DNA Barcodes and Watermarks

BEST AVAILABLE COPY

20040809 028

MITRE

DISTRIBUTION STATEMENT A:
Approved for Public Release -
Distribution Unlimited

DNA Barcodes and Watermarks

Study Leader:
Steven M. Block

Study Participants:
David Donoho
Terry Hwa
Gerald Joyce
David Nelson
Tim Stearns
Peter Weinberger
Ellen Williams

June 2004

JSR-03-305

Approved for public release; distribution unlimited

JASON
The MITRE Corporation
7515 Colshire Drive
McLean, Virginia 22102-7508
(703) 883-6997

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information estimated to average 1 hour per response, including the time for review instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget. Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 2004	

4. TITLE AND SUBTITLE

DNA Barcodes and Watermarks

5. FUNDING NUMBERS

13049022-DC

6. AUTHOR(S)

Steven Block, et al.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The MITRE Corporation
JASON Program Office
7515 Colshire Drive
McLean, Virginia 22102

8. PERFORMING ORGANIZATION REPORT NUMBER

JSR-03-305

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Department of Energy
Office of Science
Washington, DC 20585

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

JSR-03-305

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

Distribution Statement A

13. ABSTRACT *(Maximum 200 words)*

This study was conducted on behalf of the DOE during the summer of 2003. The JASON Study explored the feasibility of a program to tag genetically the microorganisms used for bioremediation, for the purpose of identification.

14. SUBJECT TERMS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

Contents

1 EXECUTIVE SUMMARY 1
The Bioremediation Problem 1
The Barcode Concept 1
Barcode and Watermark Implementation 1
Recommendations 2

2 JASON STUDY (JSR-03-305) DNA BARCODES & WATERMARKS 5
Briefers We Heard 5
Introduction: Statement of the Problem 7
Multiple Types of Bioremediation 8
The DNA Barcode Concept 11
What a DNA Barcode is *NOT* 12
What a DNA Barcode *IS* 13
What is a DNA Watermark? 14
The Barcode/Watermark Tagging System 15
Introduction to the Bacterial Chromosome 15
Where to Place Barcodes & Watermarks 16
Anatomy of a Barcode 18
A Practical Barcode is Shown 20
How Many Practical Barcodes Exist? 20
Constructing a DNA Watermark 22
Introducing Barcodes & Watermarks 23
Placing Barcodes & Watermarks by Recommendation 24
Placing Barcodes by Group II intron "Retro-homing" 25
Reading Barcodes & Watermarks 27
A Worked Example:
Barcoding & Watermarking the Model Organism, *D. radiodurans* 29
Selecting the Barcode Insertion Site 30
Choosing the Watermark Sites 31
An Alternative Barcode Scheme 32
A JASON Idea: Fast, Adjustable Molecular Clocks 33
A Clock Based on Tandem Repeats 34
A Better Clock Based on a Directed, Error-Prone Polymerase 35
Barcode and Watermark Considerations 36
Barcodes and Bioremediation 37
JASON Recommendations 409

3 APPENDICES
Appendix A: Construction and Analysis of DNA Barcode Libraries 43
Appendix B: Placing Barcodes and Watermarks by
Recombination within the Genome 50

1. EXECUTIVE SUMMARY

The Bioremediation Problem

It is estimated that 60% of DOE facilities contain groundwater contaminated by heavy metals, radionuclides, or chlorinated hydrocarbons. Moreover, 50% of the topsoil or sediment at DOE facilities is now contaminated. The DOE faces a massive remediation problem, with a need to treat 1.7 trillion gallons of polluted water and 40 million cubic meters of contaminated soil. To assist in this considerable effort, a "genomic approach to waste cleanup" is currently being explored, harnessing the power of microbial biochemistry (in communities of living bacteria and fungi) to degrade complex organic molecules, and to reduce or sequester certain chemicals, particularly heavy metals. *Bioremediation* can take on several forms, ranging from natural degradation ('intrinsic bioremediation'), to encouraging the growth of endogenous organisms *in situ* ('enhanced bioremediation'), to the introduction of non-native microbial species ('bioaugmentation'), to the application of sophisticated bioengineering to generate novel strains optimized for the specific remediation task at hand. *Irrespective of the origin of the microbes used, however, it will be vital to establish a socially and legally acceptable means of following the growth and ecology of all species used for bioremediation.*

The Barcode Concept

This JASON Study explored the feasibility of a program to tag genetically the microorganisms used for bioremediation, for the purpose of identification. Such DNA-based tags would be fully heritable, but carefully designed to convey no phenotype to the organisms being labeled ("genotype without phenotype"). Tags would be structured so that they could be specifically amplified by PCR (using a universal set of primers) and rapidly read in the field or in the laboratory. The system we contemplate would support the sensitive, multiplexed readout of mixed populations of strains, such as those typically found in complex microbial communities. Moreover, DNA tags would be designed to be robust against most mutations and certain kinds of intentional interference. Properly implemented, these tags would constitute an effective 'barcode' system for tracking microorganisms in the wild. DNA barcode tracking could be used quantitatively to monitor microbial growth, dispersal, transport, blooms, die-outs, ecological niches, and more. All DNA barcode labels would be registered in a public database, and they would be designed and introduced following uniform standards established by the DOE. Barcode integrity could be further protected against tampering by a system of DNA 'watermarks'— a covert set of minor, distributed genomic changes — whose implementation details would remain proprietary.

Barcode and Watermark Implementation

A useful barcode system could be developed quite economically, adapting many of the methods currently available in genetics and molecular biology (such as PCR and site-specific recombination), along with straightforward adaptations of existing or planned instrumentation

from the biotechnology sector (such as hybridization arrays). We envision that such a system would be composed of several distinct elements, as follows:

1) A public database designed for recording and registering barcodes and all associated information (organism, strain and variant, full genotype, release date, etc.)
2) A private database designed for recording and registering watermarks against their associated barcodes
3) A design strategy to generate thousands of unique, robust DNA sequences appropriate for use as barcodes and watermarks in microbial organisms, and the 'universal' PCR primers intended for these
4) Bioinformatic methods to identify suitable and appropriate target locations in microbial genomes that could receive DNA barcodes and watermarks
5) A practical means of inserting both barcodes and watermarks into microbial genomes using site-specific recombination approaches
6) Sensitive, efficient, multiplexed ways to amplify and read out both barcodes and watermarks in a (potentially) mixed population of microbial organisms

Items (3), (4), (5) and (6), in particular, warrant additional consideration, and are covered in detail in the full report. There, we present some practical approaches for realizing each of these elements using existing biotechnologies, and discuss the use of various alternative methods. We also supply *a worked example* of how to barcode and watermark the genome of a microorganism that has been fully sequenced, *Deinococcus radiodurans* (a bacterial strain of interest to the DOE because it can tolerate high levels of radiation), which has recently been engineered to reduce metabolize mercury, as well as to digest toxic compounds such as toluene and chlorobenzene. Finally, we suggest ways that barcodes and watermarks might be enhanced, in principle, in second-generation designs, including the incorporation of a fast mutational 'clock' into a special portion of the bacterial genome that would allow the number of generations since the environmental release of tagged organism to be estimated.

Recommendations

JASON recognizes that introducing any heritable DNA label into a microbe — *even one that does not impart a phenotype and has no impact on its natural fitness* — technically constitutes a 'genetic modification.' Despite the popular stigma now being attached to genetically-modified organisms (GMOs), particularly those intended for food, we believe the numerous advantages conferred by barcode labels in monitoring the spread of microbes outweigh such concerns, *provided* that barcodes perform benignly, as designed. However, the deployment would have to be preceded by an educational outreach program backed up by valid scientific data from actual tests, before the general public will likely accept barcoding as a means of tracking microorganisms in the environment. That said, significant levels of support for a barcoding program may develop from members of the scientific community, especially biologists whose voices are now strong in support of environmental concerns, assuming that the safety and efficacy of microbial barcodes can be demonstrated in pilot studies.

We conclude that a program for barcoding the microorganisms used in bioremediation is not only feasible, but advisable. The Report offers a set of specific recommendations for a

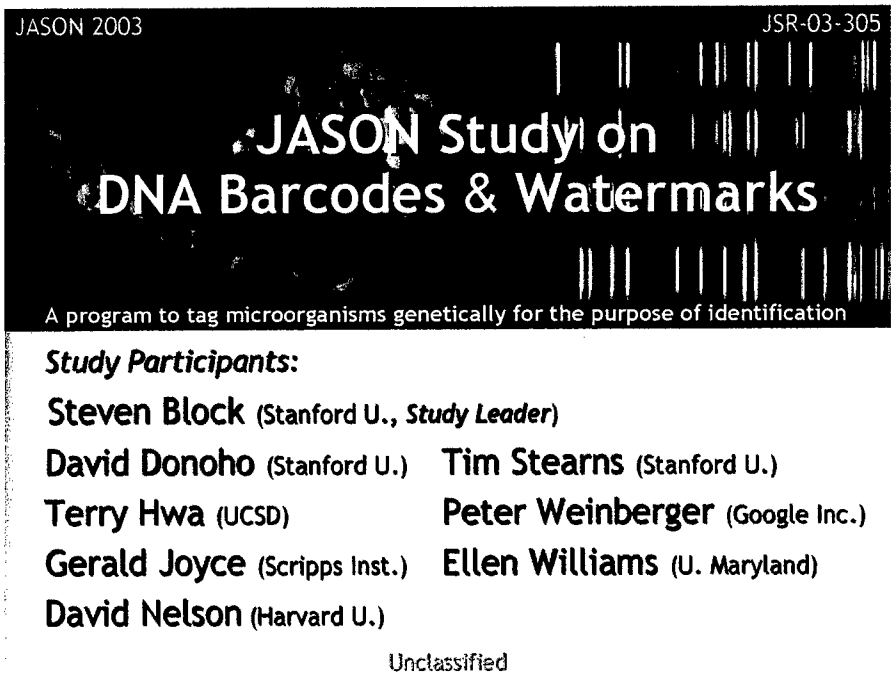
comprehensive program aimed at developing and testing barcodes and watermarks. In brief, we urge the DOE to consider the following:

- Look into ways of adapting existing biotechnological instrumentation for use in key aspects of barcode synthesis, insertion, and readout. Develop specific hybridization arrays (DNA chips) for conventional barcode readout and for barcode readout accompanied by watermark readout.
- Sponsor applied research into adapting site-specific recombination methods to insert barcodes and watermarks into bacterial genomes, including approaches based on (1) homologous recombination, (2) 'retrohoming' insertion by Group II intron vectors, and (3) excision-deficient bacteriophages. Sponsor basic research aimed at identifying alternative and improved methods.
- Establish a working group to formulate uniform standards for microbial barcoding, which would standardize the database structures, as well as specify rules for the coding, design and disbursement of the barcodes and watermarks used.
- Perform feasibility studies to measure the actual stability of barcodes and watermarks introduced by any of the various methods, their effect (if any) on natural fitness, their mutation and loss rates, etc.
- Develop a barcoding and watermarking 'testbed' program using one or a few model organisms, which would be barcoded, watermarked, grown up, and monitored in a trial release program (performance monitoring in a microcosm study).
- If the above developments prove successful, implement DNA barcoding and watermarking standards for performance monitoring of all bacterial and fungal strains used in DOE programs, and promote the global use of barcoded organisms in bioremediation.

We feel that this is a leadership opportunity for the DOE. Success in this arena may encourage the exploration of barcodes for other purposes, such as tagging the microorganisms in the American Type Culture Collection (ATCC), tagging the bioengineered organisms proposed for commercial bioremediation purposes (e.g. digesting oil slicks), and tagging bacterial and fungal pathogens on the CDC List of Select Agents.

2. JASON Study JSR-03-305 DNA Barcodes & Watermarks

This study was conducted on behalf of the DOE during the summer of 2003. The names of and institutional affiliations of the eight JASON Members participating in the study are shown. Their fields of academic expertise include biology, medicine, physics, mathematics, statistics, and computer science.



Briefers We Heard:

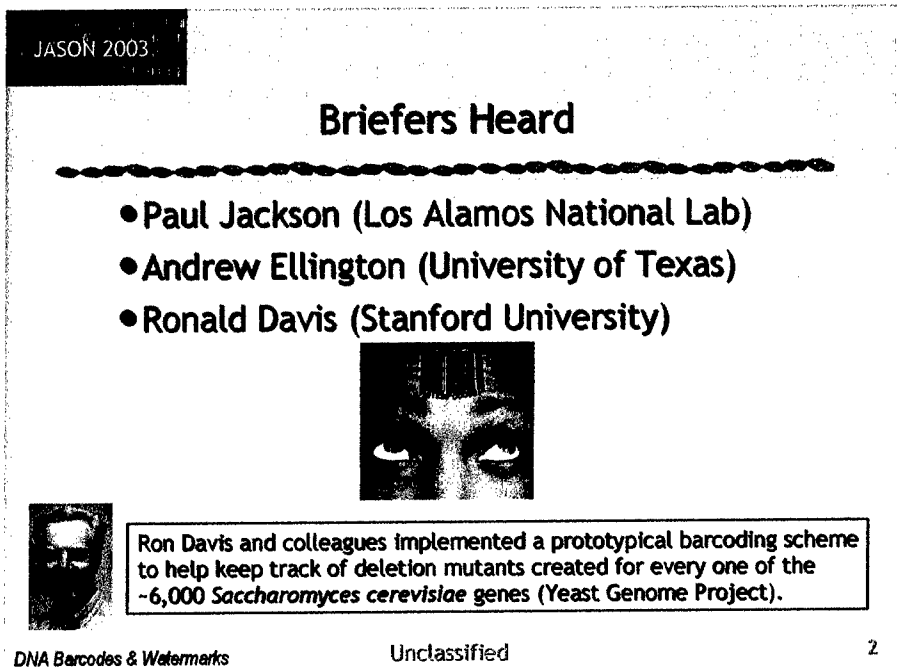
Three briefers were invited to discuss their work relevant to this study. Each briefer made a unique and valuable contribution to this study, and JASON is indebted to them for their input.

Dr. Paul Jackson is a Laboratory Fellow and Technical Staff member of the Biosciences Division at Los Alamos National Laboratory. His work is on developing a basic understanding of biological threat pathogens, and he has been responsible for developing "the most mature methods and associated reagents developed in the past several years for strain and species identification...provided to the CDC and FBI and...currently in use."[1] Dr. Jackson has helped to pioneer the use of intrinsic genetic information, including natural sequence variations, tandem DNA repeats, and single nucleotide polymorphisms, in tracking and identifying pathogen species.

[1] http://t8web.lanl.gov/people/rajan/CT2002/BIO/jackson.html

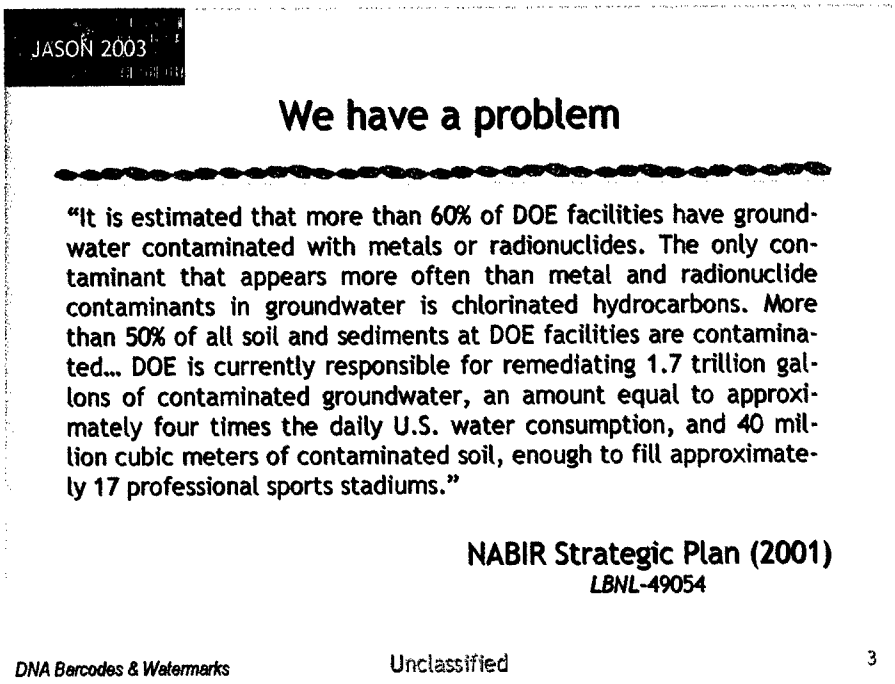
Dr. Andrew Ellington is the Wilson & Kathryn Fraser Research Professor in Biochemistry at the University of Texas, Austin. His lab works on the evolutionary engineering of molecules, metabolism, and organisms, including the design and selection of catalytically-active RNA molecules. Dr. Ellington has developed catalytic RNAs derived from the Group II Intron of *Lactococcus lactis* to target the introduction of sequences at specific, pre-defined sites in bacterial DNA. This technology holds particular promise as a way of introducing genetic sequence tags into bacterial genomes.

Dr. Ronald Davis is a Professor of Biochemistry and Genetics and the Director of the Stanford Genome Technology Center. His group is using yeast (*Saccharomyces cerevisiae*) to conduct whole genome analysis projects. Dr. Davis' group has made a nearly complete set of haploid and diploid strains (21,000 in all), each containing a deletion of one of the ~6,000 yeast genes. To facilitate whole genome analysis, each deletion is molecularly tagged with a unique 20-mer DNA sequence. This sequence acts as a molecular barcode and makes it easy to identify the presence of each deletion. Dr. Davis' pioneering work provides an existence proof that genetic barcodes can be made to work, and these have already shown their considerable utility in research context.



INTRODUCTION

Statement of the Problem

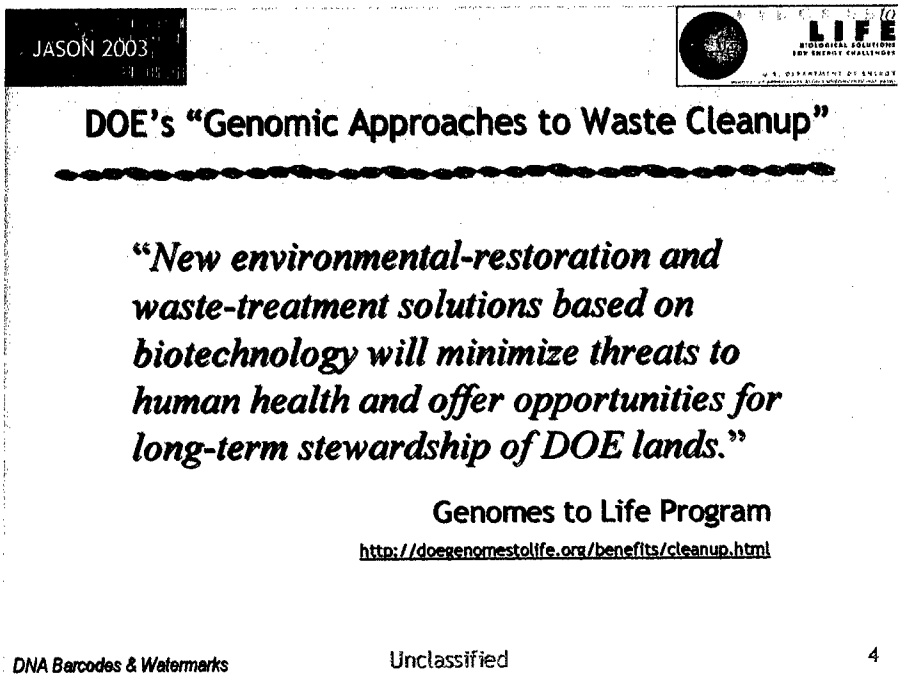


The Department of Energy "has a 50-year legacy of environmental problems resulting from the production of nuclear weapons"[2] and faces a massive cleanup problem on the lands contained within its facilities. Primary contaminants are found in both the soil and groundwater at most DOE sites, and include halogenated hydrocarbons, acids, chelating agents, radionuclides, and heavy metals. It is estimated that 60% of DOE facilities contain contaminated groundwater, and that 50% of the topsoil or sediment is presently contaminated. *The scope of the problem is truly immense*, with a need to treat 1.7 trillion gallons of polluted water and 40 million cubic meters of contaminated soil. According to NABIR (Natural and Accelerated Bioremediation Research Program), a DOE-funded interdisciplinary effort to explore the scientific basis for strategies to reduce the risk of contaminants to humans and to the environment,

> "[M]any of the contaminated soils, sediments, and groundwater are believed to be impossible to remediate with existing technology. Examples of such intractable problems include the Snake River Aquifer in Idaho, contaminated groundwater at the '100' and '200' areas at Hanford, Washington, contaminated sediments in the Columbia River, and groundwater at the Nevada Test Site (DOE, 1995). The huge cost, long duration, and technical challenges associated with remediating DOE facilities present a significant opportunity for science to contribute cost-effective solutions."

