Path to Effective Recovering of DNA from Formalin-Fixed Biological Samples in Natural History Collections

WORKSHOP SUMMARY

Evonne P.Y. Tang

Board on Life Sciences
Division on Earth and Life Studies

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS
500 Fifth Street, NW Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by the Consortium for the Barcode of Life, Museum of Comparative Zoology of Harvard University, the National Evolutionary Synthesis Center, New England Biolabs, Inc., Sigma-Aldrich Company, the U.S. Department of Agriculture, and the U.S. Environmental Protection Agency's Environmental Monitoring and Assessment Program. The content of this publication does not necessarily reflect the views or policies of the organizations or agencies that provide support for the project, nor does mention of trade names, commercial products or organizations imply endorsement by the agencies or organizations.

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, NW, Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Printed in the United States of America.

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Wm. Wulf is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Wm. Wulf are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

STEEERING COMMITTEE FOR THE WORKSHOP ON RECOVERING DNA FROM FORMALIN-FIXED BIOLOGICAL SAMPLES

ANN C. BUCKLIN *(Cochair)*, University of Connecticut, Groton
DONALD M. CROTHERS *(Cochair)*, Yale University, New Haven, Connecticut
TIMOTHY O'LEARY, U.S. Department of Veterans Affairs, Washington, D.C.
CHRISTOFFER SCHANDER, University of Bergen, Norway
ALISON WILLIAMS, Princeton University, New Jersey

Staff

EVONNE P.Y. TANG, Study Director
FRANCES E. SHARPLES, Director, Board on Life Sciences
TOVA JACOBOVITS, Senior Program Assistant
ANNE F. JURKOWSKI, Senior Program Assistant
KATE KELLY, Editor

Preface

Museums catalogue our knowledge of the Earth's biodiversity, and their collections represent many decades of work by experts. Access to DNA sequence information in archival specimens would greatly extend knowledge of the genetic relationships within our biosphere. However, molecular genetic analysis of museum specimens has been slowed by the usual practice of fixation of samples in formalin and storage in alcohol or formalin. The fixation and storage induce changes in DNA that are not fully understood. With more frequent use of morphological and molecular characters for taxonomic and systematic analysis, the "formalin problem" has grown in significance. For example, a global effort to determine DNA barcodes (short DNA sequences for species recognition and discovery) for life on earth could be markedly expedited by sequencing DNA from specimens in museum collections. Molecular analysis of formalin-fixed tissue would allow biologists to address retrospective questions about how patterns in the genetic diversity of plant and animal species have changed over time. With application of molecular genetic analysis to formalin-fixed museum specimens, new insights about biodiversity, population dynamics, and ecosystem function can be gained from collections sampled years and perhaps decades ago.

There are many technical challenges to solving the "formalin problem," beginning with the wide variation in curatorial practices for specimen storage. Some organisms are fixed in formalin only for a short time and then transferred to alcohol for long-term storage; others are fixed and stored in formalin permanently. The rapid reactions of formalin with double helical DNA generally are reversible, but over the long term, especially with denaturation of the DNA, a variety of reactions can occur, many of which have not been characterized. Those reactions can be irreversible, and they can either mask the nature of the modified nucleotide in enzymatic replication, or they can block chain elongation altogether, resulting in failure of polymerase chain reaction amplification. To further complicate matters, oxidation of formaldehyde in formalin to formic acid produces an acidic solution in which depurination reactions and subsequent chain scission are greatly accelerated. Given the variation in preservation practices, and the variable age of the samples, it is unlikely that the "formalin problem" can be solved for all samples. However, with better understanding of the chemistry of formalin reactions and their effect on DNA integrity, and with better knowledge of curatorial history and practices, it should

be possible to select likely candidates for intensive DNA isolation and sequencing experiments, with the goal of reconstructing significant portions of the genome.

On May 8-9, 2006, the Board on Life Sciences of the National Academies convened a workshop, "Recovering DNA from Formalin-Fixed Biological Samples." Participants were experts—biophysicists, chemists, molecular biologists and bioinformaticists—who discussed the path to successfully obtaining DNA sequence information from formalin-fixed biological samples. Unlike study committees of the National Research Council, workshops do not reach conclusions or present recommendations. However, participants in this workshop spent much time considering the research and experimentation that could be done to advance retrieval of genomic information from formalin-fixed samples. We thank all of the workshop participants for sharing their expertise and experience and for the stimulating discussions and insightful suggestions.

Ann C. Bucklin
Donald M. Crothers
Cochairs, Steering Committee for the
Workshop on Recovering DNA from
Formalin-Fixed Biological Samples

Acknowledgments

This document presents the author's summary of the workshop discussion and does not necessarily reflect the views of the roundtable members or other participants.

This summary has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council Report Review Committee. This independent review is intended to provide candid and critical comments that will assist the institution in making its published workshop summary as sound as possible and to ensure that the workshop summary meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this workshop summary:

Robert DeSalle, American Museum of Natural History
Neil Hall, The Institute for Genomic Research
Jack Lichy, The Veterans Affairs Medical Center
Stephen Quake, Stanford University
Gary Rosenberg, Academy of Natural Sciences

Although the reviewers have provided many constructive comments and suggestions, they were not asked to endorse the content nor did they see the final draft of the workshop summary before its release. The review of this workshop summary was overseen by Marvalee Wake, University of California, Berkeley. Appointed by the National Research Council, she was responsible for making certain that an independent examination of this workshop summary was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this workshop summary rests entirely with the institution.

Contents

Introduction

APPLICATIONS OF DNA SEQUENCE INFORMATION

In the past two decades, advancements in DNA-sequencing have opened many new doors that range from evolutionary biology to biomedical sciences to forensics. Taxonomists and systematists use genome analysis to decipher relationships within species and in the branching patterns of the tree of life (Eisen and Fraser, 2003). Diseases, such as breast and prostate cancer, can be caused by genetic alteration and DNA sequencing allows the identification of genetic biomarkers for those diseases (Rubin et al., 2002; King et al., 2003). Forensic scientists use nucleotide variants that are characteristic of an individual as the person's identifier or DNA "fingerprint" (Gill et al., 1985).

Taxonomists and systematists have developed another application for DNA sequencing: DNA barcoding is a technique for characterizing species of organisms using a short DNA sequence from a standard and agreed-upon position in the genome (CBOL, 2006). Barcoding could be a universally applicable diagnostic tool for species identification. In fact, the Consortium for the Barcode of Life (CBOL), which is devoted to developing DNA barcoding as a global standard in taxonomy, has begun to construct a reference library that links species' names with DNA barcode sequences (CBOL, 2006).

NATURAL HISTORY COLLECTIONS AS A SOURCE OF DNA

Natural history collections in museums and academic institutions contain a wealth of specimens that could be used to construct a DNA reference library. The specimens are invaluable; collectively, they constitute a partial documentation of biodiversity and they serve as the tools for evolutionary and comparative physiology and for many other disciplines. More important, many of those specimens are irreplaceable. Many natural history collection specimens are fixed and sometimes stored in formalin, which is inexpensive, widely available and effective, although it is an environmental toxin. A saturated solution of formaldehyde (CH_2O) in water, formalin is about 37 percent formaldehyde by weight, and a standard fixation solution is 10 percent formalin in water, buffered to about pH 7. Formalin prevents degradation of specimens by microorganisms,

and because it stabilizes and maintains the fine structure of soft tissue, it is still a widely used fixative.

Aside from its toxic properties, formalin has another shortcoming—its use alters the DNA in samples. When they are exposed directly to formalin, mammalian cells undergo genetic and chromosomal alterations. Pathologists and other biomedical researchers who use archival tissue samples taken during epidemics, for example, or from victims of rare diseases have had some success in extracting DNA from formalin-fixed samples, but those samples have been embedded in paraffin rather than suspended in aqueous formalin or ethanol. Few of the many attempts to obtain and sequence DNA from formalin-fixed specimens stored in aqueous formalin or ethanol have been successful (Shedlock et al., 1997; Schander and Halanych, 2003). All of the protocols are slow, difficult, and often expensive, and few produce DNA fragments longer than 500 base pairs. Development of an effective protocol for recovering DNA sequence information from specimens fixed in formalin and stored in formalin or alcohol will give access to sequence information for thousands of species that are extinct, rare, or difficult to re-collect.

WORKSHOP DESCRIPTION

At the request of the Consortium for the Barcode of Life, the Museum of Comparative Zoology of Harvard University, the National Evolutionary Synthesis Center, New England Biolabs, Inc., Sigma-Aldrich Company, the U.S. Department of Agriculture's Agriculture Research Service, and the U.S. Environmental Protection Agency's Environmental Monitoring and Assessment Program, the National Research Council convened a one and a half day advanced workshop to discuss the future of DNA recovery from formalin-fixed specimens in museums or other natural history collections. The workshop's participants were chemists, biophysicists, biochemists, molecular biologists, bioinformaticists, and researchers and managers of natural history collections interested in obtaining DNA from their specimens, all of whom participated actively. They examined attempts at DNA recovery on formalin-fixed specimens and discussed the research to advance the development of similar but more efficient and cost-effective protocols (Box 1-1).

BOX 1-1
Statement of Task for the Workshop on Recovering DNA from Formalin-Fixed Biological Samples

The workshop was to bring together chemists, biophysicists, biochemists, geneticists, and bioinformaticians to examine past attempts on DNA recovery from formalin-preserved biological specimens and discuss the research needed to advance the development of similar but more efficient and cost-effective protocols. The goal of the workshop was to develop a research agenda that will shed new light on the problem and lead to a solution. Among the questions to be discussed at the workshop are:

- What is the state of preservation of DNA in the presence of formalin? Are the

DNA chains intact or broken? Does formalin denature DNA or is it the process of extraction that is fragmenting the DNA?

- Are the nucleotides at each site being preserved or altered?
- How can the physical and chemical states of the DNA-formalin cross-linkages be better characterized? What additional information on these cross-linkages is needed?
- What new chemical and physical methods for DNA extraction should be tested, beyond those that have already been applied to formalin-fixed tissue?
- In what ways and to what extent can fragmented DNA be repaired physically and chemically after extraction from formalin?
- Can bioinformatics techniques be used to reconstruct the original sequence in silico from the DNA fragments recovered from formalin?

Because it is known that formalin induces DNA fragmentation and nucleotide alteration, the workshop focused primarily on the extent and process of DNA damage and on DNA recovery from formalin-fixed samples stored in formalin or alcohol. The briefing material distributed to participants specified that some of the lessons learned from the protocol development for DNA recovery from formalin-fixed, paraffin-embedded biomedical specimens may be examined to assess whether they could be applicable to museum specimens. However, discussions on how to enhance protocols for extracting DNA from formalin-fixed, paraffin-embedded samples and discussions on extraction of ancient DNA were beyond the scope of the meeting.

To plan the workshop, the National Research Council appointed a steering committee of experts in nucleic acid chemistry, structural chemistry, biomedical sciences, molecular biology, and biodiversity (Appendix B). The steering committee had several teleconferences to discuss the goals of the workshop with the sponsors, to identify workshop participants, and to discuss the workshop format. The workshop convened May 8-9, 2006, at the Keck Center of the National Academies. The participants were the steering committee members; representatives of sponsoring agencies; and others invited because of their expertise in biochemistry and biophysics of nucleic acids, organic chemistry, DNA repair proteins, optimization of DNA extraction, single-molecule sequencing, bioinformatics, mass spectrometry, DNA damage and repair, molecular biology, and taxonomy and systematics (Appendix B). The group included experts who could shed light on why DNA extraction from formalin-fixed samples had been mostly ineffective or unsuccessful, those who knew the methodologies for studying DNA damage, and "end users" who were attempting to obtain sequence information from formalin-fixed samples for their research. Participants discussed the questions outlined in the statement of task, and as a group developed a list of suggestions for how to move toward effective recovery of DNA from formalin-fixed samples in natural history collections.

