dr. Compton said that users will do those analyses and provide the data. NCI and the National Human Genome Research Institute are working together on the Cancer Genome Atlas (TCGA), a major attempt to catalog histopathology in different types of cancer. It would have been wonderful if caHUB had been launched in time to serve TCGA from its start. The concept of building human knowledge is huge if all participants use the same

biomarkers and all generated data could be directly compared. That was the vision, and TCGA was a major source of inspiration for caHUB. For TCGA, investigators polled the entire country for three different types of cancer samples. They wanted 500 samples for each type, but were unable to fill the quota for any but ovarian cancers because quality metrics were not available for platforms. They found out in a rude way that they did not have the science they needed. caHUB still plans to coordinate with the needs of TCGA. Dr. Veltri commented that the genomics and proteomics of benign areas adjacent to tumors should also be categorized and defined, because changes in benign areas can predict cancers.

Clive Taylor, M.D., Ph.D., professor of pathology, Keck School of Medicine, University of Southern California, commented that this workshop probably could not have occurred at an earlier time. He does not think that the need to qualify specimens is ignored. The biggest problem is obtaining reproducible data. It is very difficult for pathologists to process tissues in a rigorous, standardized manner. The reason is manpower and he does not believe it will ever be possible. Blood specimens can be standardized, but not tissue. Internal reference standards are needed and should be the true focus. Development of internal reference standards could take much of the noise out of the system.

Dr. Compton replied that she believes change is possible if sufficient evidence of the need is demonstrated and funding can be found. The enormous costs downstream if this is not done dwarf the developmental costs. Dr. Taylor said that unfortunately medicine does what it can afford to do. To process tissues in a standard way requires manpower that is not affordable. Dr. Robb added that from the perspective of the NCI Community Cancer Centers Program, it is clear that pathology must control the process. The value of increased patient care and safety will save money. Dr. Taylor said he agreed that standards should be sought, but questioned the possibility of success. Dr. Locascio said that NIST is now looking at people solutions and will turn to technology solutions in the future.

Tianhong Li, M.D., Ph.D. a medical oncologist from the University of California, Davis, said that her focus is on the patient and pathology is the key to offer treatment options to patients. She asked two questions to the panelists. The first questions was: Could newly-developed NIST STR analysis, as used in human cell line authentication and forensic science be applied to identifications of specimens in biobanking? Also, does a standard exist for *in vivo* testing in animal models? Dr. Compton responded that animal models do not exist for biospecimen standards. Animal researchers have their own needs for standards and have approached NCI to talk about that. Starting with humans and human samples is the most difficult approach; usually researchers work up from animal and substitute models.

Dr. Locascio said that NIST has a robust forensics program using the STR 26plex assay, and it easily could be extended to patient identification. Dr. Robb said that DNA identification is done on each aliquot of banked biospecimens, and identification on the DNA level will be possible. Dr. Compton added that NCI is trying to use a new system of universal identifiers for every daughter and granddaughter aliquot of banked biospecimens. Dr. Robb said that it is particularly important to apply these technologies for specimens ordered from foreign countries.

Dr. Li's second question was: Does a standard exist for *in vivo* drug testing in animal models engrafted with human tumor specimens? Dr. Compton responded that there are currently no standards for testing these animal models. Animal researchers have their own needs for

developing standards and have approached NCI to talk about that. Starting with humans and human samples is the most difficult approach; usually researchers work up from animal and substitute models.

Paul Kimmel, M.D., director of the Chronic Kidney Disease Biomarker Program, National Institute of Diabetes and Digestive and Kidney Diseases, said that he heard a great deal of uncertainty in the presentations. What is the certainty that new practices will be incorporated? Are multiple collection strategies being considered? What is good for proteomics might not be good for metabolomics or genomics or clinical practice. Dr. Compton said that is true, but the field will be standardized to one-size-fits-all, with best practices applied. She appreciates that best practices are woefully evidence-deficient at this point.

Dr. Vaught said that the NCI best practices currently are a conglomeration of NCI's own best practices and observations and was updated recently. The public comment period just ended. The update represents an improvement over best practices from 3 years ago, but still sorely lacks an evidence base. It is mostly observational and has a long way to go. Dr. Compton said that the update underscored the burning need for an evidence base in biobanking. The state of the science is weak. Standard operating procedures will be updated as evidence comes in. Best practices are now based on what best practitioners have dictated and can be improved with objective scientific evidence.

Dr. Robb said that it is necessary to determine how much variance can be tolerated. A quantitative basis is needed as a starting point, with tools in place to do it right from the start. This is on the horizon, but not yet in place. Dr. Veltri asked whether biopsy versus surgically-obtained tissues would be included, and Dr. Robb replied that needle biopsies will be compared to surgery samples. Dr. Rimm said that core biopsy also will be compared to surgery. The core can be transferred directly to formalin, with no lag time. For surgery samples, the lag can be from 15 minutes to 6 hours, and it will be possible to analyze the differences.

Christopher Otis, M.D., director of surgical pathology at Baystate Medical Center and professor of pathology at Tufts University School of Medicine, said that he is on several cancer committees, including digital pathology, and that work seems prehistoric compared to what has been suggested here. How can digital pathology be validated in the assessment of images? Could NIST representatives be invited into that sphere? Will diagnostic criteria become available?

Dr. Compton said that she would invite any partners. OBBR is a public resource that exists only to help the public reach its goals. It is a big challenge. CAP represents the workforce, the individuals who must implement whatever improvements can be made. Pathologists and surgeons are two key professional groups who need to focus on metrics and molecular quality. Pathology is now a qualitative field, which leads to diagnostic differences. It must move to a more objective place. Objective standards are also needed in radiology. Dr. Locascio said that NIST is involved with radiology, and Dr. Compton said that NCI could provide images.

Dr. Robb noted that CAP had budgeted $40 million to move pathology to molecular imaging, to move from the microscope to digitalized images. This could be merged with radiology. He is excited that NIST has the capabilities to help with this. Good instrumentation and good algorithms that can be trusted are needed. Radiology has transitioned, but pathology is 20 years behind. Dr. Compton said that there is some fear in pathology about technology replacing the

human element. That is the wrong approach. The ability to derive information from specimens is spectacularly more powerful than ever before, and pathologists must become comfortable with analysis as a new tool. She added that pathologists have never before needed more finely distinguished categories. A morphological marker for an underlying molecular change would preserve some of the economics and efficiency of pathology and could be used to supplement molecular underpinnings.

Introduction to Case Studies for Afternoon Session

Tony Dickherber, Ph.D.
AAAS Science & Technology Policy Fellow
OBBR

Dr. Dickherber explained the procedure for the afternoon session. Participants were to break into smaller groups and discuss the different case studies that had been distributed. Distributed case studies can be found in Appendix G. The discussions were to be interrupted by two presentations, one about the importance of experimental design and the other about regulatory concerns.

The breakout sessions were to continue the following day. Dr. Dickherber asked participants to consider the following topics in their discussions:

- Points within the specimen collection and processing continuum where sample integrity issues might arise
- Appropriate quality assessment (QA) metrics for determining the integrity of biospecimens at those critical points
- Unmet need for standards to ensure accuracy and comparability of data from biospecimen quality assessment assays and to accelerate their development and use
- Scientific questions that need to be answered before a standard could be developed (i.e., identify the missing fundamental science from the literature)
- Routes to a properly designed plan that ensures the quality of collected, processed and stored biospecimens and that data from sample quality analysis assays are interpreted appropriately
- Phenotype of a scientific advisory group to plan the projects, interpret the findings, and develop standards for use of fitness-for-analysis assays and their reference standards
- Impact assessment for development of such a standard(s) within the biospecimen procurement enterprise on the quality of research results and subsequent translation into clinical practice
- Potential for changing current research and development and clinical practice

Importance of Experimental Design, Bias, and Rigorous Statistical Methods

Terrence Paul Speed, Ph.D., B.Sc.
Walter and Eliza Hall Institute of Medical Research
Parkville, Victoria, Australia

Dr. Speed thanked NCI, NIST, and FDA for stepping up to address the topics of this workshop. About 3 years ago he attended a workshop on designing experiments for blood. Variability,

which has been discussed at this workshop, was an underlying theme, and he will use a real example to illustrate problems that can arise from variability.

A number of buzzwords are used to describe the relevant statistical principles of experimental design. Replication means taking more than one measurement with all other variables held constant. Dr. Speed said that he is aware of an amazing number of experiments that took only one measurement. Some researchers say replication is waste of time, but no basis for statistical analysis exists without it.

Randomization is another general statistical principle, but it can be anathema to biologists. Lack of randomization is one of major shortcomings of the experiment Dr. Speed will describe. Two papers about an experiment without randomization ran cases on day 1 and controls on day 2 with no hint of randomization, and differences arose from running assays on different days. The quality of the assay degraded over two days; investigators said that they corrected for it, but they did not randomize, which invalidated any conclusion. It was an irreproducible result, although the paper claimed 100 percent sensitivity and specificity.

Other general statistical principles are local control and the factorial principle, which apply to efficiency.

Special statistical principles involve designs with confounding, so that it is possible to measure several things at once. Another principle is two phases. Examples of experiments with two phases are what is done before and after a sample is taken (e.g., whether the subject is sitting or standing, how long the sample remains in the freezer after it is taken). External validity is another principle.

The first blood experiment that Dr. Speed described has not been written up and was not successful, but it might have been useful to write it up. Blood plasma was prepared simply by spinning a tube of fresh blood in a centrifuge until the blood cells fell to the bottom of the tube. The aim of the first experiment was to assess the difference between processing plasma from ethylenediaminetetraacetic acid (EDTA)-preserved whole blood at room temperature (24°C) compared to plasma at 4°C on proteomic analysis. (Metabolomic analyses were carried out.) The investigators who preferred processing at 4°C were concerned about reducing protease and peptidase activity, while the proponents of room temperature processing were trying to prohibit platelet activation and platelet protein contamination. All of the other variables were fixed.

The study design began with four healthy male subjects. Blood was drawn from each arm of each subject, beginning in all cases with the right arm, for a total of five 6-ml K2-EDTA tubes from each arm. The blood was taken in series from one position on the arm. The first draws from each arm were discarded (tubes 1 and 6) and the others moved forward. Tubes 2, 4, 7, and 9 were processed to plasma at 24°C, and tubes 3, 5, 8, and 10 at 4°C. The 24°C samples were spun on the Sorvall RT7 Plus centrifuge and the 4°C samples on the Sorvall RT 6000D centrifuge.

In response to a question about the speed of the centrifuges, Dr. Speed said that he would address that later in his presentation, but the speeds were not relevant.

Dr. Speed simplified the actual experiment somewhat for this presentation. The 32 samples of whole blood were converted to plasma at 4°C or 24°C by spinning 1,500 grams (2,588 rpm) for 15 minutes, followed by 2,000 grams (2,988 rpm) for a second 15 minutes. Approximately 400

metabolomic measurements were taken on every sample by liquid and gas chromatography-mass spectrometry, but those results are not relevant to the point here, which is the design of the comparative experiment focused on the temperature differences.

In response to questions and suggestions about problems with the study design, Dr. Speed said that the design did not include replication of data. Some samples were processed immediately while others were held for 24 hours. Women were not used as subjects because their blood values might be unstable due to hormonal influences. Samples were all drawn from the same spot on the subjects' arms. The plasma was stored after the second centrifugation, probably in plastic tubes.

Dr. Speed then offered his comments on the study design. The four subjects were too few to allow major conclusions, and they were not a random sample of any well-defined population. This refers to external validity, which includes generalizability and whether results of this experiment could have broader use.

Second, the draws from each arm began in all cases with the right arm, which appears promising with the possibility of replication, but a better design would have been to choose the first arm at random. Investigators must get into habit of doing this. The duplicate samples and ability to estimate measurement error and comparisons within a subject and possibility of local control also appear promising, but systematic assignments of units to treatments should be avoided. Assignment should be done randomly using different patterns, and there are many different ways to do it.

The two different centrifuges represent a striking and potentially fatal design flaw. Both machines have the capability for different temperatures and might have different degrees of spin. Temperature differences are completely confounded by the two centrifuges. Some form of randomization is essential at this step, possibly restricted to achieve balance.

In discussion of randomization, Dr. Speed said that he understands why biologists do not favor it and it can introduce errors. But those problems will be overcome in time. While randomization does induce the possibility of mistakes, the failure to randomize ensures it.

Dr. Speed summarized that replication, randomization, local control, and correctly measuring the relevant variability must be built into the study design and were not in this experiment. Relevant variability can be tricky and can be technical or biological, but investigators must consider it. It is relevant to the statements they will want to make.

Finally, the blood was collected and converted to plasma in Frederick, Maryland and the assays were conducted and partly analyzed in Durham, North Carolina. The data were sent for further statistical analysis to Berkeley, California. There was no close three-way collaboration. Any statistician knows that this is not the way to get the best interdisciplinary science.

Question and Answer Panel

Dr. Speed; Lisa McShane, Ph.D., mathematical biostatistician, NCI

Marc Salit, Ph.D., of NIST, asked about the possibility of introducing bias systematically. For example, if centrifuges were randomized, it might induce a time change between sampling and

processing. The effects of randomization must be factored and balanced. Dr. McShane responded that Dr. Speed is not suggesting complete randomization. Dr. Speed counseled investigators to control what they know about and randomize the rest.

P. Mickey Williams, Ph.D., director of the Patient Characterization Center, Clinical Assay Development Center, SAIC-Frederick, NCI, commented that the assay is extremely important. Time invested up front could have minimized problems in the experiment. Dr. McShane agreed that the assay could build robustness into the experiment. Dr. Quick added that the assay was novel. It is not a good idea to do an experiment with a novel assay because investigators did not know what to expect. Even with randomization, it is possible to explore differences between two centrifuges, with half of the samples on one centrifuge and half on the other.

Dr. Robb said that the lesson from the previous presentation is to keep experiments very simple and focus on the most important relevant variable. Dr. Speed said that is true, but not completely. Once investigators have some confidence in their design, they can vary two or three variables. That's what factorial experiments and confounding are about.

A participant asked how it would be possible to determine downstream implications from multiple effects. The processes must interrelate and it is not clear how to apply some of these results to complex biology. Dr. McShane replied that she does not claim that design would answer all questions. It is necessary to analyze the data and take correlations into account. Dr. Quick said that often a component is assay-dependent. Different variables will matter for RNA and DNA. He cautioned against broad brush oversimplification, but getting too detailed might lead to a realm from which it is difficult to return. It is plausible that individual measurements might have different designs. Sometimes some averaging is necessary.

In response to a question about how many measurements are necessary for certainty on a clinical specimen, for example, for a PSA sample, Dr. Robb said that clinical labs have standards and there are reference standards and NIST values. The questioner asked whether some instruments could collect samples for all the sites. If samples were sent to Quest, LabCorp, and three local hospitals, would all of the results be the same? Dr. Speed said that external validity must bring together observational studies and other data. Participants commented that information must be linked, and even without a reference standard, it is necessary to accept that one lab might be very precise in one area and another lab's precision might be in another area. Lab-to-lab variability cannot be controlled and must be accommodated.

Karen Phinney, Ph.D., a NIST research chemist, commented that when samples are taken on different days with a small number of samples, comparisons might appear to be a waste of time because the sample pool is so small. Dr. McShane said that it depends on what is being checked and the sources of variability. How big are the effects? Dr. Speed said that those are excellent sources of data for replication and cross-replication and ideally double blind. Dr. Phinney said that one of challenges is that many samples were taken long ago and it is impossible to return to patients.

Dr. Veltri asked about a patient whose PSA increases from .01 to .04 nanograms per milliliter. Does that matter? Perhaps the level depends on the assay. Eventually the job of the physician or surgeon is to help the individual patient. When is the clinical lab result so reliable that it adjusts for biological, technical, and personal variations? Dr. Quick said that in the final analysis

everything is statistical with levels of variability. Dr. Locascio asked what percentage of published biological studies has good statistical reliability. Dr. Quick noted that variability across individuals is relevant. Many statistics become confused across studies. Variations in storage also have an effect.

A participant observed that radioimmunoassay technologies run samples in duplicate or triplicate. Dr. McShane replied that even with replicates, it is important to be careful that replicates represent the full range in the samples. Replicates are useful, but they must be representative. Another participant commented that some tests that affect patient well-being are repeated, but 60 percent of labs no longer repeat extreme values. Dr. McShane said that that kind of replication is not fully protective. It might be that the real value *was* extreme.

Dr. Speed concluded that at the end of the day, an assay is an assay, whether it is done once or three times. Dr. Robb has suggested the individual measures might be less precise, but that all balances out. A procedure can always be improved by doing it twice. Dr. Robb added that cutoffs are not always definitive.

Measurement Concerns from a Regulatory Perspective

Elizabeth Mansfield, Ph.D.
Senior Policy Analyst
Office of In Vitro Diagnostic Devices, Center for Devices and Radiological Health
Food and Drug Administration

The FDA looks at issues of reproducibility and repeatability when clearing devices. Biospecimen quality really does matter. To the FDA, it matters in development of a test and in discovery, validation, and instructions for use so that it can be repeated. Clinically, adequate collection and preparation (i.e., of the patient, of the specimen) are important, as are correct measurement and interpretation.

The FDA takes the approach that intended use is indication for use. It is looking, first, for a target condition to be measured—a disease, condition, or state of interest. Then the purpose of the test must be determined. Is it diagnostic, prognostic, for risk assessment, or for some other purpose? The analytes tests measured need to be specified—RNA, DNA, protein, metabolite, or something else. The target population, on whom and when a test will be used, is important; a test should be developed and tested on the population it will be used on. Another consideration is whether a specimen is primary tumor, fine needle aspiration, bone marrow aspirate, or biopsy. Is the matrix whole blood, FFPE, serum, or another matrix? Will the result type be primarily qualitative or quantitative? Who will be the user—the patient, a lab tech, a nurse, someone else? Where in the clinical algorithm does the test fit? Will it be used to screen a patient and refer to a specialist or to make a final diagnosis? All of these questions must be answered.

For validation, it is necessary to show that a test works on the specimen type to be used. Access is needed to appropriate specimen types such as biopsy. Evidence is needed demonstrating that the specimen source meets the intended use population, for example, in the areas of age, sex, and disease state. Specimens must be handled and stored in a controlled manner, and often treatment history such as prior therapy is needed.

For validation, it is necessary to account for all claimed matrices. It is important to understand how uniform the matrix is—for FFPE processing, for example, what was the fixative and its quality and how long was the sample left in fixative? This will have a huge impact on the results of a test. How long is the analyte stable in the matrix? What storage conditions are required? Is purification or concentration required? Does the matrix interfere with measurement?