[2] Natural and Accelerated Bioremediation Research (NABIR) Program Mission statement, available at http://www.er.doe.gov/production/ober/nabir/mission.html.

To assist in this considerable effort, a "genomic approach to waste cleanup" is currently being explored, harnessing the considerable potential of microbial biochemistry—in communities of living bacteria or fungi—to degrade complex organic molecules, and to reduce or sequester chemicals, particularly heavy metals, thereby rendering the cleanup process more tractable. This, in essence, defines the process of *bioremediation*. 'Bioremediation' can take on many definitions, so it is worthwhile examining some of these at the outset.



Multiple Types of Bioremediation

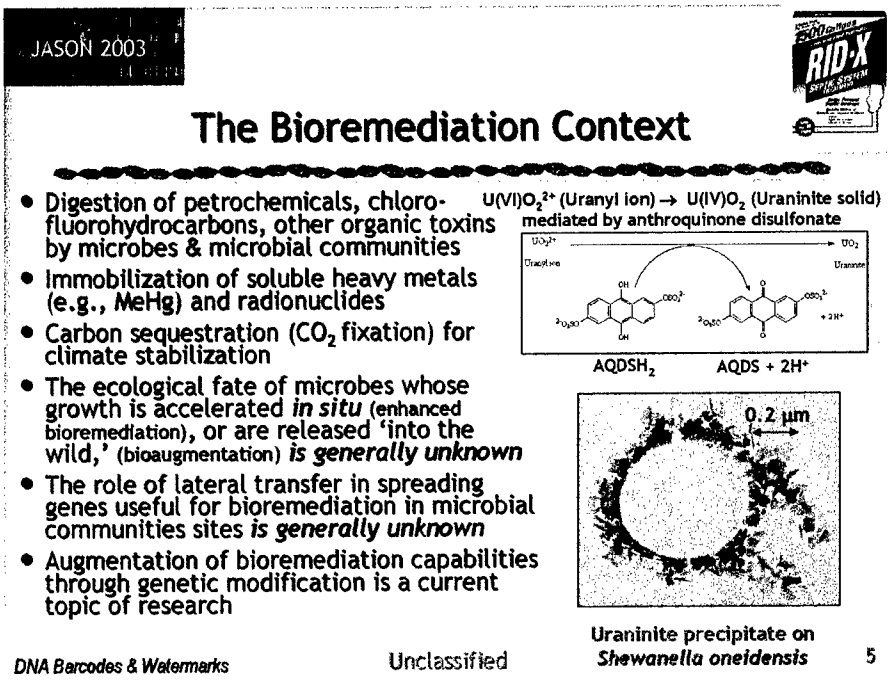
In its simplest form, known as '*intrinsic bioremediation*,' populations of microorganisms naturally present in the water or soil adjust adaptively to accommodate any new stress in the chemical environment. While some indigenous microbial species experience toxic effects and die off, others may be able to better cope with the changing chemistry, and possibly even derive metabolic benefit from it. Those species that—individually or in combination—successfully metabolize the pollutant will prosper. Over time, the resistant species will tend to occupy a larger niche in the environment, until such time as their food sources (i.e., the pollutants or related compounds) give out. In essence, then, *intrinsic bioremediation* is just an alternative phrase describing the process of natural biodegradation.

A straightforward way to accelerate biodegradation is to foster the growth of specific indigenous organisms: namely, those which are the most effective in metabolizing the pollutants of concern. This can be done, in principle, by supplying certain limiting nutrients (or gases, such as oxygen, or surfactants) that encourage the growth of the desired organisms, either directly or indirectly. Conversely, it is also possible to eliminate certain nutrients essential to any competing organisms that limit the growth of the desired species. In either case, altering the

environmental abundance of chemicals and nutrients for the purpose of promoting specific types of growth and degradation is known as '*enhanced bioremediation.*'

Finally, it is also possible to introduce non-native microbial species into the environment to carry out the desired degradation process more efficiently. Ideally, these exogenous species would have improved biochemical characteristics relative to those of native species (and possibly other attributes as well). This approach is referred to as '*bioaugmentation.*' Bioaugmentation is by no means a new or untried concept. For example, augmentation is routinely performed by homeowners who use products such as Rid-X™ to improve the performance of their septic tanks and cesspools. Rid-X™ has been a successful commercial product for over 50 years, and advertises itself as being "100% Natural." Its active ingredients consist largely of dried bacteria, yeasts, plus some nutrients that help to colonize septic tanks and promote fermentation. Products like Rid-X™ alter the natural balance of flora and fauna in the immediate environment of the tank, but are nevertheless widely considered to be far more 'friendly' to the environment than harsh chemicals, like acids, bases and detergents, which may unclog pipes or tanks but frequently cause further damage (and waste treatment problems) downstream.

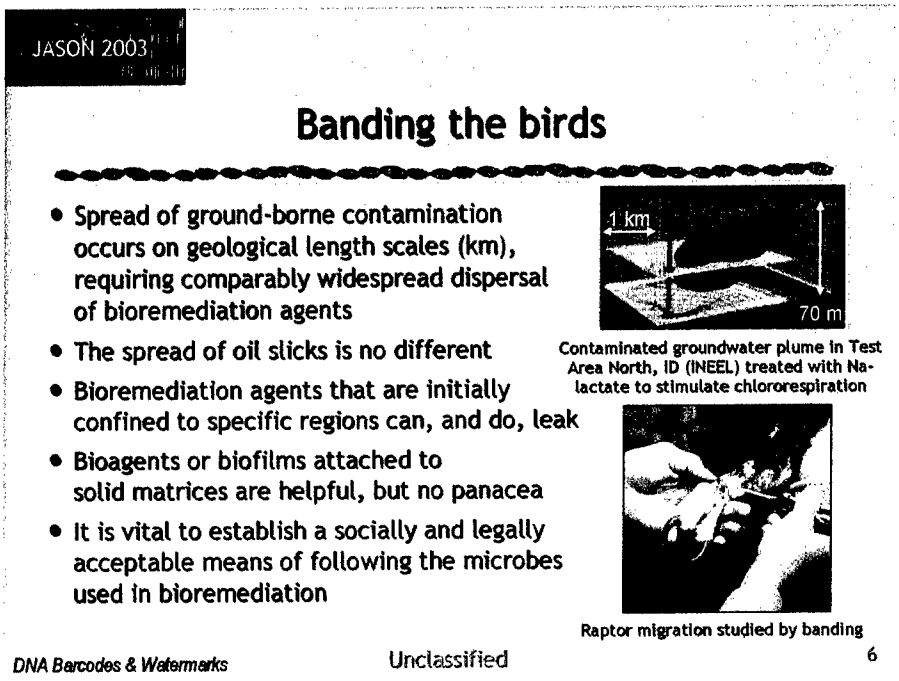
In its most advanced form, bioremediation involves the introduction of microorganisms that have been deliberately altered to have desirable properties, such as the genes for metabolizing certain pollutants, or the genes to tolerate extreme environments. Such strains may be produced by a traditional process of selection and breeding, or they may be created by genetically engineering. In the future, it seems likely that genetically-engineered strains of bacteria will play an increasingly important role in bioremediation.



Irrespective of the origin of the microbes used, however, all bioremediation efforts lead to environmental changes – hopefully, beneficial ones. Ideally, microorganisms whose growth is fostered during the remediation process die off naturally once their job is complete, as the nutrients become limiting, thereby restoring the environment to a more pristine, original state.

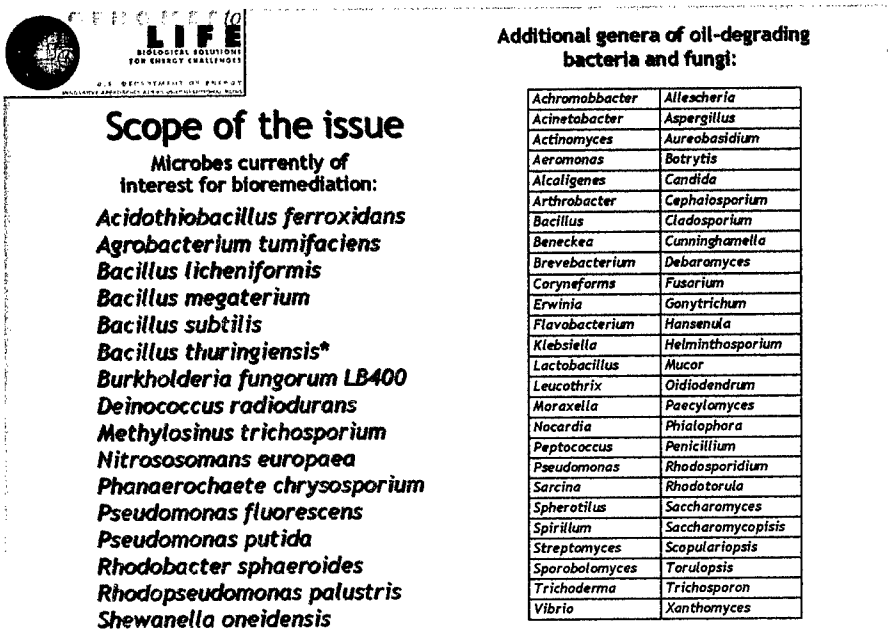
However both introduced organisms and indigenous species whose growth has been enhanced may linger in the immediate environment or spread beyond their original site, producing unintended consequences.

To monitor the process of bioremediation, it will be vital to establish a socially and legally acceptable means of following the growth and ecology of all species used for bioremediation. Towards that end, JASON explored the possibility of developing a robust system of genetic tags that would facilitate microorganism identification and tracking. This system would be, in many ways, analogous to the process of banding birds (and other fauna) in wildlife studies, except that the tag would be incorporated into the genome itself.



If such a genetic 'barcoding' system could be developed for the microorganisms used in DOE bioremediation efforts and demonstrated to work successfully in the field, it may eventually have much broader utility. Perhaps the most significant application beyond the bioremediation context might be the use of an analogous barcode system to track the spread of genetically-modified organisms (GMOs) in the environment, a topic of considerable national and international interest.

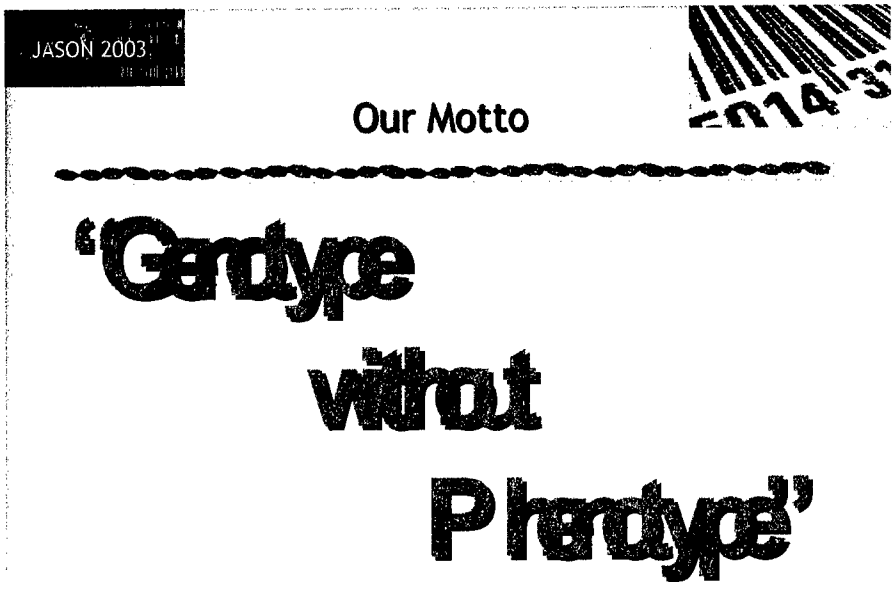
An astounding a number of microbial species have unusual metabolic capabilities that enable them to accomplish such feats as degrading petrochemicals, metabolizing halogenated organic compounds, reducing heavy metals such as mercury and uranium, and so on. The following list shows 16 promising bacterial species currently under investigation in projects sponsored by the DOE, alongside another 52 genera of bacteria and fungi that are capable of degrading oil (either individually or in combination):



*1st GM plant containing *B.t.* toxin (*Cry*) registered with EPA, 1995

The DNA Barcode Concept

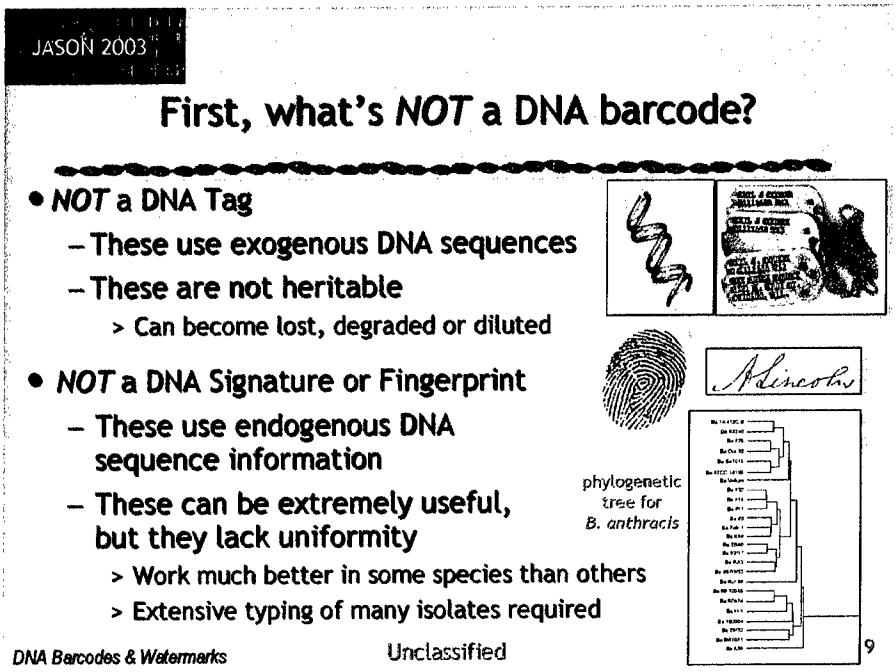
A desired feature for tagging a microbial species intended for release into the environment is that any new sequence introduced into the genome for the express purpose of labeling it should be detectable (with appropriate technology), *but entirely silent*, conveying no discernible change in the metabolism, behavior, or fitness of the organism itself. In a phrase, the ideal barcode tag would convey "genotype without phenotype." This became, in effect, the motto for the study.



What a DNA Barcode is *NOT*

The phrase "DNA barcode" has already been used by some researchers to refer to some other things. To alleviate any possible confusion, we therefore begin by drawing a clear distinction between DNA barcodes, in the context of this study, and these alternative uses of the phrase. First, a barcode is *not* a DNA tag. DNA tags involve the use of prepared DNA molecules containing known, amplifiable sequences as a means of tagging and identification. DNA tags are in widespread use today, and are commercially used to mark objects of value for anti-counterfeiting purposes. For example, DNA tags were used on tickets and merchandise associated with the 2000 Sydney Olympics, and also to mark footballs used in the NFL Super Bowl and hockey pucks used in the NHL[3]. Such tags consist of exogenous DNA sequences mixed with chemical stabilizers. They are not heritable and they can become lost, degraded, or diluted over time — although they are surprisingly robust. DNA tags are not suitable for marking self-reproducing organisms in a bioremediation context.

'DNA barcode' has also been used, unfortunately, to refer any convenient, variable DNA sequence pattern already found in Nature as a consequence of evolution. Thanks to rapid and powerful DNA hybridization techniques, such patterns can now be exploited, for example, to distinguish among otherwise similar organisms, to identify the genus and species of DNA from an unknown source, or to catalog many diverse species. Since higher organisms all possess mitochondria, the variable portions of the mitochondrial genome sequence have proved especially useful for categorizing eukaryotes[4].



[3] http://www.dnatechnologies.com

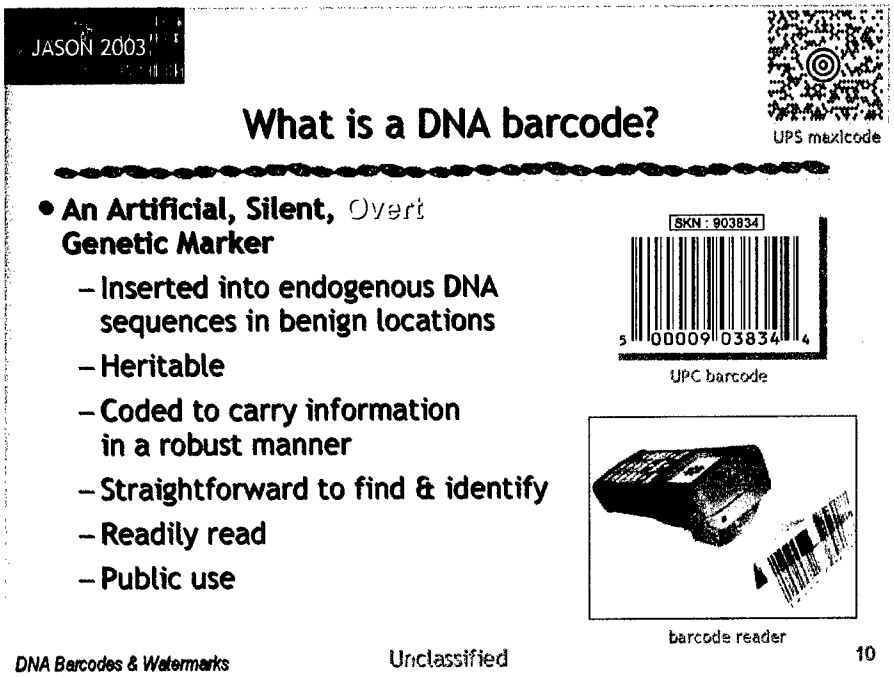
[4] Hebert, P.D.N., Ratsingham, S. & Dewaard, J.R. Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. Roy. Soc. Lond. B.* 270;1524: Suppl. 1: S96-99 (2003).

In a similar fashion, most prokaryotes can be categorized by means of the sequences of their 23S ribosomal RNAs. However, such natural genomic sequences are not analogous to "barcodes" in any strict sense, because they do not represent manmade labels that have been manufactured and introduced into the organism for the express purpose of identification, as DNA barcodes are. Such natural sequence variations are therefore best described as "*DNA signatures*" or "*DNA fingerprints*."