Workshop Proceedings

The workshop was divided into four sessions (Appendix C). The first focused on properties of DNA after formalin fixation. The second examined ways to obtain sequence information from formalin-fixed samples. In the third session, participants discussed applications of bioinformatics for reconstructing DNA sequences from formalin-fixed samples. Each session began with brief presentations by participants with relevant expertise, followed by open discussion. The challenge to participants was to identify a path to successful recovery of DNA sequence information from formalin-fixed samples stored in either alcohol or formalin. In the final session, participants made suggestions on areas of research and experimentation needed to investigate the mechanisms and kinetics of DNA damage by formalin fixation and on how to develop ways to repair DNA that would make it useful for study.

Workshop cochair, Donald M. Crothers (Yale University) acknowledged the enormous potential for advancing science that could accrue to DNA sequence information from museum specimens. Biologists in various disciplines, for example, would like to use natural history specimens collected over 100 year spans for evolutionary, molecular, and genetic studies. However, users encounter major problems in obtaining both the DNA and the sequence information from those samples because of interference from the formalin fixation. Although the proximate goal is to obtain sequence information of the cytochrome c oxidase subunit 1 (COI) gene for DNA barcoding,[1] the ultimate goal is to obtain genome sequences for other studies.

Biologists have used different methods for extracting DNA from formalin-fixed samples, and some have yielded DNA sequence information, but only under narrow conditions. However, those nominal successes suggest that the problem of recovering DNA sequence information from formalin-fixed samples is solvable. This workshop brought a group of experts together to discuss potential solutions and alternative methods.

[1] "DNA barcoding is a technique for characterizing species of organisms using a short DNA sequence from a standard and agreed-upon position in the genome. DNA barcode sequences are very short relative to the entire genome and they can be obtained reasonably quickly and cheaply. The cytochrome c oxidase subunit 1 mitochondrial region (COI) is emerging as the standard barcode region for higher animals. It is 648 nucleotide base pairs long in most groups, a very short sequence relative to 3 billion base pairs in the human genome, for example" ("DNA barcoding," Consortium for the Barcode of Life, http://barcoding.si.edu/DNABarCoding.htm).

Crothers clarified that the specimens in question had been fixed and stored in 5 to 10 percent formalin solution or had been fixed in formalin solution for a few days and then preserved in ethanol. Workshop cochair, Ann Bucklin (University of Connecticut), explained that in some cases the formalin for preservation or storage was unbuffered and therefore acidic.

Mark Rubin (Brigham and Women's Hospital) and David Schindel (Consortium for the Barcode of Life; CBOL) suggested that, although the workshop's focus was on biological samples stored in aqueous solution, much could be learned from protocol development for DNA extraction from formalin-fixed and paraffin-embedded samples. Marvin Caruthers (University of Colorado) said that paraffin embedding creates a more stable environment for the biological sample than storage in aqueous formalin or alcohol. For example, whereas the pH of the paraffin does not change, the formaldehyde in formalin can be oxidized to formic acid by exposure to atmospheric oxygen, thereby reducing its pH.

DNA IN SAMPLES EXPOSED TO FORMALIN

Reactions of DNA and Formaldehyde

To begin the discussion on the effect of formalin exposure on DNA, Crothers showed a slide that sums up the reactions that occur in formaldehyde fixation of a drug, adriamycin (Figure 1). During fixation, formaldehyde reacts with amino groups of guanine (G), adenine (A) and cytosine (C), and a guanine residue is a typical reaction product. The formaldehyde forms covalent linkages with the amino groups, which then cross-link with proteins, so the drug is linked to the guanine moiety. The other guanine moiety has a strong hydrogen bond with adriamycin, which produces a tight, stable complex. However, the drug must be kept cold to maintain the stability of the fixation; heat can cause the disassociation of the entire complex. The cross-links are labile to the aromatic amines of DNA, and the cross-links or the reaction to formaldehyde are stable.

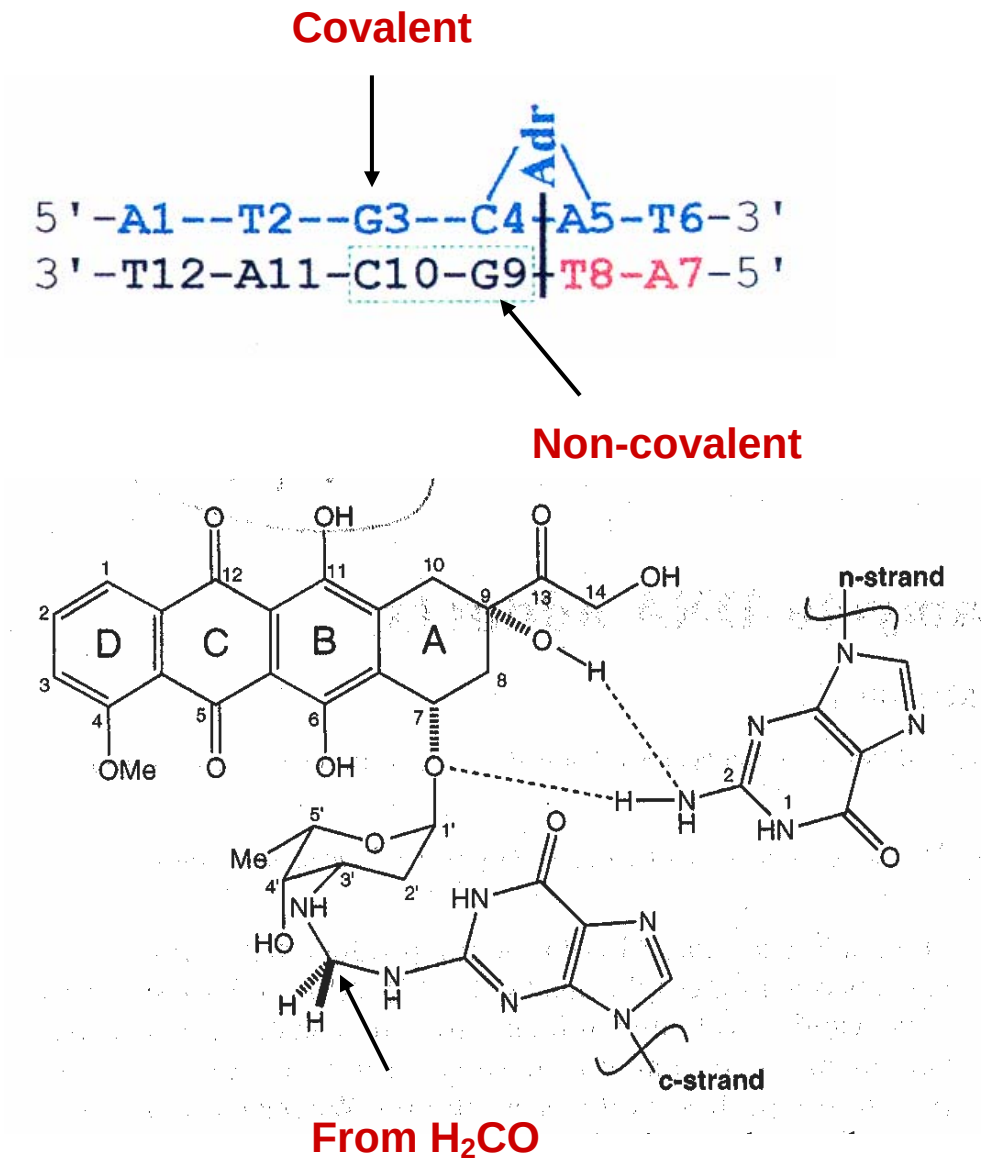

Figure 1 Covalent adriamycin-DNA adduct. SOURCE: Zeman et al., 1998.

The process of formaldehyde fixation alters DNA in three ways; through fragmentation, sequence modification, and cross-linking. Cross-linking is not destructive to nucleic acids, and is reversible. Using ^{13}C-labeled formaldehyde, Crothers said, it is possible to see that the methylene carbon came from the formaldehyde (Figure 1). Nuclear magnetic resonance (NMR) spectroscopy can be used to show what happens to the formaldehyde carbon when it reacts with DNA in different circumstances so that the kinetics of the reactions between formaldehyde and different kinds of nucleotides over time can be revealed. Using modern NMR methods to study formaldehyde's reactions with double-helical or single-stranded oligonucleotides could reveal information on the kinetics.

If formaldehyde is left unbuffered, it is oxidized to form formic acid, which has destabilizing properties. The formic acid depurinates DNA, that is the cleavage of the N

glycosidic link between purine bases and deoxyribose in DNA resulting in the loss of purine from the DNA backbone, and the degradation is likely to be irreversible. Crothers mentioned a paper by Quach and colleagues (2004) that assessed sequence modifications due to formalin fixation. That group reported that formalin fixation speeds sequence modification, but that the rate does not depend on the duration of formalin fixation. The ability to make a longer amplicon using polymerase chain reaction (PCR) analysis decreases dramatically with increasing duration of formalin fixation. Crothers said he suspected that the formaldehyde used for fixation is oxidized to formic acid over time which causes denaturation of DNA and more cross-linking reactions. Storage of samples in unbuffered formalin for prolonged periods is likely to produce DNA that is so degraded it cannot be used for PCR analysis.

Crothers cautioned that, in some cases, even if PCR did not produce amplified DNA, the lack of an amplicon does not imply an absence of DNA. Rather, DNA purification reagents could contain PCR inhibitors. In response to that comment, Charles Cantor (Sequenom, Inc.) suggested the use of an internal control (that is, adding copies of a standard that is known to amplify with its primers) to ensure that PCR was not inhibited. Crothers suggested mass spectrometry or single-molecule sequencing for small DNA fragments as an alternative to PCR. Because single-molecule sequencing can be done on multiple molecules, the resulting sequences could be compared to locate the damage in each sequenced molecule.

Caruthers agreed with Crothers that sequencing DNA from formalin-fixed samples is comparable to sequencing apurinic acid with small stretches of pyrimidines. He suggested that sequence information can be recovered from those small stretches but the informatics involved would be challenging. Mitochondrial DNA (MtDNA) has many adenine-thymine base pairs (that is, it is A-T rich) so that the method that Caruthers suggested is not likely to work, said Robert DeSalle (American Museum of Natural History). Timothy O'Leary (U.S. Department of Veterans Affairs) added that MtDNA is less accessible than is nuclear DNA, probably because of the abundance of adenine-thymine base pairs.

Caruthers stated that the major problem is in the cross-linkages. First, there are cross-linkages between bases—adenine with adenine or adenine with cytosine. Second, there are cross-linkages with proteins—for example, histones with a lot of lysine. The samples would require protease treatment to yield DNA. Protease K is commonly used to degrade proteins, but Caruthers said that the cross-linkages would halt DNA polymerization. He suspected that the polymerase would be stopped by lysine-adenosine or adenine-adenine linkages. Aside from cross-linkages, Caruthers said it is not clear how DNA would be degraded by formalin unless a solution were acidic and acid hydrolysis were causing depurination.

Daniel Ryan (Agilent Technologies, Inc.) suggested a possible solution to the depurination problem. In DNA, purines are base paired with pyrimidines so that all that remains in depurinated DNA is the pyrimidines—absent their complementary bases. Ryan and his colleagues are working with microarrays and their addressable beads. They have seen a single nucleic acid bound to a spot, and they have detected one molecule or one of those spots. He suggested a technique to isolate depurinated DNA using the DNA's remaining binding energy. Crothers questioned whether there is a hybridization

system that can recognize depurinated DNA. Ryan suggested that it is an astringency problem.

Cantor suggested a method complementary to Ryan's. He said that if a universal base, such as inosine, could be added to those apurinic sites, the fragments would become nucleic acids again, and working with nucleic acids is simpler. Caruthers and Timothy Harris, (Helicos BioSciences Corporation) agreed that sequence information could be obtained from nucleic acids reconstructed from fragments of depurinated DNA by single-molecule sequencing. However, Tom Evans (New England Biolabs, Inc.) said the proposed repair method would work only with double-stranded DNA.