Regulators need to know the history of stored specimens. Are they annotated with information such as demographics, diagnosis, and treatment history? Do they match the intended use population in geography, age, disease stage, and other descriptors? As practice changes, particularly in oncology, treatment paradigms change and the treatment histories of current patients might differ from those from whom the specimens were taken, making comparisons difficult. Do samples come with informed consent that allows use? Was there bias in sample collection? Bias almost always exists, Dr. Mansfield noted, based on where a sample was collected, the age and of size of the specimen, or how many samples have been lost. Knowing storage history is important; it could predict future use. For example, urine samples from large population studies for longitudinal studies might be ambiguous because they were stored in bisphenol A (BPA) with no way to distinguish endogenous BPA from the BPA that entered the specimen during storage.

When testing, it is necessary to determine adequate instructions for use that the user will be able to follow. Variables that must be controlled include patient preparation and concurrent exposures; these depend on what the test is for and what is being measured. Are procedures required prior to testing? What are the controls? How are the assays done? The answers to these questions must be clear.

A societal cost of poor measurements is failed discovery. Specification of the intended use population is often inadequate, with erroneous specification of specimen parameters, control materials, and availability of useful specimens that actually meet the needs for development.

A patient cost of poor measurements is inconsistent test results. Once a test is on the market, users might not achieve the same type of results as the test developer did. This can be the result of poor instructions for use; poor mechanisms to control or manage specimen collection variables, which HER2 has suffered from; lack of recognition that variables matter; and lack of standards (materials or methods) to trace collection, handling, and storage history. Ultimately, patients can die from these harms.

This workshop offers a number of possibilities, Dr. Mansfield said. In an ideal discovery and product development world, the well-curated specimen collection will have a complete demographic cross-section, complete handling and storage history from moment of collection, complete patient history, matched specimens from the same patients, and broad informed consent.

The FDA needs tests to be developed and validated in a way that supports clinical diagnostic use in the intended population. Investigators must think about who they are going to test, not simply the best samples they can get. Studies to validate tests must be controlled analytically and for patient safety, informed consent, institutional review board (IRB) oversight, and investigational use. Test instructions must correspond to the circumstances in which the test actually will be used.

Dr. Mansfield concluded that the benefits of improved biospecimens will be better discovery, better tests, patient benefit, and good science.

Question and Answer Panel

Dr. Mansfield
Dr. Robinowitz, Medical Officer, OIVD, CDRH, FDA

Dr. Rimm asked about using retrospective samples. Dr. Mansfield said that retrospective samples are acceptable as long as conditions are accounted for. Kristen Meier, Ph.D. statistician for the Center for Devices and Radiological Health, FDA, added that this depends on the use of a test. In the case of a claim of absolute risk, it is difficult to use retrospective studies. Dr. Robinowitz noted that with growing interest in biomarker-targeted therapy, if a specimen is used to estimate the effect of a drug, it will have to be prospectively collected, a true clinical trial experiment. Dr. Mansfield cautioned about bias and use of retrospective samples.

Dr. Compton commented that the FDA has taken an increasing interest in advancing regulatory science. She asked about direct applications of the standards being discussed at this workshop to regulatory science. Dr. Mansfield replied that the standards allow regulators to judge the quality of a sample, which is regulatory science. Dr. Robinowitz observed that the almost revolutionary difference between the FDA of 1990 and the current FDA is that it now uses the largest group of biostatisticians in the world in both drugs and devices. Analytic and clinical pre-market approval is robust and in-depth.

Dr. Rainen said that from the standpoint of the manufacturer in pre-analytical procedures and tool development, the innovative technology or tool that the future needs might not be used by an FDA assay. The role of the manufacturer is to develop generic tools that will have downstream analytic methodology, which is evolving. She asked for advice for pre-analytical tool manufacturers to help them provide tools in a timely manner.

Dr. Mansfield said that many variables are involved and it is not simple. The most successful method is to get a product into research use. This is where PCR technology came from—scientists saw how it worked, saw the performance parameters, and were able to develop assays that could predict performance, and PCR evolved from research into *in vitro* diagnostic use. Researchers bypassed the FDA to get to the market, and it was possible to sell the technology for clinical use without going through the FDA. The FDA is working to close that loophole. Validation of an instrument with an assay will be built into a test.

Dr Rainen said that the process is circular, and if a pre-analytic test is not approved, a market for it will not exist. Her company does produce some products for research use only. It can be difficult for a researcher to be aware of pre-analytical tools. Dr. Mansfield said that is a business problem, not a regulatory problem. The business must market its tool. The FDA does not approve tools that do not have a use, just as they do not approve something that is not safe and effective.

Dr. Taylor asked how to get to the point of FDA approval for a product, and Dr. Robinowitz said that if the developers claim that the test works best in a patient who is positive for a certain

biomarker, they must produce an experiment that shows the correlation between the drug and the biomarker status. An analytic test can be done on archival tissues, but effectiveness must be tested on a patient. Further discussion considered the difficulty of testing on patients. Dr. Mansfield commented that collected and archived tissues blocks were used to test HER2 and then those patients were enrolled. New testing for HER2 was on old tissue. Once the drug was on the market, tumor biopsy tissue was used. She added that for most clinical trials, sample collection issues are not that critical. If investigators want to collect samples at the beginning of a trial and test later, they can be treated the same as samples tested when they are collected.

A participant asked about K-ras testing in Europe and whether the method was FDA-approved. Dr. Mansfield said that she understood that only a small number of samples were available to test so that the chance of bias was high. Then K-ras testing developed very quickly and it was impossible to run a new trial.

In response to a question of who actually owns tissue outside the context of consented clinical trials, Dr. Mansfield said that there have been court opinions about this question. Dr. Robb added that three court opinions have concerned biological material. Informed consent is a contract with the investigator using the tissue. CAP has ruled that pathologists are custodians of the tissue, but has issued no decision on who owns it. Some clinical trials are not taking samples because of pathologists' concerns about legal issues. Samples are supposed to be released for research after 2 years, and there is no restriction on releasing remnant diagnostic material to a clinical trial. He advised careful documentation of all activities with biological material. Dr. Mansfield said this is now documented in case-law, and Dr. Taylor noted that the requirement is 10 years for paraffin and 5 years for slides.

Report Out from Each Table

Included in this section is a report summary of the recommendations from each of the breakout tables for the afternoon of the first day. For a more detailed reporting of the discussions captured from some of these groups, please refer to Appendix B.

Table 1: RNA Case Study

The case study concerned archival breast cancer samples retrieved from a site near the equator. Microarray analysis found little increased expression of several signaling molecules relevant to oncogenesis, but later successful therapeutics are developed by other groups based on such increased expression.

Participants discussed some common housekeeping genes, but did not look at the genes known to be more unstable. RNA results are likely to be assay dependent. Some perform better with lesser RNA, and some require superior RNA to work at all.

The field could benefit from better reference materials, xenografts, and items that correlate from one assay to another.

Table 2: DNA Case Study

Michael Hogan, Ph.D., chief scientific officer for GenVault Corporation, reported.

The case study scenario proposed that specimens could have mutations due to the amplifying process in formalin fixation. The group discussed whether it is likely to be true, whether anything can be done about it, and assays to address the problem. The take-home message was that assays have grossly different specimen requirements and whether the specimen was good enough depends on the assay.

More data are needed about the high artifactual mutation rates. The literature does not specify rates. A two-dimensional experiment would select certain tissues in which to assess formalin fixation through time and time rate of change. Specimens could be examined at a couple of points in storage time to differentiate what happened because of tissue or storage time. It is necessary to dig deep to address the question of mutations in DNA samples during fixation. A simple quality control test and an enhanced STR test could be useful and are directions that NIST could pursue.

Table 3: Serum Protein Integrity Case Study

The group discussed key variables involved with assessment of serum proteins. These include the age of the sample and other factors. A three-tiered screening process was suggested to assess whether a specimen is good for use. An inexpensive initial screen would be based on a multiplex assay using an automated machine to assess the integrity of a blood sample and could be conducted from multiple tests such as lipid or hepatic tests. A second level would be based on immunoassay or mass spectrometry to assess protein concentration, proteolysis, oxidation, and other factors. A third tier could use spike-in materials for specialized cases. When looking for phosphorylated materials, a cocktail of spike-in materials would be used to assess reaction.

Table 4: Formalin-Fixed Tissue Case Study

Dr. Li reported.

Diagnostic tests on formalin-fixed tissues guide clinical decisions about whether to treat or not. But clinicians have learned from pathologists that there is no single perfect fixative for all biospecimens. Fixation of a biospecimen is a potentially very complex issue, depending on needs for morphological assessments or requirements of diagnostic tests for individual markers. The gold standard for assessing the effectiveness of a fixative should be a comparison of a diagnostic test on fixed tissue with that of unmanipulated, resected fresh tumor from patients rather than fresh frozen tissue. But this is not practical. Ideally, the fixative that works at room temperature is better able to control for pre-analytic variables. Formalin is the most popular and familiar tool that pathologists have been used for diagnostic tests for the past 100 years. To develop guidelines about the use of FFPE, the three most important issues to address

scientifically are identification of control molecule(s) that can serve as a reference, pre-analytic ischemia time, and time of fixation in formalin.

Dr. Robb added that Dr. Plant, of NIST, had some helpful suggestions, including beginning investigators with a cell or cell line and quickly moving into a mouse model. This cannot take 5 years—it must be completed in months.

Table 5: RNA Case Study

Dr. Mansfield reported.

The metrics in current practice are insufficient to characterize RNA quality for a repository. Group members discussed a range of possible concerns that impact RNA quality. They proposed that new metrics of RNA quality be developed. All individual metrics should be reported in addition to a composite score. Possible metrics would include pH of tissue; variety of transcript integrity measures; ratios from PCR assays; a variety of different RNA degradation mechanisms and representative genes for each enzyme activity, especially RNases; the relative abundance of different transcript species; the amount of RNA isolated from a given number of cells; and apoptotic measures such as intactness of RNA.

A multivariate equation could rate the metrics from best to worst.

Table 6: Serum Protein Integrity Case Study

Participants discussed examining post-translational modifications to derive measures of degradation and how to tell whether a sample was suitable for a new analyte. Multiple-reaction monitoring and mass spectrometry could be used to at least discover regions of plasma that could be used as markers of stability or lability. It would be necessary to qualify samples for a biorepository.

Table 7: Formalin-Fixed Tissue Case Study

Dr. Taylor

This was a challenging and punitive case study in which rival investigators arrived at different outcomes.

The points where degradation might arise mark the biggest problem in achieving reproducible results. A core biopsy could circumvent the ischemia problem. However, if diagnoses are not frank, biopsies are irrelevant. This needs to be recognized as an issue, particularly when looking at subtle differences.

The group then discussed types of materials that might represent reference standards, which have three uses: for quality assurance, to assure the quality of the assay itself; for control or validation of the specimen need to go beyond quality for intrinsic assays; and for a perfect reference

standard, which should be present at all steps of the process. The scientists represented in this process should parallel the composition of this workshop with the addition of histotechnologists. The group considered how much difference a reference standard would make to practice and decided that it would add at least 20 percent to diagnostic accuracy.

Adjournment

Dr. Dickherber reminded participants that NIST would use their experience and recommendations to direct their efforts to determine a project plan. The workshop will continue tomorrow with four separate breakout groups for each case study topic.

Dr. Compton thanked all participants for their efforts.

Dr. Dickherber adjourned the session at 5:35 p.m. EDT.

Day 2: Friday, October 22

The second day of the workshop required all meeting participants report to dedicated breakout working groups to discuss project proposals for NIST regarding the previously mentioned four focal areas. This section provides the summary reports from each group at the close of the day. For in-depth coverage of the conversations from each working group, please refer to Appendices C-F.

Report Out from Working Groups

DNA Integrity Group

Dr. Hogan

Dr. Hogan reported that group members were excited to learn how much work is being done in the area of DNA quantitation and DNA quality analysis at NIST. The work has been primarily in the forensics world and is now increasing in pathology, with the development of standards for quantification. A Venn diagram for quantitation and quality analysis needs in forensics vs biomedical work already overlap considerably, and this should provide multiple opportunities for finding low hanging fruit for extension of NIST forensics technology validation into medical genetics and companion diagnostics.

Project ideas, in order of priority, are:

1. DNA quantitation
    - NIST to provide fitness-for-purpose guidelines with existing products

    - NIST to assist in promoting the Quantifiler RT-PCR assay in medical genetics research and genetic diagnostics

2. DNA fragmentation and spike-in material
    - Modify and expand the Identifiler to become a general tool for DNA QA
    - Potentially use ERCC reference materials
    - Specimen identification and reporters for degradation

One proposal is to use existing STR-based human identification assays validated at NIST, as the basis for a new generation of high-throughput DNA quality analysis. These STR assays were, to a large extent, developed at NIST. If amended to be more sensitive to biospecimen integrity and DNA quality needs, they could show how DNA quality manifests over sizes up to a kilobase, which is the range required by most downstream assays. These projects could be direct spinoffs from what NIST has been doing quite elegantly in the forensics world.

For spike-in material, NIST has been supplying materials that could be used for internal quantitation standards and serve as the basis for analysis to assess sequencing damage. NIST has already provided reference material that could be spiked in to determine internal quality standards.

3. Purification guidelines
    - Assess impact of collection and storage variables, especially as they apply to multi-molecule extraction methods.

DNA cannot be used unless it is purified, and the purification process itself can induce additional contamination. DNA can be taken from plasma, whole blood, or tissue.

The group described the optimal team makeup:

- Statistician/bioinformatician
- Specimen acquisition expert
- Laboratory clinician
- Pathologist
- Technology expert
    - One per platform assessed

The group also considered impact assessment, with the following conclusions:

- Quantitation
    - Most platforms are extremely sensitive to input quantities, and quantitation requires standardization.
- Fragmentation/spike-in
    - Most lab errors are a result of swapping specimens.
    - Use STR Identifiler from NIST plus unique spike-in sequence ID.
- Purification guidelines
    - A mechanism is needed for quality assurance and performing comparative effectiveness/proficiency testing.

Serum Protein Group

Gordon Whiteley, Ph.D., director, Antibody Characterization Laboratory, SAIC-Frederick, NCI

The group considered "degradomics" as a process of discovery. Measures of sample integrity are not known and cannot be discovered without research, followed by validation in a trial. Several clues are available, however.

A number of pre-analytical factors are well known and well defined. These include:

- Needle gauge
- Patient position (sitting or standing)
- Venipuncture or port
- Blood tubes as a variable—lots, etc.
- Protease inhibitor
- Fasting
- Time of day
- Use of tourniquet
- Mixing
- Time
- Temperature
- Pre- and post-centrifugation
- With or without gel
- Centrifugation speed
- Freeze/thaws

Three levels of investigation were described:

- Currently used lab tests to see if existing well defined markers can be used
- Physical measurements on samples
- Discovery of new markers

Current practices include:

- Barcodes for sample tracking
- Radio frequency identification (RFID)
- Specimen identification and misidentification

The experience of nucleic acid studies can be used. A special preservative was developed to study nucleic acid and it might be possible to do the same for proteins.

Known lab tests to be considered include:

- Metabolic panels
- Potassium
- Electrophoresis/immunoelectrophoresis—LD1,2 CK
- Any tests in the LabCorp catalog that require quick measurement or transport on ice
- Possible immunoassay panels for both sensitive and robust markers
- Coagulation cascade

- Kallikreins

For physical and chemical measurements, the group recommended using the expertise of NIST and others to assess:

- pH
- Viscosity and density
- Infrared, light, fluorescence, Raman spectroscopy
- Nephelometry
- Isoelectric point
- Measurement of carbohydrates, phosphates, etc.
- Sialic acid, neuraminic acid

Liquid chromatography/mass spectroscopy are other important tools that can be used for discovery. Samples can be used as a discovery platform in carefully collected and abused samples (matching from same patients) that have been validated by a statistician. Multiple reaction monitoring is useful as a quantitative verification tool, although it is impractical because of poor throughput. The discovered markers would be converted to a higher throughput platform such as matrix-assisted laser desorption/ionization (MALDI) mass spec, immunoassay, or chemical test. Sample size is a consideration. Spiking may not be a good idea because of possible interactions with other serum proteins.

New technologies for the future might include:

- Development of a chip in a tube to monitor the conditions of storage from the time the sample enters the tube
- Multiplexing technologies for degradome solution
- New computational solutions because of the probable complexities

The group did not discuss required personnel for these tasks, except to suggest that technicians as well as scientists and clinicians will be needed.

The work remains in early stages.

Formalin-Fixed Tissue Group

Stephen M. Hewitt, M.D., Ph.D., clinical investigator, Center for Cancer Research, Laboratory of Pathology, Advanced Technology Center, NCI

Dr. Hewitt said that the FFPE group considered three primary issues. Comparability is an important and necessary measure. Morphology is embedded in anatomic pathology. And formalin is here to stay. That does not mean it won’t continue to evolve, but it will remain the main fixative. It has undergone enormous changes in fit-for-purpose, back to the 1900s, and slides from 2000 and later should be of higher quality.

A challenge is that most pathologists dealing with fixed tissue believe it is substantially divorced from frozen tissue. Frozen tissue does not necessarily have reference tissue. Well preserved archival tissue in paraffin is better than frozen tissue.

Biologic noise increases the difficulties. Tools that pathology has used for the past 20 to 30 years try to reduce the noise, but analytic tools must be improved.

Spiking is probably not a viable option, and the group did not think any extrinsic dyes or chemicals might be useful for measuring fixation.

One large controversy is whether to use surrogate measures or analyte-specific measures. Will surrogates be biochemical measures or assays that measure analytes of integrity? This question can be viewed as an opportunity rather than a problem. It allows a two-pronged approach to develop chemical measures and better measures of quality. The current issues are related to HER2 and estrogen receptors. The real problem the field faces is not having a specimen.

A related issue is whether surrogate biochemical measures can be used. When using molecular measures, RNA and protein should be assayed separately. The group agreed that protein and RNA should be investigated separately until it is proven that they track closely. Current evidence is that they do not always track closely; an excellent example is hypoxia.

The group ultimately identified three potential approaches:

- A highly calibrated approach using phosphoproteins, for which a bit of knowledge exists, but the reaction is antibody-mediated.
- Fluorescent probes for potential markers of protein.
- An analyte-specific approach to define the optimum process for specific markers. Much of the group's discussion related to HER2. The challenge of this approach is how to proceed if an optimum is found that differs from clinical practice. That would force optimizing a specimen for a specific assay or leveraging the assay.

The last issue is clinical applicability. This raises questions of cost and utility. A set of tools is needed that will allow the pathologist to hold a biospecimen in hand and provide a quality measure that can be applied beyond the highest level labs.

RNA Quality Group

Marc Salit, Ph.D., group leader, Multiplexed Biomolecular Science group, NIST

The group had a good discussion about how to address questions related to RNA quality, but decided that it was necessary to bound the scope. The effect of handling must be separated from chemical degradation. An important issue is how transcripts can be used as proxies for a biological state. How would the biological activity in a healthy sample be changed in a compromised or diseased sample? How do anesthesia, surgery, and storage time affect the RNA profile?