More importantly, DNA fingerprints (when used as barcodes) do not represent a universal form of label: they lack uniformity because they rely on variable, endogenous information. In order to use DNA fingerprinting, unique sequences identifying each variant of the organism must first be identified and characterized, and this requires extensive collection and typing of specimens. Furthermore, some species mutate more slowly than others. As a result, on occasion it can be quite difficult to identify sequences that reliably encode variation. A notorious example here is the anthrax bacterium: most strains of *B. anthracis* differ remarkably little, with as few as 8 point differences among all the recent "Ames" strain variants.

What a DNA Barcode *IS*

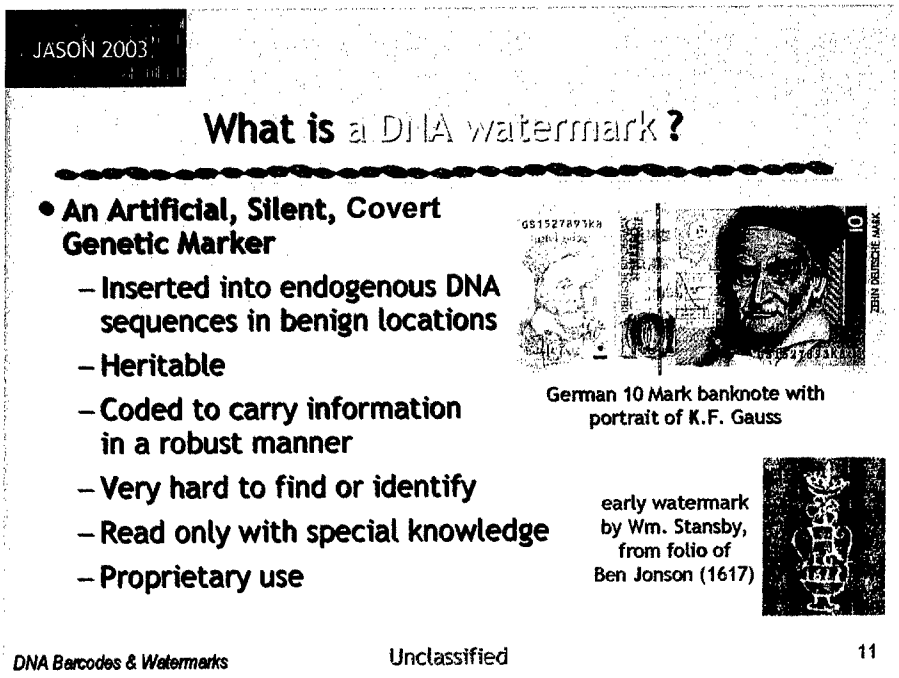
A DNA barcode is an artificial, silent genetic marker consisting of a relatively short stretch of DNA carrying a designed sequence. In practice, a DNA barcode would be introduced benignly into the genome of the host organism as a sequence tag, in such a fashion that it conferred no phenotype. It would be heritable, so that all progeny carry the identical barcode. The sequence of the barcode would carry encoded information in a robust manner. For example, the barcode could encode a serial number corresponding to an entry in a public database carrying further information about the strain. The barcode would be constructed in such a way that it would be straightforward to locate and identify in the genome, using established methods for DNA amplification and hybridization. A DNA barcode is *overt*, in the sense that its presence is not



deliberately hidden or encrypted: it should be readily detected using standard methods, and it is intended for general use in the public domain.

What is a DNA Watermark?

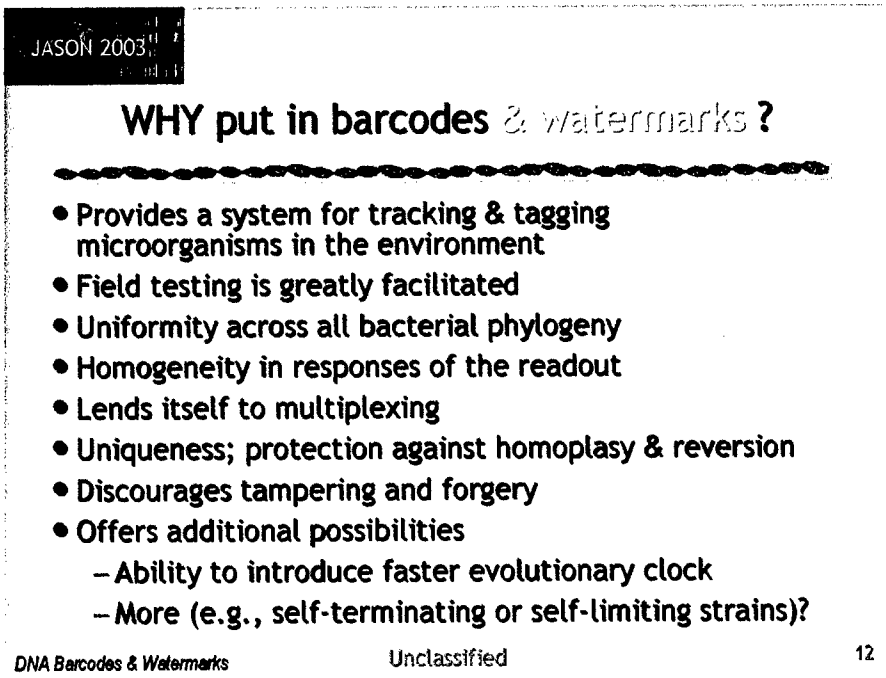
In conjunction with DNA barcodes, it is useful to develop a closely-related concept: a DNA watermark. Unlike the barcode sequence, which is a continuous length of DNA designed for easy detection, a DNA watermark carries information in a more distributed (holistic) manner. As a direct consequence, it is difficult-to-impossible to locate and read a watermark without special knowledge. Like a barcode, however, a DNA watermark should be heritable and convey no measurable phenotype.



A DNA watermark, then, is *covert*, and it serves as a cryptic form of identifier. By analogy with traditional watermarks, it can function as a unique identifying characteristic and as an anti-counterfeiting measure. We imagine that DNA watermarks will have proprietary uses, chief among these being to protect the DNA barcode (found elsewhere in the genome) against tampering. Whereas barcode information is placed in the public domain, watermarks are closely held. *Bona fide* tagged organisms will carry both a barcode and its associated watermark. An organism discovered subsequently with a barcode in it but no corresponding watermark is likely to be counterfeit. In a similar vein, an organism discovered with a watermark but missing its corresponding barcode is likely to have been tampered with. The watermark, therefore, can be used to protect the integrity of the barcode system. We imagine that DNA watermarks, by their very nature, will be comparatively more difficult to implement and construct than barcodes, and will therefore be used to encode less information (i.e., have a lower coding capacity).

The Barcode/Watermark Tagging System

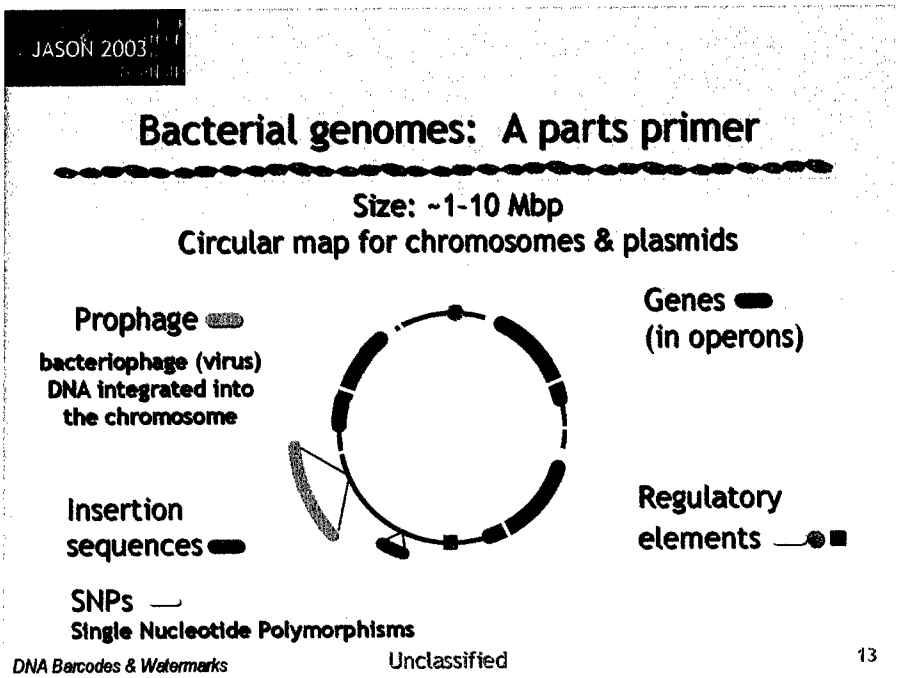
A correctly implemented system combining both barcodes and watermarks would confer multiple advantages for tagging and tracking microorganisms in the environment. Since all barcode tags will carry common sequences alongside additional sequences unique to the tagged organism, rapid field testing for the presence or absence of labeled organisms is greatly facilitated. The uniformity of the barcode system across phylogeny finesses the problem of limited natural sequence differences associated with traditional identification systems based on natural variation (fingerprinting). Other advantages of a barcode/watermark system include robustness and homogeneity in the readout, immunity from several kinds of ambiguity arising from sequence homoplasy (that is, convergent evolution towards the same sequence) and reversion (accidental return to the same sequence), as well as protection against tampering and forgery. Finally, it may even be possible to enhance the basic barcode/watermark system by introducing special barcode sequences deliberately designed to mutate over time, but at rates much faster than normal evolution. Such hyper-mutating sequences could provide a kind of 'ticking clock' that changed some region of the barcode on a regular basis. In principle, hypervariable sequences might be used to determine the time lapse, measured in generations, from the release of the original tagged strain to the capture of the target strain.



Introduction to the Bacterial Chromosome

To set the stage for a practical plan for the implementation of barcodes and watermarks, we begin by describing salient aspects of the bacterial chromosome. Most bacteria have one, or at most a few, circular chromosomes made of DNA, each containing a unique origin of replication. In addition to the chromosome, a bacterium may have one or more much smaller, circular DNA elements called plasmids, each with its own origin of replication. (Indeed, the

distinction between a plasmid and a chromosome is mainly one of size, not of function. Some plasmids carry supplementary genes and are dispensable; some are not.)

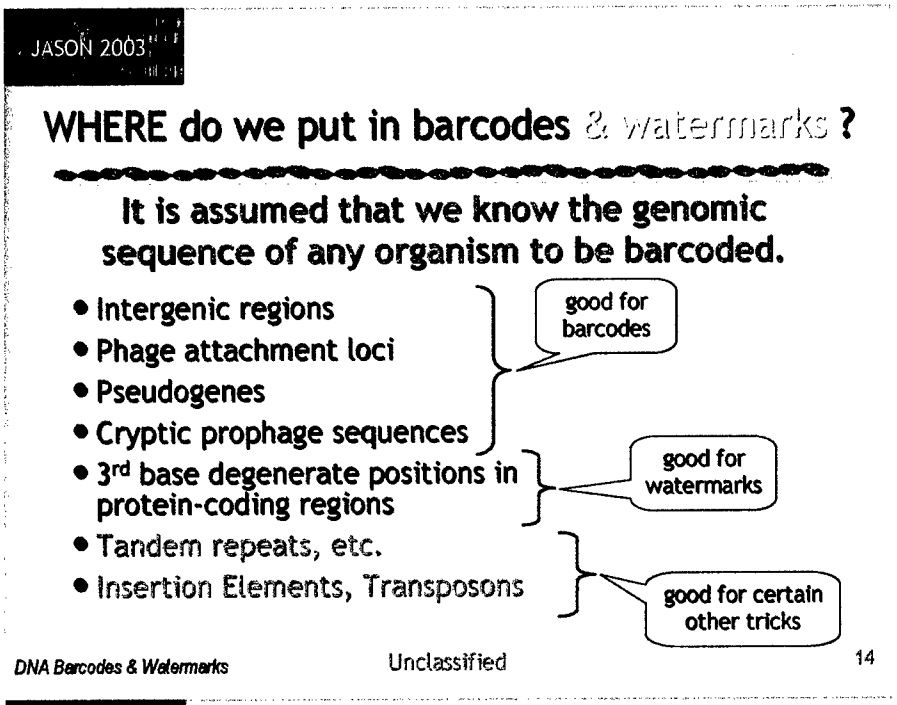


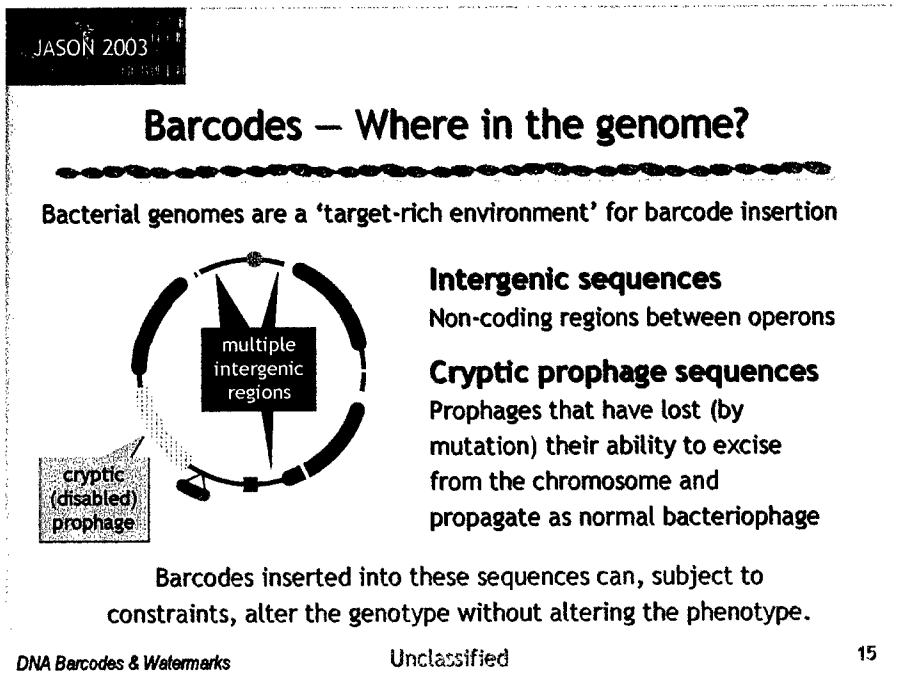
The bacterial chromosome consists mainly of genes (i.e., sequences that code for proteins) organized loosely into co-expressed groups, called operons, each with its own gene control region. But there are other notable features as well: see the figure above. These include regulatory elements, such as the aforementioned control sequences, as well as a replication origin, short intergenic regions, and insertion sequences such as transposable elements. Also notable are DNA sequences where a bacteriophage (virus) has inserted itself into the chromosome and generally lies dormant. Such a continuous segment of intact viral DNA is termed a "prophage" sequence. Imbedded prophages can, with appropriate stimulation, excise from the chromosome, then replicate and proceed to form live virus particles, eventually lysing the host cell. Hence, dormant phage sequences are said to be 'lysogenic.' DNA sequences that have lost by mutation their ability to excise from the genome are called 'cryptic prophages.' Also illustrated in the figure are examples of locations where the genomic sequence may have mutated away from the wild type by point changes to single bases in the DNA . These are called SNPs (Single Nucleotide Polymorphisms).

Where to Place Barcodes & Watermarks

To discover places where barcodes might be placed in a candidate organism, it is extremely helpful to know its complete DNA sequence. Fortunately, the full genomic sequences for many species of important bacteria are now known, and plans are already in place to sequence all major human pathogens, as well as most of the bacteria vital to agriculture and industry. The DOE is also sequencing several of the major organisms of interest for

bioremediation purposes. The sequencing for one of these, *D. radiourans*, is now complete, and several more are in the works. In the near future, whole-genome shotgun sequencing of any organism with a genome as small as a bacterium should not present much of an obstacle, and could be accomplished in a matter of days. We will therefore assume that the genomic sequence of any microorganism selected for DNA barcoding is known in advance.



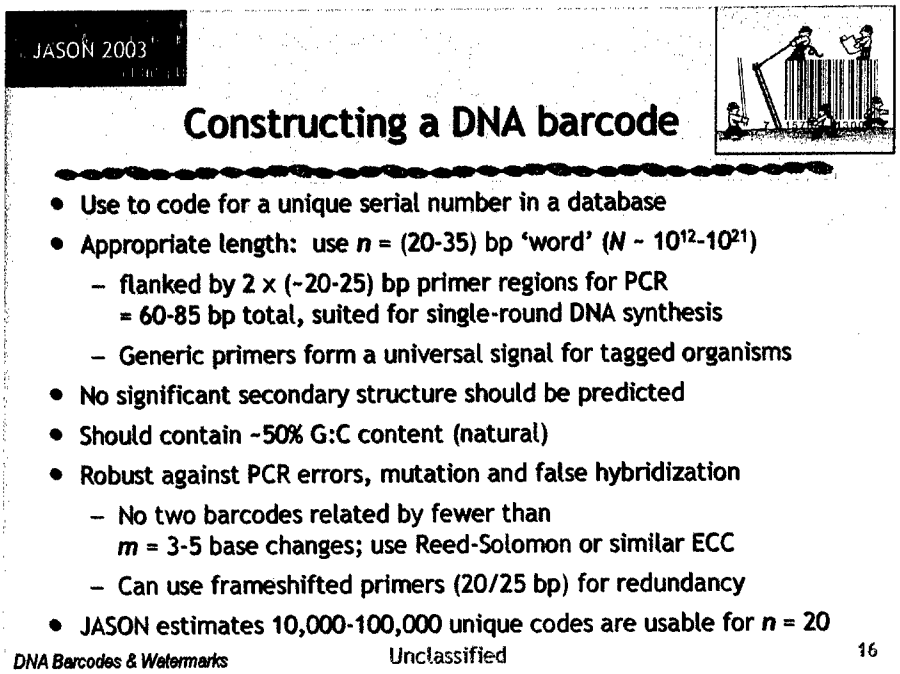


Clearly, it is not desirable to insert a DNA sequence into the genome in any place where it will disrupt either genes or genes expression, and thereby confer a phenotype. There are several specialized types of chromosomal region, however, that may accommodate the introduction of exogenous DNA sequences without producing a measurable phenotype. Some amount of experience and experimentation may be required to identify the most suitable regions for the placement of barcodes and watermarks, in practice. Generally speaking, barcodes are best placed in some of the following locations: intergenic regions (regions between genes, particularly those situated between operons of opposite polarity), phage attachment loci (positions where lysogenic phage insert), pseudogenes (genes that have lost their function and are no longer expressed as full-length, functional proteins), or inside cryptic prophage sequences. For watermarks, a scattered set of silent, single-base changes (SNPs) represent attractive candidates, particularly when such changes occur in the degenerate, 3rd-base positions of codons for amino acids. This approach is described in greater detail later in this study report. For purposes beyond barcodes or watermarks, tandem repeats and insertion elements offer other opportunities: these, too, are described in more detail in a later section of this report.

Despite the comparatively high density of protein-coding regions found in prokaryotes as compared with eukaryotes, the bacterial chromosome is still a fairly "target-rich environment" for the insertion of DNA barcodes.