O'Leary showed the reactions of formaldehyde with nucleotides and nucleic acids over a short period (under 24 hours) (Equations 1 and 2).

$$=NH + CH_2O \leftrightarrow =NCH_2OH \qquad \text{[Equation 1]}$$

$$-NH_2 + 2CH_2O \leftrightarrow -N(CH_2OH)_2 \qquad \text{[Equation 2]}$$

He reported that the reactions are reversible. The reaction between formaldehyde and nucleic acid carried out at 24 ^{O}C could be reversed by incubation at 70 ^{O}C or by dialysis. However, if the sample were transferred to alcohol after fixation, other reactions and molecular alterations would occur. Transferring the formalin-fixed sample to ethanol triggers depurination of DNA in the sample and formation of ethanol adducts. The ethanol adduct could be cleaved at pH 4, which is not acidic enough to cause depurination. Even a short period of formalin fixation followed by dehydration can cause damage but the nature of that reaction is not well-studied. Investigating the reaction with single nucleotides could prove useful.

Reliability of Sequence Information

Some participants questioned the reliability of the sequence information obtained from formalin-fixed samples. For example, would certain DNA strands be more susceptible to sequence modification as a result of formalin fixation? If so, the DNA sequence obtained would not be a true representation of the original specimen. DeSalle suggested that multiple clones could be examined to see whether there is consensus in the DNA sequences among the clones. Harris pointed out that examination of multiple clones would be only useful if the sequence alterations that result from fixation are random. By comparing multiple clones and looking for overlap in sequences, a true sequence can be deciphered. However, if the alteration is systematically biased—that is, if sequence modification occurs in the same region of every clone—then there is no way to determine where the alteration occurs unless the formalin-fixed sample can be compared with a fresh sample.

Rubin pointed out that conducting a systematic comparison of fresh and preserved samples could lead to a better understanding of the problems associated with recovering DNA from formalin-fixed samples. The comparative study would allow documentation of alterations in formalin-fixed samples. Then, more research could determine whether any of those alterations hampers the determination of the original sequence.

Oxidative Damage

Miral Dizdaroglu (National Institute of Standards and Technology) discussed his work on oxidative stress and damage. His laboratory uses mass spectrometric techniques to observe oxidative damage in DNA isolated from animal tissues between 10,000 and 20,000 years old. Oxidative stress causes the formation of highly reactive hydroxyl radicals, which react with DNA bases and with the sugar moiety of DNA, possibly to cause base damage, sugar damage, DNA protein cross-links, and single- and double-strand breaks in DNA. The damage occurs because a hydroxyl radical can add to the double bonds, forming other intermediate radicals that react further. Cytosine and purine lesions, for example form as a result of reactions with hydroxyl radicals. Most lesions are mutagenic, which means that polymerase will not stop but will go onto the wrong base across from those lesions. Some lesions are lethal; they stop polymerase, and DNA cannot be synthesized from that point on. Dizdaroglu has found thymidine dimers in samples exposed to ultraviolet radiation, but he had not compared lesions in DNA extracted from fresh samples with that extracted from formalin-fixed samples. Harris questioned whether a vial of DNA that sustains oxidative damage could be repaired by an enzyme cocktail. Several participants said that it could be repaired to some extent. Basic repair, however, involves many enzymes, said Dizdaroglu. Crothers added that DNA had to remain double stranded for successful enzymatic repair. Furthermore, hydroxyl damage has some minimal sequence preference, so lesions might not always occur in the same region.

Variations in Curatorial Treatments

Participants who work with natural history collections discussed the variations in the curatorial processing of biological samples. Buffered and unbuffered formalin, for example, have been used for fixation and storage. Although most zooplankton samples are fixed and stored in formalin, others often are fixed in formalin and transferred to a 70 percent ethanol solution after fixation. The duration of formalin fixation varies widely among samples stored in ethanol.

Ryan asked whether anyone had determined the size of DNA extracted from formalin-fixed specimens. Specifically, he was wondering whether a specific formalin treatment or curatorial treatment of a specimen leads to increased fragmentation of DNA. Crothers asked whether anyone had obtained PCR products from specimens preserved in aqueous formalin. Bucklin replied that she and her colleagues had examined DNA sequences for northern krill, *Meganyctiphanes norvegica* (Crustacea, Euphausiacea), fixed and stored in formalin for 2, 3, 18, and 15 years (Bucklin and Allen, 2004). When they amplified the DNA to determine the size of fragments, they found that the longer the specimen had been preserved in formalin, the shorter DNA fragments. Bucklin and colleagues had not been able to obtain DNA from specimens stored in unbuffered formalin. Crothers reiterated that obtaining DNA from samples that had been stored in pH 2 formalin is fruitless because the formic acid is likely to have irreversibly depurinated the DNA. Caruthers stressed that an attempt to restore damaged DNA by neutralizing pH 2 formalin might incur more damage, at least to the sample's RNA.

Christoffer Schander (University of Bergen) also has compared the success of different protocols for extracting DNA from tissues fixed in formalin for different durations (Schander and Halanych, 2003). He reported that the success of DNA extraction depends not only on the protocol, but on the tissue from which the DNA was extracted. O'Leary suggested that bones and teeth could be useful alternatives to soft tissue for obtaining DNA. Bones and teeth might be better protected from damage by formalin. However, Evon Hekkala (U.S. Environmental Protection Agency) and Gonzalo Giribet (Museum of Comparative Zoology, Harvard University) pointed out that the strategy would be useless for many organisms that have neither bones nor teeth.

Some participants asked how many repeats of the sequencing process would be needed to obtain reliable information. The number, according to Ernie Mueller (Sigma-Aldrich Company) would depend on the size of the DNA fragments. The shorter the fragment, the more repeats are needed. One participant indicated that fragments of 500-600 base pairs were the longest that had been obtained from formalin-fixed samples stored in aqueous solution.

Rubin suggested that information about the curatorial history of samples is critical to developing an optimal protocol. He mentioned that he chaired a task force at the National Cancer Institute to devise an optimal protocol for obtaining DNA from archival human tissues. That group reported that researchers in different laboratories use different protocols for sample preservation, and sometimes, even a slight variation can make a big difference in the success of DNA extraction. Determining which variation in a preservation or processing protocol has the largest effect is an important step in identifying optimal protocols for DNA extraction. Hekkala thought that Rubin's approach might be useful for developing a matrix that could be used to predict whether particular specimens would be useful for recovering sequence information. Participants agreed that a survey of curatorial practices could be useful for determining which other factors should be considered in identifying specimens that have the potential for DNA extraction.

Giribet recalled a project by Bhadury and colleagues (2005) that examined nematodes preserved in formalin. When the nematodes were fixed for 7 days, the extracted and amplified DNA showed clear bands on a gel. But when the nematodes were fixed for 11 days, the gel showed smears instead of distinct bands, indicating that much of the usable material was lost. Bucklin said that she had obtain good DNA from samples fixed for a week or less, regardless of whether the formalin was buffered or not. Steven Hofstadler (Isis Pharmaceuticals) questioned whether Bhadury's group quantified the extracted DNA. If not, the results could be interpreted as a lower yield in extracted DNA from samples fixed for longer duration instead of as a decrease in amplification.

Schindel asked whether there is a way to detect the DNA's integrity without extraction. Crothers said that would be a true analytical challenge. Cantor mentioned that the DNA sequences with multiple thymines in a row (called poly-T tracts) are more stable than others. Those sequences preserve well because they do not react with formaldehyde, said Crothers. One of Cantor's students found that poly-T tracts in closely related bacteria can be distinguished if they are measured precisely. However, Cantor does not know the variability of poly-T tracks in higher organisms.

Schindel asked about the relative importance of DNA degradation and PCR inhibition when DNA amplification has not been observed. There is a potential for small

molecules to block PCR, especially if there are only few copies of the DNA to be amplified, said Crothers. Therefore, an internal control would ensure that PCR was not inhibited. Hofstadler cautioned that the amount of internal standard could mask the DNA to be amplified if the DNA is present only in a low concentration. In addition to small molecule inhibitors, a lesion in DNA also can inhibit PCR, said Crothers. A lesion is an inhibitor in a sense that even though a molecule is of a given length, it cannot be amplified because the enzymes cannot get through it. That kind of PCR inhibition is more difficult to control for. The larger the DNA fragment, the more likely it is to have a lesion or protein bound that blocks PCR.

The discussion turns to questions of how to assess the integrity of DNA before extraction. Alison Williams (Princeton University) suggested that capillary electrophoresis might be sensitive enough, and that perhaps a few bases could be observed from one or two kilobases. Cantor suggested ethidium bromide and 4',6-diamidino-2-phenylindole (commonly known as DAPI), and DeSalle suggested spectroscopic analysis. Ryan suggested reversing cross-linking by hydrolysis with water at 65 ^{O}C. Intercalating dyes that have specific fluorescence signatures, such as ethidium bromide or the more sensitive PicoGreen, can be added (Ahn et al., 1996).

Schindel asked whether heating the samples would enhance extraction. O'Leary explained that the aqueous heating might not be useful for samples from natural history collections because those samples are likely to sustain more extensive damage than are paraffin-embedded, formalin-fixed samples. Hekkala said she had followed the critical drying method proposed by Fang et al. (2002) but did not obtain any amplifiable DNA from her samples. Crothers noted that some participants had not shared information about failed attempts at DNA extraction from formalin-fixed samples until this workshop. The user community has no means of communicating the protocols that they have tried to use to extract DNA and failed. Sharing that information is important because the collective information could shed light on why some attempts succeed. Crothers suggested a Web site be setup for that purpose.

The development of a set of standardized reference samples to test DNA extraction protocols would be useful for comparing the protocols to determine what works under which conditions, said Schander. Scott Miller (Smithsonian Institution) explained that the Smithsonian has a set of specimens—goldfish exposed to a series of different formalin treatments—that could be used as the standards. The specimen set includes samples of frozen tissue that has not been exposed to formalin. Schindel said he would like to see the goldfish standard held in reserve until an acceptable approach to extraction protocol testing is developed. For example, Schindel had hoped that chemists could elucidate the degradation processes in formalin fixation that block DNA extraction or hamper PCR analysis, and then identify or develop better protocols. "There is no gold standard method at the moment", Crothers said. Because of variations in curatorial processes, the chemistry of DNA degradation in formalin-fixed specimens is largely unknown.

In summation, Crothers listed possible alterations or damage to DNA exposed to formalin. They include irreversible depurination caused by acidification (if formaldehyde is unbuffered), cross-linking, oxidation from reactions with minor content of formaldehyde, cytosine deamination, and minor adducts. The chemistry of cross-

linking is not well understood, and some cross-linkages are rather stable. Cytosine deanimation is enzymatically reversible in samples that contain double-stranded DNA.

OPTIMIZATION OF DNA SEQUENCE INFORMATION

This section explored how sequence information can be optimized after DNA is extracted from samples. Several participants explained methods that could be used to assemble sequence information from DNA.

PCR-Mass Spectrometry

Cantor explained a PCR-mass spectrometry method developed at Sequenom. The method's advantages are in its sensitivity, precision, and cost-effectiveness. Its sensitivity is better than what is possible in conventional sequencing methods because mass spectrometry produces no background noise, and it is less expensive than real-time PCR analysis. Although the technology cannot be used to survey an entire genome, it is a cost-effective method for examining hundreds of loci in thousands of samples. PCR-mass spectrometry is fully automated, and it can process about 3000 samples per day. Some 160 entities are using the PCR-mass spectrometry technology; many are organizations that provide the service for a fee.