These limitations and case study design flaws were presented from yesterday's table discussions:

- No control over cold ischemia time
- Need for specific fixation protocol
  - Biopsy immediately to formalin
- RIN not sufficient

    - Expression from a set of genes
    - More than housekeeping high-expression genes appropriate
        - Not representative
- Kinetics of RNA degradation not considered
    - Need to select appropriate biomarkers for the study of interest
    - Design sample handling as needed for study
- Microarrays
    - Multiple genes solve problem for assessing biomarker
        - Robustness
- Need to engage multiple disciplines
    - Pathologists
    - Assay developers
    - Statisticians
    - QA personnel

The following responses were presented:

- Problem summary
    - Metrics in current practice are insufficient to characterize RNA "quality"
    - A range of possible concerns that impact RNA "quality" were discussed
    - Proposal: that new metrics of RNA quality be developed
        - Report all metrics, as well as a composite score
- Possible metrics
    - pH of tissue
    - Variety of transcript integrity measures
        - 3'/5' ratios from real-time quantitative reverse transcription PCR assays
        - variety of different RNA degradation mechanisms; find groups of representative genes for each
    - Enzyme activity
        - RNases
    - Relative abundances of different transcript species
    - Amount of RNA isolated from cells
    - Apoptotic measures

The group decided that a score card would be a better end product than a single measurement. A variety of possible metrics were discussed, and group members converged on a study that would provide a measure of RNA intactness. They proposed degrading samples purposefully and then systematically creating a library of degraded tissue samples, using RNA sequencing to obtain a measure of degradation. Signatures of intactness would:

- Expect groupings of genes
    - Subject to different modes of degradation
    - Different robustness
    - Different abundances
- Develop a dimension reduction as a composite score
    - Principal component analysis—?
- Feed with XX-platform data

- Validate on multiple platforms, multiple sites, multiple tissues
- Variety of stressors

A systematically degraded set of RNAs from tissue samples would be valuable and a resource to investigators. An empirically selected dataset that fed on multiple platforms would confirm that they repeat and are not platform dependent.

Possible stressors are:

- Time
- Temperature
- Type of stabilization
- Freeze/thaw cycles
- Duration of stabilization process
- Age of sample
- RNases

The above list is merely a sample. The objective is to measure RNA integrity and link it to mechanistically relevant modes of degradation. The ultimate goal, however, is to not have to measure every RNA sample, but rather to measure some of them. A score would be one gauge for how much confidence the user should have in the integrity of the RNA.

The group also discussed a plan to move forward with activities related to RNA integrity.

Adjournment

Dr. Compton thanked participants for their work in the difficult subjects that this workshop addressed. She has learned a great deal, and many smart scientists are focused on the topic. The workshop marks a first step forward and she anticipates many more steps. Addressing the subject of biospecimen integrity will be a world changer when the field is able to grapple with it in the right way. It is complex, and the complexities have permeated the workshop. Ultimately techniques must be useable in real world and reduced to something elegant, useable, and affordable.

Dr. Compton dedicated the workshop to Michael Amos, Ph.D., NIST bioscience adviser, who made the program possible. Between the first discussions and this meeting, Dr. Amos was diagnosed and began treatment for cancer. He has shared his experience in a public way.

Dr. Compton thanked participants for this excellent start and said she is looking forward to more to come.

Dr. Compton adjourned the workshop at 3:57 p.m. EDT.

Summary of workshop products and recommendations

Workshop participants developed a process roadmap with the following series of projects for NIST that would lead to the development of timely, fit-for-purpose, quality assessment metrics in the four focal areas:

- **RNA molecular quality assessment** – Deep sequencing of identified targets with known stressor variations in known tissues to derive a scorecard for RNA integrity
- **DNA molecular quality assessment** – Quantifier materials to address inter-lab variation and spike-in materials to address specimen mix-up and degradation
- **Formalin-fixed tissue quality assessment** – quantitative assessment of general phosphorylation, variability inherent to tissue staining, and optimal collections for specific biomarkers
- **Serum-level protein molecular quality assessment –** employ existing technologies for *in situ* storage conditions monitoring of storage tubes and a standardized multiplexed technology approach for serum quality assessment.

These critical assessment materials and methodologies were identified as being both important for the research community and achievable in a reasonable timeframe. A full report of workshop deliberations and presentations will be available before the end of the year, and will detail the projects envisioned. Immediate next steps involve development of the management plan and commitment of resources. The remainder of this section is dedicated to describing the high-level concepts for the projects in each area.

RNA Molecular Quality

The transcriptome of a tissue sample is a readily measured proxy for biological activity; gene expression profiling is part of routine assessment of cancer pathology samples. RNA degradation is a common challenge when measuring an expression profile from preserved tissue samples, as RNASEs are ubiquitous and active. RNA quality is commonly assessed with a proprietary heuristic measure, the "RNA Integrity Number," or RIN value. RIN is measured for total RNA, and is based on electrophoretic measurement of ribosomal RNA, degradation products of the ribosomal RNA, and the size population profile of mRNA. There is a wide range of experience with RIN, but little authoritative validation of it as a measure of the "soundness" or quality of a transcriptome. Development of a quantitative, objective measure of RNA integrity as an alternative to RIN is appropriate for high value samples, such as those to be deposited in a specimen bank, and might also serve to independently validate the use of RIN. Work to develop such a measure is consistent with programs underway in metrology for transcriptome measurements, and would leverage the availability of standards and techniques already in place.

A research path to develop a new RNA integrity measure was outlined in the breakout sessions addressing RNA integrity. The proposed measure would assess the integrity of the transcriptome with respect to chemical and biophysical degradation; effects on the transcriptome from perturbation (arising from patient treatment) would not be readily addressed in the proposed work.

The proposed work is to develop a multiple-gene "signature of intactness." Selecting a subset of

genes is predicated on the hypothesis that a limited number of RNA degradation mechanisms exist (on the order of 6-10), and that measuring the degree of degradation of genes that represent those mechanisms would permit a composite "intactness" score to be economically measured.

The research to test this hypothesis and develop the composite intactness assay would assess a variety of degradation mechanisms on the transcriptomes of a variety of tissue samples. The experiment design would include sufficient replication to permit a robust signature to be created. Various "stressors" of the transcriptome would be evaluated, including: storage temperature, time before stabilization, freeze-thaw cycles, age of sample, enzymatic degradation (RNases), chemical degradation, physical degradation (i.e. ultrasonic). Deep Ultra High Throughput sequencing of the RNA would be employed to directly measure the intactness of the RNA – both the messages, and the small RNA fractions.

The RNA intactness would be classified, initially using principal components analysis, and groups (i.e. genes subject to the same degradation pathways) would be identified. Representative genes from these groups would be selected to cover the dynamic range of relative abundance, and ultimately multiple qRT-PCR assays would be developed along the lengths of the RNA molecules, to provide a measure of the degree of intactness. These assay results would be aggregated into a composite intactness score that should establish the utility of the sample for further RNA analysis.

DNA Molecular Quality

The DNA quality integrity working group identified three primary projects, enlisted below. They are prioritized by increasing difficulty in achieving and the requirement for collaboration with investigators beyond the NIST. These were all identified as being of immediate and significant importance for the research community.

A quantitation standard reference material. This was identified as being a reasonably straightforward as it would involve the reissuance of the existing standard reference material (SRM) 2372. SRM 2372 is a Human DNA Quantitation Standard, which was developed for forensics laboratory applications. Coupled with robust fitness-for-purpose guidelines for using NIST products, it was felt that this would be achievable in a very short timeframe and would be of significant utility to the research community, especially for those engaged in team science projects.

A DNA fragmentation and spike-in material. It was felt that a longer-term project would be to modify and expand the Identifiler STR product from NIST to become a general tool for DNA quality assessment. This product could further inform the development a spike-in material that could be used as both an *in situ* specimen identifier as well as reporters for different kinds of degradation. It was thought that findings from the External RNA Control Consortium (ERCC) and materials toward that effort produced by NIST could potentially be leveraged for this end.

Robust DNA purification guidelines. It was thought that this project would require collaboration with other larger efforts across the research enterprise to understand which purification methods were most commonly used, especially those methods that involved the extraction and

purification of multiple molecular types. This effort would involve assessing the impact of collection and storage variables.

Formalin-Fixed Tissue Quality

There currently exists no widely used metric for assessing the quality of FFPE tissue. One of the challenges in developing such a metric is the wide range of experimental techniques that are applied to FFPE material. Consequently, the FFPE Working Group reached a consensus that no universal standard or reference can be developed to address quality for all purposes. However, there was enthusiasm for having a metric to rate FFPE tissue that could then be applied to "fit-for-purpose" use. Using this metric, individual experimenters would determine the value that will predict the success of a particular assay using FFPE tissue. Within this framework, three possible strategies for the development of standards or reference materials by NIST were discussed, with general but not unanimous agreement that further pursuit of these strategies is worthwhile.

Analysis of marker expression. One proposal to create a reference standard involves the use of one or several markers that are commonly expressed in many types of cancers. HER2 was discussed as a representative example. HER2 levels will be analyzed after varying time to fixation and other collection parameters, and compared to the expression of a marker that is stable to degradation. The identity of this second marker will need to be determined. It was also agreed that two independent assays of HER2 expression will be required. The ratio of HER2 expression to the invariant marker will be a measure of tissue quality. This methodology requires the availability of specific antibodies and a method for quantitative analysis of immunohistochemical signal in FFPE material, all of which should currently exist. Additionally, a reference material will need to be developed (such as purified recombinant HER2 protein) that could be used as a standard in this assay.

Analysis of phosphoprotein expression. This proposal is similar to that involving analysis of markers such as HER2, but instead relies on the idea that the expression of some phosphoproteins will be dynamic, a key feature of any analyte used as an indicator of quality.

Analysis of general tissue quality using fluorescently-conjugated reagents. A more general approach to assessing FFPE tissue quality involves the use of chemical probes to visualize features of cellular integrity and molecular organization. Examples of features of interest include protein and organelle distribution, levels of reactive chemical groups, and lipid concentrations. To illustrate this concept, the following scenario considers the measurement of the ratio of free amines to thiols using fluorescently conjugated molecules with appropriate reactivities. In this scenario, several potential factors that are likely to affect tissue quality, include degree of formalin fixation, protein degradation, and protein denaturation, are expected to significantly alter the number of free amines in the sample while leaving the number of free thiols relatively

constant. Quantitative measures of free amines normalized to free thiols will be compared in FFPE tissue collected after varying fixation time and other tissue collection parameters. Such comparisons will also be made after purposeful degradation of tissue quality using heat, protease, or nuclease treatment. The advantages of this approach include available reagents, and the relative ease of assay performance and quantitation. These features are likely to result in reduced variability when relative to immunohistochemical methods. However, it is unknown if amines will vary in a predictable manner or thiol levels will remain constant as tissue quality is altered.

For any of these quality assessment assays to provide meaningful information, they must be able to guide the selection of tissue for a given assay or measurement; ideally by providing a numerical cutoff above which that assay can be reliable used. Therefore, tissue of varying quality, as determined by any of the assays that are developed, will be interrogated for expression of biomarkers using a range of common techniques (FISH, immunohistochemistry, in situ hybridization, etc.).

Note: The NCI Biospecimen Research Network (BRN) program is currently funding Dr. David Rimm (Yale) to investigate analysis of marker and phosphoprotein expression as a measure of quality. NIST can support this effort by developing a methodology or reference material to allow standardization and normalization of this assay in different laboratories. Specific project plans will be developed in coordination with Dr. Rimm (Yale) and Dr. Helen Moore (Program Director of BRN).

Serum-Level Protein Molecular Quality

Serum is an important biological sample in clinical medicine because it reflects the overall proteomic profile of an individual and is readily accessible. It is also generally accepted that new diagnostic (or prognostic) tests will be based on detection of disease markers in blood or other body fluids. Serum makes up approximately 55% of blood by volume. To isolate serum from raw plasma, fibrinogen and other clotting factors are removed. Serum is composed of approximately 92% water by weight, 7% protein and 1% other solutes. Moreover, approximately 95% of the protein content is represented by fewer than 25 proteins, which would seem to indicate that the protein component is relatively simple. However, the dynamic range of serum has been estimated to span some 12 orders of magnitude, containing hundreds to thousands of proteins, and indicating that the remaining 5% of the protein fraction contributes a great deal of complexity. It is within this latter fraction that most discovery assays are focused.

Mass spectrometry is the method of choice for sensitive identification of protein analytes in complex biological samples, including serum. While these instruments provide sensitive identification and quantitative capabilities, they are not immune to analytical and pre-analytical variability such as that caused by sample degradation.

NIST proposes a two-phase study aimed at identifying a panel of protein degradation markers, and qualifying those markers given a biologically diverse sample set. While our initial investigations will focus on identifying peptide markers, metabolites that respond to treatment conditions will also be sought. The goals of this proposal are to *(1) identify new markers in serum that reproducibly correlate with protein degradation, and (2) determine the sensitivity and*

specificity of these markers, and existing markers (e.g., bradykinin), in a population of samples where biological variability is a significant factor.

Phase 1 - Discovery and measurement reproducibility

NIST proposes a series of discovery experiments aimed at identifying a panel of protein degradation markers that can be reproducibly measured. Part of this phase will include an iterative process, during which methods and metrics are refined. In Phase 1, a small number of serum and/or plasma samples (3-5) will be subjected to conditions known to affect sample integrity (see below). All samples will be procured from the caHUB repository which are collected under highly-controlled conditions. The experimental assumptions are as follows:

1. Samples obtained from caHUB are 'best case scenario' and therefore represent degradation point zero with respect to pre-storage protein degradation.
2. *If the identification of degradation markers and patterns of decay are not reproducible in a small number of samples, they will not be reproducible in a large number of samples.* We therefore limit screening sample size to <5.

Common sample storage and handling (e.g., increased temperatures, freeze/thaw cycles, etc.) as well as processing conditions (e.g., no digestion, partial digestion, mass filtration, etc.) will be varied on specimens collected from selected individuals (3-5 individuals, as indicated above) to generate a matrix of conditions. Once samples have been received at NIST, they will be processed according to the experimental design matrix. All discovery-based measurements will be made using LTQ or Orbitrap mass spectrometers following short LC separations. Quantitative validation, including that of bradykinin (see below), will be performed using multiple reaction monitoring methods on triple quadrupole mass spectrometers. It is expected that Phase 1 will generate many hundreds of data files.

Data analysis will include the following:

1. QC analysis according to methods published by the participating NIST divisions for proteomics samples
2. Ion current-based quantitative analysis for differential testing (discovery)
3. Development of new metrics for monitoring protein degradation
4. Statistical ranking of storage, handling and processing variables

Each sample will be run with enough technical replication to ensure adequate statistical power.

Phase 2 - Validation in the face of biological variability

While the logical progression of this proposal lends itself to using only the degradation markers identified in Phase I in validation studies, we propose adding at least three additional markers to these follow-up studies: (1) human serum albumin (HSA), (2) bradykinin, and (3) prostate specific antigen (PSA). For the past two years, NIST has been performing extensive experimentation on the digestion properties of HSA and, through these studies, has amassed a

large knowledgebase of experimental analysis of this protein that indicate that some HSA peptides are differentially susceptible to analytical sample treatments. We propose targeting some of these HSA peptides in validation studies. Additionally, Becton Dickinson (BD) (http://www.bd.com/proteomics/pdfs/ASMS06-Yi_11x17_prnt.pdf) has presented data indicating that bradykinin is specifically affected by a host protease cascade, and that it might therefore serve as a sensitive marker for protein degradation. PSA was also proposed by the NIST advisory working group as a marker to consider because it exhibits differential stability based on state (free, bound or truncated.) Together, with markers discovered in Phase 1, the following experimental approach will be taken:

1. Procure ~50 serum samples from caHUB covering a broad demographical distribution
2. Design MRM assays to measure peptide or metabolite targets
3. Apply degradation conditions from Phase I that cause significant changes
4. Apply sample processing conditions that maximize sample integrity from Phase I
5. Perform multiplexed MRM assays
6. Analyze data and calculate the value of each degradation marker across the sample cohort

At the end of Phases I and II, the following expectations should be met:

1. A good understanding of the storage, handling and processing variables affecting protein degradation will be established
2. A good understanding of the difficulties in discovering protein degradation markers (or lack thereof) in serum will be established
3. A good understanding of the relative impact of biological variability on protein stability testing in serum will be established
4. If Phase I successfully produces a qualified panel of degradation markers, next steps will include examining ways to engineer a high-throughput assay for protein degradation in serum

Appendix A: Workshop Case Studies

RNA Integrity Case Study

You wish to identify genes that exhibit altered expression in c-Kit-positive breast tumors after treatment with imantinib. Due to recent cutbacks in NIH funding, you need to save money. Therefore, as a cost-saving measure you obtain archival tissue from an organization that collects breast tumor biopsies from a tent located on the equator before shipping them by Fedex Ground to a facility in the Mojave Desert for fixation or freezing. Upon obtaining your tissue, you extract RNA and find that the RIN number is good. In addition, you check the expression of several standard RNAs (HPRT, GAPDH, etc.) by PCR, and find that they all are expressed. You then perform microarrays analysis, and focus on the expression patterns of several signaling molecules with relevance to oncogenesis. While previous studies reported that these molecules exhibited increased expression in tumors relative to normal controls, you find that you are barely able to detect any expression of these genes at all in your tumors, even though your microarrays passed all other quality control measures. You release a high profile paper in which you claim to have discredited the previous studies, only to be embarrassed when the competing groups develop highly successful therapeutics based on their studies.

On considering the matter, you decide that your tissue was the source of the problem. From a tissue quality standpoint, you consider the following questions:

What variables during sample preparation may have influenced biospecimen quality? How might these variables have affected the results of this study?

What information regarding tissue quality would need to be available to avoid the pitfalls encountered in each study?

In assessing RNA, what are advantages and disadvantages of RIN number? What are the advantages and disadvantages of housekeeping gene expression analysis? What alternative methods exist to evaluate RNA quality?

How would understanding the mechanism and kinetics of RNA degradation in tissue aid in assessing tissue quality?

Would it be possible to assess tissue quality by expanding the panel of genes analyzed to about 15 or 20 molecular targets? If so, what features should this panel encompass?

Assuming that “high quality” tissue is not available, are there alternative experimental approaches to address biomarker expression?

What expertise is required to manage and execute this project in order to develop a useful result (report, methodology, or even physical standard) within a 16-24 month timeframe? Given successful development of a useful result, could one estimate the potential impact on both research practices and clinical practice?

DNA Case Study

False mutation targets are wasting the valuable resources of our research labs, especially for those engaged in deep sequencing of the genome. It is believed that some of the origin for this has to do with amplifying formalin-fixed samples (in which the formalin-induced crosslinking of cytosine causes erroneous substitution of a guanine for an adenine resulting in an artifactual mutation, for example). Artifactual mutation rates of 1:500 have been reported in genomic studies of archival tissue. Other concerns that focus on DNA involve molecular integrity for epigenomic studies, such as identifying methylation sites as markers for disease and the stability of those conditions. Given the investment of time, money and valuable specimens in pursuing everything from basic discovery of genomic markers of disease or disease susceptibility to targets for therapeutics, it is critical that specimens are assessed for their fitness for purpose.