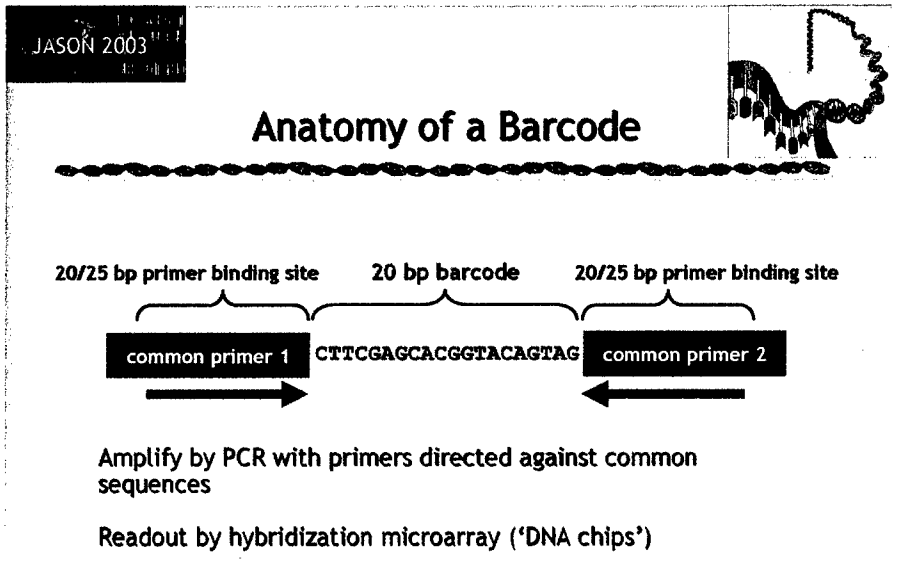
Anatomy of a Barcode

We now consider the essential elements of a prototypical DNA barcode. As a practical matter, the full barcode sequence should be short enough that it can be produced with few errors during a single round of DNA synthesis at a reasonable cost. As of today, DNA oligos can be ordered from commercial suppliers in lengths of 80-100 basepairs, so this represents the current practical upper size limit. However, one may anticipate that somewhat longer barcodes may become feasible in the future, should these be required.



The overall barcode sequence is conceived to consist of three parts: a single, central region containing a unique DNA code sequence, flanked on either side by two shorter regions, designed to serve as primer sites for the PCR amplification of the central region. The central portion would code for a unique number corresponding to a serial number entry in the barcode database. A sequence of 20-35 bp could, in principle, generate from 10^{12}-10^{21} unique serial numbers, but there are practical considerations that severely restrict this limit (see below). The flanking regions would be complementary to a single common set of '*universal barcode primers*' that would be used to initiate a PCR reaction to amplify (and thereby to pull out of a complex reaction mixture) any DNA fragments carrying these flanking regions. The existence of an amplicon derived from the common primer set would, therefore, constitute a generic signal for the presence of a barcoded organism, irrespective of its serial number. This could serve as the basis for a rapid field test for barcoded organisms. The sequences of the central regions of any amplicons would indicate the organisms themselves. Common flanking primers would be from 20 to 25 bp long apiece (see the illustration below), so the overall length of the barcode insertion might range from 60-85 bp, which falls readily within current manufacturing limits.

There are further criteria that are important to meet in the design of a practical barcode. To prevent self-priming and false amplification during the PCR reaction, a barcode should not have any propensity to form a DNA hairpin structure in single-stranded form (i.e., it cannot be self-complementary). Furthermore, to remain biocompatible, the G:C content of the barcode should be roughly similar to that found in living organisms, i.e., ~50%. Finally, to make the system robust, it is advantageous to encode the central region in such a way that no two serial numbers are related by fewer than 3 (or more) base changes. In that way, barcode serial numbers may develop up to 3 random mutations but still be unambiguously identified. Finally, by using common primer regions containing 25 bases, in which just 20 contiguous bases are actually used for amplification in the PCR reaction, it is possible to use multiple sets of redundant, 'frame-shifted' primers. Frame-shifted primers can be used to overcome potential problems of false priming at alternative sites that might occur by chance in some species.

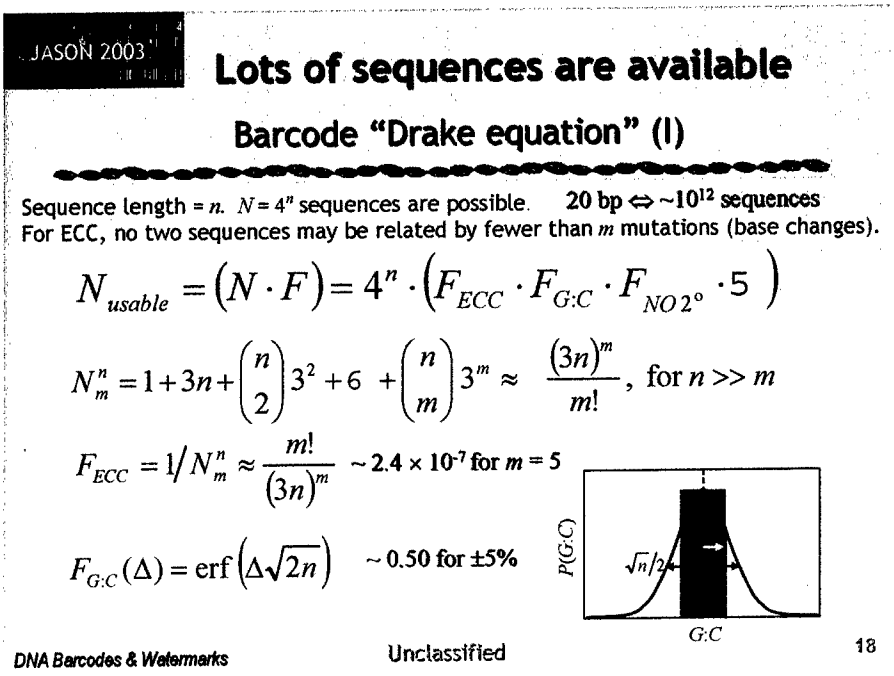


A Practical Barcode is Shown (figure above)

In this implementation, the serial number region is encoded as a 20 bp region, while the two flanking primer regions are each 25 bp long, of which 20 bp are used to prime the PCR reaction, as indicated by the arrows. The overall barcode length is 70 bp. Once amplified by PCR, DNA barcodes could be 'read' without a need for direct sequencing by using a suitable hybridization arrays carrying complementary sequences for all barcodes found in the database. This process is described in more detail later in this report.

How Many Practical Barcodes Exist?

We now turn to the question of how many unique serial numbers exist, from among the $\sim 10^{12}$ possible choices offered by $N = 20$ bases, once all the design criteria are met for G:C content, lack of self-priming DNA secondary structure, and the requirement that all codes be separated by several mutations. Are there enough codes to go around? JASON looked into this question in detail. *The bottom line is that there are sufficient codes to tag all the organisms of interest for the foreseeable future, provided that a central coding region of 20-24 bp (or longer) is used.*



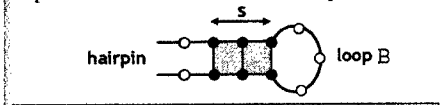
Sequence availability is examined in detail in **Appendix A**. However, a simple way to think about the numbers is to construct an analog of the "Drake Equation", which was used to guesstimate the chance of finding intelligent life elsewhere in the universe by multiplying out all the independent probabilities that underlay its likelihood. To form an estimate of the numbers of useful barcodes, we assume that the design criteria are statistically independent, and therefore multiply the fraction of codes with normal G:C content ($F_{G:C}$) by the fraction that don't form hairpins (F_{No2}) by the fraction of codes with sequences separated by m or more point changes

(F_{ECC}). This is then multiplied by the number of possible codes, $N = 4^{20}$, to estimate the usable number, N_{usable}. For random sequences, it turns out that $F_{G:C}$ is close to 50%, assuming that one accepts sequences in the range of 45% to 55% G:C content. Furthermore, F_{No2} is found to be somewhere around 30%, assuming that a hairpin will form from three or more complementary bases whenever these are separated by a loop of three or more bases. (A more accurate estimate of this fraction can be obtained by running the M-FOLD program against all sequence candidates to calculate the free energies of possible secondary structures and rejecting all those sequences that form stable hairpins). By far the most stringent criterion for rejection, as one might easily anticipate, comes from F_{ECC}, which is estimated at 2 x 10^{-7} or less. Accepting these values, one winds up with the rough estimate $N(20)_{usable} \approx 40{,}000$.

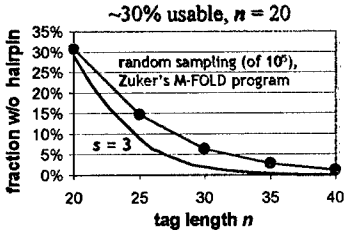
JASON 2003

Barcode "Drake equation" (II)

Effect of secondary structure: No hairpins or 'primer dimers' desired. Min. loop size = 3.



loop length	1	2	3	4	5	10	15	20
free energy (kcal/mole)	∞	∞	3.5	3.6	3.3	4.3	4.6	5.1



$$M = \sum_{B=3}^{n-2s} n - 2s - B + 1 = \tfrac{1}{2}(n-2s-2)(n-2s-1)$$

$$F_{NO\,2^{\circ}} = (1-q)^{M};\quad q = 1/4^{s} \qquad M(s=3) = \tfrac{1}{2}(n-7)(n-8) = 78$$

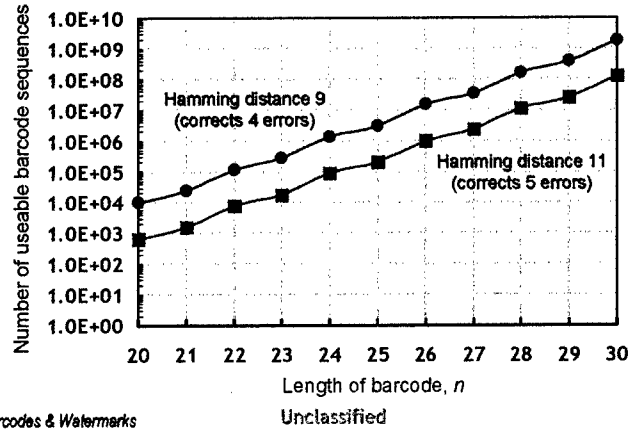
$$N_{usable} = (N \cdot F) = 4^{n} \cdot (F_{ECC} \cdot F_{G:C} \cdot F_{NO\,2^{\circ}})$$

$$N_{usable} = (N = 1.1\times10^{12})(F_{ECC} \approx 2.4\times10^{-7})(F_{G:C} \approx 0.50)(F_{NO\,2^{\circ}} \approx 0.3)$$

$N(20)_{usable} \approx 40{,}000$ = more than we'll need. Or, choose n>20.

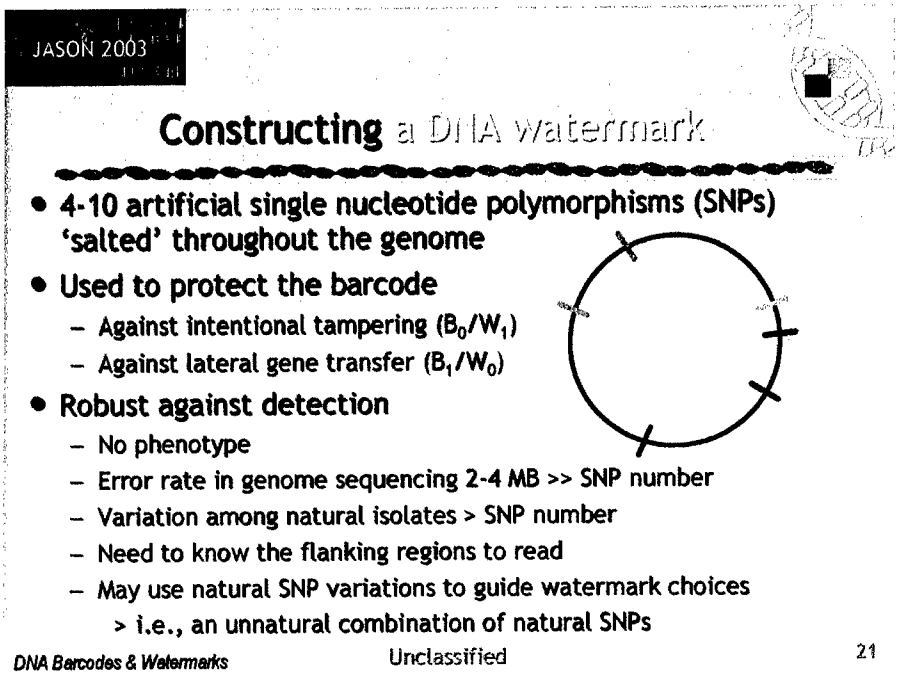
JASON 2003

Barcode Drake Equation (III)



A more careful estimate can be prepared by computer simulation, and simply enumerating (either exhaustively or by statistical sampling) the fraction of randomly-selected sequences matching all the design criteria, leading to the graph shown above. If it is a requirement to correct 4 or fewer base errors, then there are of order 10,000 barcode sequences available for a 20-mer barcode, ranging up to 100,000 codes for a 24-mer. To correct 5 or fewer errors, one typically finds 10-fold fewer sequences than for the case of 4 or fewer errors.

Constructing a DNA Watermark

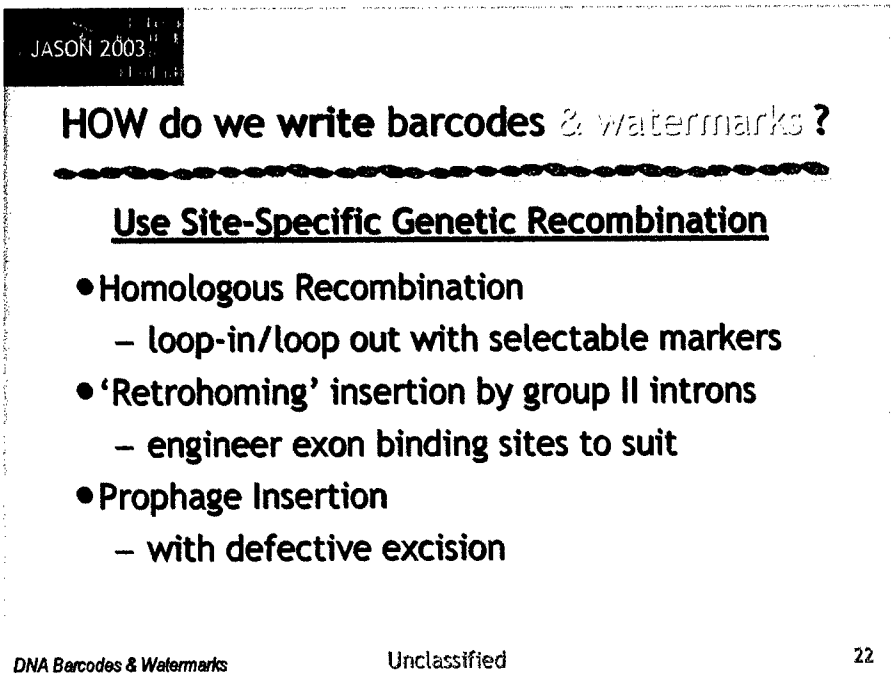


As discussed, a DNA watermark is a cryptic series of changes to the genomic sequence that produce no discernible phenotype. The watermark is designed to be difficult-to-impossible to detect and read without special (proprietary) information. We envisage the use of DNA watermarks directly in conjunction with barcodes, as a means to protect the integrity of the barcode system for marking organisms slated for bioremediation purposes. [Clearly, however, watermarks could equally well be developed for applications independent of a barcode system, whenever clandestine genetic marking alone is desired. We will not pursue this possibility here.]

There are many ways, in principle, to generate and encode DNA watermarks: some of these schemes, by analogy to digital watermarking solutions in the computer world, can be quite sophisticated. However, one particularly straightforward system would be to introduce several single-nucleotide polymorphisms (SNPs) at scattered locations throughout the genome. A simple watermark might consist of 4-10 such SNPs. All barcoded bacteria would initially be watermarked (B=1/W=1). Organisms subsequently recovered without their barcode but nevertheless carrying the watermark (B=0/W=1) would be candidates for strains that may have been tampered with. Conversely, organisms carrying a barcode but missing the watermark (B=1/W=0) would be candidates for strains that picked up the barcode by some form of gene

transfer, either accidentally (by lateral transfer) or as result of deliberate tampering (engineering). A series of just 4-10 SNPs distributed throughout a typical bacterial chromosome may be exceedingly hard to detect without detailed advance knowledge of their identities and chromosomal locations. The error rate for whole-genome sequencing is such that the number of mistakes made while sequencing the entire organism would exceed the total SNP number. Furthermore, the number of DNA changes associated with the watermarking process is likely to be smaller than the number of SNPs found among natural sequence variants of any given bacterial species. Finally, if desired, a very cryptic watermark could be created by producing some unnatural combination of naturally-occurring SNP variants.

Introducing Barcodes & Watermarks



JASON considered a number of mechanisms for inserting barcodes into a bacterial genome. Broadly speaking, three such mechanisms seemed particularly promising: (1) using homologous genetic recombination in conjunction with selectable markers to target barcode sequences to selected sites, (2) 'retrohoming' by genetically-engineered Group II Introns carrying barcode information, again targeted to selected sites, and (3) insertion of modified lysogenic bacteriophage into insertion sites as prophages, carrying barcodes inside phages designed with defective excision.

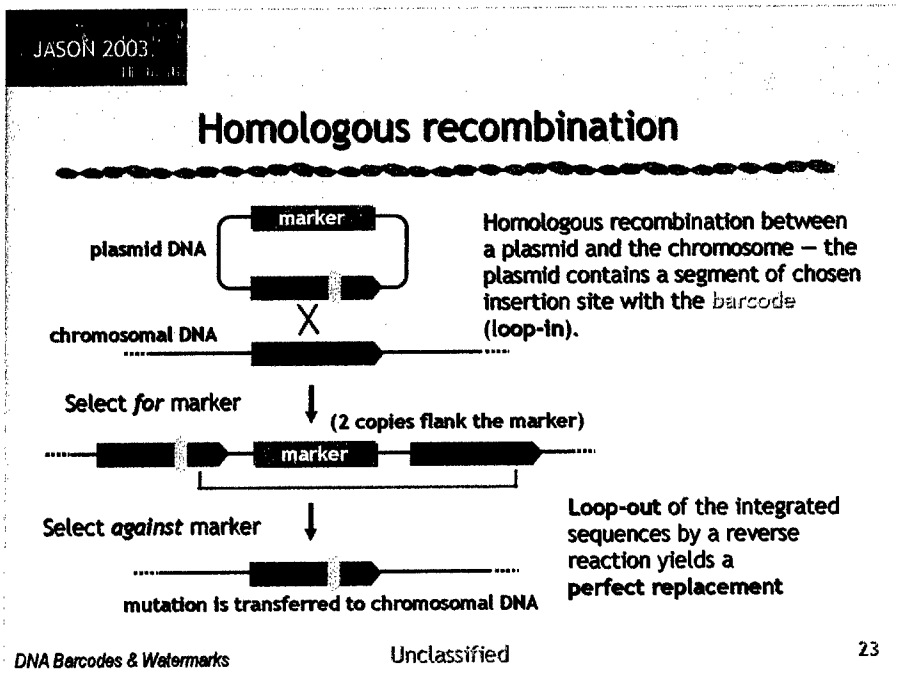
The last of these three mechanisms, which takes advantage of process by which lysogenic phage insert at specific attachment sites in the chromosome, is the least general and arguably makes the greatest demands on a would-be fabricator. For every bacterial host species to be barcoded, it is first necessary to identify a suitable lysogenic phage that can infect the host (there would be no *a priori* assurance that such a virus could be found for all species of interest). The

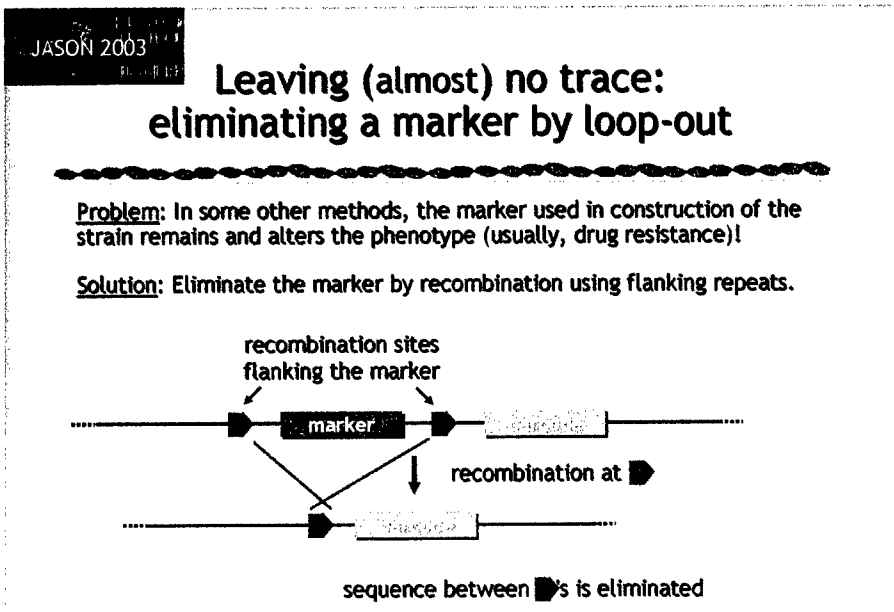
site of phage insertion into the host chromosome would then have to be mapped, and the phage would also have to be sequenced, so that the genes required for its chromosomal attachment and excision could be identified. A suitable mutant would have to be created that was deficient in excision. Furthermore, a lysogenic prophage carrying a barcode tag would introduce excess DNA into the bacterial chromosome beyond that needed purely for the barcode itself. For these reasons, our study chose to concentrate mainly on the first two alternatives, above. However, there may be instances where prophage-based barcoding is both practical and effective (and we explore one of these later).