PCR-mass spectrometry is a multiplex method that can analyze 30 genotyping or 20 gene expression samples simultaneously in one tube. Matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry requires a smaller amount of DNA as input than PCR, so that the method is optimally designed for small amplicons. The mass spectrometry method covers an unlimited dynamic range. Because mass spectrometry is expensive, it is not used for standard sequencing. Rather, nucleic acids are sequenced by base-specific cleavage reactions. Sequenom has all four single base-specific cleavage reactions working with complicated, but single-tube, technology. It is possible to quantify reactions at every locus in mixed, complicated samples, including deanimated samples. Conventional sequencing requires one continuous target. But sequencing in the mass spectrometer by fragmentation does not require a continuous target, so a discontinuous set of sequences can be sampled for the cost of a single sequencing reaction. The PCR-mass-spectrometry method yields 98-99 percent correct typing. Cantor encouraged the users to try the method because it is a mature and available technology. Caruthers asked whether multiplexing PCR is a problem. Cantor explained it is not, because the amplicons used are short and because all amplicons are close to the same size. Sequenom had experience working with short amplicons, and the multiplex is designed by software. Generally, 28 of 30 multiplexes work well.

Mass Spectrometry for Interrogating Nucleic Acid

Hofstadler explained another mass spectrometry method for obtaining base composition information from PCR products. He reiterated Cantor's point that the mass spectrometer is a powerful tool for analyzing base composition of nucleic acids. Unlike conventional sequencing, mass spectrometry as described by Hofstadler derives a "base

composition" signature, which represents the exact count of adenine, guanine, cytosine, and thymine (for example, A10 G23 C32 T17), on the basis of the precise measurements of the PCR product's mass. Unlike a sequencing technique, the approach can provide information on a mixture of nucleic acids with a dynamic range of about 100:1. Mass spectrometry is highly automated, so it can run round the clock to analyze more than 1500 samples in 24 hours. It also is more sensitive than conventional sequencing. Hofstadler said he had performed single-molecule detection at the stochastic limit of PCR to obtain PCR amplicons from 10-20 copies of a genomic template or reaction.

Hofstadler described the electrospray interface and solution conditions he used to ionize intact DNA. The procedure starts with DNA molecules in solution. Using specific interface and buffer conditions, the DNA is denatured in the gas phase into complementary strands. The mass spectrometer measurement provides independent measurements of the forward and reverse strands of the amplicon. From those strands, an unambiguous base composition can be determined for the complementary amplicon pair. Because large molecules yield many charge states upon electrospray ionization, measurement of those charge states from the same molecule facilitates the determination of molecular weight, which is in turn used to determine base composition.

Analytical chemists routinely identify the elemental composition of small molecules based on mass measurements. Knowing the weight of the molecule and the atomic mass of the elements, chemists can calculate the proportion of each element in the molecule. Similarly, base composition of DNA can be determined from the mass of the DNA strand and from the known masses of its adenine, guanine, thymine, and cytosine. The PCR products that Hofstadler typically examines are no more than 150 base pairs long. Molecular weights from both strands are used to derive base compositions because determination of base composition from a single strand is ambiguous even with high-precision measurements (for example, to one part per million [ppm]). For a DNA strand that weighs about 33,000 mass units, there typically are 80-90 base compositions that add up to the same mass, within achievable mass measurement uncertainties. However, because PCR reveals complementary products—the number of adenine on one strand equals the number of thymine on the other—the determination of base composition is simple as long as a mass measurement of 25 ppm is achievable. PCR typically is driven to saturation (35-40 cycles), so analytically useful measurements are derived routinely from low numbers of copies.

The base composition approach offers at least two potential advantages over conventional sequencing: Complex mixtures can be interrogated directly, and useful signatures can be derived from relatively short pieces of highly degraded DNA.

Single-Molecule Sequencing by Synthesis

Harris discussed single-molecule sequencing, a method designed for short-read resequencing of genomes with a reference that was first presented by Braslavsky and colleagues (2003). Single-molecule sequencing requires fragmentation of genomic DNA but, as Hofstadler pointed out, that might not be necessary because the DNA obtained from formalin-fixed tissue is already fragmented. In the method, genomic DNA is fragmented into pieces on the order of 100 base pairs long and the fragments are then melted to single strands (Figure 2). If the DNA is to be repaired, it should be repaired at

the ends, and then poly-A tails are added to the 3' end of each fragment. For convenience, a Cy-3-labeled dideoxy-nucleotide also is added at the 3' end so the strand cannot be extended in that direction. Each fragment shown at the bottom of Figure 2 represents a single molecule of DNA that will be probed as an individual target template. Slides are made with poly-T on the surface. The poly-A serves as the capture probe and as a primer for sequencing. A picture is then taken with image green to record the position of each template. A dye labeled "nucleotide X" is added to the DNA fragments enzymatically at sites where the target strand contains a nucleotide that is complementary to X. Misincorporation is negligible (less than 0.1 percent). The excess deoxyribonucleotide triphosphate is then washed out, and a picture is taken of each dye molecule.

Finally, the dye molecules are cleaved from the DNA and the sequencing reaction is repeated for nucleotide Y (X, Y, and so on correspond to A, T, G, C). Each piece is a single molecule, sequenced by synthesis. The read length is limited to fewer than 50 nucleotides (50mers), usually about 30 nucleotides (30mers), because the sequence is synthesized one base at a time and because of unknown chemical failures. Harris asked the bioinformaticians present whether they could assemble 700 base pairs (1400 bases) of complementary sequence from 30 polymers. Steven Salzberg (University of Maryland, College Park) stated that would be possible if the sequence were only 700 bases long, but not if the sequence has 7 million bases.

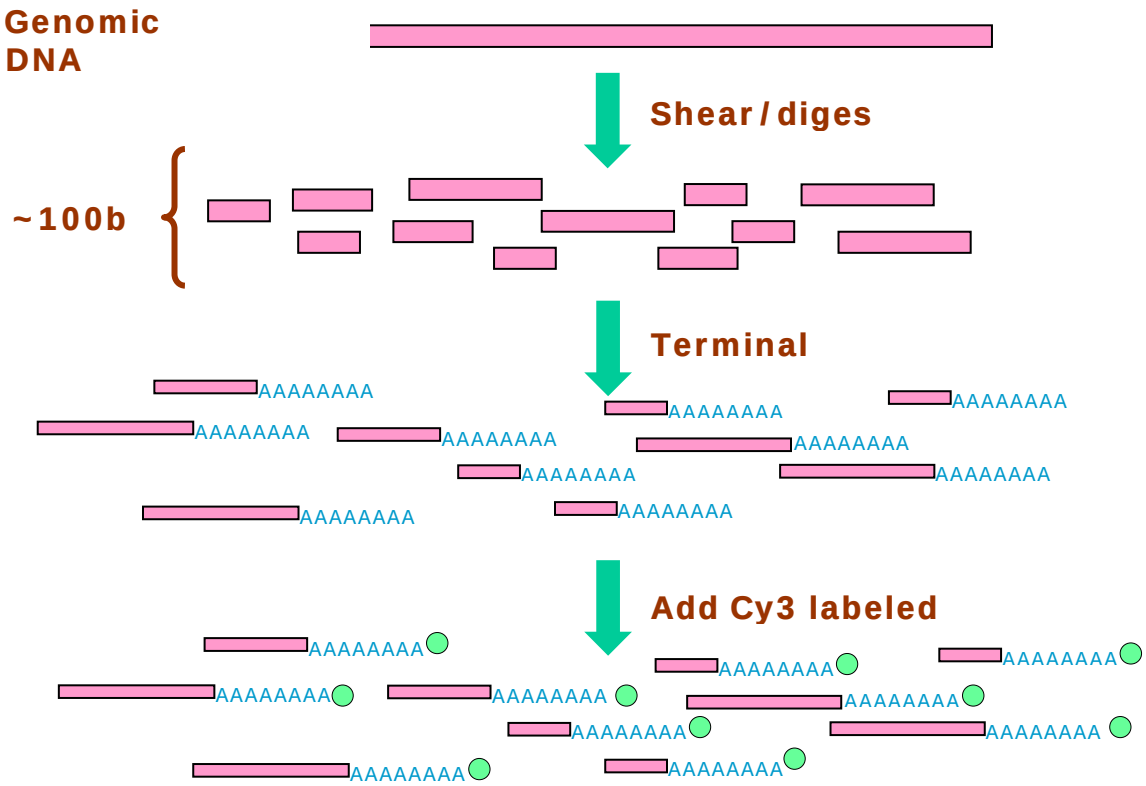

Figure 2 The first three steps in single-molecule sequencing. SOURCE: T. Harris, Helicos BioSciences Corporation.

DeSalle asked whether Harris had encountered problems with normalization. Harris replied that although no problems had arisen yet, the representation of sequence is not uniform. Some sequences appear overrepresented (excess coverage), others are underrepresented. This nonuniform variation of sequence coverage was found in single-molecule sequencing for viruses. It is not known whether the coverage bias arises from biased sample fragmentation, from bias in the sequence generation process, or both.

In Vitro Repair

Evans discussed a "PreCR" method for in vitro DNA repair that involves treating damaged DNA in vitro with a mixture of DNA repair enzymes before PCR. The enzyme mixture emulates base excision repair, the core enzymes are a DNA polymerase and a DNA ligase. A broad spectrum of DNA damage is repaired by including other DNA repair enzymes that act in concert with the ligase and polymerase. At the time of the workshop, the PreCR enzyme mix Evans had been using could repair abasic sites, nicks, gaps, ultraviolet radiation damage, deaminated cytosine, and some forms of oxidative damage. DNA cross-links or highly fragmented DNA could not be repaired effectively, and the enzyme mix cannot repair highly damaged DNA.

The PreCR repair mixture was tested against templates purposely damaged by methylene blue, which causes oxidative damage, and by low pH, heat, and ultraviolet radiation. 8-oxo-guanine is reported to be a common oxidative product that causes cytosine-to-thymidine changes in the PCR product. If oxidatively damaged DNA is a template for PCR, the amplicon contains many more mutagenic lesions than generally would be found in damaged DNA treated with FPG (formamidopyrimidine [fapy]-DNA glycosylase). FPG excises 8-oxo-guanine so that the number of mutagenic lesions decreases, but the DNA template is fragmented in the process, and the PCR results are affected. If FPG is used in conjunction with other enzymes, specifically a ligase, polymerase, and endonuclease IV, the mutagenic lesions are effectively removed and an amplicon is obtained.

The repair mix was designed to repair abasic sites, thymine dimers, gaps, nicks, and deaminated cytosine. A purposely damaged template was used during development to optimize the repair. However, PreCR was not successful when tested on formalin-fixed samples. The DNA damage needs to be better defined so that researchers can devise more effective solutions. Evans suggested the need to identify markers that indicate different types of DNA damage, so that he and his colleagues could design specific enzyme mixes for use in DNA repair. The lack of success could result from the presence of unrecognized DNA lesions that were not repaired by the enzymes currently in the PreCR mix.

Whole Genome Amplification

Mueller presented an overview of GenomePlex®, a Sigma-Aldrich whole genome-amplification system for use on damaged DNA and possibly on formalin-fixed tissue. The system starts with partially fragmenting genomic DNA, a step that would be unnecessary with the fragmented DNA found in formalin-fixed samples. Mueller, however, cautioned that the system would not work with DNA that is too fragmented.

The fragments are efficiently primed and amplified to make a library of fragments that have a known sequence at each end, and thus are PCR-amplifiable units. PCR is then performed on those fragments with universal primers to create an amplified Omniplex® library (Figure 3).