Could the high mutation rates of DNA derived from formalin-fixed tissue affect methodology related to studies of the genome? Please consider techniques related to whole genome sequence determination, genotyping, and other epigenetic modifications (e.g. methylation). Is it necessary to catalog the actual mutation rates? Are there other measures of DNA integrity that would address the quality of DNA derived from such biospecimens?

If mutation rates can be determined for relevant study methods, how could these errors be accounted for and possibly corrected? What platforms should be used to assess this? What specimen types and specimen parameters would need to be varied in order to understand the extent and dynamic range of these errors?

Is it possible to identify housekeeping genes or sequences that might act as reporters for mutations or degradation due to fixation or improper amplification? Can other targets be identified to serve as reporters for artifactual mutations? Would a spike-in synthetic molecule be an appropriate way to gauge these mutations?

What expertise is required to manage and execute this project in order to develop a useful result (report, methodology, or even physical standard) within a 16-24 month timeframe? Given successful development of a useful result, could one estimate the potential impact on both research practices and clinical practice?

Formalin-Fixed Tissue Case Study

You wish to determine if measurement of c-Kit expression levels in breast tumor biopsies is a viable way of selecting patients who respond successfully to imantinib. To perform this study, you collect archival FFPE breast tumor biopsies from various hospitals, and perform c-Kit immunohistochemsitry using a commercially available Perfecto c-Kit kit. You also include as a positive control tissue in which you had successfully detected c-Kit expression using the Perfecto kit. After processing for immunohistochemistry, you analyze levels of c-Kit in each nucleus using newly developed imaging software that can quantitatively and accurately measure the immunohistochemistry signal. By plotting c-Kit activity against survival time after imantinib treatment, you fail to find a correlation between c-Kit levels and imantinib effectiveness, such that it is not possible to predict for any given level of c-Kit/nucleus whether a patient will respond positively to imantinib. You publish this study in *Failed Drug Trial Letters.* Only 2 local journalists attend your press conference announcing these results; one of these is your nephew, who is reporting for his high school paper.

As it turns out, your study is incorrect. Your major scientific rival carries out a study proving conclusively that c-Kit levels above a certain threshold predict with 100% accuracy which patients will exhibit positive responses to imantinib. This study is published in *High Profile Drug Trial Reports*, and is extensively covered by the national media. Your spouse leaves you for the senior author of this study, and your children are no longer willing to be seen with you in public.

Please consider the following questions.

What variables during sample preparation may have influenced biospecimen quality? How might these variables have affected the results of each study?

For both incorrect studies, how would poor tissue quality lead to misleading results?

What information regarding tissue quality would need to be available to avoid the pitfalls encountered in each study?

Assume that “high quality” FFPE archival material processed with specific protocols tailored for preservation of one or more specific biomarkers is not available. What experimental approaches could be used to determine if the tissue is suitable for accurate detection of a particular biomarker, such as c-Kit? What alternative experimental approaches could be used to determine analyze c-Kit in the absence of “high quality” FFPE biospecimens?

What expertise is required to manage and execute this project in order to develop a useful result (report, methodology, or even physical standard) within a 16-24 month timeframe? Given successful development of a useful result, could one estimate the potential impact on both research practices and clinical practice?

Serum Protein Integrity Case Study

The emergence of various high throughput mass spectrometry platforms, protein microarray technologies, and the perpetually emergent field of innovative IHC screening of archival tissue have launched clinical proteomics and protein biomarker discovery methodologies requiring a virtual doubling of the bioinformaticist needs to manage the incoming data. One particular target of these investigations is identification and quantitation of large proteins and post-translational modifications (PTMs). Deciphering a disease-relevant PTM versus an artifactual modification that is a result of biospecimen reaction to environmental stresses is a significant challenge.

We therefore ask you to consider whether it is possible to identify reporter peptides from common proteins as specific indicators of sample degradation or environmental stress. Could human serum albumin be an appropriate candidate? What other ubiquitous proteins might serve as useful reporters? What specimen types and environmental stressors would need to be varied in order to understand the extent and dynamic range of the impact of these stressors?

Could quantitative discovery proteomics approaches be employed to identify peptides with differential responses to controlled environmental stress? Perhaps multiple-reaction monitoring (MRM) could be applied to variety of specimens collected under controlled conditions in order to study status over time and across storage conditions. Would MRM methods using a triple-quadrupole mass spectrometer be an appropriate platform for assessment?

What expertise is required to manage and execute this project in order to develop a useful result (report, methodology, or even physical standard) within a 16-24 month timeframe? Given successful development of a useful result, could one estimate the potential impact on both research practices and clinical practice?

Appendix B: Day 1 Case Study Working Groups

RNA Integrity Case Study Discussion

Participants from two tables discussed the case study and proposed a number of answers to the questions.

Question 1

What are the variables that influence the specimen quality and results?

Answer 1

Variables include:

- Diagnosis and how it was established.
- Collection, processing, storage, and shipment conditions.
- Temperature. Temperature at the equator is near that of body temperature—does that affect tissue?
- How the sample was treated after removal. Was it fixed in formalin or snap frozen? Formalin affects gene expression and mRNA content.
- Cold ischemic time (time from ligation and clamping to sample removal to preservation in fixative). Per research by David Hicks, narrowing of cold ischemic time from 4.5 hours to 2 hours increases accuracy of quality prediction.
- Temperature at which the tissue was kept, processed, and stored.
- Time in fixative. The longer the time in fixative, the greater the RNA degradation.
- Use of microwaving. Microwaving should never be used because it ruins RNA.
- Tracking of collection, processing, shipment, and storage data.

Question 2

What information regarding tissue quality would need to be available to avoid pitfalls encountered in this case study?

Answer 2

- A protocol for collection, processing, storage, and shipment is needed.
- Each process—collection, processing, storage, and shipment—should be controlled and documented.

Question 3

What are the advantages and disadvantages of the RNA integrity number (RIN)? What are the advantages and disadvantages of the housekeeping gene? What are some alternative methods to evaluate RNA quality?

Answer 3

- Advantage: The RIN can distinguish between good and bad RNA samples.
- Disadvantage: The RIN is a surrogate measure and measures ribosomal RNA that is not the analyte.
- Disadvantage: Changes in bioanalyzer affect the RIN.
- Disadvantage: Individual RNA vulnerabilities are variable.
- Disadvantage: The housekeeping gene is not a good control. It can be a low abundance gene or changed through sample handling. If the RNA sample is changed by sample handling, it cannot be linked to the housekeeping gene. Hematoxylin and eosin (H & E) stain contains RNase in reagents; water; and contamination from other samples via hands, coughing, breathing, fingerprints, slides, microtome blades, and other instruments.

Alternative methods to determine RNA quality include:

- Real-time polymerase chain reaction (rtPCR)
- Range of internal controls
- Length of RNA molecule
- Amplification of nucleic acids
- pH of tissue. pH and temperature of formalin are not standardized between labs, and FDA guidance or a standard for fit-for-purpose is needed, regardless of assay.
- Relative abundance of transcript species
- Amount of isolated RNA
- Apoptotic measurements of RNA intactness
- Individual metrics plus composite score
- Use of xenografts, cell lines, or mouse RNA to obtain quality measures, time studies, temperature studies, formalin penetration studies, or pH studies
- Use of large lymphomas from human tissue to obtain a large number of replicates
- Use of cryo-preserved lymphocytes

Question 4

How would understanding the mechanism and kinetics of RNA degradation in tissue aid in assessing tissue quality?

Answer 4

- Formalin fixation time—RNA quality at 8 hours compared to 24 hours compared to 48 hours fixation time
- Tissue necrosis. Longer than 24 hours in formalin increases necrosis.

Question 5

Would it be possible to assess tissue quality by expanding the panel of genes analyzed to about 15 or 20 molecular targets?

Answer 5

Participants cautioned that it is necessary to pick genes carefully for an expanded panel, and the results would be a statistical extrapolation at best.

Question 6

Are alternative experimental approaches available to address biomarker expression?

Answer 6

Biomarkers are not impacted by sample handling.

Question 7

What expertise is required to manage and execute the project?

Answer 7

A team is needed and should include:

- Molecular pathologist
- Anatomical pathologist to work on the correlation between morphology and connection to biomarkers; interact with surgeons and pathology technicians to decrease cold ischemic time; and develop standard operating procedures for sample handling and processing.
- Scientist/assay developer to understand the technology and clinical implications
- Biostatistician

DNA Integrity Case Study Discussion

Discussion chair: Joe Vockley, Ph.D., director, the Cancer Genome Atlas (TCGA) Program, NCI.

Participants commented that they did not like the way the case study was worded, contrasting technology and sample quality focus. Both approaches have limits.

Research view

The research view focuses on how to get DNA out of paraffin tissue without mutation. Instead, the focus should first be on the quality of the sample, then on the technology (the how-to of extracting DNA). Formalin is the standard. That won't change overnight.

But is the formalin technology fit for the purpose? Must chemical damage be taken for granted? Formalin formulation might change shortly. Perhaps it will become possible to determine the percent of expected damage. The question is how to measure the damage of a sample as accurately as possible. It is not clear what platform should be used—the most forgiving technologies or the least forgiving? Can a standard be developed for the formalin-induced-DNA damage platform?

Is there a better preservation technology than formalin? Other substances are superior, but they cannot be used for pathology. It might be possible to propose a fixative other than formalin, but it is necessary to acknowledge that all methods will cause some damage. How can the mutations be measured and compensated for? The difficulty is not knowing how bad the formalin-induced DNA damage problem is.

Even with DNA damage, assays could often be replicated, meeting one requirement. But the percentage of error remains to be discovered. Sensitivity is another requirement of assays. Whatever method is developed has to consume only a small amount of the specimen, so that the remainder of the specimen can be used for replication.

The group concluded that a method should be developed to evaluate the formalin-fixed, paraffin-embedded (FFPE) process of preservation and the damage it causes. A caveat is that FFPE is not the best preservation method for long-term storage of DNA.

Industry view

The group suggested creation of a PCR-based test of DNA quality, since the DNA complement of FFPE is very often PCR-amplified anyway as part of routine genetic analysis. The low-hanging fruit would then be to sample all parts of the DNA sequence at low resolution to see the damage and arrive at an assay. The assay should:

- Measure the degree of fragmentation of DNA and the lesions before the assay
- Measure the effect of such strand breakage and DNA lesions—in terms of artifactual errors—after the assay

For DNA damage to be “clinically relevant,” it is necessary to determine what constitutes the maximal acceptable lesion density, which will in general be dependent on the sort of test to be performed. In general, this might be in the <1/100 base pairs (bp) range for simple tests such as RT-PCR, but might be closer to <1/1,000 bp for sequencing and microarray applications. Other important information would include the type of damage incurred, including: single nucleotide polymorphisms (SNPs), deletions, base-plane chemical modifications and strand breaks.

After ‘standard’ formalin fixation, current data would suggest that DNA has incurred strand breakage in the range of one per 200–500 bp in a typical sample, with the inclusion of additional DNA damage at a somewhat higher density than that. Thus, any DNA analysis technology to be employed must be able to accommodate the possibility of overt strand breakage and occult DNA damage of that magnitude. The Illumina, Inc., technology also has drawbacks. The technology does not match the desired output.

NIST would like to develop a standard test to determine the length of the DNA chain fragments resulting from FFPE treatment and a test that could identify DNA damage that is independent of such fragmentation. Having performed such standardized tests for length and occult DNA damage, investigators could determine which technology could be applied to samples with observed fragmentation and damage profile.

The possibility of freezing FFPE blocks or the DNA extracted in them in nitrogen for up to 10 years was discussed, but it was concluded that a significant fraction of the observed damage may have occurred during fixation, and thus could not be mitigated after fixation by subsequent cooling.

Tests that need to be done on DNA to detect errors are:

- Length of DNA chain
- Covalent modification of the DNA bases

A number of questions were raised about FFPE samples:

- Should FFPE samples be used for discovery?
- Should FFPE samples be used for large populations?
- Should FFPE samples be used to search for a particular DNA pair? Will the damage affect this type of research?
    - Participants thought that it probably would not, but perhaps the same type of sample shouldn’t be used for all purposes.
- Can NIST develop a standardized test to evaluate the quality of FFPE samples?

The group also considered what the Office of Biorepositories and Biospecimen Research (OBBR) needs to do in the future for DNA testing. What sort of tests can OBBR use to evaluate the quality of paraffin samples? Suggestions were to test the results of exposure to formalin at different time points to see how many mutations are created (at 0 hours, 2 hours, 4 hours, 6 hours, 8 hours, etc., up to 1 week, 2 weeks, 5 weeks) and to test different tissue types—normal as well as tumors.

The group then tackled specific questions specified in the DNA case study handout:

Question 1

Could the high mutation levels of DNA damage found in formalin-fixed tissue affect methodology related to studies of the genome?

Answer 1

Yes.

Question 2

Is it necessary to catalogue the actual rate of damage formation, as a function of fixation time and storage time after fixation?

Answer 2

Yes, but only on prospective samples, since such studies would be impossible on retrospective materials. It would be important to catalogue where such damage might occur in the genome; such as specific base types or classes of base sequence.

Question 3

Are there other measures of DNA integrity that would address the quality of DNA derived from such biospecimens?

Answer 3

Yes. The induction of measured base change, manifest as apparent base changes that would be detected after sequencing, should be the last criterion to measure, since it is costly and therefore not as appropriate as a QA test.

Question 4

If DNA damage rates, vis-à-vis apparent mutations incurred upon sequencing, can be determined for relevant study methods, how could these errors be accounted for and possibly corrected

Answer 4

The group said that accounting for “hot spots” of DNA damage, and hence apparent mutations, should be done—a mutational spectrum.

Question 5

What platforms should be used to assess this?

Answer 5

All major platforms.

Question 6

What specimen types and specimen parameters would need to be varied in order to understand the extent and dynamic range of these errors?

Answer 6

All specimen types *and* normal tissue.

Question 7

Is it possible to identify housekeeping genes or sequences that might act as “reporters” for mutations or degradation due to fixation or improper amplification?

Answer 7

Yes.

Question 8

Can other targets be identified to serve as reporters for artifactual mutations?

Answer 8

If hot spots are found and sufficient analysis is conducted to determine specificity and other parameters, then targets can be identified to serve as reporters for artifactual mutations.

Question 9

Would a spike-in synthetic molecule be an appropriate way to gauge these mutations?

Answer 9

No.

Question 10

What expertise is required to manage and execute this project in order to develop a useful result within a 16–24-month timeframe?

Answer 10

The group determined a need for appropriate samples, a surgeon and a pathologist on the scene, a technologist to run samples, a bioinformatics expert to analyze and compare sites of damage, bioethicist, a microbiologist, and a chemist to interpret the DNA changes measured.

Question 11

Given successful development of a useful result, could one estimate the potential impact on both research practices and clinical practice?

Answer 11

The group felt that it would be an excellent idea to conduct the research but that it would not be possible to predict the impact on clinical practice or pharmaceutical developments.

The discussion concluded with a suggestion to conduct a parallel test, with and without fixation on the sample, to examine the next generation sequencing with greater clarity. In a 12-month study, investigators could measure fixation time and storage time (in paraffin) on the mutation of the sample. In all of the time gradations of the various tests—1 hour, 24 hours, 2 days, 5 weeks, etc.—they could determine whether significant enough damage is being done to the samples to justify more tests.

The group concluded that the final aim of the case study was to develop an effective, low-cost method of testing DNA integrity.

Formalin-Fixed Tissue Integrity Discussion

Group members:

Gary D. Fletcher, Ph.D., technology leader, Becton, Dickinson and Company Preanalytical Systems

Tianhong Li, M.D., Ph.D., assistant professor, School of Medicine, University of California, Davis

Anne Plant, Ph.D., group leader, NIST

James A. Robb, M.D., Governor of College of American Pathologists (CAP), consulting pathologist, OBBR

Max Robinowitz, M.D., medical officer, FDA

Donald Coppock, Ph.D., scientific review officer, NCI

Participants began with a review of what they do and the questions they would like to answer.

- As an FDA pathologist, Dr. Robinowitz is interested in immunohistochemistry and drug trials. Many debates are ongoing about models for appropriate drugs and design of clinical trials.
- Dr. Plant would like to evaluate image analysis algorithms and predict whether they will help solve image analysis problems such as biologic variability and organization of data.
- Dr. Li does translational research on lung cancer and wants to integrate research into clinical care. She is looking for a universal fixative.
- Dr. Fletcher works on a wide variety of products related to anatomical and surgical pathology. Even if formalin is not changed, tools to record time of collection and other variables would make it easier to use. Products must be developed to automate and standardize collection of tissue.
- Dr. Robb said that the focus of the workshop is to tell OBBR how to collect and tell NIST what standards to use. A common denominator is needed, a constant, with a variable—e.g., time in formalin—that changes. He is working with a group on a universal fixative.
- Dr. Coppock wants to know more about biospecimens in reviews for grants and contracts.

Participants agreed that sharing samples is a challenge.

The case study involved measurement of c-Kit expression levels in breast tumor biopsies to select patients who will respond successfully to imantinib.

Question 1

What variables during sample preparation may have influenced biospecimen quality? How might these variables have affected the results?

Answer 1

- It is important to find whether the target sought in the tumor exists in normal cells. Human epidermal growth factor receptor 2 (HER2) is expressed normally at a lower level and is overexpressed in positive sample.
- The control should not vary or vary at a low rate. Measurement must be proportionate to concentration.
- Understandings about deterioration are advancing. It was recently discovered that formalin mutates DNA. Formalin takes 6 hours to fixate.
- Fixation at low temperature probably would not cause DNA mutation.
- Even frozen, however, living tissue mutates and SNPs change in chemical exchanges. Validating frozen tissue for clinical use can be a problem.
- While the changes that occur in formalin-fixed tissue are known, it is not clear where in the DNA they can be found.

Question 2

For both incorrect studies, how would poor tissue quality lead to misleading results?

Answer 2

- Bookkeeping and reporting are the basis of valid results. Full documentation, clinical annotation, and full specimen history annotation are essential.
- Clarify how markers work.

Question 3

Assume that high quality FFPE archival material is not available. What experimental approaches would be used to determine if tissue is suitable for accurate detection of a particular biomarker?

Answer 3

- The usual controls are external controls to ensure that the analytic process is working. But that does not provide information about the pre-analytic condition.

Question 4

What expertise is required to manage and execute this project in order to develop a useful result within 16 to 24 months?

Answer 4

- It depends on the resources that are available.
- Fluorescence in situ hybridization (FISH) and HER2 assays must be conducted within 1 hour before the samples begin to decay.
- The biggest issue is that fixation requires an overnight waiting period. Microwaves are being investigated to accelerate fixation.
- Rapid processing has not been validated. Huge variability in tissues has been documented.
- Knowledge remains incomplete about how to freeze tissue. Snap-frozen tissue (liquid nitrogen) cannot be qualified.