Placing Barcodes & Watermarks by Recombination

Provided that the sequence in the immediate target region is known and that a suitable genetic marker can be identified that can be selected both *for* and *against*, it is possible to insert an arbitrary region of DNA into the chromosome, by taking advantage of the process of homologous recombination using a "loop-in, loop-out" procedure. Bacterial species differ widely in the degree with which they support genetic recombination, but many species can be induced to recombine when the selection pressure is high, as with (for example) antibiotic selection.

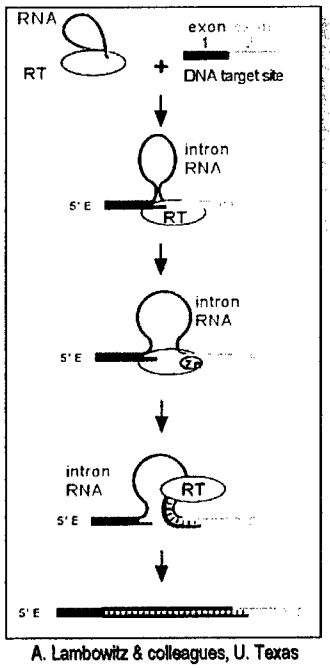
The following two graphics illustrate how homologous recombination is used to introduce a sequence, and **Appendix B** describes the entire process in greater detail. Because homologous recombination can insert a sequence without introducing any flanking DNA (i.e., it generates no additional remnants), it is also suitable for generating the SNPs required for watermarking, as well as for introducing barcodes.





Placing Barcodes by Group II intron "Retro-homing"

A promising method for inserting a barcode at a targeted location is to take advantage of the natural ability of certain catalytic RNA sequences (ribozymes) to self-splice into DNA and RNA. The following three graphics outline the procedure, which is described in greater detail in **Appendix B**. This approach has the advantage of being 'universal,' in the sense that it is



A. Lambowitz & colleagues, U. Texas

Group II intron-based barcode insertion

Problem: Homologous recombination does not always work efficiently in all strains

Solution: Use engineered group II introns (catalytic RNA/enzyme complexes)

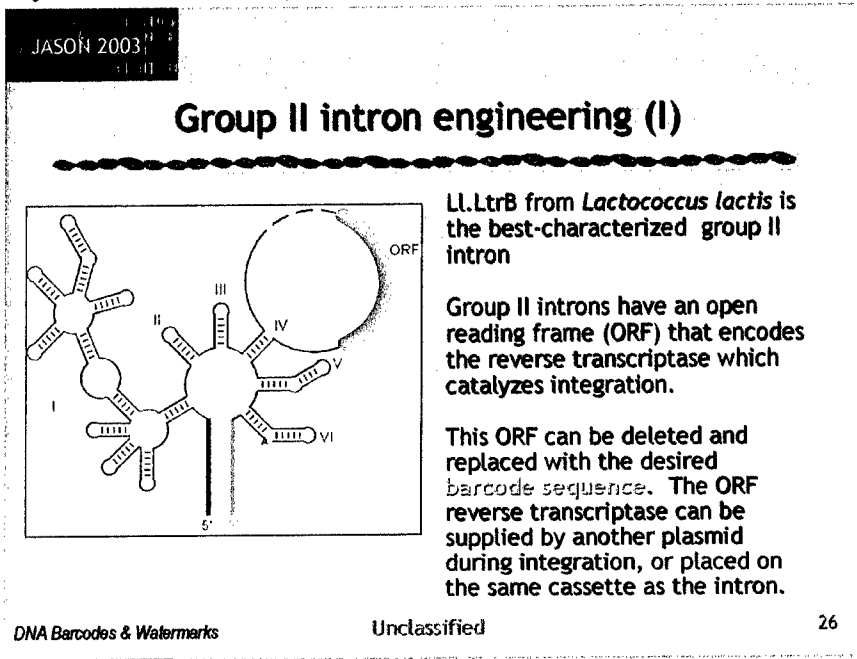
Group II introns use **"retrohoming"** to insert into specific DNA sites. The active element includes the intron RNA plus an intron-encoded reverse transcriptase (RT).

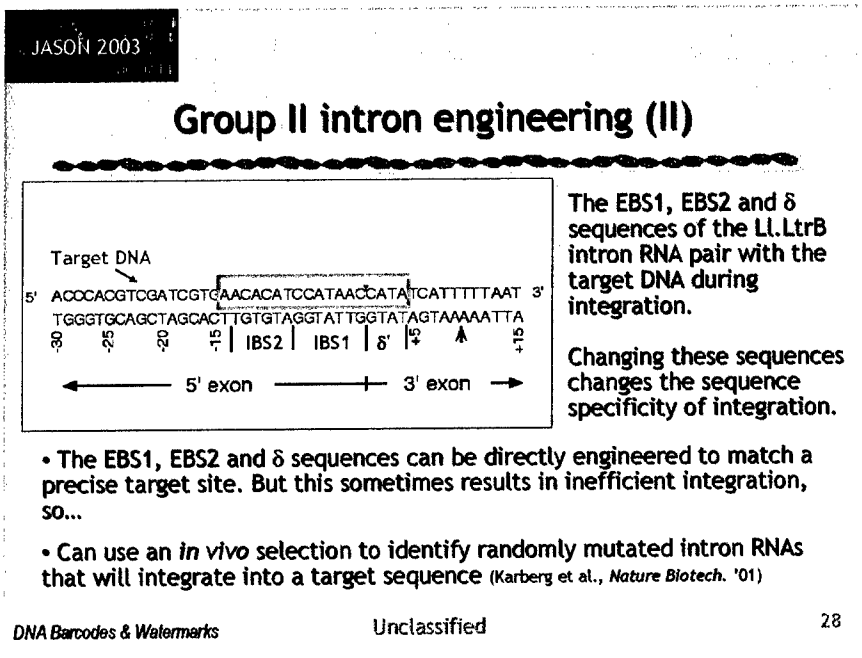
Advantage: Universality

Can be engineered to target any DNA sequence

- Minimal dependence on host factors for integration
- Fully active in recombination-deficient hosts

independent of the properties of the host organism, and it bypasses any need for homologous recombination, which may not work efficiently in certain bacterial strains. Dr. A. Lambowitz's research group (U. Texas) has engineered Group II Intron ribozyme sequences derived from *Lactococcus lactis* to serve a vectors for targeted DNA insertion[5]. Such vectors also conveniently carry a gene for reverse transcriptase, which is required to synthesize DNA complementary to the RNA that's spliced into the chromosome.

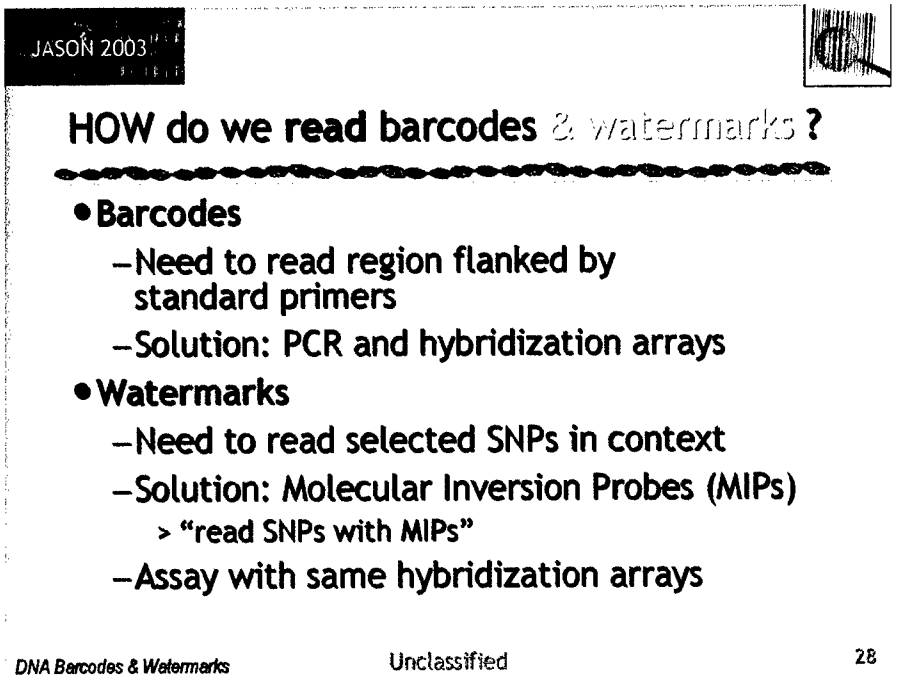




[5] M. Karberg et al. Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. *Nature Biotech.* **19**:1162-7 (2001).

Reading Barcodes & Watermarks

Using appropriate technology, it should be possible to read DNA barcodes and watermarks by PCR amplification followed by DNA hybridization, without any need to sequence the host genome. This should make rapid field testing possible, for example using hand-held devices. Moreover, the entire dataset of barcoded organisms could be scored at the same time with a single round of hybridization, using DNA extracted from environmental samples containing potentially complex mixtures of organisms, both tagged and untagged.



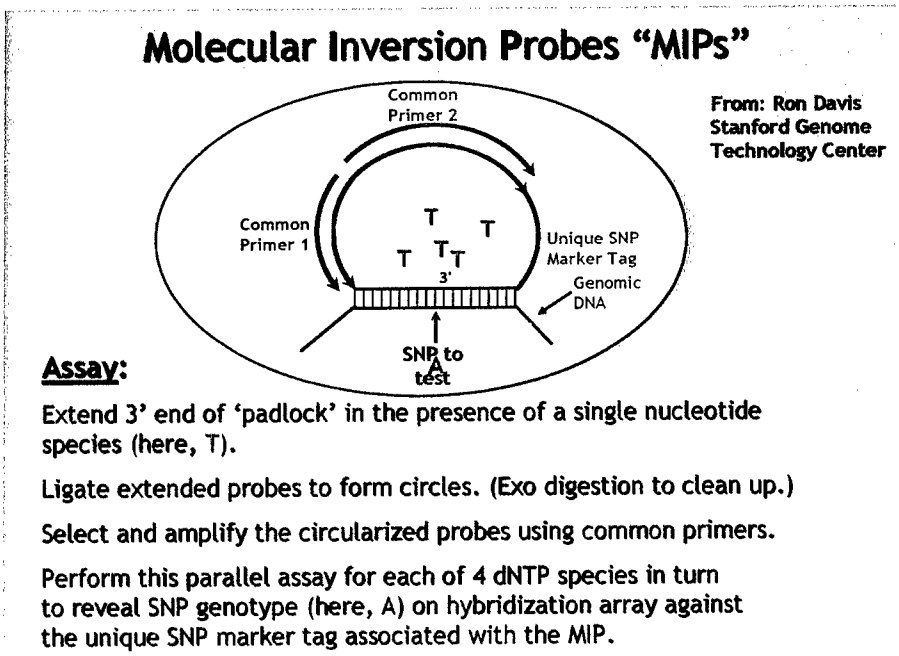
For barcodes, the readout process would first consist of extracting and partially purifying DNA within an environmental sample, using protocols and sample-handling procedures similar to those currently employed for DNA field testing. The DNA would then be mixed with a set of standard primers (20-mers) complementary to the common flanking regions of the barcodes,[6] followed by PCR amplification. The presence of amplicons in the reaction mixture would indicate the presence of a barcoded organisms. Specific barcodes would be read out using a hybridization microarray that carried sequences complementary to each barcode.

In most instances, DNA microarrays would be designed to check for barcodes but not watermarks. Such readout devices would remain in the public domain. Under certain circumstances, however, it is desirable to check both the barcode and its associated watermark (e.g., to check for counterfeit organisms). It turns out that it is equally feasible to use hybridization technology to read the SNPs used to watermark bacterial genomes, once again bypassing any need for direct sequencing. One practical approach to amplify and identify targeted SNPs by

[6] An overlapping set of frame-shifted primers could be used to improve the signal to noise of this step, if necessary, as discussed earlier. Provision for frame-shifted primers was made by increasing the common flanking regions to 25 basepairs, of which any 20 contiguous bases may be used for priming the PCR reaction.

hybridization was developed by Dr. R. Davis (Stanford Genome Technology Center), as is called 'MIPs.' (Molecular Inversion Probes)[7].

MIPs work as illustrated in the following graphic. In brief, a DNA oligo is designed that forms a *nearly complete circle*, hybridizing directly with genomic DNA on either side of the SNP to be tested, but leaving the base opposite the SNP unpaired. The oligo is composed of three different types of engineered sequence: (1) regions designed to pair with the genomic DNA flanking the SNP to be tested (shown in black), (2) generic sequences that are complementary to each of two common primers used for later PCR amplification (red and blue), and (3) a SNP marker tag that uniquely marks the particular oligo (purple). The oligo is then allowed to pair with the genome, and the reaction mixture is divided into 4 aliquots. Each aliquot is then incubated separately with a single nucleotide (A,T,C or G) plus ligase. Closed circles are formed only in the reaction mixture that's been incubated with the complementary nucleotide. An exonuclease digestion follows to degrade all DNA not formed into circles. The 4 reaction mixtures are then incubated with common primers, which are arranged back-to-back as shown, and amplified by PCR. Finally, the identities of the amplified circles are scored by hybridization against an array that detects the unique marker probe. Knowledge of both the hybridizing marker probe and the single nucleotide leading to successful amplification (circularization) uniquely maps the identity of the SNP.

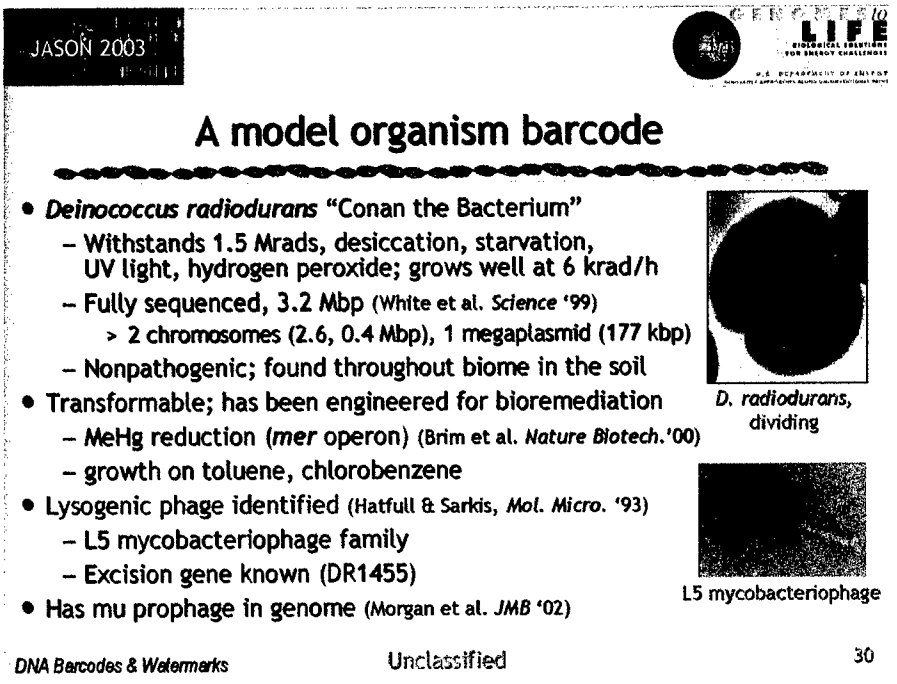


The technology for MIPs-based detection of SNPs used for watermarking and those needed for barcode detection can be combined into a single DNA hybridization array, which would have proprietary uses.

[7] P. Hardenbol et al. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature Biotech.* **21**:673-678 (2003).

A Worked Example: Barcoding & Watermarking the Model Organism, *D. radiodurans*

To frame the discussion somewhat more concretely, JASON explored a specific example of barcoding and watermarking. The strain we selected for further study was *Deinococcus radiodurans*, sometimes jokingly referred to as "Conan the Bacterium" for its astounding ability to tolerate stress and punishment. It was first discovered in 1956 in Oregon in canned meat that had spoiled despite exposure to X-rays. *D. radiodurans* is among the most radiation-resistant of all known organisms: it can withstand up to 1.5 Megarad acute dose and can grow continuously in the presence of radiation levels of 6 kilorad/hour. Moreover, it can cope with extended periods of desiccation, starvation, UV light, and high levels of peroxides. It is nonpathogenic, it is transformable, and it is quite commonly found throughout the biome in the soil. Much attention has recently focused on *D. radiodurans* as a candidate organism for use in bioremediation, particularly at DOE wastewater sites where radiation levels can reach 10 mCi/liter or more. The genome of *D. radiodurans* has been fully sequenced[8] and analyzed[9], and the organism has recently been genetically engineered to carry genes from *E. coli* that allow it to reduce Hg(II) and certain other heavy metals[10].



[8] O. White et al. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286**:1571-1577 (1999).

[9] K.S. Makarova et al. Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. Microbiol. & Mol. Biol. Rev. **65**:44-79 (2001).

[10] H. Brim et al., Engineering *Deinococcus radiodurans* for metal remediation in radioactive mixed waste environments. *Nature Biotechnology* **18**:85-90 (2000).

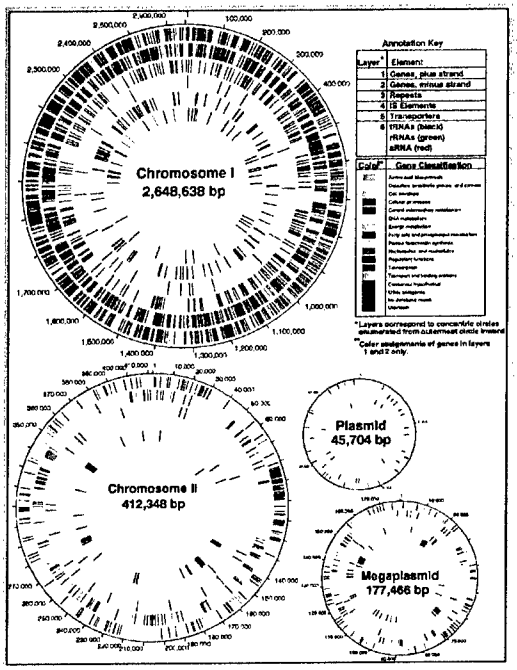
The *D. radiodurans* R1 genome (3.2 Mbp) consists of two chromosomes and two episomes, a normal-sized plasmid and much larger 'megaplasmid'. Depending on the chromosome or episome, between 81% and 91% of the DNA codes directly for protein; the rest consists of either control sequences or regions containing repeat content, ranging from 1.4% to 13%:

JASON 2003

The *Deinococcus radiodurans* genome

Table 1. General features of the *D. radiodurans* genome.