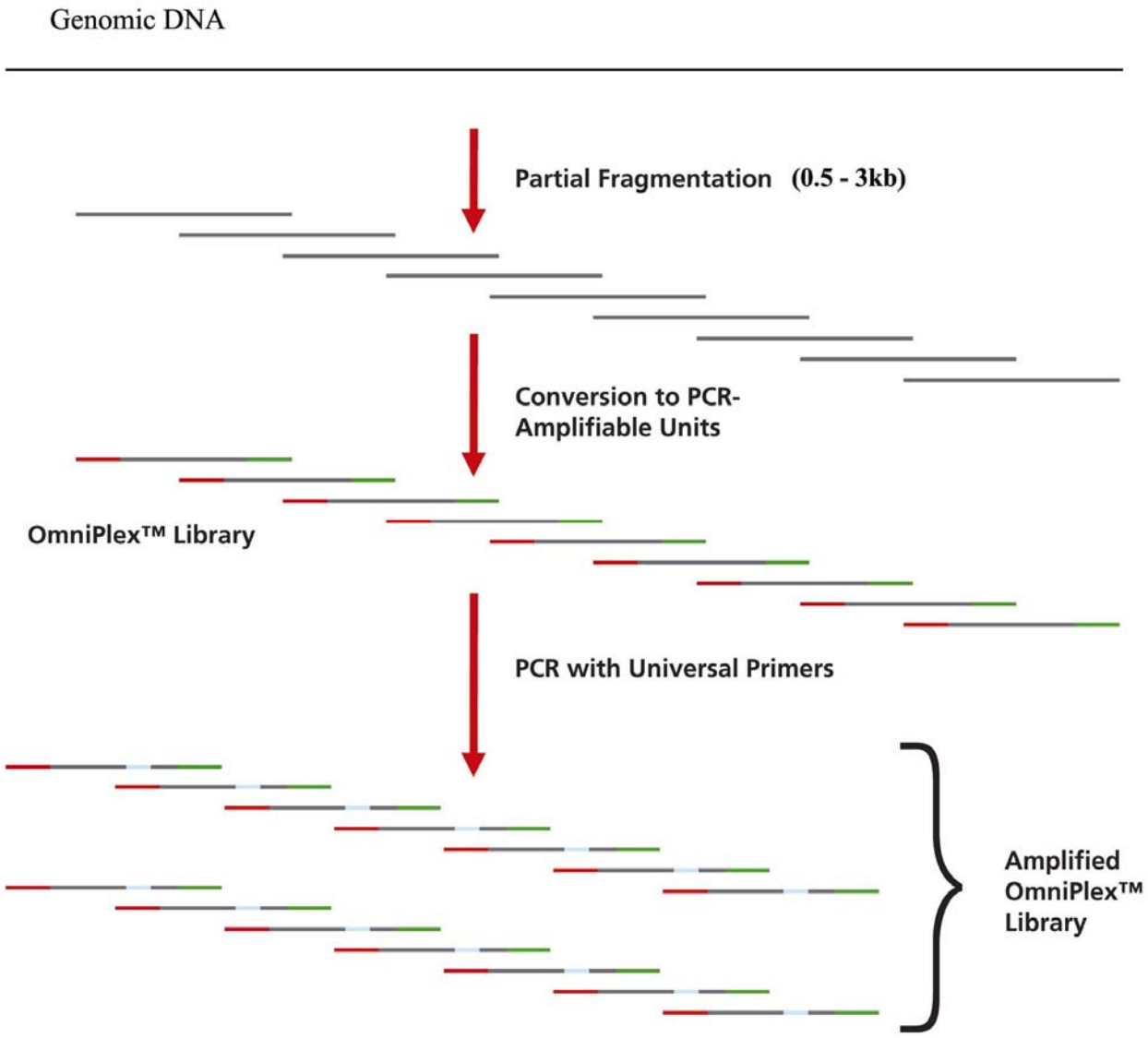

Figure 3 Overview of GenomePlex® whole genome amplification. SOURCE: E. Mueller, Sigma-Aldrich Company.

Before amplification, Mueller tested the amount of DNA extracted from formalin-fixed and fresh tissues using two protocols, the Leed's protocol, which uses proteinase K (Jackson et al., 1990), and a detergent-mediated lysis, CelLytic Y. His results showed a consistently lower amount (30 times less) of amplifiable DNA in formalin-fixed tissue than was available from fresh tissue. DNA yield varied slightly with the type of tissue used in the extraction. Because extraction gave consistent yields, as shown by quantitative PCR, Mueller and his colleagues presumed that the limited amplification was attributable to the quality of the extracted DNA. Large additions of extracted lysate

carried over from the extraction process often resulted in PCR inhibition. There are two issues associated with formalin-fixed samples. First, a substantial amount of DNA is lost in the fixation process—only three to five percent of the DNA can be extracted for PCR. Second, the DNA extraction process itself could introduce PCR inhibitors.

Mueller also mentioned that Sigma-Aldrich has been pursuing the development of a strand-based technique for DNA repair, similar to the one Evans presented. Mueller and his colleagues found a model system that repairs abasic sites. They sent the system for testing by M. Hajibabaei at the University of Guelph, Canada. Hajibabaei was working with the COI gene (cytochrome c oxidase subunit 1) in formalin-fixed samples. The model system worked for one batch of sample but failed for a second batch. Mueller said that the Smithsonian goldfish specimens might provide a good standard test for elucidating those inconsistent results.

Protocols for Obtaining Sequence Information from Formalin-Fixed Samples

Hekkala described her successes and failures in experimenting with various protocols for obtaining sequence information from formalin-fixed samples. She used crocodile tissue—brain, heart, and muscle—in her tests. She obtained different yields by various extraction protocols, although the quality of DNA did not differ. She had about 20 percent success in obtaining sequence information with the Shedlock protocol (Shedlock et al., 1997). Using identical tissue, she produced no sequence information with the critical point drying method (Fang et al., 2002) or with the Qiagen protocol (Wickham et al., 2001). The Shedlock protocol uses a glycine buffer to "soak out" the formalin. In that test, Hekkala ran 40 samples, 10 of which produced amplicons. She ran the extraction again on those 10 samples, and 8 yielded amplicons. There was no change from the 30 samples that produced no amplicons either on the first run or in subsequent runs. The 40 samples were obtained from different individual specimens provided by different museums. The 10 samples that yielded amplicons were taken from specimens provided by the American Museum of Natural History and the California Academy of Science. Twelve specimens provided by the Field Museum of Natural History yielded no amplicons. Curatorial processes at different museums could have affected the ability to extract and amplify the DNA.

Choosing the Path to Optimize Sequence Information

After hearing about different methods for optimizing sequence information from extracted DNA, Schander raised the question of what path could lead to sequence information from formalin-fixed samples. Should the effort focus on developing DNA extraction protocols suitable for formalin-fixed samples? Should it work at identifying enzymes for DNA repair? Should bioinformatics be used to assemble short fragments of DNA? Schindel said he favored research on improving DNA extraction, but DeSalle disagreed. Based on the presentations and discussions in the workshop, DeSalle thought the success of obtaining DNA sequence information would depend on whether the formalin-fixed samples in fact contained extractable DNA and not on the availability of extraction protocol or the suitability of an existing one. Extraction might produce long or short fragments, but sequence information can be obtained using some of the methods

already discussed. Therefore, he said, repairing DNA to obtain longer fragments might not be necessary, and identifying the samples with extractable DNA is more important than is refining extraction protocols. Evans countered that it is not known which aspect is the largest problem—damage to DNA, inefficient DNA extraction, or the presence of PCR inhibitors—in blocking retrieval of DNA sequence information from formalin-fixed samples. Therefore, DNA repair should not be discounted as part of the solution.

Schander echoed DeSalle's thought that it would be useful to determine whether a formalin-fixed sample contains any extractable DNA—samples left in unbuffered formalin are not likely to have extractable DNA—because extraction, PCR, and sequencing are time-consuming. Hekkala agreed and suggested experiments be designed to correlate traits of biological samples (curatorial history, pH, taxonomic group) with the potential for obtaining DNA from those samples. The correlations could inform the choice of samples for sequencing. Participants agreed that developing a framework to identify samples most likely to yield usable DNA would be useful. For example, samples kept in unbuffered formalin are not likely to be usable because of formic acid damage, and samples with free purines, which could be detected by mass spectrometry, also are unlikely to be usable. A formal assessment of the physical and chemical conditions and the curatorial history of formalin-fixed samples in natural history collections coupled with controlled DNA extraction experiments would be necessary to develop the framework. Crothers suggested a pilot explanatory assessment because the DNA degradation processes in formalin-fixed samples are largely unknown, although controlled experiments could shed light on that. He said that characterizing the quantity and quality of DNA extracted would be important, and that no one, at least at the workshop, seemed to have done that.

Determining the mechanism of DNA degradation (for example, depurination or formation of formaldehyde adducts), the efficiency of DNA extraction (whether extraction yields any DNA at all), and the efficiency of PCR amplification (whether damaged sites are amplified or PCR inhibitors are present) could help explain why sequence information cannot be obtained from particular formalin-fixed samples. Some participants said that the type of tissue used for extraction also affects the success of extraction or PCR amplification. Ryan suggested having a set of samples that included different tissue types sent to different laboratories to test the same DNA extraction protocols. The outcome of systematic, simultaneous, independent testing of extraction protocols would provide information about protocol success, and whether some tissue types yield larger amounts of extracted DNA than others. Schindel and Hofstadler also said that the Smithsonian's goldfish specimens would be useful for systematic testing.

BIOINFORMATICS FOR RECONSTRUCTING DNA SEQUENCES

Can sequence information be obtained from the ordinarily short DNA fragments obtained from formalin-fixed samples? Schander asked whether sequence information obtained from short fragments can be considered reliable. Salzburg had indicated earlier that assembling a 700 base region from 30mers is possible, but assembling an entire genome from 30mers is impossible. Neil Hall (The Institute for Genomic Research)

added that if random 30mers of a 700 base pair region are generated, then the 30mers can be assembled into larger contigs, but the problem is to pull out the right regions.

Hall said that the reliability of sequence information depends on what is done with the PCR product. Assuming that the PCR analysis is working well and the DNA is amplified to generate long sequences, there is no way to verify whether the reconstructed sequence is representative of the original if sequencing is performed directly on the PCR products. Bioinformatics also would be unable to determine the reliability of the sequence. However, if resequencing were done by cloning the PCR products so that the entire population of DNA in the sample were analyzed, it would be possible to identify and correct the sequence modifications by assembling those reads.

Salzburg emphasized that depth of coverage is necessary to distinguish between sequencing errors, sequence modifications and true mutations. A minimum of four-times coverage could be necessary, especially if the DNA is likely to have been damaged; the cost of sequencing 700 bases is within reason for most laboratories. Schander agreed that it would be worthwhile to increase depth of coverage to examine sequence modifications caused by formalin fixation. If sequence modification is induced by formalin, and if that is a common phenomenon, then the more that is known about it, the effects could be better predicted, said Hall. He reiterated that error screening requires sequencing of the cloned PCR products, and not the direct sequencing of a PCR product.

Participants discussed how many replicate sequences of COI gene would be needed to create a reference library for DNA barcoding. Schindel said the goal of the barcoding project is to create a reference library with bidirectional reads of five specimens per species, but no replicate per individual specimen. Hekkala said that sequencing several individuals of the same species is important so that sequence variation within a group can be observed. More important, if one formalin-fixed specimen is used for sequencing, it would be necessary to compare it against fresh or frozen tissue to ensure that sequence variation is not an artifact of formalin fixation. Schander questioned the likelihood of sequence error attributable to properties of formalin fixation. O'Leary stated that cloning would be more likely to introduce artifacts than would cycle sequencing. He suggested that data collection on the likelihood of sequence errors as a result of formalin fixation could help determine whether 5 replicates per 10,000 or more specimens would be necessary. Error introduction in PCR sequences is not unheard of, said Hall. Bioinformatics could be used to reassemble the short sequences and to identify errors in sequences if many small PCR products with overlapping regions were being assembled. The overlapping regions show where the sequence information differs. However, Hall said he did not have a good sense of the magnitude of error that would be introduced to a PCR sequence as a result of formalin fixation; so it would be difficult to decide whether it is worth investigating. Turning the question around, Schindel asked how many times a formalin-fixed specimen that produces DNA fragments would need to be resequenced to ensure that a correct sequence was assembled. Salzburg said the number of replications would depend on the confidence level sought. Bioinformaticians can quantify the likelihood of an error in a sequence if they are given a particular number of raw sequences. On the basis of that information, they can estimate the number of replicates necessary to achieve a given confidence level.