Dr. Robb addressed a number of questions about qualifying samples. The mechanism of study is less important than the actual sample. A living cell sample is not the same as a piece of human tissue. The best tissues are seen through H & E morphology. Formalin is clinically correlated with outcomes. Dr. Robinowitz said that formalin is standardized artifact, turning tissue into leather and staining it and manipulating it in a standard way.

Participants made several other points about use of FFPE tissue:

- Concern was expressed about the use of formalin-fixed tissue for diagnosis.
- Clinicians need biomarker tests to make treatment decisions.
- Information from a fixed tumor specimen must be in the context of tumor morphology.

- There is no perfect fixative, but formalin is the most popular tool for pathologists. Standards must be developed, but many internal issues need to be addressed including internal control of stable molecules, ischemia time to process tissue, and how long tissue can be fixed
- The type of target molecule is also important. This relates to fitness for purpose. FFPE is known to be problematic for phosphoproteins. Is the target protein DNA or RNA? Information presented today indicated that formalin can cause mutation.
- It is difficult to quantify the effects of FFPE.
- Time or temperature might be critical variables.
- About 80 different parameters are possible for FFPE. The group will have to pick a few for a starting point.
- Dr. Li commented that about one-third of a gene can by changed from the fixative, but genes do not change from fresh to frozen. She added that the Onco*type* DX Breast Cancer Assay is a useful test, particularly for estrogen receptor (ER) status. MammaPrint is used more in Europe than in the U.S.
- An animal model might not be as useful as human samples.
- Breast cancer tumors at diagnosis are becoming smaller and smaller, which is why Dr. Robb is conducting a study of breast tumors from women in Latin America.
- Discussions of FFPE can go in many different directions.

Appendix C: RNA Integrity Working Group

Group members:

Sherilyn J. Sawyer, Ph.D., program manager, Office of Biorepositories and Biospecimen Research (OBBR), NCI (facilitator)

Curt Hagedorn, M.D., professor of medicine, Huntsman Cancer Center, University of Utah

Christos Hatzis, Ph.D., vice president of technology, Nuvera Biosciences

Irina Lubensky, M.D., branch chief, Cancer Diagnosis Program, NCI

Elizabeth Mansfield, Ph.D., senior policy analyst, Office of In Vitro Diagnostic Devices, Center for Devices and Radiological Health, FDA

Lynne Rainen, Ph.D., scientific director, Preanalytix, Becton, Dickinson and Company

Marc Salit, Ph.D., NIST

James Sorace, M.D., M.S., medical officer, Assistant Secretary for Planning and Evaluation,

U.S. Department of Health and Human Services

Terrence Paul Speed, Ph.D., B.Sc., Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia

P. Mickey Williams, Ph.D. director, Patient Characterization Center, Clinical Assay Development Center, SAIC-Frederick, Inc., NCI

The group listed the goals of today's sessions:

- Merge ideas from yesterday's groups and develop into a 16–18-month pilot program.
- Design a roadmap for the pilot.
- Collaborate among NCI, NIST, and FDA to develop standards for RNA integrity.
- Consider experimental variability to assess RNA integrity
- Identify the group of experts who would work on this project.
- Discuss the impact of the pilot on clinical medicine.
- Address the question of whether formalin, fresh frozen, and paraffin-embedded all are options/solutions.
- Address the question of how to measure and have standards in stabilizing RNA. Actual stabilization is a more long-term goal, not the ultimate goal of today's considerations.

Current Issues with RNA Integrity and Measurement of Quality

<u>Microarray variability</u>

One current issue related to RNA integrity and measurement of its quality is microarray variability. Dr. Salit questioned what is needed to obtain a good gene expression profile. Dr. Speed commented that an assay that summarizes the entire microarray data is superior to other assays. Dr. Williams noted that all microarrays are not equal and considerable variation exists among different arrays. Dr. Sorace asked whether recommendations for how assays are done would be helpful.

Fit-for-purpose

Dr. Salit asked the reason for measuring RNA, and Dr. Speed questioned how good the data are for the task. Dr. Mansfield commented that measures must align with how the specimen will be used.

Degradation

Degradation is different in various tissues, making measurement very complex, Dr. Hagedorn said. Dr. Rainen added that penetration rates also differ in different tissues. Dr. Sorace suggested literature review and focused experiments on tissue variabilities. Dr. Hagedorn asked whether liquid nitrogen could stop degradation and gene induction. Dr. Salit said that further understandings of the mechanisms of RNA degradation are needed.

Gene induction

Dr. Rainen commented that gene induction occurs within 12 minutes in whole blood studies. She added that gene induction should be considered in addition to degradation in blood and lymphocytes, even those that are cryopreserved. Dr. Hagedorn said that liquid nitrogen can stop degradation and gene induction.

Bias

Dr. Williams observed that bias is introduced *ex vivo* and there is a need to understand, qualify, and quantify bias. Also, multiple instruments come with bias. Dr. Hatzis said that the tradeoff might be between an assay that has low noise but could introduce bias and conditions that would increase noise and would not have bias but would have too much noise. He added that bias exists against low abundance probes.

Animal versus human studies

Dr. Rainen said that studies of formalin penetration and cold ischemic time are done in animal models and perhaps an animal model could be considered for these purposes. Metrics are not available, however, and trying to determine metrics with variable handling techniques is difficult in humans. Dr. Sorace noted that many procedures are cheaper and faster in an animal setting. An example is the development of microRNA extraction techniques in animals. It is difficult to obtain quality samples in humans.

Dr. Sawyer said that the goal is to perform in humans. Work with formalin in human tissues is ongoing in hospital centers. Dr. Lubensky asked whether pristine animal samples and associated metrics could be relatable to humans because time zero conditions will not be achieved. Dr. Rainen commented on the value of an animal model, adding that animal metrics are relatable to humans. Animal models are needed for time testing, and only animal models will provide sufficient tissue for paired testing. Dr. Sorace agreed. Dr. Rainen added that animal tissue experiments will enhance human metrics. Many questions can be tested on an animal model without wasting precious human samples.

Dr. Hagedorn asked how quickly metrics are needed. Would testing in animals and then translating markers to humans expedite progress toward the goal? Dr. Salit said that animal and human testing can be conducted concurrently—they are not mutually exclusive. Knowledge is needed of tightly controlled animal studies. Dr. Sawyer asked whether animal studies have been done. Beckon Dickinson has done some studies and will try to make data publicly available, but some parameters (e.g., cold ischemic time, time in formalin) were not investigated. European trials are ongoing. Dr. Rainen said that it is necessary to choose the best animal metrics and mesh animal and human study results.

Small tumor samples

Dr. Rainen commented that samples are becoming smaller and smaller because tumors are detected at smaller sizes.

Fixative

Dr. Lubensky spoke of the need to concentrate on fresh frozen versus paraffin specimens. Dr. Hagedorn noted that formalin alters hydroxyl groups in RNA and cleaves it into smaller fragments. Dr. Rainen added that temperatures of FFPE are a factor and lower temperatures can reduce RNA degradation.

RNA extraction versus whole tissue

Dr. Sorace said research and clinical decisionmaking have different needs. He asked whether RNA and protein extracts should be stored rather than whole tissues. Dr. Mansfield said that storing intact tissues is important and suggested pairing RNA extract and tissues. Real tissue samples are needed, she added, and extraction methods might evolve. Dr. Salit said that many steps are involved in extraction of RNA.

Research versus clinical diagnostics

Dr. Sorace observed that inherent differences exist between research and clinical diagnostics. Dr. Mansfield said there is a floor for diagnostics to meet quality parameters that allow an assay to work but not a top. She asked whether it is necessary to find a floor for metrics.

Prioritize cancer tissues

Dr. Hagedorn asked which cancer tissues would be prioritized. Dr. Williams suggested picking the top five tissues.

RNA Integrity Measurements

Dr. Williams proposed rtPCR of adjacent H & E RNA extraction and aliquot. Real-time PCR provides an indicator of RNA and DNA yield and is a simple assay to estimate. It is possible to aim for low, intermediate, and high level transcriptases. He added that sustaining a level of eosin intensity usually yields poor RNA.

Other techniques to measure RNA integrity could include:

- Chemical activity level of RNA degradation to homogenize tissue to obtain pH
- Yield of RNA per gram of tissue
- RIN used in conjunction with other metrics
- Length of RNA
- RNA deep sequencing

Project Proposal Critical Questions

- RNA Isolation methods
  - Snap freezing
  - dT priming
  - Random hexamer primer
- Tissue specificity
  - Tissue types
- Stabilization method
- Methodology
  - Method based on method/platform
- Effort/time
- Bias versus low abundance probes
- Bioinformation heavy
- What are critical insults?
- What is the gold standard?
- Information on amplicon length for meta-analysis

Roadmap Steps and Scope

Literature search and development of compendium

The first step is to gather known metrics of RNA integrity and types of insult on RNA and let the literature search be the guide to what is known and unknown. This includes information about:

- Amounts of RNases
- Decay rate per time on tissue/blood
- Collection and storage parameters and how they affect RNA
- Metadata on samples of different cancer tissues from the Gene Expression Omnibus (GEO)

Deep RNA sequencing is needed and requires at least 10–20 million reads. Non-coding RNAs are also captured with RNA sequencing with 5-prime cap. It is possible to obtain aggregate measures of degradation from deep RNA sequencing because of biological effects and gene induction.

Ideas for developing metrics

Group members suggested a number of different approaches to develop metrics related to RNA integrity, including:

- Triage metrics from literature.
- Design experiment utilizing agreed-upon metrics.
- Develop a continuum with metrics using a wide variety of gene panels.
- Use paired samples from caHUB.
- Assign a composite score of RNA integrity.
- Use of archived samples with proven biomarkers is a possible approach, but insults to these samples are unknown.
- Move into future biomarkers to develop a scorecard.
- Use pH, staining intensity, and RIN.
- A measure might be needed for each category of loss (e.g., loss of RNA size).
- A standard curve/database from next generation RNA sequencing is needed.
- For baseline, use flash frozen samples from quick autopsy from OBBR.
- Quality is multidimensional and required multidimensional measures.
- Pay attention to amplicon length and its relationship to insult.
- Different assays measure different places of RNA.
- Use quantitative rtPCR to assess peripheral blood; see gene induction with rtPCR.
- Replicate sampling within platforms; multi-platforms will add robustness and power.
- Know baseline prior to sample stress; then, with adequate controls in place, perturb the sample.
- Develop a scale of stress and a matrix of stresses and alter them one at a time.
- Develop groups of metrics in two to three phases.
  - Phase I: sample collection, measurement of cold ischemic time
  - Phase II: after RNA extraction, determine composition of tissue, e.g., percentage of tumor, necrosis

- Compare the amount of RNA abundance (low, intermediate, or high) in samples to matched controls.
- Develop panels that provide information about stability and focus biomarkers on stable items.
- In the paraffin-embedding process, use a low melting temperature for the paraffin, which does not degrades RNA as much as a higher melting temperature does.

Select a panel of genes

The following considerations and questions arise when selecting a gene panel:

- What gene expression set can be used across all types of tissue or specific to particular tissues?
- Mine GEO to select genes that are variable across tumor types and identify levels of gene expression tissue by tissue.
- An alternative to sequencing might be measuring gene expression. Perhaps data already exist for key tissues.
- Select a cancer metabolism panel. The NanoString company has useful products. GEO could also be mined for this purpose.
- Develop a list to use in parallel with sequencing.
- A potential starting point would be identification of 40 genes, which could be indexed.
- Search for studies of tissue-specific panels.

Obtain Tissues

Group members suggested a number of topics related to how tissues can be obtained:

- Organ procurement organization (need for tissues to be drug- and anesthesia-negative).
- Need for rapid autopsy.
- Surgical sites.
- OBBR has staff and standard operating procedures in place. Liquid nitrogen flash freezing can be used for stabilizing.
- Compare formalin or universal fixatives with flash freezing.
- Paired box (PAX) gene fixation is another approach.
- NIST must have raw tissue.
- Conduct a pilot study with blood or obtain preserved organs that are “alive” from quick autopsy. Dr. Compton indicated a preference for this approach where NIST has tissues and can apply stressors/insults to develop metrics.
- Is a time-zero snapped frozen control needed?
- Dr. Compton suggested starting with a simple system, eliminating hard-to-control variables from biospecimens, and using blood. She also suggested beginning with homogeneous solid tissue such as liver, which demonstrates uniformity of cell type that is reproducible from person to person.
- It will be important to document the total micrograms of RNA in addition to RIN.

- Measure the degradation of the RNA. It is also important to obtain a differential of the best case sample components. This is best done by snap freezing tissue in liquid nitrogen and reading the exons.
- Prospective caHUB samples need to have access to block, fresh frozen, and FFPE tissue and slides. Tissues collected by caHUB should have access to a standard-of-care block from the pathologist.

Develop conditions/insults/stressors

- Quantify the stress/insult.
- Develop a continuous quality score that is calibrated.
- Develop a database of what stressors do to transcripts.

Stressors

Stressors include:

- Time to stabilization
- Type of stabilizer (e.g., neutral buffered formalin, liquid nitrogen)
- Temperature
- Duration of stabilization
- RNase-free environments
- Age of stabilized sample
- Number of freeze/thaw cycles
- Anesthesia
- Ischemia
- Degradation

Dr. Hagedorn suggested that stressors can differ for different tumor types. Dr. Rainen cautioned that working with organ procurement organizations can result in inadequate samples (RIN, yield per mg of tissue, gene expression panel) even with explicit instructions.

Dr. Hagedorn presented to the group an example of how RNA deep sequencing to assess the impact of one stressor using preliminary data from his own lab. He presented RNA sequencing data where human liver biospecimens were placed at room temperature for increasing lengths of time prior to RNA-seq analysis and the impact this had on results as an example of the research projects he thought appropriate for standards development.

Develop RNA Integrity Metrics Based on Insults

A multi-sample, multi-tissue study could aid in the development of a metric to assess systematic degradation. A signature of intactness could be developed related to the robustness of stressors. This could allow RNA to be isolated and divided out to labs to do their preferred panels. Data could then be returned.

Dr. Salit commented that some measures of quality are known, such as in paired samples of fresh frozen tissue. He recommended development of a component composite quality score. Deep sequencing for 100 samples could demonstrate the effect of differential conditions on RNA.

RNA could be classified according to abundance. Selecting a group of genes from each class could be an inexpensive way to measure rtPCR.

Correlates could be defined to show the amount of degradation specimens have been exposed to. For example, microRNAs could provide a multi-gene assay of intactness. Degradation could be plotted on a calibrated axis that also showed amount of insult to the sample. It is necessary to determine which genes are induced and which miRNAs are degraded, thus demonstrating the genes most likely to be degraded or induced because of time, temperature, and fixative variables.

Necessary Collaborators

- MicroArray Quality Control (MAQC) project/FDA
- Residents from MAQC for correlation
- Coordinate with biostatistics/bioinformatics for primary and metadata analysis

Appendix D: DNA Integrity Working Group

Group members:

Tony Dickherber, Ph.D., AAAS Science Policy Fellow, OBBR, NCI (facilitator)

Michael Coble, Ph.D., research biologist, NIST

David Eberhard, M.D., Ph.D., director of clinical trials pathology services, LabCorp, and adjunct associate professor, University of North Carolina

Margaret Gulley, M.D., professor of pathology, University of North Carolina at Chapel Hill

Michael Hogan, Ph.D. chief scientific officer, GenVault Corporation

Marianne Henderson,M.S., Chief of Office of Division Operations and Analysis, Division of Cancer Epidemiology and Genetics, NCI

Kenna Mills Shaw, Ph.D., scientific projects manager for the Cancer Genome Atlas, NCI

Leming Shi, Ph.D., associate director, Center for Bioinformatics, National Center for Toxicological Research, FDA

Morning Session: Overview

Dr. Dickherber said that NCI, NIST, and OBBR are collaborating on a pilot project that they hope will become a long-term commitment to develop tools and standards to quality control the human biospecimens used for research and will seek additional funding. The focus is on the big questions that need to be asked about the DNA molecule. What are the quality issues? What are the benchmarks that will satisfy the regulatory requirements? Today's focus is to explore the primary issues in DNA research beyond formalin-fixing, which was discussed yesterday. What are the research issues? What are the clinical practice issues?"

Dr. Coble's group has focused on forensics standards for STR (short tandem repeats) in genomic DNA, and sequencing standards for mtDNA (mitochondrial DNA) sequencing and DNA Y-strands. Coble's group has also been involved in the use of the Quantifiler assay and the supply of DNA quantitation standards for the Quantifiler. They note that too much DNA will create artifacts in the Quantifiler and thus it is essential to target the amount of DNA put into an experiment.

NIST, especially Dr. Margaret Kline, has engaged in systematic studies of long-term DNA storage, mainly from a forensics perspective, and shown that the standard identifiler assay can be used to evaluate DNA damage over time. The test has a biometrics component, which is used to help solve kinship issues. NIST is working on this test with the U.S. Department of Homeland Security to help identify immigrants and their kin.

NIST is also focused on cell line certification and authentication, using STR to help ensure that researchers are working with the cell line that they intended to work with.

Dr. Dickherber noted that yesterday's discussion focused on both the process of developing appropriate recommendations that could support this effort and what the relevant considerations were. He expressed the desire to continue to focus on both of these components today, but to now venture beyond the process and identify the biggest issues affecting DNA quality and its effects on genetic analysis.

Dr. Hogan asked about older DNA samples, archived for 20 years by NCI. How do these older specimens perform over time? Retrospective experiments are desirable to test for DNA stability. Ms. Henderson commented that NCI has samples that date back 40 years. She said it was important to know what was "junk." She noted that NCI does not collect tissue and FFPE samples. However, she said, it might be inappropriate to focus only on tissue (as caHUB does). She pinpointed several issues:

- Phenol-chloroform extraction interferes with DNA extraction, particularly the longer strands. Can such residue be detected and removed?
- Buccal swab collections are easier to obtain and study. The question is how to get more out of them (i.e., to look at more than just "junk" bacterial and food contamination).
- Prequalification of samples is important. It was suggested that a version of the Identifiler assay could be developed as a generic QC test, ahead of time. Such QC costs more upfront, but is more cost-effective than using poor samples. Aliquoting now makes less sense than combining.

Dr. Gulley stated that NIST can help ensure the quality of specimens and test results. She urges more exploration of older samples, to see how long the DNA is viable.

Dr. Shaw said it was difficult to obtain "good" DNA. She would like to see NIST develop metrics for DNA. Dr. Eberhard asked about the meaning of tumor content in a sample. Is it the area occupied by a tumor? What is the percentage of tumor tissue?

With regard to the discussion of day 1, a group member asked about the meaning of the loss of integrity. Is it the loss of bases? Or base modification? Or the degree of fragmentation (leading to abandonment of sample)?