Molecule	Length	Average ORF length (bp)	Protein coding regions	GC content	Repeatcontent
Chromosome I	2,648,638	913	90.8%	67.0%	1.8%
Chromosome II	412,348	1,044	93.5%	66.7%	1.4%
Megaplasmid	177,466	1,100	90.4%	63.2%	9.2%
Plasmid	45,704	928	80.9%	56.1%	13.0%
All	3,284,156	937	90.9%	66.6%	3.8%



Choosing the *Deinococcus* barcode site

Site must tolerate insertion without altering phenotype, and be genetically stable.

Top choice: **Intergenic regions**

We select candidate sites by:

- Identification of phage attachment sites
- Bioinformatic analysis of the sequence

Selecting the Barcode Insertion Site

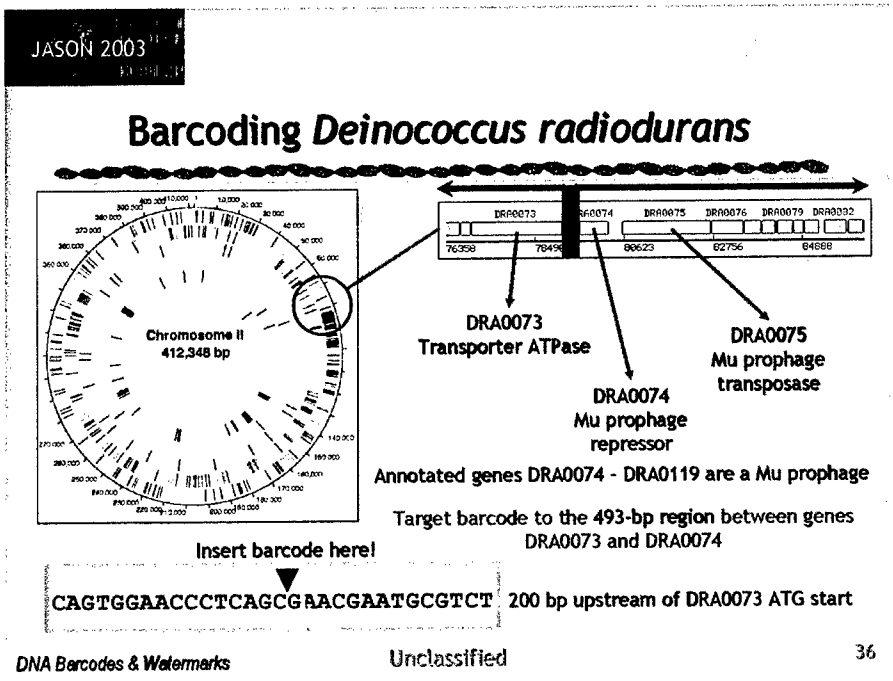
The bioinformatic content of *D. radiodurans* genome offers ample opportunity for barcode insertion. Candidate locations include both phage attachments sites and intergenic regions, as discussed. A lysogenic prophage has been identified in the *D. radiodurans* sequence

from the L5 myocobacteriophage family whose excision genes are already known[11], as well as a complete mu prophage[12]

A promising site for barcode insertion would be in the region immediately adjacent to the insertion site of the resident mu prophage (illustrated in the graphic below). Between gene DRA0074, which codes for the mu repressor, and gene DRA0073, which codes for a transporter ATPase used by *R. radiodurans*, lies a short, 493-bp region that is almost certainly noncoding. This supposition is reinforced by the fact that the ATPase and repressor genes are transcribed with opposite polarities, as shown. Selecting a specific location comfortably within in this region, we pick a site between the C and G bases in the intergenic sequence

...AACCCTCAGC↓GAACGAATGC...

located 200 bp upstream of the start codon of DRA0073.



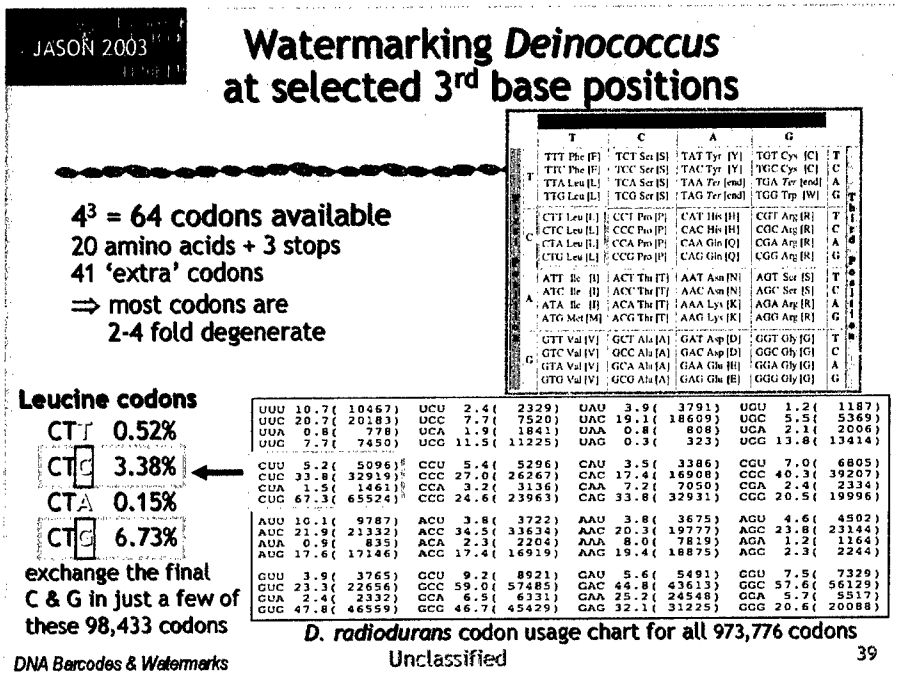
Choosing the Watermark Sites

To watermark *D. radiodurans*, one takes advantage of the intrinsic degeneracy of the 3-letter amino acid code. There are $4^3 = 64$ possible 3-letter codons, but only 20 amino acids. Three codons code for stop sequences (TAA, TAG, and TGA), leaving 61 codes for amino acids (i.e., 61-20 = 41 codes are 'extra'). As a result, all but two amino acids (methionine and tryptophan) have more than one corresponding codon, and some have as many as 6 (serine). We

[11] G.F. Hatfull and G.J. Sarkis. DNA sequence, structure and gene expression of mycobacteriophage L5: a phage system for mycobacterial genetics. *Mol. Microbiol.* **7**:395-405 (1993)

[12] G.J. Morgan et al. Bacteriophage mu genome sequence: analysis and comparison with mu-like prophages in *Haemophilus*, *Neisseria* and *Deinococcus*. *J. Mol. Biol.* **317**:337-359 (2002).

begin by consulting the codon usage table for this organism: the chart below shows both the proportion and number of each of the *D. radiodurans* codons. There are 973,776 codons in all, and 105,000 (~11%) of these happen to code for the amino acid leucine. There are four possible codons for leucine (CTT, CTC, CTA, CTG). Of these, two are used very infrequently (CTT, CTA), and the remaining two account for 94% of all the leucine codons (CTC, CTG).

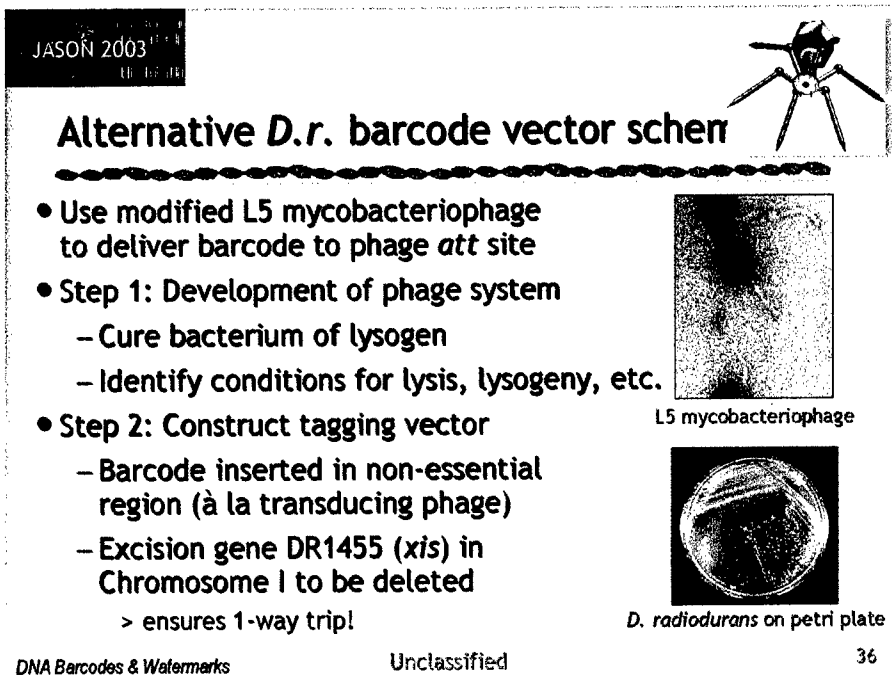


To introduce a watermark, one simply changes selected CTC codons to CTG, or *vice versa*, by transposing the identity of the base found in the third position (this is often referred to as the "wobble base"). Almost any two frequently-used codons with comparable usage statistics that code for the identical amino acid should work for this purpose, of course. Additional candidates include CCC⇔CCG (proline), GAA⇔GAG (glutamate), GCC⇔GCG (alanine), GTC⇔GTG (valine), and TTT⇔TTC (phenylalanine).

An Alternative Barcode Scheme

The discovery of the complete genome for the bacteriophage L5, found imbedded within the DNA of *D. radiodurans* after it was sequenced, presents an opportunity to explore the alternative possibility of introducing a DNA barcode using the phage insertion mechanism discussed earlier. In broad strokes, one proceeds more-or-less as follows. First, it will be necessary to establish the phage-host system of *D. radiodurans* and L5, defining appropriate physiological conditions under which the phage becomes either lytic or lysogenic. Then, one 'cures' the bacterium of its lysogen, re-generating the bacterial chromosome with a pristine phage attachment site (L5 *att*), but free of the phage sequence. The L5 bacteriophage would then need to be modified appropriately to produce a barcode 'tagging' vector. The excision gene critical to the phage lytic cycle (the L5 *xis* gene, which corresponds to open reading frame DR1455 of the *D. radiodurans* sequence) would need to be deleted or mutated to ensure a 'one-

way' trip into the host chromosome. Also, the barcode sequence would need to be inserted somewhere into the phage chromosome, presumably at a location that is non-essential (possibly within the disrupted *xis* gene, for example). This overall process would be quite analogous to that involved in making a generalized transducing phage, such as λgt11, and inserting the barcode sequence as the payload.

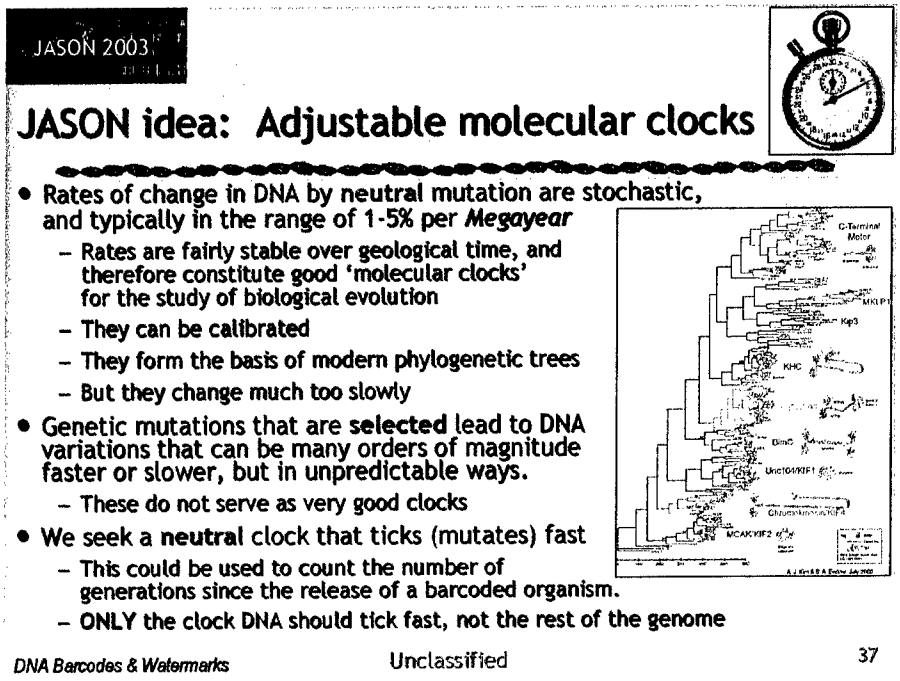


A JASON Idea: Fast, Adjustable Molecular Clocks

Beyond DNA barcodes and watermarks, which constitute fixed sequence tags for tracking and identifying organisms, it might be useful to imbed manmade DNA sequences in organisms for additional purposes. One intriguing possibility entertained by JASON was to insert a special DNA sequence that (somehow) mutated harmlessly, but at a much faster rate than the normal genomic DNA. In principle, if the rate at which such a sequence accumulated mutations were sufficiently high, it might be possible to score the number of mutations acquired, and thereby to estimate the number of generations that had passed since the organism was created and released. In effect, this hyper-mutating DNA sequence would function as a stochastically ticking, molecular timer.

Slower molecular 'clocks' based on DNA sequence comparisons have long been used to establish phylogenetic trees relating organisms on evolutionary time scales. Such comparisons form the experimental basis for modern molecular approaches to evolution, in fact. However, the rate at which natural sequences diverge by a process of neutral mutation is quite slow, about 1% to 5% per million years. Sequences diverge more quickly when subjected to selection pressures, but they tend to do so in non-random and unpredictable ways. To compute generation number on a timescale of months-to-years using information from sequence divergence would require that the DNA mutate several orders of magnitude faster than is normal. Furthermore,

these high rates of mutation would need to be confined to the clock sequence region, allowing the rest of the organism to mutate normally.



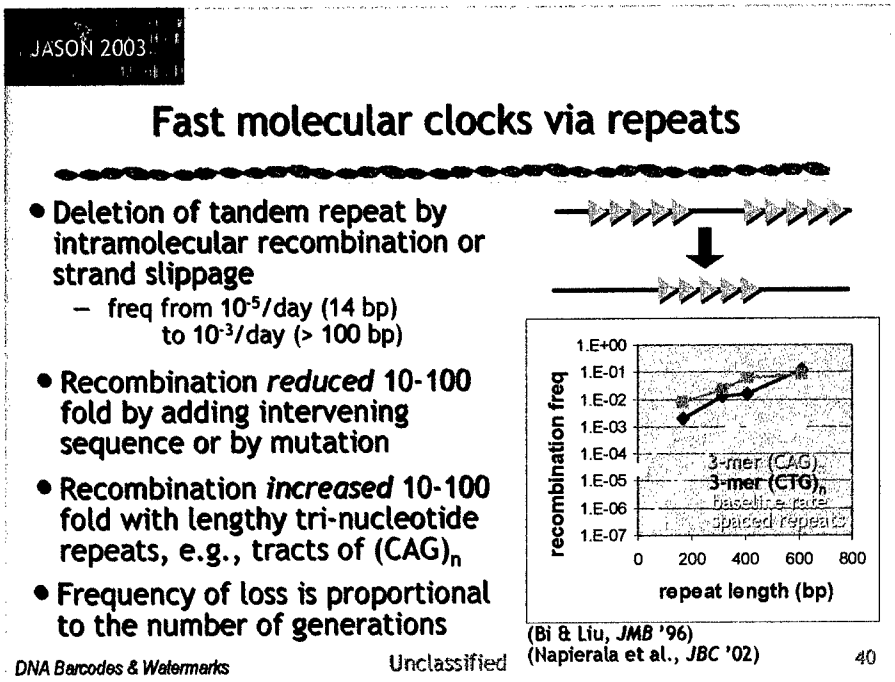
Could some kind of hyper-mutating sequence be developed based on current biotechnology? And could the accelerated mutagenesis rates be confined as required within the special sequence? Recent experimental findings suggest that such a thing may indeed be possible, and we present two approaches that offer some degree of promise along these lines.

A Clock Based on Tandem Repeats

First, it may be possible to take advantage of the fact that intermolecular recombination occurs much more rapidly between homologous regions of DNA than between unrelated regions. Specifically, recombination rates have been found to increase from 10-100 fold within direct stretches of short sequence repeats[13]. DNA rearrangements by direct repeats represent a leading cause of genomic instability in *E. coli*[14]. Lengthy direct repeats of the trinucleotide sequence $(CTG)_n$, with $n = 165$, have a recombination rate 60-fold higher than shorter stretches with $n = 17$. [Trinucleotide repeat sequences are of particular experimental interest because these also appear in certain human genetic diseases, such as myotonic dystrophy.] Furthermore, some short sequence repeats appear to be more active in promoting intramolecular recombination than others, with long stretches of $(CTG)_n$, in particular, presenting hot spots for genetic recombination. It might be possible to place such hot spots into a bacterial chromosome for the purpose of introducing a molecular clock. However, it is by no means clear whether such

[13] M. Napierala et al. Long CTG·CAG repeat sequences markedly stimulate intermolecular recombination. *J. Biol. Chem.* **277**:34087-34100 (2002).
[14] X. Bi and L.F. Liu, A replicational model for DNA recombination between direct repeats. *J. Mol. Biol.* **256**:849-858 (1996).

introduced sequences would have the desired properties. A considerable amount of exploratory work would be required to determine the actual levels of instability of specific repeat sequences (Do they mutate fast enough? Are they lost as well as gained?) as well as any possible side effects (Do they affect DNA repair mechanisms or other functions vital to the cell?) before they could be considered as candidate for a molecular clock.



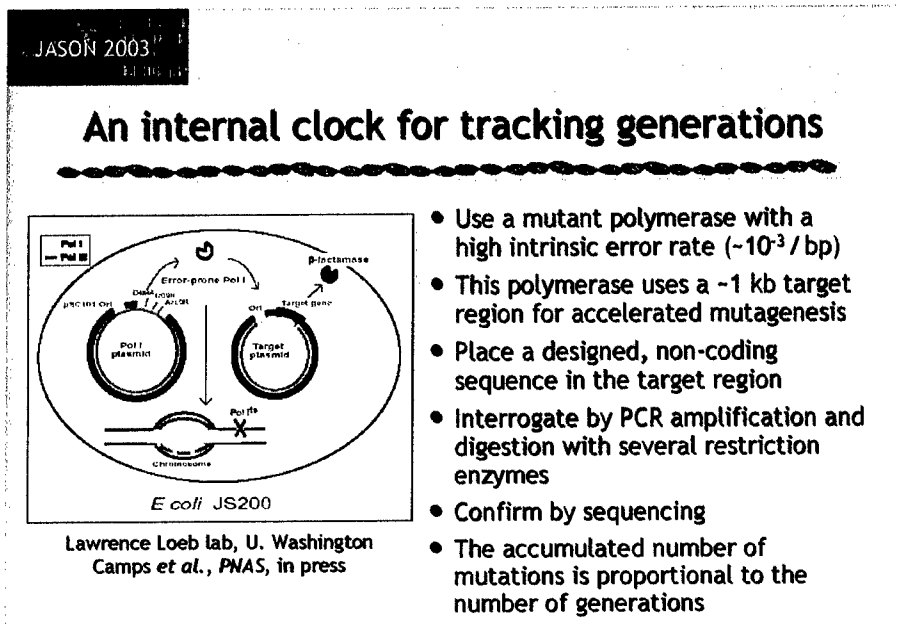
A Better Clock Based on a Directed, Error-Prone Polymerase

Perhaps a better possibility for developing a fast molecular clock was presented by the recent discovery by L. Loeb and coworkers[15] (Univ. Washington) that an error-prone mutant of polymerase I (Pol I) will introduce mutations at a high rate into a specific, predefined target region in *E. coli.* The measured rate of introduction of errors is particularly well-suited for clock purposes, because it is ~8 x 10^{-4} mutations/bp, i.e., every generation will introduce approximately one mutation into each kilobase of the target DNA region. Moreover, mutations were found to be quite evenly distributed throughout the target zone. The target is defined by its location on a plasmid downstream from a specific type of origin of replication (ColE1 *Ori*), which differs from the origin found on the main chromosome, and also from the origin found on the plasmid vector encoding the mutant Pol I. The plasmid region affected by mutagenesis can be up to 4 kb long or more (which was a surprising finding, since it was previously believed that replication by a different, more faithful polymerase, Pol III, should begin to take over from Pol I after 400-500 nt from the origin). Evidently, neither the host chromosome nor the plasmid with the non-ColE1 origin suffers an enhanced rate of mutation, because Pol I is not involved in the replication process for either of these DNA molecules.