SUMMING UP

This section reviews the questions in the charge to the workshop participants and their answers to the questions as presented by the rapporteurs in their summary on the second day of the meeting. The questions are listed in boldface type.

What is the state of preservation of DNA in the presence of formalin? Are the DNA chains intact or broken? Does formalin denature DNA or is it the process of extraction that is fragmenting the DNA? Are the nucleotides at each site being preserved or altered?

The quality of DNA in a sample, the percentage of recoverable or amplifiable DNA, the length of the fragments, and whether the DNA is well preserved or nucleated in formalin-fixed samples are largely unknown. The variations in processing of formalin-fixed samples partly contribute to that lack of knowledge. For example, some samples are stored in unbuffered formalin, others are fixed in formalin for different durations and some are transferred to ethanol after fixation. Because of those variations in curatorial treatment, the kinetics of formaldehyde and DNA reactions and the byproducts of different reactions are largely unknown. DNA damages and degradation that can occur in formalin-fixed samples include:

- Cross-linking with formaldehyde.
- Fragmentation.
- Sequence modification.
- Modifications to adenosine, including methylol adduct formation and depurination.
- Formation of oxidative adducts that lead to mutagenic lesions.
- Modification of bases, including adduct formation, if the sample is stored in ethanol.

It is not known whether extraction processes fragment DNA, but they could introduce PCR inhibitors that prevent amplification of the extracted product. Therefore, quantification and characterization of DNA extracted from formalin-fixed samples would reveal both whether DNA can be obtained from those samples and what the quality of the extracted DNA would be. Amplification of an internal control sequence would verify that PCR inhibitors are not the source of the problem.

How can the physical and chemical states of the DNA-formalin cross-linkages be better characterized? What additional information on these cross-linkages is needed?

The condition of the DNA obtained from formalin-fixed tissue can be characterized by NMR spectroscopy and by mass spectrometry. In addition to characterizing the cross-linkages and other damage to DNA, it is important to correlate the type of damage and degradation attributable to different curatorial practices. Detailed knowledge of curatorial history might signal the likely damage or degradation. For

example, if a specimen were kept in unbuffered formalin, its DNA is likely to have become depurinated and useless for sequencing. Additional important information includes data on the kinetic stability of formaldehyde adducts and cross-linkages, whether the stable products are read as mutations by DNA polymerase, and whether they serve to block polymerase altogether. Mass spectroscopy and NMR on small, single-stranded and duplex DNA samples would aid in characterizing the structure and the stability of the formaldehyde reaction products. Additional work also could focus on the reactions that ensue when a sample is exposed to ethanol.

What new chemical and physical methods for DNA extraction should be tested, beyond those that have already been applied to formalin-fixed tissue?

Participants agreed that testing new methods for DNA extraction will not be fruitful if the condition of the DNA in formalin-fixed tissue is largely unknown because a failure to obtain sequence cannot be attributed unamibiguously to a failure of extraction protocol, or the absence of usable DNA in formalin-fixed samples, or the presence of PCR inhibitors. Some participants, including Schander, Bucklin, and Hekkela, reported that published protocols have led to some success in obtaining sequence information from formalin-fixed tissue. Instead of testing new protocols, they said testing existing protocols with a set of standardized samples could provide greater insight. Those samples would include several tissue samples from one organism fixed in formalin for different periods; frozen or fresh tissue would be used as a control. Testing different protocols on samples that had been fixed and preserved with standardized curatorial methods could shed light on which extraction protocol is optimal for each type of sample. Once DNA is successfully extracted from formalin-fixed samples, different methods—mass spectrometry, single-molecule sequencing, or whole-genome amplification—could be used to obtain sequence information. As with the extraction process, the optimal sequencing method might depend on the quality of the extracted DNA.

In what ways and to what extent can fragmented DNA be repaired physically and chemically after extraction from formalin?

In some cases, fragmented DNA can be repaired with excision enzymes mixed with other enzymes and with polymerase. But without more information about the type of damage sustained as a result of formalin fixing, designing the appropriate mix of enzymes for the repair will be difficult.

Can bioinformatics techniques be used to reconstruct the original sequence in silico from the DNA fragments recovered from formalin?

Bioinformatics can be used to construct large, contiguous, consensus sequences from short fragments of DNA. One complication is that formalin fixation could cause sequence modification and a sequence obtained from a formalin-fixed sample might not accurately represent the original sequence of the untreated sample. Whether formalin fixation introduces random or systematic error into a DNA sequence is not known, but it is worth investigating. The potential for error introduction could be studied by repeated sequencing of cloned PCR products and by repeating PCR analysis from replicated DNA preparations. The repeated sequencing and PCR from replicated DNA preparations could reveal whether there are overlapping consensus sequences. Based on the overlapping

sequences, random errors could be corrected accordingly. Systematic errors would be more difficult to correct. They tend to occur consistently in the same place in the sequence, thereby appearing to be a correct base. In that case, the only way to determine whether formalin fixation alters the sample's sequence would be to compare it with a sequence from a fresh sample.

In his summary, Cantor stressed that without knowledge of the quality of the DNA, finding a solution to the problem of obtaining sequence information from formalin-fixed samples is difficult. Rubin agreed and suggested initiating a process to determine which formalin-fixed samples could be used for DNA extraction and how. He listed three elements: First, it would be necessary to screen the specimen before DNA extraction to assess its usability. Screening could be done with mass spectrometry to detect free purines or by testing sample pH. If the DNA damage appeared reversible, a repair could be attempted. The second phase would involve using different protocols to extract DNA from samples that have been subjected to various curatorial treatments. That test would provide a framework for predicting which specimen would be most likely to yield high-quality DNA in each protocol. The last phase would test how well the framework developed in the second phase could predict success in DNA extraction with a whole new set of specimens. To reveal practical limitations, O'Leary said, the process should be iterative and cover a spectrum of samples representing various species curatorial treatments.

THE PATH TO EFFECTIVE RETRIEVAL OF GENOMIC INFORMATION FROM FORMALIN-FIXED SAMPLES

A better understanding of the quality of DNA in samples and of how quality relates to the success of DNA extraction will be needed to inform solutions for effective retrieval of genomic information from formalin-fixed samples. To conclude the workshop, Crothers urged participants to suggest action items for advancing the retrieval of genomic information from formalin-fixed samples. This section compiles the participants' suggestions.

Properly characterize formalin-fixed samples for DNA extraction.

Discussion during the workshop involved the difficulty of deriving effective ways to obtain sequence information from formalin-fixed samples—especially when there is little information about the causes of the problems. To help identify cause-and-effect relationships, participants developed a table with columns of curatorial treatments and rows of problems caused by each (Table 1). The information to be collected would include curatorial history (duration of formalin fixation and whether the formalin was in a buffered solution) and information about the quality of the DNA in the sample (presence of free purines or adducts). The quantity of DNA and its ability to be amplified would be assessed after extraction and PCR would be conducted on highly conserved sequences. The information collected would be used to fill in the table's rows and columns, and those data in turn would serve as a guide for determining whether sequence information could be obtained from a given sample.

Table 1 Curatorial Treatments of Formalin-Fixed Samples and Factors that may Impair DNA Extraction or PCR Amplification.

	Curatorial Treatments					
Factors prohibiting DNA extraction or PCR amplification	Excessive fixation	Excessive heat	Impurities in alcohol	Low alcohol level	Unbuffered formalin	Other treatments
Cross-linking						
Cytosine deamination						
Denaturation						
Depurination						
Formalin-ethanol interaction						
Oxidative damage						
Point sequence modification						
Presence of PCR inhibitors						
Other factors						

The matrix is designed to identify classes of samples in natural history collections that should not be used for DNA and sequence information on the basis of their curatorial history. The table provides some examples of curatorial treatments that could affect the quality of DNA and PCR amplification. Others that could affect the quality samples also should be considered.

A survey could be sent to curators to obtain information on the variety of curatorial treatments in natural history collections. Establishing a network also would be useful to facilitate communication among researchers who want to obtain DNA and sequence information from formalin-fixed samples.

Table 1 presents only the curatorial treatments discussed during the workshop. Other factors could inhibit DNA extraction or PCR amplification, especially given the multiplicity of treatments used to preserve natural history specimens. Participants suggested designing a survey—in consultation with experts at institutions that have natural history collections—to gather information on curatorial history and on any successes or failures in the extraction of DNA or PCR amplification.

Participants also noted that testing of DNA extraction protocols has been done mostly by groups interested in obtaining sequence information from formalin-fixed samples. Although occasional successful attempts are reported in the literature, failed attempts are not reported. Yet comparison of successful and failed attempts could provide clues about determining factors. Thus, the establishment of a Web forum was suggested to allow researchers to pool information on their work with formalin-fixed samples.

In addition to retrospective assessment, controlled experiments on standardized, formalin-fixed samples could be used to examine the mechanisms and kinetics of chemical and physical reactions that could hamper efficient DNA extraction or PCR amplification.

Several experiments could begin immediately to reveal the mechanisms prohibiting efficient DNA extraction:

- Preliminary studies in selected laboratories could be conducted to assess the effect of formalin and alcohol on the integrity of duplex DNA and RNA. The time course of DNA degradation and the efficacy of DNA repair also could be examined.
- The effects of extraction versus fixation could be examined by studying the properties of freshly fixed tissue, oligonucleotides, and DNA samples mixed with protein.
- The standard goldfish samples that had been subjected to different curatorial treatments can be used for testing DNA extraction protocols and for elucidating the effects of different treatments on DNA. In addition, each goldfish sample could be sent to different institutions for independent testing. Each institution could then test its own extraction method by quantifying DNA yield (using an Agilent analyzer, for example) and then using the extracted DNA for PCR amplification. Because the fixation and extraction processes could introduce PCR inhibitors, PCR analysis would be conducted with an internal control to ensure that the reaction is not blocked. Results from the multiple-institution protocol testing would be used to develop a standard protocol for DNA extraction from formalin-fixed samples. Repeating the experiment with an invertebrate standard (for example, a flatworm) could provide useful information.
- Genomic library clones could be established for samples from which DNA had been successfully extracted. The DNA could then be correlated to the fixation process and to sample history and properties.
- Cloning PCR products for selected formalin-fixed samples that have a known gene sequence or an equivalent frozen or fresh sample could reveal whether formalin

fixation induces sequence modification. A comparison of cloned PCR products from fixed tissue with products from fresh or frozen samples—or with the known sequence—would help to quantify mutations. Replicating the experiment on different samples and different species could lead to characterization of patterns and level of sequence modifications attributable to fixation.

A database can be established to collate the information as collected in retrospective and experimental studies. Information in the database could guide the determination of whether particular formalin-fixed specimens could be used for DNA sequencing on the basis of the specimens' chemical and physical properties.

Participants drafted an example of how the data collected from retrospective and experimental studies could be organized (Table 2). Because institutions with natural history collections have so many formalin-fixed specimens, an assessment of curatorial history and of the chemical and physical properties of the specimens would help identify those that are still useful for DNA sequencing and help prioritize sequencing and barcoding efforts.

Table 2 An example how data collected from the retrospective and experimental studies on DNA extraction and sequencing from formalin-fixed biological samples could be organized. Such database could serve as a tool for assessing the feasibility of using certain specimens for DNA sequencing.