Major points in the search for standards in DNA samples were outlined and discussed:

<u>DNA quantitation</u>

How much DNA is in a sample? It was argued that the ability to correctly measure the amount of DNA in any sample was a crucial parameter that is often performed incorrectly. Two key parameters are needed:

- Sample size (i.e. the mass concentration and total mass of sample)
- The heterogeneity of a sample (i.e. the fraction that is tumor vs normal in collected tissue, or human vs bacterial in a swab)

This is a general heterogeneity problem—fit for use. For example, as in FFPE, the amount of tumor DNA can be so small as to be indistinguishable for the contaminating normal tissue in the

specimen. At that point, mutation in a small DNA sample cannot be detected. This could be unsafe for the patient and compromise the meaning of research studies on such tissue.

Dr. Eberhard stated that he needs standards in order to be able to say to a clinician, “This patient sample is not adequate to determine the degree of mutation.” The question for NIST is: What is the number that should be associated with the ratio of mutant to wild that is good enough to produce a finding for treating the patient?

DNA heterogeneity

Dr. Shaw discussed SNP assay purity. She would like a better view of purity in the context of heterogeneity.

Standards for spiking

Dr. Gulley advocated monitoring everything that tracks what happens to DNA that is spiked-in. She urged spiking DNA to know its quality, sequence it, measuring it in the assay, and tracking damage to the molecule.

Fragmentation assay standards

Dr. Gulley urged measuring the DNA fragment in all samples — via an assay off the shelf assay to measure whether DNA fragments for all specimens to be studied.

It would be useful to determine how to prove which methods are better than others. CSA Proficiency Standards measure this, by subscription.

Labs that use different methods from each other and single labs that use different methods within the lab should be surveyed and the results made known. It is possible that certain platforms have more errors than others. Dr. Dickherber noted that NIST has already done an assessment of which methods and labs are better than others.

Once this information exists, it is essential to disseminate it and educate people. Dr. Coble noted that the starBase Web site associated with his group has a great deal of this information online.

DNA purification (FFPE)

NIST needs to develop surveys that can be used to identify which methods are best for determining DNA quality, fragmentation, and other factors in various biospecimen types (e.g. FFPE, buccal swabs, etc). Multiple analytes are necessary.

Testing comparability of data—platform proficiency

Two approaches are necessary: (1) to identify platforms, methods, and products for which the qualified DNA are fit for purpose and (2) to ensure that all potential users know how to use the NIST products that are developed. Dr. Gulley suggested a study of how to use NIST products. Ms. Henderson asked whether there is a recommended temperature and a recommended duration for freeze-storing DNA. The effect on aliquoting is the issue. Is colder better for DNA? Is it worth spending money on equipment to keep DNA at -80 or at LN2 temperatures?

Specimen storage

Storage is a function of the specimen type.

Dr. Coble discussed the need for validation of assays. Is there a standard method of validation? Some labs try 5 samples, other 5,000, before moving forward with an assay. Another member of the group noted that lab inspectors check validation. There is a recommended number of samples, depending of what is being tested.

Dr. Hogan commented that caHUB wants to be the model for biobanking. caHUB should subject itself to testing, as well as those who provide samples to them.

Dr. Gulley noted that many lab standards already exist. He expressed a wish for worldwide lab standards. The International Society for Biological and Environmental Repositories would like to make this happen but does not have funding.

Dr. Gulley suggested that it would be beneficial to look at the same series of questions about DNA, RNA, and proteins at the same time. Combining the perspectives could be productive. She would also like to know more about new extraction methods to use on different analytes. Some analytes are extracted more effectively one by one. Or sometimes a lab wants the same tests to be run on exactly the same DNA/RNA sample.

Dr. Shaw asked whether tests should be conducted to evaluate DNA assays in order to help with the next generation of sequencing. The Clinical and Laboratory Standards Institute publishes guidelines for labs.

FDA Perspective

Dr. Shi responded to a number of news stories stating that the FDA is dealing with very few applications. He noted that the number of proposals submitted to the FDA has declined because it takes so much time (12–15 years) and money (approximately $1 billion) to go through the FDA system. In addition, a limited number of patients are in clinical trials.

Dr. Shi commented that each platform has its own types of mistakes. Drug manufacturers do not want the information to get out. The frightening thing is that the high rate of mistakes is for clinical results rather than research. It will hurt patients. He concluded that independent verification is absolutely essential.

Dr. Gulley recommended approving a drug based on the best tests that can be conducted currently. Then the drug can be refined, based on new technologies that emerge. But she urged that drugs not be held back.

Ms. Henderson expressed concern that the field is not ready to move forward on DNA quantitation because of the multitude of ways that molecules are now prepared. NIST needs to provide guidance. Does every way of evaluating DNA produce the same results? Probably not. Her goal is to remove chloroform from the picture.

Afternoon Session: Top Three Project Ideas for NIST to Undertake

The discussion focused on tasks or projects for NIST to undertake that would produce a near-term effect on both research and development and clinical practice. Participants reviewed and pruned the morning's list of priorities, placing the lowest-hanging fruit (most doable tasks in the short term) first. The top three project ideas recommended below dovetail with work that NIST has been doing for many years in the forensics world. That is the reason that these will give the greatest value in the near term. In order of priority, the top three tasks that NIST should undertake are:

1. DNA Quantitation—looking at the quantity and quality of DNA in samples

NIST has helped develop the Quantifiler assay, which is a world-wide standard for measuring the quantity of DNA in forensics samples, but is not widely used outside of forensics labs. Perhaps the FDA and NIST need to disseminate or mandate use of the Quantifiler and train personnel in the research community to use it, so that its use would be extended. The FDA also could teach standard protocols or issue guidelines for how to run PicoGreen, the primary method of small sample quantitation in non-forensics applications, since at present the way people use PicGreen differs, creating widely varying results. It could also be helpful to conduct a training session via satellite or webcast. Benefits would include enhancing the ability to connect three samples from the same patient (e.g., blood, urine, liver tissue).

2. DNA Fragmentation Assessment

It was proposed that NIST could modify its highly-standardized STR assay, which was originally developed for forensic ID, to become a powerful "off-the-shelf" tool to measure DNA fragmentation (a key component of DNA quality) in small, valuable DNA samples: of the sort used in biobanking, clinical diagnostics, or research. This Identifiler technology is well known and easily used as a high throughput test, therefore such STR-based DNA QC could become a world-wide standard.

In addition, a synthetic spike-in could be incorporated to act as an identifier as well as to enhance quality. Perhaps an External RNA Control Consortium spike-in could be used for this. (It is necessary to verify that the spike-in is inert to DNA.) Add sequences with susceptibility to specific stressors. This could also address quantitation—help determine how much is lost while profiling.

All of this potentially could be done by using NIST materials that are already in hand.

3. Guidelines for Purification and Other Pre-analytical Variables

Variables include the duration of storage, the temperature, and the method of purification. This would take the longest to accomplish of the three NIST goals. Participants urged that the project include application to multi-molecular (DNA, RNA, protein, etc.) extraction methods.

Considerations for the Three Tasks

Researchers much decide which specimens to collect and test. Possibilities are:

- Blood and blood products (serum, etc.)
- BD vacuum tubes
- PAX gene
- Buccal swabs and collection methods (cyto-brush)
- Biopsy products (including fine needle aspiration)
- Other body fluids

Some relevant collection parameters such as the technology platform being used still need to be explored.

Also, the need to understand contamination and other impacts of chemical storage is acute.

In addition to the NIST group, the DNA integrity team should include:

- A statistician/bioinformatician.
- A specimen acquisition expert. It is important to keep in mind the breadth of data elements that will be included in the sample collection (e.g., gender, age, smoking history).
- A laboratory clinician.
- A pathologist with training specific to the sample or disease type or gene type being tested.
- A technology expert; one expert per platform should be included (i.e. PCR, the various microarrays. sequencing platforms, etc)
- A chemist, with deep expertise in the physical biochemistry of nucleic acids, to be provided by NIST or an outside collaborator.

Modern molecular assessment technologies may require going back over collection and processing variables (e.g., freeze/thaw cycles). The Biospecimen Research Network (BRN) should do this rather than NIST.

It is important to investigate long-range/long-term damage inflicted on large DNA fragments (100 kbps). Along with this should come longer-term investigations into subtle alterations, such as methylation.

The group concluded its session with a consideration of impact and accountability:

- *Quantitation*: High-profile platform technologies (e.g., next generation sequencers) are unforgiving and extremely sensitive to input quantities and quantitation standards. It is proposed that the Quantifiler assay, first validated by NIST for forensic DNA quantitation, could become a new standard for DNA quantitation in biobanking and clinical genetics.
- *DNA fragmentation assessment/spike-in material*: The state of DNA fragmentation, especially in FFPE and stored samples, is poorly understood and no simple, fast methods are available to quantify such fragmentation, especially on small samples. A method is proposed to measure such fragmentation, that could become a world-wide standard:
  - Modify the Identifiler STR from NIST to become a high-throughput DNA fragmentation test. The range should be extended from <500 bp to < 100 bp detection
- *DNA Identification and tracking*: Many lab errors are due to accidental swapping of specimens or mislabeling. The same two methods proposed for DNA QC, above, would serve as a tool for sample tracking
  - Use the modified Identifiler to confirm sample ID
  - Spike-in a unique sequence identifier
- *Purification guidelines*: It is essential to secure specimen quality assurance to make comparisons of samples effective.

Group members acknowledged that for NIST to perform everything outlined here would be an overwhelming undertaking, given the magnitude of the work and the time that it will take to achieve results. NIST needs to share the load with BRN and other consortia.

Appendix E: Fomalin-Fixed Tissue Integrity Working Group

Group members:

David Litwack, Ph.D. (facilitator), AAAS Fellow, OBBR, NCI

Jeff Cossman, M.D. chief executive officer, U.S. Diagnostic Standards, Inc.

John Elliott, Ph.D., research scientist, Biochemical Sciences Division, NIST

Gary D. Fletcher, Ph.D., technology leader, Becton, Dickinson and Company, Preanalytical Systems

Stephen M. Hewitt, M.D., Ph.D., clinical investigator, Center for Cancer Research, Laboratory of Pathology, Advanced Technology Center, NCI

Allison Hubel, Ph.D., M.S., professor of mechanical engineering, University of Minnesota

Shyam Kalavar, M.P.H., scientific reviewer, FDA

Piotr Kulesza, M.D., Ph.D., assistant professor of pathology, Northwestern University School of Medicine

Tianhong Li, M.D., Ph.D., assistant professor of hematology and oncology, School of Medicine, University of California, Davis

Laurie E. Locascio, Ph.D., chief, Biomedical Science Division, NIST

Lisa McShane, Ph.D., mathematical statistician, NCI

Christopher N. Otis, M.D., professor of pathology, Tufts University School of Medicine

Anne Plant, Ph.D., group leader, NIST

David L. Rimm, M.D., Ph.D., professor of pathology, Yale University

James A. Robb, M.D., consulting pathologist, OBBR, NCI

Max Robinowitz, M.D., medical officer, FDA

Clive R. Taylor, M.D., Ph.D., professor of pathology, Keck School of Medicine, University of Southern California

Eric Walk, M.D., chief medical officer, Ventana Medical Systems, Roche Diagnostics

Morning Session: Overview of Expectations

Dr. Litwack summarized that yesterday's conversations about qualifying FFPE tissue were useful in working toward the goal for OBBR and NIST to produce a plan to provide metrics in this area. This workshop will provide a list of goals for pre-analytic variables. caHUB is set up to produce tissue with systematic variation (e.g., time to fixation) and address other important variables. It is also important to hear from NIST about its perspective.

Dr. McShane noted that NCI has not yet defined how it will know when it finds optimal settings. Before discussing setting up experiments, it is necessary to decide what the endpoint of the experiments will be. One difference between NIST and NCI is that NIST begins with measurements. A good endpoint is something that can be measured with certainty.

In discussion about the value of a fluorescent probe inside a cell for measurement, Dr. Rimm noted that a fluorescent probe is not very reproducible and depends on the marker added. Some markers may be stable, but others change dramatically. Fluorescence can pose problems, but it can be measured by a microscope. Dr. Plant added that a fluorescent probe would provide a relatively easy real time measure, so that conditions of fixation could be measured fairly easily.

In a test of a large cell fixation process, investigators screened though a number of different fixatives to find conditions in which there was no change in live and fixed cells. They assessed before and after fixation to see how much fluorescence was lost and were able to begin to look at the mechanisms of fixation on the cell. Addition of formaldehyde produced blebs that protruded from the edges of cells, suggesting that the cell contained proteins other than green fluorescent protein (GFP). A probe inside a cell can be followed to fixation state, which represents a good measurement system.

Dr. Hewitt emphasized the importance of considering the tissue as a full package and finding a balance between nucleic acids and proteins. Protein assays are generally weaker. Use of antibodies can introduce species issues, because antibodies are often from mice. Antibody mediated detection methods have their limitations. Probe design must be done in tissue, because cells have limitations. Cell probes or a permeable reagent must be used. Investigators must use the tools they have or create new ones.

Dr. Taylor remarked that replacing formalin with more genetic-friendly fixatives is not realistic. Formalin-fixed tissue is what is available, and that is what the system must use. The control must be as simple as possible.

Dr. Hewitt asked whether the goal is to build a tool that can be widely applied or a research tool that leaders in the field can use for research and development. Dr. Robb said that the field would like a means to qualify FFPE as quickly as possible—within months to 1 year—and not have to change standard operating procedures every month. Dr. Hewitt replied that it might be that two tools are needed—the tool the Dr. Robb wants *and* a chemical tool, which would be a more long-term goal. Dr. Robb agreed with that approach. Perhaps there is a brand new fixative down the road. Group members agreed that a "holy grail" will be found in 5 years or so. Dr. Hewitt said that formalin has been evolving.

Group members summarized the goals, both specific and long-term, for this breakout group:

- To establish a formalin fixation method that preserves tissue morphology, protein, and nucleic acid.

- A fixation time of less that 1 hour cold ischemia time and 24 hours (plus or minus 8) fixative time.
- Uniformly fixed and processed specimens.
- Apply the science of molecular measurements. Select an axis and a buffer time, and then measure.
- Develop a robust procedure to compare fixed tissue with non-fixed comparative samples.
- Further understanding of mechanisms.

Dr. Hewitt advocated chemical fixation to avoid homogenization. Also, datasets can be too large and certain tradeoffs will be necessary. Dr. Rimm agreed, noting that a reductionist approach does not provide all the necessary assays. Dr. Taylor commented on the need to balance the delight of having a measure that can be done against the practicality of having one that is useful. Because something is difficult does not mean it should not be done. Information remains lacking about how proteins behave. It is important to know more precisely how formalin affects tissues. For example, how does it affect HER2?

Dr. Robb said that more information is needed about phosphoproteins. Dr. Hewitt said that the solution is in mass spectrometry and the chemistry of modifications in protein. Currently, imaging mass spectrometry does not give him the resolution he needs.

Dr. Plant asked how important it is to measure the chemistry going on during formalin fixation and how accurately it can be measured. Dr. Taylor said that better measures of how much protein is present in fixed tissue are needed as well as the ability to extrapolate back how much was present at the moment the tissue was moved from the body. Dr. Plant asked about the value of flash frozen tissue, and Dr. Robb said that controls are needed.

Dr. Walk asked if anyone has explored molecular imaging. Molecular imaging remains decades away from clinical use, but in animals HER2 can be examined *in vivo*. Dr. Hewitt said that some investigators have studied use of confocal Raman, but the limitation has been penetration. Dr. Hubel said this can be improved. A contrast agent is needed. The current technique is appropriate for animal use but not human application. Human application is possible, but issues of specificity and purity of binding of contrast agent still exist. Dr. Hewitt added that magnetic resonance (MR) technologies are beginning to become available, a marriage of molecular imaging and histology. Dr. Hubel commented that these technologies present signal-to-noise problems. With a low-density signal, scans must be repeated and scanning times are long. High resolution scanning would be an excellent approach.

Dr. Litwack asked group members to consider what caHUB could do. MR imaging will not be a high throughput process.

Dr. Robb said that freezing tissue also presents problems. Colder is not necessarily better, and tissue will go through a brittle phase. Dr. Hubel said that she is very happy to hear this discussion. Freezing to -80° C will result in intracellular material that is degradative but still active. What happens to cells and proteins at extremely low temperatures is not known, although it can be imaged. Adjacent to ice crystals, proteins unfold and an aggregation occurs. It is not clear whether how much of this is reversible. Perhaps some of these problems can be mitigated. Enzymes hang on to activity at very low temperatures. Dr. Rimm noted that a huge variable is thawing time. Dr. Hubel added that it is just as dangerous for biological tissues to be thawed as it

is for them to be frozen. Are there engineering solutions to overcome the problem? She is not aware that the necessary tools exist. No one has looked at the downstream effects of different speeds of thawing. Dr. Hewitt said that he explored research into freezing with a food science partner but could not find funding.

Dr. Robb suggested that NIST and caHUB investigate freezing techniques. Dr. Hubel noted that this task is not insurmountable—measurement of temperatures is well established as is rate of change of temperatures. The process has been worked out in the cell therapy field and is very doable but must be done well.

Dr. Robb asked whether a sample can be put directly into liquid nitrogen. That question is essential, Dr. Kulesza responded, and it is critical to ask about the criteria needed. If a sample shatters into peptides, it will not be measurable. What does NIST need to measure? Dr. Locascio asked whether measurements should be intrinsic or whether they should consider spike-ins.

Dr. Taylor asked Dr. Robb for more details on how he collects tissue. Dr. Robb said that it can be done simply using a 1½ x 3 cm module with a piece of tissue on one side for standard of care and a 5 mm piece for molecular characterization. Across the middle are two formalin fixed slices and another slice over the subculture.

Dr. Hewitt said that his experience is that modifications to proteins can be measured globally. This includes methylation of DNA and phosphorylation. If those are close enough surrogates, they are available. Dr. Locascio asked whether the matrix matters. Dr. Hewitt said that for chemical imaging, external imaging could provide all of those parameters. Dr. Locascio asked about enzyme degradation. Dr. Hubel observed that quantitative information on nucleic acids does not exist and it would be a tremendous addition to the field.

Dr. Robb illustrated the method of his lab with a diagram of a three-dimensional block divided into multiple sections. Standard operating procedures are in place to apply to all parts of the body. In the cancer module, two standard of care blocks are on left, then two sets of two frozen modules, and paraffin models as controls. The right side is biomarker fixatives. There is a total of 12 aliquots.

Dr. Huber said that she has some tips on how to process frozen tissue to prevent fracture in the container or fracture in the tissue, and she and Dr. Robb will continue that conversation. He added that the diagram has a place for peri-tumor normal organ, normal tissue from as far away as possible so that margins are not involved, and pre- and post-surgery blood. In response to a comment that the method is exclusionary to certain tumor systems because of insufficient available tissue—for example, breast cancer because of the small size of tumors currently being diagnosed—Dr. Robb said that is why he is forming a study to collect breast tumor tissue from women in Latin America. Another requirement is that the specimens cannot be pretreated, which excludes patients who have had neoadjuvant therapy in advance of surgery.