---

[15] M. Camps et al. Targeted gene evolution in *Escherichia coli* using a highly error-prone DNA polymerase I. *Proc. Natl. Acad. Sci. U.S.A.* **100**:9727-9732 (2003)

A molecular clock based on this approach would work roughly as follows. The gene for the mutant pol I would be introduced into the species of interest, perhaps on a plasmid vector. An engineered, non-coding target sequence DNA would be introduced on a small, self-replicating plasmid containing the ColE1 origin of replication. In addition, the target sequence could be flanked by convenient PCR primer regions, just as barcode sequences are, to facilitate their subsequent amplification and interrogation.



Barcode and Watermark Considerations

DNA barcodes and watermarks offer a unique, biotechnology-based approach for following the growth, dispersal, transport, and the ecological relationships of tagged bacterial species. Clearly, an ability to track and monitor species using molecular genetics will not solve the problem of bioremediation in any way. However, barcoding should be viewed as an important adjunct to this process. Regardless of whether organisms used for bioremediation purposes are naturally-occurring or genetically-enhanced, and regardless of whether such organisms are confined tightly within holding areas or allowed to roam freely in the environment, it will be vital to follow their progress. *Unexpected things can, and do, happen.* Examples include unwanted releases of organisms into certain environments, unanticipated population spikes or crashes, the development of new mutations, lateral gene transfer, undesired effects of selection pressures, emergence of new variants and species, new symbioses, and a host of other possibilities. It therefore seems prudent to develop a means of closely following the progress of bioremediation at the level of the very organisms involved.

Barcodes not only offer a powerful new way to track the ecology and species diffusion of microorganisms, but also introduce a useful level of transparency into the bioremediation process itself. This is desirable on many grounds, including for regulatory compliance and from a public

relations perspective. Barcodes can supply data that would next-to-impossible to obtain by alternative means. Finally, the combined use of barcodes with watermarks could potentially offer a level of legal protection for the DOE, for example, against false accusations that microorganisms used for bioremediation purposes had somehow escaped and contaminated other environments (since the origins of any tagged organisms could be unambiguously traced).

Political Considerations

Hester Prynne, with A tag

- **Pros**
  - Supplies 'species labels' for scientific tracking
    - > Growth, dispersal, transport, niches, blooms/die-outs
    - > Monitor rates of gene transfer, mutation, & evolution
  - Assists in regulatory compliance (Sunshine Laws)
    - > Provides open evidence of species/genetic diffusion – or lack thereof
    - > Detailed information on release available (type, date, origin) via the open database of serial numbers
    - > Supplies ground truth – figuratively and literally!
  - Liability reduced against any falsification of origins
    - > Watermarks provide protection
- **Cons**
  - The stigma of genetically-modified organisms (GMOs)
  - Irony: Introducing a barcode tag alone renders the organism 'genetically engineered.'
- **Bottom line:** ***We believe the tradeoff is worthwhile.***

The irony here, which was not lost on JASON, is that the act of introducing a barcode into an otherwise wild-type species constitutes, *in-and-of itself*, a form of genetic modification—even though the tag is designed purely for informational purposes, and produces no change in the behavior or fitness of the organism whatsoever. Therefore, *all* barcoded organisms, regardless of their phenotype, technically become GMOs, to which a certain stigma is attached in the politics of the day. However, we believe that the tradeoff in tagging otherwise wild-type organisms used for intrinsic or augmented bioremediation is worthwhile, provided it is done correctly, i.e., that no phenotype is truly conveyed. The GMO issue is moot for organisms that have already been engineered specifically for enhanced bioremediation purposes, because these have been genetically modified in any case.

Could barcoding (or watermarking) an organism introduce secondary effects that pose additional risks for the environment? For example, would barcoded organisms, despite — by virtue of their construction — initially bearing no phenotype to distinguish them from the wild type, become more susceptible to certain types of natural mutation, and possibly acquire undesirable properties after release? Or, more generally, could barcoded organisms behave or interact or evolve *any* differently than their unlabeled counterparts in the real world, producing unforeseen consequences? This sort of question is best addressed by carefully controlled experimentation (under realistic settings) and practical experience, and not by informed speculation based on theory or laboratory findings alone. Indeed, similar considerations apply to

the deployment of any form of bioremediation strategy, regardless of the organisms involved. The JASON recommendations (found in the final section of this report) suggest taking a staged, deliberate, and cautious approach to such issues, including the development of testbed systems and performing a careful series of feasibility studies.

Barcodes and Bioremediation

Properly implemented, the barcoding of microorganisms could make a significant, positive contribution to DOE's nascent bioremediation efforts. The cleanup of toxic wastes from our federal lands, and from the National Labs in particular, has now assumed a high priority. By any standard the job is enormous, and will no doubt require innovative approaches that challenge our current technology. "Classic" approaches to bioremediation — intrinsic or enhanced bioremediation, using those microorganisms found *in situ* — will likely need to be supplemented by next-generation bioaugmentation agents that have been genetically engineered and optimized to the task at hand. It will be vital to keep careful inventory of all microorganisms involved in the bioremediation process (using engineered microbes or otherwise). In JASON's opinion, barcode tagging affords the most practical, robust and reliable approach to this task. DOE's strong commitment to bioremediation and biotechnology development was underscored by recent testimony of Undersecretary Robert Card:

> "In FY04, the Office of Biological and Environmental Research will continue to explore new clean-up strategies, including bioremediation and treatment of radioactive wastes. The goals of the Environmental Management Science Program, transferred in FY 2003 from Environmental Management, are to develop and validate technical solutions to complex problems, provide innovative technical solutions where there are none, and lead to future risk reduction and cost and time savings."[16]

In addition, DOE has an interest in exploring the potential of microorganisms to affect the balance of global carbon cycling, thereby reducing greenhouse gases and further stabilizing against climatic change. This undertaking has been described by DOE Secretary Spencer Abraham:

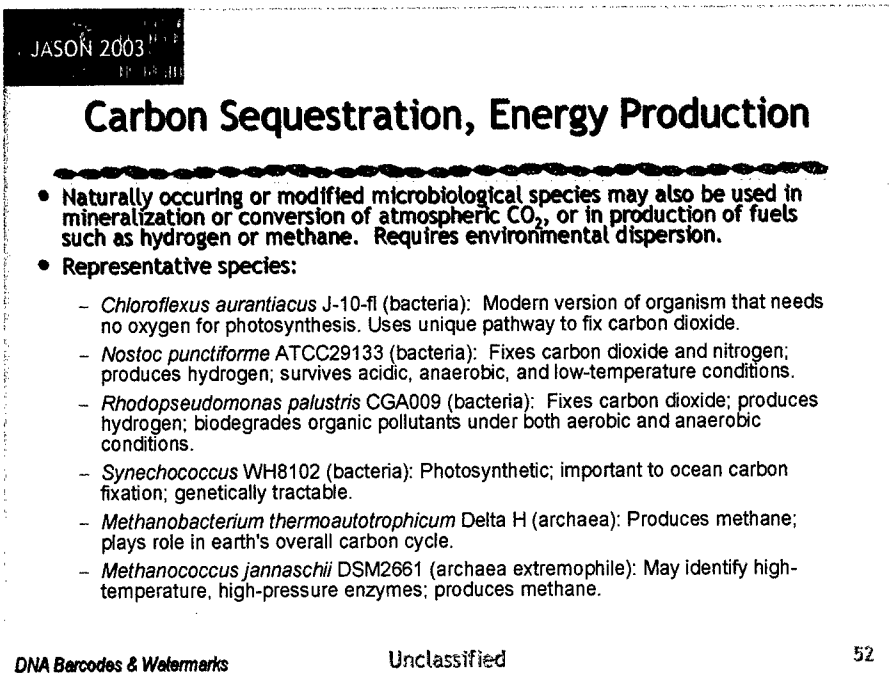
> "In FY 2002, the Global Climate Change program will conduct research designed to reduce uncertainty in predicting the effect of greenhouse gases on future climates. Carbon cycle and sequestration research will help to assess current carbon sinks and to develop methods of enhancing natural processes for terrestrial and ocean sequestration of carbon."[17]

---

[16] Testimony of Robert Card, Under Secretary of Energy, to the House Committee on Science, Feb. 13, 2003. http://www.energy.gov/engine/content.do?PUBLIC_ID=13616&BT_CODE=PR_CONGRESSTEST&TT_CODE=PRESSSPEECH

[17] Statement of Spencer Abraham, Secretary of Energy, before the House Committee on Appropriations, Subcommittee on Energy and Water Development, May 2, 2001. http://www.energy.gov/engine/content.do?PUBLIC_ID=13956&BT_CODE=PR_CONGRESSTEST&TT_CODE=PRESSSPEECH

Here again, the deployment of microbial species will likely involve their dispersion in the environment, necessitating some means of tracking and monitoring organisms. Finally, the DOE is exploring the potential of microbes for possible use in the economical production of fuels, such as methane and hydrogen, for direct solar energy conversion, using photosynthesis-linked biochemistry, and for biomineralization. Once more, genetic barcoding would seem to offer the most practical, robust and reliable means of tracking and monitoring the relevant microbial species.



JASON Recommendations

We conclude that a program for barcoding the microorganisms used in bioremediation is not only feasible, but advisable. JASON feels that the DOE should consider the establishment of an exploratory program for barcoding and watermarking bacteria, for eventual deployment in conjunction with its ongoing bioremediation efforts. Such a program could investigate the practical feasibility of implementing a barcoding system, fill in whatever gaps that remain in the science needed for such a system, and lay the regulatory groundwork needed for establishing future barcoding standards and practices. Not only would barcoded organisms be useful in the immediate context of bioremediation on contaminated DOE lands, but they may eventually have a variety of other uses. In principle, barcoded organisms may offer novel ways to:

- Track the distribution of microorganisms used for toxic cleanups such as oil spills
- Follow the spread of GMOs in our ecosystem
- Monitor the dispersion of microbes in the biosphere for scientific purposes (driven, for example, by ocean currents or by the atmosphere)
- Diminish biowarfare and bioterror threats, by making possible attribution through genetic tagging of the pathogens involved.

Success by the DOE in the bioremediation arena may therefore encourage the use of barcodes for some of these other purposes. We therefore urge the DOE to consider some of the following steps towards the implementation of a comprehensive barcode system:

- Look into ways of adapting existing biotechnological instrumentation for use in key aspects of barcode synthesis, insertion, and readout. Develop specific hybridization arrays (DNA chips) for conventional barcode readout and for barcode readout accompanied by watermark readout.
- Sponsor applied research into adapting site-specific recombination methods to insert barcodes and watermarks into bacterial genomes, including approaches based on (1) homologous recombination, (2) 'retrohoming' insertion by Group II intron vectors, and (3) excision-deficient bacteriophages. Sponsor basic research aimed at identifying alternative and improved methods.
- Establish a working group to formulate uniform standards for microbial barcoding, which would standardize the database structures, as well as specify rules for the coding, design and disbursement of the barcodes and watermarks used.
- Perform feasibility studies to measure the actual stability of barcodes and watermarks introduced by any of the various methods, their effect (if any) on natural fitness, their mutation and loss rates, etc.
- Develop a barcoding and watermarking 'testbed' program using one or a few model organisms, which would be barcoded, watermarked, grown up, and monitored in a trial release program (performance monitoring in a microcosm study).
- If the above developments prove successful, implement DNA barcoding and watermarking standards for performance monitoring of all bacterial and fungal strains used in DOE programs, and promote the global use of barcoded organisms in bioremediation.

We believe that this represents a leadership opportunity for the DOE.

JASON 2003

JASON Recommendations I

- DNA barcodes and watermarks are A Good Thing
- Look into ways of adapting existing instrumentation for barcode/watermark fabrication, insertion, and readout
- Adaptation of biotechniques
  - Barcode and watermark insertion by:
    - > homologous recombination
    - > retrohoming introns (group II)
    - > excision-deficient phage
  - SNP readout by MIPs
  - Explore alternative insertion and readout approaches
  - Fast mutational clock development
- Forge useful standards
  - Design the prototype barcode and watermark database
  - Form group to establish uniform standards for barcoding (ISO...)
  - Formulate rules for 'clean tagging'
    - > E.g., barcoded strains should not contain superfluous markers

JASON 2003

JASON Recommendations II

- Perform feasibility & stability studies
  - Measure the impact on natural fitness
    - > competition expts. in fermenters, microenvironmental trials, etc.
  - Confirm barcode stability
  - Monitor mutation rates
- Develop a trial barcoding and watermarking program (microcosm)
  - Tag a model organism
  - Grow, study, trial release
  - Readout
- Explore further uses of barcodes
  - Microbes for remediation of oil slicks, EPA Superfund sites
  - American Type Culture Collection (ATCC) strains
  - CDC Select Agents, etc.
- This is a leadership opportunity for DOE

3. APPENDICES

APPENDIX A: Construction and Analysis of DNA Barcode Libraries

In this appendix we provide support for our claim that it is possible to construct libraries of DNA Barcodes which will:

(a) supply distinct barcodes for tens of thousands or more types of organism;
(b) provide unique identification even after suffering a small number of mutations;
(c) be compatible with current DNA readback technology;
(d) correspond to a genetically biocompatible sequence, once inserted in the genome.

Construction of a Barcode Library

We begin by describing a simple procedure for building a library of good barcodes. It has four steps:

(a) *Specify Parameters.* We pick n, the number of base pairs in the sequence, and m, the number of mutations we can accommodate while maintaining unique identification. For concreteness, think for now of $n = 20$ and $m = 5$. In the appendix, we will use e in place of m; think of "the number of errors we can tolerate while correctly identifying the original organism."

(b) *Obtain an Error-Correcting Code.* We obtain an existing Error-Correcting Code (ECC) over Z mod 4 compatible with our parameters. ECCs (see [Pless]) are usually described in terms of triples $[n,k,d]$, where n is the number of letters in a codeword, $d = 2e$ or $2e+1$ is called the minimum distance of the code, and k is an index of the size of the library. In our setting of 4-symbol alphabet (AGCT), an $[n,k,d]$ ECC will then have at least 4^k codewords, where each codeword is at least d-distant from every other codeword. There are only certain k's possible for a given n and d; generally speaking, for a given n, the number of codewords is 4^n when $d = 0$ and decays rapidly with d. There is a large experience in the ECC literature showing what $[n,k,d]$ triples are possible and how to build concrete codes attaining a specific $[n,k,d]$ triple.

(c) *Remove Sequences with Bioincompatible G:C content.* Starting with an $[n,k,d]$ ECC, we then look at each codeword in the code and remove those codewords where the G+C fraction lies outside the limits [0.45,0.55]. Biologically, this means we are removing sequences that depart significantly from the normal G:C content of most eubacteria, which is typically close to 50%. As a practical matter, artificial sequences synthesized with normal G:C content tend to be more stable when inserted into the genome. The operation of enumerating the codewords in an ECC is straightforward given the so-called generator matrix of the code and the operation of measuring the G+C fraction is, of course, trivial.

(d) *Remove Sequences Prone to Hairpin Formation.* We once again enumerate the codewords, identifying those sequences which may have a tendency to form 'hairpins' (stem-loop regions) when unzipped. A hairpin occurs where a sequence contains a short block which is matched elsewhere on the same strand of the sequence by a complementary sequence that could base pair with it when the DNA is in single-stranded form and can fold back on itself. The complementary regions must be physically separated by a short loop region that constitutes the bend in the hairpin, typically 3 or more unpaired bases. The operation of

checking for hairpins is nontrivial and involves applying energy-minimization programs that compute the possible DNA secondary structure, like Primer3 or M-Fold. We remove hairpin-prone sequences from the library.

The codewords remaining after the above steps form our proposed barcode library.

Heuristic Analysis of Barcode Library Size

The central question is of course: will any codewords remain after the above winnowing process; and if so, how many will remain? We developed a heuristic calculation that has been validated through computer experiments. This was called (tongue-in-cheek) a "Drake Equation" in the body of the Report. Although it has nothing to do with Frank Drake's original equation to estimate the number of planets hosting life elsewhere in the universe, it *does* involve multiplying together several unrelated probabilities to form a crude estimate. The equation takes the form

$$N_{\text{usable}} \approx 4^n \, (F_{ECC} \times F_{G:C} \times F_{No2}) \, . \qquad \text{(A.1)}$$

Here the term on the left gives the number of codewords remaining in our library after winnowing, while each of the terms in parentheses on the right side is a fraction between 0 and 1. The approximate equality sign $\approx$ should for now be interpreted as saying that the two terms have similar orders of magnitude.

The individual fractions F on the right side all depend implicitly on the specific ECC we start from and the $[n,k,d]$ parameters that are involved. Thus, F_{ECC} is defined as 4^{k-n} for an $[n,k,d]$ code, so that $4^n F_{ECC} = 4^k$, the number of codewords at mutual distance at least d in our code. $F_{G:C}$ is the fraction of these sequences passing our melting test; empirically, this number is about 49% for codes of length 20. F_{No2} depends on the definition of hairpin-prone actually used; different standards are proposed in the programs Primer3 and M-Fold. If we take M-Fold as the authoritative source for hairpin testing, a test of many randomly-generated 20-mers showed that about 30% are NOT hairpin-prone.

Putting all the terms together, in the case $n = 20$, we note that there exist [20,8,10] codes, and we get, taking M-Fold as authoritative:

$$N_{\text{usable}} \approx 4^n \, (F_{ECC} \times F_{G:C} \times F_{No2}) \; = 4^8 \times 0.49 \times 0.3 = 256\text{K} \times 0.15 \approx 37.6\text{K}$$

Hence, the equation gives a dictionary of many thousand viable DNA barcodes, exhibiting

(a) fair error-correcting properties – correcting up to 4 errors, detecting up to 5 errors
(b) good melting properties – always G+C fraction between 0.45 and 0.55
(c) hairpin avoidance – always passing either the M-Fold or Primer3 tests

We focused here on $n = 20$, and $e = 4$ and got thousands of viable barcodes.

The same heuristic can be applied at slightly larger n, or slightly smaller e, giving a rapid increase in the number of viable barcodes. Thus, if we reduce the error tolerance to correcting 3 and detecting 4 errors, we get that at $n = 20$, $d = 8$, there are [20,10,8] codes, and so, taking M-Fold as authoritative:

$$N_{usable} \approx 4^n \ (F_{ECC} \times F_{G:C} \times F_{No2}) \approx 4^{10} \times 0.49 \times 0.3 = 1M \times 0.15 \approx 157{,}000$$

On the other hand, if we retain $e = 5$ but grow the word length to $n = 24$, we get that there exist [24,10,11] codes, and so, taking M-Fold as authoritative:

$$N_{usable} \approx 4^n \ (F_{ECC} \times F_{G:C} \times F_{No2}) \approx 4^{10} \times 0.46 \times 0.3 = 1M \times 0.14 \approx 145{,}000.$$

Either way, the heuristic tells us that for n slightly larger than 20 and with a mutation tolerance of around 5, we can construct suitably large libraries of viable DNA barcodes.