Fixation type	Tissue type	Duration of formalin fixation	Quantity of DNA obtained, by protocols					
			Shedlock protocol	Leeds	Qiagen	Chelex	Critical drying point	Other
Formalin-fixed, paraffin embedded								
Formalin-fixed, stored								
Formalin-fixed, stored in ethanol								
Other fixation type								

After potentially useful DNA extraction protocols are identified from the preliminary experiments, random and focused sampling of PCR products could be conducted on a variety of sample types (from different taxa or fixed and preserved with different curatorial treatments) to identify the best protocol for each type. The issues to be addressed involve the recoverability of DNA, including sequences other than polypyrimidine tracts. Samples with euchromatic and heterochromatic DNA also could be considered.

In the long term, consideration of high-throughput processing of formalin-fixed samples for DNA barcoding and genomic studies would be appropriate, given the large number of samples in museum collections. Some individuals in institutions with collections are identifying specimens in their collections or taxa suitable for high-throughput processing, but a systematic and collaborative effort could facilitate and speed up the process.

Ideas and suggestions from the workshop participants could further the efficient extraction of DNA from formalin-fixed samples, and that in turn could improve access to the sequence information of many rare or difficult-to-collect species in natural history collections. Action by the Consortium for the Barcode of Life to follow up on the workshop participants' ideas and suggestions could facilitate the effective recovery of sequence information from formalin-fixed biological samples.

References

Ahn, S.J., J. Costa, and J.R. Emanuel. 1996. PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. *Nucleic Acids Research* 24(13):2623-2625.

Bhadury, P., C. Austen, D.T. Bilton, P.J.D. Lambshead, A.D. Rogers, and G.R. Smerdon. 2005. Combined morphological and molecular analysis of individual nematodes through short-term preservation in formalin. *Molecular Ecology Notes* 5:965-968.

Braslavsky, I., B. Hebert, E. Kartalov, and S.R. Quake. 2003. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences U.S.A.* 100:3960-3964.

Bucklin, A. and L.D. Allen. 2004. MtDNA sequencing form zooplankton after long-term preservation in buffered formalin. *Molecular Phylogenetics* 30:879-882.

CBOL (Consortium for the Barcode of Life). 2006. DNA Barcoding. Accessed on 05/11/06 at http://barcoding.si.edu.

Eisen, E.A. and C.M. Fraser. 2003. Phylogenomics: Intersection of evolution and genomics. *Science* 300:1706-1707.

Fang, S.-G., Q.-H. Wan, and N. Fujihara. 2002. Formalin removal from archival tissue by critical point drying. *Biotechniques* 33:604-611.

Gill, P., A.J. Jeffreys, and D.J. Werrett. 1985. Forensic application of DNA 'fingerprints'. *Nature* 318:577-579.

Jackson, D.P., F.A. Lewis, G.R. Taylor, A.W. Boylston, and P. Quirke. 1990. Tissue extraction of DNA and RNA analysis by the polymerase chain reaction. *Journal of Clinical Pathology* 43:499-504.

King M-C., J.R. Marks, J.B. Mandell, and the New York Breast Cancer Study Group. 2003. Risks of breast and ovarian cancer due to inherited mutations in BRCA1 and BRCA2. *Science* 302:643-636.

Quach, N., M.F. Goodman, and D. Shibata. 2004. In vitro mutation artifacts after formalin fixation and error prone translesion synthesis during PCR. *BMC Clinical Pathology* 4(1):1.

Rubin, M.A., R. Dunn, M. Strawderman, and K.J. Pienta. 2002. Tissue microarray sampling strategy for prostate cancer biomarker analysis. *American Journal of Surgical Pathology* 26(3):312-319.

Schander, C. and K.M. Halanych. 2003. DNA, PCR and formalinized animal tissue—A short review and protocols. *Organisms, Diversity and Evolution* 3(3):195-205.

Shedlock, A.M., M.G. Haygood, T.W. Pietsch, and P. Bentzen. 1997. Enhanced DNA extraction and PCR amplification of mitochondrial genes from formalin-fixed museum specimens. *Biotechniques* 22:394-396.

Wickham, C.L., M. Boyce, M.V. Joyner, P. Sarsfield, B.S. Wilkins, D.B. Jones, and S. Ellard. 2001. Longe PCR products from paraffin-embedded tissue. *Qiagen News* 3:15-17.

Zeman, S. M., D.R. Phillips, and D.M. Crothers. 1998. Characterization of covalent adriamycin DNA adducts. *Proceedings of the National Academy of Sciences* 95:11561-11565.

Appendix A

Glossary

Adduct	A chemical compound formed by the addition of two or more substances.
Adriamycin	An antibiotic.
Amplicon	A piece of DNA synthesized by an amplification technique.
Bioinformatics	The study of genetic or other biological information using computer, mathematics and statistical techniques.
Contig	A group of clones or sequences that represent overlapping regions of a genome.
Euchromatin	Chromosomal region that is genetically active.
Fixation	The use of a chemical agent to prevent autolysis and degradation of tissue by coagulating cell contents into insoluble substances.
Genome	The entire chromosomal genetic material of an organism.
Heterochromatin	Chromosomal region that is condensed during interphase and at the time of nuclear division.
High throughput	The rapid (and simultaneous) processing of large sample sets.

Microarrays	A microscope slide or other solid support on which many distinct complementary DNA or DNA oligonucleotides are patterned at high density in an addressable array. Microarrays are interrogated by hybridization to fluorescently labeled complementary DNA or RNA to identify actively transcribed genes.
Mutagenesis	The development of a mutation.
PCR	Polymerase chain reaction; a fast and inexpensive technique for amplifying a piece of DNA.
Polymerase	An enzyme whose function is associated with polymers of nucleic acids. A DNA polymerase assists in DNA replication; an RNA polymerase assists the making of RNA from a DNA template.
Preservation	Means of protecting specimens from decay (by use of chemical agents, drying, or freezing, for example) and retaining them in collections for long periods.
Sequencing	Determination of the order of nucleotides (base sequences) in a DNA or RNA molecule or the order of amino acids in a protein.

Sources: Science Vol. 291, http://www.biochem.northwestern.edu/holmgren/glossary/

Appendix B

Participant Biographies

STEERING COMMITTEE MEMBERS

Ann C. Bucklin (cochair) is a professor in and head of the Department of Marine Sciences and director of the Marine Sciences and Technology Center at the University of Connecticut. Before joining the University of Connecticut, she was a professor at the University of New Hampshire. Dr. Bucklin was the program manager of the Oceanic Biology Program at the Office of Naval Research (1988-1991). She was a Fulbright Senior Scholar in Norway (1992-1993), and she was elected a fellow of the American Association for the Advancement for Science in 1995. She served as director of the New Hampshire Sea Grant Program from 1993 to 2005. She is currently the USA Academic Delegate to the International Council for the Exploration of the Sea. Dr. Bucklin holds a B.A. in biology from Oberlin College and a Ph.D. in zoology from the University of California, Berkeley. Her research interest in spatial and temporal patterns of molecular genetic variation in marine organisms developed from her early interest in sea anemones and continues in her current work on planktonic crustaceans. She leads the Census of Marine Zooplankton, a Census of Marine Life ocean realm field project began in 2004, and she is a member of the International Committee for the Ocean Biogeographical Information System.

Donald M. Crothers (cochair) is Sterling Professor Emeritus of Chemistry, professor emeritus of molecular biophysics and biochemistry, and senior research scientist in chemistry at Yale University. Dr. Crothers received his doctorate in chemistry from the University of California, San Diego, and completed a National Science Foundation postdoctoral fellowship at the Max Planck Institute in Göttingen, Germany. His research focuses on the structure, dynamics, and protein-binding properties of nucleic acids. Current projects include characterization of the sequence dependence of curvature and flexibility, and the role of those variables in determining the affinity for core histones and regulatory proteins. Another focus is on the dynamics of protein-DNA complexes, using nuclear magnetic resonance spectroscopy and stopped-flow methods. His research interests include the structure, properties, and function of nucleic acids, protein-DNA interactions, mechanisms of control of gene expression, and the

effects of antitumor compounds on DNA, the dependence on the activation energy for DNA bending by proteins on DNA sequence, and the physical chemistry of biological polymers, particularly nucleic acids. Dr. Crothers has done theoretical and experimental work on the physical properties of nucleic acids and on their interactions with other molecules. Dr. Crothers has received numerous awards, including the Alfred P. Sloan and Guggenheim fellowships and the Alexander von Humboldt Senior Scientist Prize. He was elected into the American Academy of Arts and Sciences in 1986, inducted into the National Academy of Sciences in 1987, and named a fellow of the American Association for the Advancement of Science in 1992. He also has served on the editorial boards of numerous journals.

Timothy O'Leary is director of the Biomedical Laboratory Research and Development (BLR&D) and the Clinical Science Research and Development Service (CSR&D) of the U.S. Department of Veterans Affairs. He holds a doctorate in physical chemistry from Stanford University and a medical degree from the University of Michigan. As BLR&D director, he oversees all research basic biological or physiological research in humans or in animals that involves animal models or tissues, blood, or other specimens from humans. As CSR&D director, he supervises research on intact human beings as the unit of examination. He is certified in anatomic pathology and in molecular genetic pathology by the American Board of Pathology and the American Board of Medical Genetics. For more than 15 years before joining the Department of Veterans Affairs, he chaired the Department of Cellular Pathology at the Armed Forces Institute of pathology. He also is a reserve member of the Public Health Service Commissioned Corps. His research involves molecular changes in gastric tumors, ultrasensitive detection of biological toxins, and mechanisms of formaldehyde fixation. He is credited with expanding the capacity of the Armed Forces Pathology Institute to include molecular genetics and tissue magnetic resonance microscopy. Dr. O'Leary has published more than 130 scientific papers since 1973 and written numerous book chapters. He edited the 2002 text *Advanced Diagnostic Methods in Pathology: Principles, Practice and Protocols*. He has received many awards for his work, including the 2004 Armed Forces Institute of Pathology Medallion.

Christoffer Schander is professor of marine biodiversity at the University of Bergen, Norway. He received B.Sc., Ph.D., and docent degrees from Göteborg University, Sweden. His research on evolutionary forces and phylogeny in creating organism diversity uses phylogenetic analyses that integrate morphological and molecular data. His research focuses on molluscs, specifically the ectoparasitic pyramidellid gastropods and the shell-less, primary deep-sea aplacophorans.

Alison Williams is a research staff member and lecturer in chemistry in the graduate program in molecular biophysics at Princeton University. She holds a B.S. from Wesleyan University and she received her M.S. and Ph.D. from the University of Rochester. Dr. Williams has been a member of the Princeton Chemistry Department since 2003. She joined the department after two years as director of studies at Princeton's Wilson College. She leads a research team investigating the properties of nucleic acids in terms of their local chemical structure and environment. Dr. Williams is a

long-time member of the National Organization of Black Chemists and Chemical Engineers, and she has spoken extensively on enhancing the role of women and minorities in science.

INVITED PARTICIPANTS

Charles Cantor is a founder, chief scientific officer, and member of the board of directors at Sequenom, Inc. He also is founder of SelectX Pharmaceuticals, a drug-discovery company based in the Boston area, and is codirector of the Center for Advanced Biotechnology at Boston University, where he also is professor of biomedical engineering. Dr. Cantor has held positions at Columbia University and the University of California, Berkeley, and he was director of the Human Genome Center of the U.S. Department of Energy at Lawrence Berkeley National Laboratory. He has published more than 400 peer-reviewed articles and is coauthor of a three-volume textbook on biophysical chemistry and the first textbook on genomics, *Science and Technology of the Human Genome Project*. He holds more than 60 patents. He sits on the advisory boards of more than 20 national and international organizations and is a member of the National Academy of Sciences.