Dr. Kalavar said that the process assumes the same tumor content in different parts of the tissue, and Dr. Robb replied that everyone knows a specimen is heterogeneous. Dr. Kulesza asked whether comparison could be done for each analyte, and Dr. Robb said scientists could do all the tests they wanted. Dr. McShane asked about markers such as proliferation that might be up-regulated. Dr. Robb said that samples would be from the edge of the tumor. Dr. Plant commented

that a number of uncertainties appear to be involved and asked whether sufficient information would be available know variability from procedure to procedure. Dr. Robb said that an enormous amount of annotation is required.

Dr. Plant asked whether Dr. Robb wants NIST to provide a series of annotated protocols to evaluate for variance. Dr. Robb said that is not what he is saying; he would like NIST to help distinguish among PA1, PA2, etc., and whether they can be used. Dr. Litwack said that one option might be a sort of rating scale of acceptability. End users will employ different assays anyway. A scale, using whatever metric NIST decided on, would address fit-for-purpose.

Dr. Plant said that she was having trouble with the concept of living tissue as a starting point. Another suggestion was a xenograft model, but Dr. Hewitt said that has been very difficult to accomplish, with 1,500 xenografts needed for one experiment. Dr. Taylor suggested liquid nitrogen as a standard benchmark. Dr. Robb proposed measuring protein in frozen tissue and comparing it to FFPE tissue, taking known heterogeneity into account. Dr. Taylor commented that the measure would need to be automated.

Dr. Hewitt said that a perfect-world scenario would include a “magic” tissue qualifier that would rate tissue from 0 to 100, a simple spectroscopic approach with a surrogate metric. Dr. Rimm noted that different measurements would require different metrics of quality. Dr. Hewitt said that he could imagine a chemical imaging method that would qualify all tissues.

Dr. Otis asked about a tissue sample with a known quantity of spiked-in protein that reflected each element examined. Would it be possible to incubate tissue to permeate with spikes and use those to measure? Dr. Locascio said that it would depend on the process. Spike-ins early in the process would follow the tissue through the entire process. Dr. Otis asked if it would be possible to do this *in vivo* during the surgical procedure, and Dr. Plant said that probably would be the best way. Dr. Hewitt said this has been done in animal models and the technology exists, but moving it into fixed tissue was not successful with low diffusion and high variability resulting.

Dr. McShane said that she objected to Dr. Robb’s cube design. Even with spike-ins, there is variable permeability. Dr. Robb acknowledged that. Biostatistical research cannot be done on humans. Pathologists use the material they can get. He sees a great deal of tissue from clear cell kidney cancers. Dr. McShane asked how Dr. Robb would implement his design, and he said there are pre-labeled kits on the tables. Dr. Otis mentioned the practical difficulty of controls. Simplicity is key to the process, and controls must be routinely integrated and as transparent as possible.

Dr. Litwack observed that spike-ins can never replicate what tissue has been through. He would like a better understanding of intrinsic readouts. Also, what is referenced for positive controls? Dr. Robb said that he is aware of the many variables involved. Using his model, he and his colleagues will catalog all of the aliquots and then decide how to use them. This is a theoretical construct focused on annotating everything quantitatively. The samples will be highly selected for researchers after their requests are reviewed. Dr. Taylor said he applauded Dr. Robb’s efforts. Dr. Rimm questioned the possible introduction of bias. Dr. Hubel suggested careful consideration of how the kits are constructed, so that it will be easy to cut the tissue and properly label it. This could help improve the quality and number of samples.

Further discussion concerned using phosphoprotein as a timer. Dr. Rimm and colleagues are working with breast cancer, lung cancer, and melanoma samples with phospho-AKT and other compounds. He said that an assay is not yet available and it will have to be specific for what is measured. Dr. Litwack said there is a need to repeat the process for each tissue and tumor type. Dr. Rimm commented that the best-case scenario would be to achieve the same results for five or six different types of cancer. He is contracted to work only on breast and lung samples.

Dr. Walk asked about the use of phosphatases and the phospho status of tissue. Dr. Rimm said that phosphorylation goes up and down; how well phosphatase and formalin enter tissue is not well understood, so he left phosphatase off the table in his study design.

Dr. Kulesza asked why freezing is the standard. Dr. Rimm responded that freezing and fixation in formalin are the current standards in pathology. Dr. Otis noted the need for techniques to be integratable. Liquid nitrogen is readily available in most operating rooms. An individual should be designated for tissue procurement. That can be incorporated into a developmental research network, Dr. Robb said. Dr. Hewitt commented that liquid nitrogen is not universally available.

Perhaps it should be promoted.

Dr. Li commented that changing pathologists' behavior will take a long time but changing the fixative is possible.

Dr. Robb asked what starting materials are needed for this discussion to move forward. Dr. Plant said that she was impressed by Dr. Rimm's assay for phosphoproteins. Dr. Rimm replied that he uses quantitative multiplex phosphorescence, and Dr. Plant said that NIST could play a role by contributing benchmarking materials. Dr. Rimm said that would address a major problem. The lab makes its own benchmarks, and all labs are different. The differences are a concern; he tries to standardize every possible aspect. Dr. Plant said that NIST might be able to look at the process and offer suggestions for physical benchmarks. Dr. Taylor noted that NIST could help the field with easily accessible standardization, and Dr. Hewitt added that generalizability is also an issue. Dr. Rimm said that his lab is now working to bind recombinant proteins to beads in order to provide a standard for a large number of assays.

Dr. Taylor compared the search for a standard to the work that was involved with the development of the enzyme-linked immunosorbent assay (ELISA). ELISA works because it is a standard instrument that is calibrated against a curve. Perhaps an external standard could be combined with a reference standard to run the curve. Working with pig samples and antibodies, the same degradational regressional curves are observed. It is necessary to know the specific regression curve for each analyte. Dr. Plant emphasized the need for physical benchmarking materials to ensure the accuracy of regression curves, and Dr. Taylor said that the basic regression curve must be common to all labs. Dr. Rimm added that it must have the right dynamic range.

Dr. Cossman noted that these incredibly interesting technologies are at least 5 years in the future. Is the ultimate vision to have panels of markers for biomolecules and their degradation rates? Dr. Robb said that technology is available today, through multi-spectral analysis and accompanying bioinformatics. Dr. Taylor said that the goal would be a panel of intrinsic reference standards to measure new analytes.

Dr. Robb reviewed his Latin American project. He is collecting breast cancer samples from south of the border because it is difficult to obtain data from Latina women in the U.S. and tumors of U.S. women are usually very small. He and his colleagues are working through the ministers of health and will teach the local investigators to conduct clinical trials and perform molecular analysis on the tissue, fixing it with the same protocols used in the U.S. Dr. Robb does not want the tissue itself, only the data that are derived from it.

Dr. Hubel said she would like to see more effort devoted to attempts to freeze tissues viably, which has tremendous clinical potential. Dr. Rimm noted that the approach has met resistance from pathologists, but CAP will move the technology into schools.

In further discussion of Dr. Rimm's fluorescence-based algorithm, Automated Quantitative Analysis (AQUA), Dr. Robb commented that it has the potential to provide quantifying markers to dictate therapy. Dr. Rimm said that pathologists resist the need to rely on a machine without actually seeing results. Dr. Hewitt said that it is an excellent analytic tool to measure biomarkers, but as a quality measure, it is expensive and not quick and will be challenging in normal tissues. Dr. Rimm replied that it works in normal breast tissue, and Dr. Hewitt said that less metabolic tissues such as heart or brain might be challenging. Dr. Hewitt has been pursuing bright field techniques but has had trouble with standardization.

Fluorescence is not more expensive than normal immunohistochemistry, Dr. Rimm said, but it takes longer. Dr. Robb said that within a few years, it will be built into the process. Dr. Otis said that he did not expect the resistance and pushback from pathologists that has greeted previous technologies, because the automation will be accepted and the programs exist and can be built into associated programs. Dr. Plant asked whether the instrument will need FDA approval, and Dr. Taylor said that is uncertain. Radiology instruments require FDA approval. Dr. Walk noted that in radiology hearings last November with FDA, the level of regulation was downgraded if appropriate data are available. Dr. Taylor said he hoped that CAP would compare the technology to a microscope, which is not regulated. Dr. Otis said that regulation is needed for digital interpretation.

Dr. Robb asked Dr. Rimm about his throughput, and Dr. Rimm said that a regular fluorescent slide takes 5 minutes. Dr. Hewitt said that the FDA plans to consider it at the pre-market approval level. Some fluorescence technologies are slower than bright field, and some effects were difficult to replicate and separate. Many corners remain to be explored.

Dr. Robb asked when quantum dot technology will replace fluorescence, and Dr. Hewitt said that penetration of quantum dots is suboptimal. Dr. Rimm added that quantum dots don't work and the field has moved on. Dr. Walk said that his team continues to work with quantum dots, and it is challenging but they have been able to achieve some results.

Dr. Elliott said that he liked the idea of using phosphorylation groups and it seems that there are other possibilities as well. Dr. Hewitt said that he has been building a panel of degradation groups and has material available to test. Protein content variability can be a problem in immunohistochemistry. He suggested a tiered approach, beginning with H & E. Dr. Litwack asked whether reagents exist, and Dr. Elliott said that they do. Dr. Hewitt said that a set of markers might be closer than people think.

Dr. Robb thanked NIST and the group members for this important discussion. The result should be something practical and useful with science to back it up. Biospecimen research has never been funded before, but funds are now available and it is important to produce unequivocal, irrefutable data to take to community pathologists to demonstrate targeted molecular therapy techniques that improve patient care. Dr. Li said that NCI's Cancer Therapy Evaluation Program (CTEP) would certainly be a user of this technology.

Dr. Litwack said that he wants to discuss a specific roadmap and timeline for how these efforts will play out. Dr. Rimm said that he is 1 year into his 3-year OBBR contract to identify phosphoproteins. He will conduct analysis in the second year and validation in the third year. Dr. Robb suggested that he contact Henry Rodriguez, Ph.D., director of NCI's Clinical Proteomic Technologies for Cancer initiative, for antibodies. Dr. Litwack noted that antibodies are not always reliable as a standardized method of quantifying, and there was further discussion about the reliability of antibodies. Dr. Locascio said that NIST has been requested to develop standard antibodies but has not been able to accomplish the task. NIST could not possibly be a supplier of standard antibodies. It is possible that the work could be contracted out. It is a big project. Dr. Robb commented that antigen panels would be extremely helpful.

Dr. Rimm observed that while the goal of the leading non-fluorescent systems is to think like a human being, Aqua thinks like a molecule. This is a profound difference in premise. Machines that think like a human have too much variability.

Afternoon Session: Moving Forward

Dr. Litwack said that some good ideas about direction were raised in the morning session. He asked for comments about what is and is not doable and what is reasonable in a timeframe of a year or two.

Dr. Elliott said that NIST would begin with a small molecule. NIST's concerns are measurement infrastructure and making labs consistent with each other. The agency has spent a great deal of time on cross-beam fluorescence. It does not have permanent solutions but does have many ideas it can bring to the project.

Dr. Kulesza questioned the value of spiking and adding small molecules to tissues. Dr. Hewitt said that it would indicate whether fixative got into the tissue, and Dr. Fletcher said that it also would show *when* fixative got into the tissue. This could be a useful tool and allow assessment of various rapid tissue processors. Dr. Hewitt said that he did not think that a spike-in analyte would work, but a spike readout tool would work.

Dr. Litwack said that fixation could be evaluated by assessing gradient of fixation in a stain. Dr. Elliott commented that many issues are related to formalin penetrating a block of tissue. Dr. Kulesza noted that staining is heterogeneous. He thought that NIST wanted a set of easily made assays in 18 months. Spiking will tell if formalin got in but not if DNA is cleaved, adducts preserved, or other attributes. Dr. Hewitt said he would expect it to take 5 years, not 18 months. Dr. Litwack asked what the important pre-analytic variables are. In addition to specific reagents, variations such as time to fixation and ischemic time should be looked at. It might be useful to develop a reagent to monitor time to fixation. Dr. Hewitt said that his impression was that under-

fixation was one of biggest problem in diagnostic chemo-pathology. Dr. Otis said that this depends on what protein is of interest. Formalin penetrates tissue but takes time to fixate. Alcohol as a fixative causes problems. It is not clear whether formalin fixation has properties that differ from others but have been overlooked because they have become so common. Dr. Hewitt said this is absolutely the case—two aspects of formalin fixation are a groove that is seen in thyroid cancers and isometric vacuolization of renal tubular epithelium.

Dr. Kulesza said that it sounds like two different missions for NIST are being discussed. Quality of sample is different from study of different analytes. Dr. Hewitt said it was necessary to define primers and what is optimal is very subjective. He would like to track down more variables and make comparisons to clinical practice. FFPE is fit for purpose, but investigators fail to appreciate that it has evolved. Formalin will continue to evolve. He can envision techniques such as adding buffer that will not detract from the process.

Dr. Otis asked about standardization of process. Do some labs add alcohol to speed processing? At what heat? What types of paraffins? A great deal of variability exists, and it is important to know what is in the black box.

Dr. Plant asked Dr. Hewitt what he compared in his preliminary studies to make the evaluation that the formalin fixative process was good. Dr. Hewitt said that he compared an RNA fixative publication and an FFPE publication. RNA was the surrogate outcome. He referenced frozen tissue and found that 30 percent of RNA got out. Four different procedures were used to qualify the tissue. Dr. Hewitt explained further details of his process. One problem with comparisons was that different RNA extraction protocols were used with fresh and FFPE tissue.

Group members disagreed about the value of frozen tissue. Dr. Hubel said that frozen tissue would have increased potential to be reproducible if researchers paid more attention to the freezing, storage, and restoring processes. Every hour that a specimen stays at -80 degrees C results in loss of information. Dr. Hewitt agreed.

Dr. Li asked what the control should be for a formalin sample, and Dr. Hewitt said there is none. Frozen tissue cannot be compared to FFPE. He would prefer to focus on improving paraffin because it is more consistent and fit for purpose. Technology has evolved so that he can do things in paraffin that he could not do 20 years ago, and he would like to do more yet, such as detecting phosphoproteins. He is looking for utility, a biomarker that he can use for a clinical decision. Dr. Plant asked how he would know that a biomarker is good for clinical decisions and accurate, and Dr. Hewitt responded that that is what he wants to determine. Dr. Plant said that more than one variable must be measured and correlated to interpret the biomarker. Dr. Hewitt said that his primary concern is reproducible artifact, that the assay he does today is the same as what he will do in 6 weeks or 6 months. Stability is critical, and freezing fails miserably.

Dr. Plant said that she would like to start with a marker that is robustly measurable and reproducible. The initial goal is metrics for reproducibility and robustness of tissue. Then it would be possible to go back and say if a marker predicts clinical outcomes. Dr Hewitt that chemical imaging seems to be the best way to accomplish that. He has seen no evidence that RNase is active in FFPE tissue. The degradation of RNA that is seen in FFPE tissue over time is not RNA being active, it is a chemical process. The current analytic tools are weak and better ones are needed.

Dr. Plant said that the role of NIST is to ask several questions: Do you have an analytic tool? If it is noisy, is there a benchmark to figure out why? What is the source of noise and can it be normalized with benchmarks? Dr. Kulesza commented that without gold-standard reference material, he was not sure it would be possible to normalize. Dr. Plant said great variability exists between samples but watching one sample over time is instructive.

Dr. Plant asked whether the speed of degradation is variable. Dr. Hewitt said there is no mechanism to prevent degradation. The question is, at what point does degradation become important?

Dr. Fletcher questioned developing molecular stability biomarkers to apply to a widely variable dataset. Dr. Plant noted that this is not a discovery project; it is an evaluation project. Dr. Hewitt agreed that the targets are moving and biologic processes are extremely noisy. For example, the baseline oxygenation of the patient is never recorded, so a baseline for hypoxia is not available. It is impossible to apply the rigor of biostatistics to these data. In fact, the No. 1 factor that predicts breast cancer behavior is estrogen receptor status.

Dr. Taylor asked about the value of the esoteric theory he has heard in this discussion. Dr. Robb's drawing represents a fait accompli, but the challenge lies in what will be done with the collected tissues. He does not think DNA, RNA, and protein can be used in the same way. Is the tissue useful for a particular protein analyte? Is it possible to develop standards for HER2 through a different process? FISH or HER2 are most likely to predict therapy. Pathologists work with paraffin tissues because that is what is available, but it would be useful to know what has been lost in the fixative process.

Dr. Kulesza urged group members to focus on specific goals about assays for NIST to work on. What happens with degradation? He agreed that comparing frozen tissue to FFPE will provide useful information. Dr. Hewitt said that it would be more valuable to step back and look for better proteins. All of the protein remains after FFPE. Dr. Taylor said that his interest is immunohistochemistry (IHC), to be able to control the stain and be narrow and focused on that. Questions about suitability of tissue are not relevant.

Dr. Hewitt said that for good downstream utility, a strong control for IHC is needed. He has worked on that, but been disappointed with his own efforts.

Dr. Taylor said that the origin of this discussion came about 6 years ago when there was trouble putting breast cancer patients in treatment or trials for HER2. The HER2 measures are not sufficiently robust. Focusing on assays for HER2 and maybe estrogen receptors is also likely to elicit other useful data. Group members agreed on the need for improved HER2 assays. An FDA-approved product exists, and no one is looking further. The only way to improve the process would require a new clinical trial and no one wants to do that. Dr. Taylor commented that a new clinical trial is not necessary, just a demonstration that specimens from patients show the same HER2. Dr. Kulesza said that these demonstrations are coming, but Dr. Taylor was dubious.

Dr. Otis agreed that the process is daunting, from the basics of acquiring tissue, to interrogating the block for integrity of tissue for purpose, to looking for discovery of new biomarkers for predictive or prognostic purposes. The major challenge is to develop tools to interrogate tissue.

Dr. Hewitt responded that the morning discussion determined a need for a quality metric that asked about protein integrity and RNA cross-linking.

Dr. Plant asked whether antigenic determinants were the ultimate target, with cross-linking another surrogate. Dr. Litwack said that to determine fitness for purpose, readouts for tissue on a qualification curve are needed. It will not be valid for all assays. Every method has different requirements. He asked whether it would be possible to develop a technique to standardize for time, fixation, storage time, and the other variables that have been discussed. Then it will be the task of end users to determine what applies.