Marvin Caruthers is a distinguished professor at the University of Colorado, Boulder. He received a Ph.D. from Northwestern University. His research interests include nucleic acid chemistry and biochemistry, and his laboratory uses nucleic acid chemistry, biochemistry, and molecular biology to study regulation and control of gene expression. Dr. Caruthers has made major contributions to the chemical synthesis of nucleic acids, culminating in development of automated DNA synthesis, and thus the availability of oligonucleotides of any desired sequence. As a result of synthesizing DNA with modified nucleotides at defined sites, he was able to measure the contribution of individual functional groups to DNA-protein interactions. Dr. Caruthers was named Guggenheim Fellow in 1981, and he was elected to the National Academy of Sciences in 1994.

Robert DeSalle is the curator in charge of the Ambrose Monell Cryo-Collection, a curator of entomology, and codirector of molecular laboratories at the American Museum of Natural History. His fields of specialization include molecular evolution, population genetics, molecular systematics, and developmental biology. His early research focused on the molecular systematics of the Drosophilidae, a family of flies. His more recent work involves on gene family evolution and comparative genomics in a variety of organisms, including pathogenic bacteria. Dr. DeSalle is among the founders of the museum's Conservation Genetics Program, which applies studies at the molecular level to the conservation of wildlife and wild lands throughout the world. In 1996, Dr. DeSalle and his colleagues developed a genetic test for caviar that helped gain protection for sturgeon in the Caspian Sea basin under the Convention on the International Trade in Endangered Species of Wild Fauna and Flora. Dr. DeSalle holds a B.A. in biological sciences from the University of Chicago and a Ph.D. from Washington University. He joined the museum in 1991. He is an adjunct professor at Columbia University, New

York University, the University of Connecticut, and Yale University. Dr. DeSalle is the coauthore of *The Science of Jurassic Park and the Lost World*, and he was curator of the museum's 1999 landmark exhibition "Epidemic! The World of Infectious Disease".

Miral Dizdaroglu is the leader of the DNA Measurements Group in the Biochemical Science Division of the National Institute of Standards and Technology. He is also an adjunct professor at the University of Maryland, Baltimore County. His research concerns oxidative stress and DNA damage and repair. Oxidative stress is produced in cells by oxygen-derived species resulting from cellular metabolism and from interaction with cells of exogenous sources, such as carcinogenic compounds, redox-cycling drugs, and ionizing radiations. DNA damage caused by oxygen-derived species, including free radicals, is the most frequent type encountered by aerobic cells. Oxidative DNA damage can produce DNA modifications that include base and sugar lesions, strand breaks, DNA-protein cross-links, and base-free sites. Accurate measurement of those modifications is essential for explaining the mechanisms and biological effects of oxidative DNA damage. Dr. Dizdaroglu received a diploma in chemical engineering from the University of Ankara, Turkey, and a Ph.D. in physical chemistry from the University of Karlsruhe, West Germany.

Catherine Fenselau is a professor at the University of Maryland, College Park. She conducts biomolecular studies using mass spectrometry, rapid characterization of microorganisms by mass spectrometry, and new methods for proteomics. She is a past president of the American Society for Mass Spectrometry and the recipient of numerous awards, including the Garvan Medal of the American Chemical Soceity and of Spectroscopy Society of Pittsburgh Award. Dr. Fenselau holds an A.B. from Bryn Mawr College and a Ph.D. from Stanford University.

Neil Hall is an assistant investigator at the Institute for Genomic Research. He has a B.S. in genetics and a Ph.D. in molecular biology from the University of Liverpool. Before joining the institute, he was a project manager in bioinformatics at the Sanger Institute. His major research is on the genomics of infectious diseases, using bioinformatics analysis to elucidate the evolution of pathogenic organisms, how they respond to selective forces acting upon them, and how the ability to infect humans originates. He recently has become interested in using genomic data to study pathogen populations and he has begun projects to develop databases and methods that will assist in that research.

Timothy Harris is the director of sequencing technology at Helicos BioSciences Corporation. Until 2004, Dr. Harris was director of research at Prealux, a wholly owned subsidiary of Amersham Biosciences, where his work involved single-molecule DNA sequencing technology. He originated the design and led the development for the Amersham IN Cell Analyzer. From 1978 to 1996, Dr. Harris worked for Bell Laboratories where his work led to the first report of single-molecule imaging, spectroscopy, and lifetimes. His work resulted in more than 80 publications, 100 invited talks, the IR Award in 1982 for the development of an intracavity laser spectrometer, and the 1992 Williams-Wright Award for his contributions to vibrational spectroscopy. Dr.

Harris holds a B.S. in chemistry from California Polytechnic State University and a Ph.D. in analytical chemistry from Purdue University. Dr. Harris holds two patents.

Matthias Hofer is a postdoctoral fellow at Harvard Medical School. He received his M.D. from the University of Ulm Medical School in Ulm, Germany. His residency was in the Department of Urology at the University Hospital of Ulm. He joined Mark Rubin's laboratory at Brigham and Women's Hospital in Boston as postdoctoral research fellow. His research is on molecular profiles of prostate cancer and on the metastasis-associated gene 1. He was a 2003 recipient of a U.S. Department of Defense postdoctoral fellowship.

Steven Hofstadler is vice president of research for the Ibis Division of Isis Pharmaceuticals. He heads the mass spectrometry program and coordinates the company's basic research activities. He also is responsible for primary high-throughput screening activities and for the development of advanced mass spectrometric instrumentation and methods to support the company's diagnostics and proteomics research. He has worked on several initiatives of the Defense Advanced Research Projects Agency and with a continuing program fro the U.S. Centers for Disease Control and Prevention. Dr. Hofstadler has a Ph.D. in analytical chemistry from the University of Texas, Austin. He is the author of more than 100 scientific publications and holds several U.S. patents. He received R&D Magazine's R&D 100 Award in 2000 for "Multitarget Affinity/Specificity Screening" and was the recipient of the Society for Biomolecular Screening's 2004 Perkin Elmer Life Sciences Award for innovations in high throughput screening. Dr. Hofstadler serves on the scientific committee of the Association for Laboratory Automation, the editorial board for the American Chemical Society journal *Analytical Chemistry*, and the National Institutes of Health Special Emphasis Review Panel.

Charles Lydeard is a program director in the Directorate for Biological Sciences at the National Science Foundation and a professor in the Department of Biological Sciences at the University of Alabama. His research is in systematics, conservation, and evolution of invertebrates. He, his students, and his collaborators have been focusing attention on freshwater mollusks, particularly unionid bivalves and pleurocerid gastropods and more recently terrestrial gastropods. Dr. Lydeard has a Ph.D. from Auburn University.

Juan Carlos Morales is a program director in the Directorate for Biological Sciences at the National Science Foundation and an associate research scientist in the Center for Environmental Research and Conservation at Columbia University. His research is in systematics, biogeography, and conservation genetics of mammals. He is particularly interested in the use of molecular markers to determine evolutionary relationships, distributional patterns, areas of genetic endemism, and population structure in several mammalian groups.

Mark Rubin is an associate professor of pathology and director of genitourinary pathology at Brigham and Women's Hospital, Harvard Medical School. He has been

closely involved in the National Institutes of Health's development of protocols for recovering high-quality DNA from formalin-fixed material.

Daniel Ryan is a senior scientist in the Genomics/Location Analysis R&D group at Agilent Technologies, Inc. His research is on improved methods for recovery of genomic DNA from formalin-fixed samples, including formalin-fixed cells for ChIP-on-chip analysis (a type of "location analysis") and formalin-fixed paraffin-embedded samples. Before joining Agilent in 2005, Dr. Ryan was a senior investigator in the Biochemistry Department at the University of California, San Francisco where he worked on integrating data from published studies of yeast and mammalian spliceosomes into a 3-dimensional molecular model of a spliceosome at the first step of chemistry. From 1995 to 2002, Dr. Ryan was a Division of Biology fellow and senior scientist at the California Institute of Technology, working to elucidate the structure-function relationships in putative catalytic RNA in spliceosomes. As a postdoctoral fellow at Columbia University, his research involved studies of the natural cross-linking products and oxidation-reduction chemistries of quinone/hydroquinone-functionalized peptides in marine animals called tunicates, or "sea squirts." He has a Ph.D. in organic chemistry from Princeton University.

Steven Salzberg is director of the Center for Bioinformatics and Computational Biology and Horvitz Professor in the Department of Computer Science at the University of Maryland, College Park. He holds a Ph.D. from Harvard University, and he is a fellow of the American Association for the Advancement of Science. His research is in bioinformatics, genomics, gene finding, and genome assembly.

REPRESENTATIVES OF SPONSORING AGENCIES

Tom Evans is a senior scientist at New England Biolabs, Inc.

Gonzalo Giribet is a curator of invertebrate zoology at the Museum of Comparative Zoology and a professor in the Department of Organismic and Evolutionary Biology at Harvard University

Evon Hekkala is a postdoctoral scientist at the U.S. Environmental Protection Agency's Environmental Monitoring and Assessment Program.

Scott Miller is a senior program officer in the Office of the Under Secretary for Science at the Smithsonian Institution and a research entomologist at the National Museum of Natural History.

Ernest Mueller is the R&D principal investigator in molecular biology at Sigma-Aldrich.

Benjamin Rosenthal is a molecular systematist in the U.S. Department of Agriculture's Agricultural Research Service.

David Schindel is executive secretary of the Consortium for the Barcode of Life.

NRC Staff
Fran Sharples, Director, Board on Life Sciences
Evonne P.Y. Tang, Senior Program Officer
Tova Jacobovits, Senior Program Assistant

Appendix C

Agenda

Keck Center
National Academies
500 Fifth Street, NW
Washington, D.C.

May 8, 2006

8:30-9:00 a.m.	**Opening Remarks and Introduction** Ann Bucklin and Donald Crothers, Cochairs
8:30-8:45	**Purpose of the Workshop** Scott Miller

Session 1: State of DNA in biological samples after exposure to formalin

- What is the state of preservation of DNA in the presence of formalin? Are the DNA chains intact or broken? Does formalin denature DNA or is it the process of extraction that is fragmenting the DNA? Are the nucleotides at each site being preserved or altered?
- How can the physical and chemical states of the DNA-formalin cross-linkages be better characterized? What additional information on these cross-linkages is needed?

Moderator: Donald Crothers
Rapporteur: Timothy O'Leary

9:00-10:00	Participant Briefings Participants share their knowledge of and experience with assessing the condition of DNA in biological samples fixed in formalin
10:00-10:30	Open Discussion

10:30-10:45	Break
10:45-11:30	Open Discussion

Session 2: Optimization of DNA Sequence Information from the Samples

- What new chemical and physical methods for DNA extraction should be tested, beyond those that have already been applied to formalin-fixed tissue?
- In what ways and to what extent can fragmented DNA be repaired physically and chemically after extraction from formalin?

Moderator: Ann Bucklin
Rapporteur: Charles Cantor

11:30-12:30 p.m.	Participant Briefings Participants share their knowledge of and experience with optimization of DNA sequence information from formalin-fixed samples
12:30-1:30	Lunch
1:30-3:45	Open Discussion
3:45-4:00	Break

Session 3: Bioinformatics For Reconstructing DNA Sequences

- Can bioinformatics techniques be used to reconstruct the original sequence in silico from the DNA fragments recovered from formalin?

Moderator: Christoffer Schander
Rapporteur: Neil Hall

4:00-4:20	Participant Briefings Participants share their knowledge of and experience with bioinformatics for restructuring DNA sequences.
4:20-5:20	Open Discussion
5:20-5:30	First-day Summation

May 9, 2006

Session 4: The path towards effective retrieval of genomic information from formalin-fixed samples

Moderators: Donald Crothers and Ann Bucklin

8:30-9:15	Summary of Sessions 1-3 Timothy O'Leary, Charles Cantor, and Neil Hall, Rapporteurs
9:15-10:30	Open Discussion
10:30-10:45	Break
10:45-11:45	Open Discussion (cont'd)
11:45-12:00	Workshop Summation