Dr. Kulesza suggested that NIST solicit expert input on how to analyze six or seven proteins, including HER2, and then develop profiler assays. Dr. Taylor emphasized the need for NIST to develop a process to improve the validity of HER2 testing. External and internal references are needed. Dr. Litwack said that has been done in the past, and Dr. Taylor replied that is has not been done with sufficient accuracy. Dr. Litwack recommended determining a roadmap for HER2 as a quality control standard for FFPE. HER2 must be understood at the front end, before work is done with the tissue. Then fixation could be tested for various time intervals, using fresh tissue to know the time brackets within which HER2 is valid, where the parameters are, and the amount of degradation. Ideally, a second molecule that does not vary much in different tissues can serve as a reference standard. Dr. Litwack added that he liked the idea of two tests, and Dr. Plant agreed.

Further discussion considered the problem of variability. Dr. Taylor said that statisticians and population registries can help with that. With lack of sufficient tissue, surrogates are needed. Dr. Litwack suggested using a pig model as a standard.

Dr. Kulesza said that standards of degradation can be developed post hoc and then applied retrospectively. That will produce answers about variability in processing and allow application of a quality standard. A separate contract will be needed to address degradation in formalin. Dr. Taylor commented that it is amazing that pathologists have been using formalin for more than 100 years and do not know what it does to proteins.

Dr. Litwack said that many HER2 assays exist—could they be graded? Dr. Kulesza replied that he has been looking for funding to study this and could not find any. Dr. Taylor agreed.

Dr. Hewitt asked what happens when HER2 and ER are optimized with the finding that what is good for one is not good for the other. It is necessary to fit the optimal assay to the optimal specimen. Formalin fixation will allow some motion. Other markers might shift in ways that are currently unknown.

Dr. Plant said that she has heard a number of conflicting ideas in this discussion and finds it appealing that the group seems to think a systematic study is warranted. NIST will need help to prepare and assess samples. She is hearing that a systematic study is needed to look at time of fixation and storage of paraffin studies over time in one analyte, and HER2 would be a good one. NIST would provide benchmarking materials for those doing IHC and Western blot.

Dr. Walk mentioned that this topic was discussed at several CAP meetings but never moved forward, so some CAP members might be interested.

Appendix F: Serum Proteins Integrity Working Group

Group members:

Chris Becker, Ph.D., program director, SRI International

Rachel Benkeser, M.S., presidential management fellow, OBBR, NCI

Andrew Brooks, Ph.D., chief operating officer, Rutgers Cell and DNA Repository, Rutgers, The State University of New Jersey

Saeed Jortani, Ph.D. associate professor of pathology, University of Louisville

Dale Larson, M.S.M.E., director, Biomedical Systems Program Office, Draper Laboratory

Mark Lowenthal, Ph.D., research chemist, NIST

Kristen Meier, Ph.D., statistician, Center for Devices and Radiological Health, FDA

Karen Phinney, Ph.D., research chemist, NIST

Paul Rudnick, Ph.D., biologist, NIST

O. John Semmes, Ph.D., director, Leroy T. Cannoles Cancer Research Center, Eastern Virginia Medical School

Stephen Stein, Ph.D., fellow, NIST

Mary Thomas, Ph.D., AAAS fellow, OBBR, NCI

Gordon Whitely, Ph.D., director, Antibody Characterization Laboratory, SAIC-Frederick, NCI

Review of Thursday Case Study Discussions

Participants shared the conclusions from the two groups that discussed the serum protein integrity case study the previous afternoon. Both groups discussed methods for monitoring blood-sample integrity in the intervals between blood draw, some period of storage, and analysis.

Group 1

The first group considered whether a marker could be found that would be common to all blood biospecimens (normal and disease) and that could serve as a marker of stability or integrity for any analytical platform, including platforms not yet developed. Unable to identify such a universal marker, the group conceived of a device such as an analysis chip that would be

activated when blood comes into contact with it. This chip should be able to monitor time, temperature prior to and during storage, freeze/thaw cycles, and any other variables that might influence biospecimen integrity between blood draw and preservation. Individuals interested in using the blood sample would be able to access the information on the chip to decide whether the biospecimen had been handled in a way to preserve the integrity necessary for the analysis of interest. The chip would measure environmental stress, not degradation per se.

When considering a universal marker, the group noted that a protein such as prostate-specific antigen (PSA) would be a valuable indicator of protein degradation within a sample. PSA exhibits different levels of stability depending on whether it is free, bound, or truncated. Although not all blood biospecimens will contain endogenous PSA, the ratios of its isoforms or another endogenous or spiked-in protein's isoforms with a similar variety of stability profiles—i.e., stable, somewhat labile, and very labile, such as kallikrein, creatine kinase, or lactate dehydrogenase—might be evaluated for the purpose of quality tracking. Multiple-reaction monitoring (MRM) mass spectrometry (MS) might be used to identify such a protein set. If a marker were spiked into samples for quality tracking purposes, it would be worth considering whether to cage the marker in a nanomolecule that would enable it to react to changing conditions while keeping it partitioned to prevent it from influencing the biology or reactivity of the sample.

Adding a monitoring chip or marker protein would be a reasonable way to monitor pre-analytical variables or degradation of prospectively collected blood samples but would not help address quality measurement of biospecimens already in storage. Neither a chip nor a protein set alone would detect all possible threats to protein integrity, such as poorly manufactured collection tubes or a bad lot of the tube's preservation matrix. Thus, a method is needed for assessing biospecimen quality immediately before storage or analysis, independent of any information concerning handling to that point.

Group 2

In the case-study discussion the previous day, the second group discussed the issue of assessing quality in a blood sample that has been exposed to unknown variables. It devised a three-tiered evidence-based system for assessing the quality of a blood sample.

The first tier would be an inexpensive, automated, multiplexed assay of readily measured clinical chemistry markers from routine complete blood count evaluations, such as potassium, lipids, hepatic function markers, or bilirubin. An existing panel, such as a metabolic panel, might serve this purpose, or a new panel might need to be devised. Panels are crucial because no single indicator is likely to be sufficient. It would be necessary to determine which of the clinical-chemistry analytes are sensitive to degradation of the sample by observing how they change after subjecting experimental samples to a variety of pre-analytical variables, such as collection-tube type and biospecimen volume in addition to those mentioned above. It will be useful to know which analytes change with mishandling as well as which markers remain stable, and analytes with a range of stabilities should be chosen for this tier of the analysis.

The next tier of analysis would be a more sophisticated evaluation of protein quality based on immunohistochemistry or MS and would include a panel of analytes to determine such characteristics as protein concentration, phosphorylation, oxidation, or proteolysis.

The third, most sensitive tier would involve evaluation of materials spiked in to the sample, and might include antibodies to phosphoproteins or a panel of phosphopeptides. Again, the analytes to include in these tiers should be determined experimentally in samples that have been exposed to a range of handling and storage conditions. Determination of fit-for-purpose from the results of this tiered evaluation would not be a pass/fail judgment but would be based on the context of the downstream application in which the sample was to be used.

The first tier as described offers the benefits of being well characterized and economical. Potassium levels, for example, are routinely evaluated in clinical chemistry panels, cost approximately 10 cents per sample, and become elevated if the biospecimen has not been separated via centrifugation in a timely manner. Other quality markers such as carbon dioxide and glucose would also be useful because concentrations at high or low extremes are not compatible with life and would indicate a clear problem with the sample.

General Discussion

Analytical platforms

The most common ways to measure proteins and peptides in blood biospecimens are via immunoassays and MS. Each offers complementary advantages: MS can measure a far greater number of proteins, whereas immune techniques detect known analytes and can distinguish isoforms, such as phosphorylated sites. Both techniques will be useful for evaluating sample quality.

The amount of sample required to perform such tests is a concern. A panel of 10 or 15 small molecules might require between 100 and 500 µl, which could be prohibitive for retrospective samples of small volumes. Immunohistochemistry evaluations also might require a significant volume of sample, although it could be possible to develop a specialized quality immunoassay that takes advantage of a recently developed dipstick-like technology that uses very little of the biospecimen.

Raman spectroscopy is a relatively new analytical technique that might be helpful in this endeavor, particularly for measuring physical changes to the sample that result from changes in conditions such as temperature. It relies on light scattering; thus it does not destroy the sample and could be employed with a microfluidic system.

Another platform that would be useful for determining varying levels of marker-protein isoforms is electrophoresis. This platform offers the advantages of using only a small volume (10 µl) of the sample and providing information on the amount and activity of the enzymes being evaluated.

A consideration for choosing any quality assay is the availability and cost of the test. Investigators in independent laboratories might not have access to MRM-MS or other costly

machinery, for example. However, it is likely that once a set of quality markers is established, some entrepreneur eventually will invent a way to offer multiplex tests cheaply and easily.

Pre-analytical variables

Work is ongoing to identify the influence on blood proteins of a variety of pre-analytical variables, including:

- Phlebotomy factors
  - Needle gauge
  - Tube type
  - Tube order
  - Tube mixing
- Patient factors
  - Position
  - Fasting
  - Draw site
- Handling factors
  - Time and temperature before, during, and after centrifugation
  - Centrifugation parameters
  - Freeze-thaw cycles

The group noted that the time between blood draw and preservation/storage is a major factor in the protein quality of a blood sample; 20 to 27 percent of the variation will occur within 20 minutes to 4 hours. It is not clear what degree of change renders a sample unusable. A time-course experiment will be needed to determine how much change occurs with time, whether that change is associated with protein degradation, and how much degradation is acceptable for the end assay.

New technology needs

In addition to understanding the changes that occur as a sample degrades, it would help researchers to have the ability to prevent any changes or degradation. PAX gene blood tubes immediately stabilize blood drawn for gene expression analysis, preventing a great deal of the alterations in immediate early gene expression that generally result from drawing blood. Species of mRNA that are not stabilized in PAX gene tubes are known and can be discounted in ensuing analyses. It would be useful to have a similar stabilizing tube for proteins that suspends proteins in the exact state they were in at the time of the draw. This has been attempted with proteases, but no combination has been identified that universally arrests all protein modifications.

Potential markers of quality

A number of analytes in a blood sample might serve as surrogate markers for sample quality, including small peptide content or levels of post-translational modifications, such as glycosylation, phosphorylation, acetylation, or sulfination. These could be evaluated in samples that have undergone known pre-analytical variations and been evaluated in the context of applicable analytical platforms. Post-translational modifications, however, will be quite variable between patients, making post-blood-draw changes difficult to quantify, although groups of post-translational modifications might prove informative.

Another idea posed was to evaluate the results of tryptic protein digestion and to evaluate the degree of truncation over time. Any of these approaches would help define scales by which known levels of enzyme activity, small proteins, post-translational modifications, or truncation could be associated with performance of a sample on a technology platform. Each class of analytes would need a separate quality assay, however, precluding the use of a high-throughput platform. It might be possible to do marker discovery on an MRM-MS platform and then migrate the best indicators to a multiplex immunoassay for large-scale quality assessment.

Proteolysis, oxidation, glycosylation, and other post-translational modifications are the biggest problems for immunoassay performance and might serve as metrics for sample quality. Free peptides are occasionally used as indicators of proteolysis; however, the assays for these are not necessarily quantitative on an absolute scale. Protein phosphorylation and other post-translational modifications have been largely unexplored to date. To identify changes associated with sample degradation would be challenging and require a great deal of discovery research. Similarly, cell membrane chemistry, such as phosphates, might hold potential as a measure of quality but also is poorly characterized to date.

Human serum albumin is the most abundant protein in blood and might have potential as a marker of sample quality. Its numerous adducts, miscleaves, and truncations might offer a desirable range of sensitivity for evaluating biospecimen quality. The complement system or Fc fragments, likewise, might do the same. Another suggestion was to identify the most labile components of blood, as defined by the elements that require special or rapid handling for accurate assays. To use any of these as a proxy for sample quality, the variations associated with degradation of the sample would have to be determined through experimentation.

Another avenue that holds potential is evaluation of enzyme substrates and products. The action of specific enzymes and the production and concentration of certain metabolites might give clues to blood/biospecimen quality. Likewise, quality might be evaluated via indirect measures of post-translational modifications, such as concentration of inorganic phosphorus or free glycans.

Physical tests also might be evaluated for their potential to indicate sample quality. Evaluation of full absorbance spectra, chemiluminescence, density, viscosity, pH absorbance, and conductivity can be obtained without destroying any portion of the sample. A library can be developed of these values in samples of varying degrees of degradation. If physical properties were determined to correlate to quality, it might be necessary to validate the result with a protein-based evaluation.

Centrifugation characteristics such as speed and temperature will factor into sample quality. Any recommendations resulting from this NCI–NIST collaboration should include optimal centrifugation conditions and suggestions about how to handle blood lipids.

One of the pre-analytical factors that is very difficult to control is variability in the technician's skills. After a blood sample is separated by centrifugation, the layers are carefully siphoned off with a pipette. Differences in the ability of the technician to draw plasma off the cell layer can have striking effects on the plasma or serum product. Another potential measure of sample quality might be a marker of cell content, such as platelet count or amino acid content from lysed cells, although a second round of centrifugation generally eliminates such issues. In general, technician variables can be minimized through the use of straightforward, well-defined protocols and education concerning the reasons behind each step and the problems that can arise if each step is not followed carefully.

Parameters of change

Any analytes that are determined to be markers of biospecimen quality will need to be evaluated to determine the baseline or normal range and the degree of change that determines a go/no-go decision. Target quality goals for precision bias or interference will need to be determined.

NIST needs

To conduct experiments identifying markers that change with sample degradation, NIST will need a solid grounding on any work that has already been done in this area of "degradomics." Recommendations were made to review the work of Hans Lilja on stability markers and the work of Edwin Joseph Cohn on fractionated plasma. Additionally, it will be important to keep NIST informed about the ongoing experimentation being conducted by BRN contractors and to coordinate programs.

When designing experiments to evaluate potential quality markers, it will be essential to determine inter-individual biological variation. It also will be necessary to involve a statistician in the experimental design to help randomize the experimental variables appropriately and determine the appropriate experimental sample sizes.

Conclusions and Recommendations

The group agreed on a number of conclusions and recommendations that were presented to workshop attendees by Dr. Whiteley. Measures of sample integrity, also known as the degradome, are largely unknown at this time. They can be discovered only through research and then validated in trials. The research will have to incorporate as many pre-analytical factors as possible, beginning with those most likely to influence a clinical sample. The research to identify markers of degradation might be approached on three levels: currently used laboratory tests, physical measurements on samples, and discovery of new quality markers. All of these will have to evaluate a large number of analytes, retaining those that prove most applicable to sample quality. Attention should be paid to development of quality metrics that can be assessed in small sample volumes so as not to destroy valuable retrospective biospecimens. Accurate biospecimen tracking and identification also will be essential. Any working group formed to develop these

tests should include not only researchers and clinicians but also medical technicians for a real-world perspective on performing the assays.

Current laboratory tests that might be employed include:

- Metabolic panels
  - Potassium assays
  - Electrophoresis and immunoelectrophoresis
- Assays that are known to be sensitive to time or temperature
- Immunoassays for sensitive and robust markers such as coagulation cascades or kallinreins

Evaluation of physical measurements might include:

- pH
- Viscosity
- Density
- Light spectra
- Fluorescence
- Raman spectroscopy
- Nephelometry
- Isoelectric point
- Measurement of small molecules such as carbohydrates, phosphates, sialic acid, or neuraminic acid.

With the assistance of a statistician in experimental design, mass spectrometry can be used as a discovery platform with carefully collected and “abused” (i.e., subjected to a variety of pre-analytical variations) patient-matched samples. Liquid-chromatography tandem MS can be used as a discovery tool and MRM-MS as a quantitative verification tool, after which the discovered markers could be converted to a higher-throughput platform such as matrix-assisted laser desorption/ionization (MALDI) MS, immunoassay, or chemical test. Sample spiking with quality markers might not be a good idea because of possible interactions with other serum proteins. New computational solutions likely will be needed to handle the complexities of these evaluations.

New technologies might be required for adequate stabilization and quality evaluation of blood biospecimens. A special preservative has been developed for use with nucleic acid studies that stabilizes nucleic acids in the sample at the time of acquisition; it might be possible to develop a similar preservative for blood proteins. Other useful technological developments include a chip in a collection tube that would monitor the conditions of storage and handling from the time the blood enters the tube and multiplexing technologies for degradome evaluation.

Appendix G: Final Workshop Agenda

Thursday – 10/21/2010

9:30 – 10AM	**Welcome & Overview** *Carolyn C. Compton, M.D., Ph.D.* Director, NCI Office of Biorepositories and Biospecimen Research Executive Director, NCI Cancer Human Biobank National Cancer Institute, NIH
10 – 10:30AM	**Current state of clinical laboratory medicine and the need for standards in context of emerging technologies** *James A. Robb, M.D.* Consulting Pathologist, Office of Biorepositories and Biospecimen Research National Cancer Institute, NIH
10:30 – 11AM	**Standards development and the role of NIST** *Laurie Locascio, Ph.D.* Chief, Biochemical Sciences Division National Institute of Standards & Technology
11 – 11:30AM	**NCI needs for specimen standards and the role of NCI/OBBR** *Carolyn C. Compton, M.D., Ph.D.* National Cancer Institute, NIH
11:30 – 12PM	**Q&A Panel for all speakers to clarify the goals and available resources**
12 – 1PM	**Buffet lunch**
1 – 1:15PM	**Description of afternoon agenda** *Tony Dickherber, Ph.D.* AAAS S&T Policy Fellow, NCI OBBR National Cancer Institute, NIH
2 – 2:30PM	**The importance of experimental design, bias, and rigorous statistical methods** *Terry Speed, Ph.D.* Professor Emeritus, University of California, Berkeley Joint with Walter & Eliza Hall Institute of Medical Research (Australia)
2:30 – 3PM	**Q&A Panel** *Terry Speed, Ph.D. – UC Berkeley* *Lisa McShane, Ph.D. – NCI/DCTD, Biometrics Research Branch*
3:30 – 4PM	**Measurement concerns from a regulatory perspective** *Elizabeth Mansfield, Ph.D.* Director for Personalized Medicine Office of In Vitro Diagnostic Device Evaluation and Safety Center for Devices and Radiological Health, FDA
4 – 4:30PM	**Q&A Panel** *Elizabeth Mansfield, Ph.D. – FDA/CDRH/OIVD* *Max Robinowitz, M.D. – FDA/CDRH/OIVD*
5 to 5:30PM	**Report out from each table** (3-5 minutes each)

Friday – 10/22/2010

Please join your respective working group, indicated by the card on the back of your badge by 9AM

9 – 9:30AM	**Overview of relevant NIST group past performance and breakdown of working group expectations** *Workgroup facilitator*
9:30 – 12PM	**Working group session**
12 – 1PM	**Buffet Lunch**
1:00 – 3PM	**Working group session**
3PM	**Working groups return to plenary room for report out**
3 – 3:40PM	**Report out from each working group (10 min each)**

1. RNA Integrity Group
2. DNA Integrity Group
3. Formalin-Fixed Tissue Group
4. Serum Protein Integrity Group

3:40 – 4PM	**Closing Statements**
	Carolyn C. Compton, M.D., Ph.D. National Cancer Institute, NIH