Justifying the 'Drake Equation' Heuristic

We now describe the basis for our heuristic formula. Let x denote a string of length n over the alphabet {AGCT}. Let C denote the collection of 4^k codewords x in a given $[n,k,d]$ code. Let X denote the result of sampling x uniformly at random from the universe of 4^n possible strings. Consider the following events:

E_{ECC} = {X belongs to codebook C}
$E_{G:C}$ = {X has between 45% and 55% G+C ratio }
E_{No2} = {X forms no hairpins (according to M-Fold)}

A sequence of letters AGCT sampled at random from the universe of 4^n possible strings will be viable as a DNA barcode if all events E_{ECC}, $E_{G:C}$, E_{No2} occur on that draw.
The correct equation for our setting is therefore

$$N_{usable} = 4^n \, P(E_{ECC} \cap E_{G:C} \cap E_{No2}) \qquad (A.2)$$

That is, we multiply the universe size 4^n by the probability of sampling a string at random and having all three properties (ECC, G:C, and No Hairpins) occur simultaneously. Now *if* the events E_{ECC}, $E_{G:C}$, and E_{No2} were statistically independent, we would be correct in rewriting this display as follows:

$$N_{usable} = 4^n \, P(E_{ECC})P(E_{G:C})P(E_{No2}), \qquad (A.3)$$

After making the substitutions

$$F_{ECC} = P(E_{ECC}) \ ; \ F_{G:C} = P(E_{G:C}) \ ; \ F_{No2} = P(E_{No2}),$$

we recover from (A.3) the Drake equation (A.1) used above. (Actually, we don't believe in strict independence, hence the approximate equality sign ≈ when we first wrote the Drake equation.) In short, our Drake equation is based on the assumption of approximate stochastic independence of the different events properties (ECC, G:C, and No Hairpins) under uniform random sampling.

To quantify approximate independence, rewrite the correct probability as

$$P(E_{ECC} \cap E_{G:C} \cap E_{No2}) = P(E_{ECC}) \cdot P(E_{G:C} \mid E_{ECC}) \cdot P(E_{No2} \mid E_{G:C} \cap E_{ECC}).$$

Define now

$$A = P(E_{G:C} \mid E_{ECC}) / P(E_{G:C}) \; ; \quad B = P(E_{No2} \mid E_{G:C} \cap E_{ECC}) / P(E_{No2}); \tag{A.4}$$

then

$$P(E_{ECC} \cap E_{G:C} \cap E_{No2}) = A \cdot B \cdot P(E_{ECC}) \cdot P(E_{G:C}) \cdot P(E_{No2}).$$

In short, the Drake equation is accurate if A and B are both close to 1. We will give evidence on the size of A and B farther below.

Probabilities in the Drake Equation.

Before studying the accuracy of the Drake equation, we review the basic elements involved in applying the Drake Equation. It involves 3 probabilities

$P(E_{ECC})$ is simply the ratio #{codewords in the given ECC}/4^n . If the given ECC is of type $[n,k,d]$, $P(E_{ECC}) = 4^{k-n}$. Sets of triples $[n,k,d]$ for which ECCs exist can be found in numerous places on the web and in books such as [Pless]. The table below (from [LinCode]) gives the best known linear codes at a given n, d combination

N	k	d
20	10	8
20	9	9
20	6	11
24	12	9
24	11	10
24	10	11
30	18	9
30	17	10
30	15	12

Table A.1: Some combinations of $[n,k,d]$ for which linear codes over GF(4) exist.

Analysis of such tables suggests that, throughout the range $20 \le n \le 30$, there exist ECCs so that with each given n and specified d=11 or 9, we have approximately

$$P(E_{ECC} \mid n,\, d = 11) \approx 4 \times 10^{-9} \,, \; P(E_{ECC} \mid n,\, d = 9) \approx 6 \times 10^{-8}.$$

The point of course is that these approximate formulas do not depend on n.

$P(E_{G:C})$ is simply the fraction of all 4-letter sequences with $0.45 \le (\#G+\#C)/n \le 0.55$. The answer depends on n and is given in the following table:

N	Min G+C	Max G+C	$P(E_{G:C})$
20	9	11	0.4966
22	10	12	0.4765
24	11	13	0.4587
26	12	14	0.4428
28	13	15	0.4284
30	14	16	0.4153

Table A.2: Fraction of all 4-letter sequences of length n having acceptable G+C ratios

Note that the probabilities depend on n, but only weakly.

The indicated probabilities are simply obtained by summing entries in the binomial table. Indeed the probability in question is just the chance that the number of 1's in a random binary sequence lies between $0.45n$ and $0.55n$. Using the binomial tables $p(k,n)$, we simply sum $p(k,n) = \text{Choose}(n,k)2^{-n}$ over the range $0.45\ n \leq k \leq 0.55\ n$, getting the values published above. Note: the entries for odd n are about 2/3 as big as adjacent entries for even n, which suggests we always *stick with even n in designing a barcode*, thereby gaining about a 50% increase in dictionary size.

$P(E_{No2})$ is obtained by applying a computer program for hairpin testing. Two programs now in wide use are Primer3 [Primer3] and M-Fold [M-Fold]. We applied M-Fold to randomly-generated strings of length n=20 and found that about 30% passed all the tests in M-Fold. Apparently, a smaller fraction of random strings at larger n would pass M-Fold tests, but the fraction decays slowly with n. We independently applied an algorithm implementing the Primer3 definition of hairpin-proneness and found that only about 10% of random strings of length 20 are hairpin-prone. Since M-Fold is significantly more pessimistic, we take M-Fold as authoritative for our design work.

Combining the elements above, we find that the Drake equation tells us that, at n =24 and e =5, we can develop a barcode dictionary containing roughly

$$N \approx 4^{24} \times (4 \times 10^{-9}) \times (0.4587) \times (0.3) \approx 150\text{K}$$

viable barcodes, allowing to correct for up to 5 errors in readout. The graph found in the figure entitled "Barcode Drake Equation (III)" on p. 20 of the main text illustrates the range of N possible as we vary n, d.

Justifying Approximate Stochastic Independence

We now explain how the heuristic behind the Drake equation can be justified. In an earlier subsection, we defined two quantities, A and B, whose product gives the ratio between the correct value of N according to (A.2) and the heuristically calculated value of N according to (A.1). We conducted numerous simulation experiments to justify that A and B are each what physicists call "O(1) terms".

In the first experiment, we studied $A = P(E_{G:C} \mid E_{ECC}) / P(E_{G:C})$. We enumerated codewords from an error-correcting code, and evaluated the G+C ratio for those elements. We calculated the fraction of

such codewords obeying $0.45n \leq \#(G+C) \leq 0.55n$. Our results were as follows. While for $n = 20$, $P(E_{G:C}) = 0.4966$ as in Table A.1 above, we found that for a linear $n = 20$, $d = 8$ code, $P(E_{G:C} \mid E_{ECC}) = 0.4608$. In short, $A \approx 0.928$. Other experiments gave similar results.

This experiment can be interpreted as follows. While an ECC is of course a highly-non random object, in certain statistical measurements, sampling from an ECC is very much like sampling a random sequence. Since $P(E_{G:C} \mid E_{ECC})$ names a simple characteristic of sampling from an ECC while $P(E_{G:C})$ names the same simple characteristic of random sampling, it is not surprising that A is near one.

In the second experiment, we studied $B = P(E_{No2} \mid E_{G:C} \cap E_{ECC}) / P(E_{No2})$. We enumerated codewords from an error-correcting code, and winnowed out all those where the G+C ratio was unfavorable. We applied the Hairpin-proneness algorithm used in Primer3 to the codewords remaining after winnowing. Our results were as follows. While for $n = 20$, $P(E_2) \approx 0.06$ by an easy Monte-Carlo simulation, we found that for a linear $n = 20$, $d = 8$ code, $P(E_{No2} \mid E_{G:C} \cap E_{ECC}) \approx 0.05$. In short $B \approx 1.2$. Other experiments gave similar results.

This experiment can be interpreted as before. While an ECC is of course a highly-non random object, in certain statistical measurements, sampling from an ECC is very much like sampling a random sequence. Since $P(E_{G:C} \cap E_{No2} \mid E_{ECC})$ names a simple characteristic of sampling from an ECC while $P(E_{G:C} \cap E_{No2})$ names the same simple characteristic of random sampling, it is again not surprising that B is near one.

Heuristics for understanding why only certain [n,k,d] codes can exist

The driving factor behind all the above results is $P(E_{ECC})$, as it depends most strongly on n. This is turn simply measures the size of the error-correcting code being used to generate the library, and so we effectively want to know: how big a codebook can exist for given n,d? While construction of ECC's is reviewed in detail in, for example [Pless], for the reader's convenience we review two basic results that indicate which $[n,k,d]$'s might possibly exist.

The first is the *volume bound*; it gives an upper bound on the number of codewords in a codebook with given n,d. One observes that if a code is able to correct e errors, then each viable codeword can be viewed as the center of a Hamming sphere of radius e, which contains no other viable codewords. The Hamming sphere of radius e contains about $\text{Choose}(n,e) \times 3^e$ sequences. If there are a total of 4^n potential codewords, and each viable codeword must exclude $\text{Choose}(n,e) \times 3^e$ potential codewords from viability, then the codebook size is constrained by

$$N \leq 4^n / [\text{Choose}(n,e) \times 3^e].$$

This bound shows, for example, that if e=5 and n=20, we can have at most

$$N = 4^{20}\, 3^{-5}\ (5\times4\times3\times2\times1\ /\ 20\times19\times18\times17\times16) \approx 285\text{K}$$

codewords in an e-error-correcting code. While this is an appealingly simple calculation and goes in the right direction, the volume bound is typically a gross overestimate of the potential size of a codebook.

Table A.1 shows that we know of [20,6,11] codes; these are 5-error correcting; but their size, at $4^6 = 4K$, is radically smaller than the volume bound of about 285K.

Going in the opposite direction is the *Gilbert-Varshamov lower bound.* [Pless, p. 29] states this as follows. *There exists a linear code over GF(4) of wordlength n, minimum distance of d or more, and dimension k, if*

$$3 \cdot \text{Choose}(n\text{-}1,1) + 3^2 \cdot \text{Choose}(n\text{-}1,2) + \ldots + 3^{d-2} \cdot \text{Choose}(n\text{-}1,d\text{-}2) < 3^{n-k} - 1.$$

This bound, while containing some elements in common with the volume upper bound, is substantially less optimistic about which $[n,k,d]$ triples can exist. This bound provides only a lower bound, because it speaks of existence of linear codes. Accordingly, when used with $n = 20$ and $d = 11$ (comparable to $e = 5$) the condition fails for each $k > 0$! When used with $n = 30$ and $d = 11$, it gives $k = 6$. Actually we know (as recounted in Table A.1) that there exist linear [30,16,11] codes, so the Gilbert-Varshamov bound is a gross underestimate in the range of interest.

There exist also nonlinear codes based on other principles, some of which might do better in certain situations. For the n,d of interest, we have found that linear codes and quadratic residue codes perform similarly. We have emphasized linear codes in our work, partially because for such codes, it is relatively easy to see that the stochastic independence analysis mentioned in the previous section should follow easily.

Our point in reviewing the above bounds, and in pointing out how widely they differ from each other, is to reinforce the point that there is no really simple way to see when an $[n,k,d]$ code exists, if n is small and we are interested in the very best k; such things are always determined by relatively sophisticated analysis, essentially on a case-by-case basis [Pless, LinCode].

References for Appendix A

[M-Fold] M-Fold Website: http://www.biology.wustl.edu/gcg/mfold.html

[Primer3] Primer-3 Website: http://frodo.wi.mit.edu/primer3/

[LinCode] Linear Codes Website: http://www.win.tue.nl/~aeb/voorlincod.html

[Pless] Pless, Vera (1998) *Introduction to the Theory of Error-Correcting Codes.* 3rd Edition, J. Wiley (New York).

APPENDIX B. Placing Barcodes and Watermarks by Recombination within the Genome.

The goal is to manipulate the genomes of microorganisms such that they contain a short, artificially constructed sequence tag (barcode), or several single nucleotide polymorphisms that would distinguish them from known isolates of a given organism (watermark). The chosen technique should be adaptable to essentially any microbe, and should be capable of engineering an insertion into virtually any site in the genome.

In this appendix, it is assumed that the full genome sequence of the microbe in question is available. This is already the case for many of the organisms that are useful for bioremediation, and the cost of microbial genome sequencing has decreased such that it should be the starting point for any organism for which complete sequence is not yet available. Also, the standard microbiology terminology will be used with respect to genetic manipulation; treating cells with DNA for the purpose of altering their genome is "transformation", and a cell that has received the DNA is a "transformant". It is assumed that a transformation system has been developed for the microbe in question, such that it is possible to introduce DNA into cells, and to select for transformants.

The barcode cassette consists of the 20-24 base pair (bp) barcode sequence, flanked by 20-25 bp priming sites that can used to amplify the barcode sequence by polymerase chain reaction (PCR). The total length of the barcode cassette then is 60-74 bp. Ideally, the barcode cassette would be inserted into a specific genome location without the addition of any other sequence. However, all genetic transformation methods require the use of a selectable marker to identify transformed cells. The selectable marker is typically an exogenous gene that confers resistance to an antibiotic; this marker should be removed after genetic manipulation to eliminate concerns about transfer of the marker to wild populations. An ideal method therefore would allow directed targeting of a barcode sequence to any location in a genome, and subsequent removal of the selectable marker to create a minimally altered microbe.

A. Homologous recombination: loop-in/loop-out

A plasmid bearing a segment of DNA that is homologous to a segment of the bacterial genome can recombine with the genome by a single crossover to yield a co-integrate of plasmid and genome (see p. 23). This is often referred to as "looping-in" a sequence, and is based on the Campbell model for integration of a bacteriophage genome into the host chromosome (Campbell, 1961). After the recombination event, the genome at this site consists of a direct repeat of the segment of DNA shared by plasmid and chromosome, with the plasmid sequences (including the selectable marker) between the direct repeats.

In most bacteria, circular plasmid DNA containing a replication origin transforms with high efficiency without recombining into the genome. However, if the plasmid either lacks a replication origin that functions in the microbe to be engineered, or has a conditional origin (Hamilton et al., 1989), then the desired homologous integration will be the predominant event recovered in transformants. For the purposes of engineering environmentally useful microbes, the plasmid used would likely be a derivative of a standard *E. coli* cloning vector, bearing an *E. coli* origin, but lacking an origin for the microbe to be engineered. The recombination event

between plasmid and genome can occur at any position in the homologous sequence shared by plasmid and genome, although in some species it is possible to direct the crossover point, and increase the frequency of the desired event, by making a double strand break in the homology region on the plasmid (Orr-Weaver and Szostak, 1983).

A. 1. Engineering a single-crossover loop-in by homologous recombination

The site for insertion of the barcode would be chosen using the described criteria (pg. XX). Leloup et al. (Leloup et al., 1997) found that single crossover integration frequency was directly proportional to the extent of homology between plasmid and chromosome, with a peak of $1.4x10^3$ transformants/μg DNA with approximately 1 kb of homology. A 1 kb genomic sequence surrounding the barcode insertion site would be amplified by PCR from genomic DNA, and cloned into an appropriate plasmid vector, lacking a replication origin for the microbe to be engineered. The barcode cassette, containing a unique barcode sequence chosen from among the available set, would then be cloned into the middle of the homology region on the plasmid using standard methods; the plasmid-borne homology sequence would thus differ from the genome sequence only by addition of the barcode cassette. Transformation of the plasmid into the microbe, and selection for the plasmid marker, would yield strains in which the recombination event could have occurred on either side of the barcode cassette, with the two possibilities differing only in which of the two copies of the direct repeat contains the barcode cassette after recombination. The genome structure at the site of integration in these strains would then be confirmed by diagnostic PCR.

A. 2. Identifying loop-out events

Once the strain bearing the duplication is isolated, the next step is to allow the reverse reaction to occur, effectively looping-out the plasmid sequence. This reaction is catalyzed by the same intracellular machinery that catalyzed the loop-in reaction, and it occurs with equal precision, removing the plasmid and exactly one copy of the repeated genomic sequence. Depending on the position of the loop-out recombination event within the homology sequence, which is randomly determined, the barcode sequence will be in either the plasmid copy or the genome copy of the repeated sequence after recombination. Since the plasmid that is looped-out lacks the ability to replicate in these cells, removal of selection would result in loss by dilution of the plasmid and its copy of the homology region. There are two considerations in this step: 1) the loop-out event is rare, so cells that have undergone the event must be distinguished from the majority that have not, and 2) among the loop-out cells, those that have retained the barcode in the genome must be distinguished from those that have not.

Cells that have undergone the loop-out event can be identified by loss of the plasmid selectable marker. The rate of such recombination events between direct repeats of 1 kb is about $1x10^{-4}$ in yeast cells, where this has been best characterized, and it is likely to be of the same order in many bacteria. Therefore, it should be possible to identify loop-out recombinants simply by screening for loss of the plasmid marker; typically this would be accomplished by growing for several generations under non-selective conditions, plating cells for single colonies, and replica-plating the single colonies to selective medium. The desired strain would be unable to grow on selective medium.

An alternative strategy to increase the frequency of the loop-out event makes use of a rare-cutter restriction site in the plasmid sequence to create a double strand break that stimulates recombination (Posfai et al., 1999). In this method the 18 bp recognition site for cleavage by the meganuclease I-*Sce*I is incorporated in the plasmid. The strain bearing the integrated plasmid is then transformed with a separate plasmid expressing I-*Sce*I, which cuts the genome uniquely at the plasmid site. In *E. coli*, the resulting double-strand break in the genome stimulates recombination between the flanking direct repeats. Another approach to the problem is to use a counterselectable marker on the plasmid - one that can be selected against - so that cells that have looped-out the plasmid can be selected directly. For example, expression of the *sacB* gene of *B. subtilis* in *E. coli* inhibits growth in medium containing 5% sucrose, and has been used as a counterselectable marker (Gay et al., 1985). However, the effectiveness of this selection is strain-dependent even within *E. coli* and it is not clear how well it would work in other microbes.

After identifying cells that have lost the plasmid marker by loop-out, those that have retained the barcode in the genome can be identified by a physical test of genome structure. Since the position of the loop-out recombination event in the repeated sequence is random with respect to the barcode, approximately 50% of the recombinants will be of the desired structure. Therefore the simplest method of identification would be to pick ten strains that have lost the plasmid marker and use PCR to identify those that have retained the barcode. This can be done with a single pair of diagnostic primers. Finally, the barcode cassette and surrounding DNA should be sequenced to confirm that the structure is as predicted.

A.3. Application of loop-in/loop-out to watermark insertion:

In addition to barcode insertion, the loop-in/loop-out method is directly applicable to the related problem of inserting a watermark in the genome. A watermark, as defined in this report, is a single base pair change in the genome sequence. To insert a watermark, the procedure would be the same as above, with the exception that the plasmid vector would bear a 1 kb homology region that is identical to the genome sequence, with the desired single base change in the middle of the sequence. This base change would be created by standard methods of site-directed mutagenesis. Presence of the single base change in the genome after loop-in/loop-out is more difficult to detect than a barcode cassette insertion, but can be achieved by diagnostic PCR with primers that are specific for the wt and mutant alleles (De Milito et al., 1995). Because the loop-out event precisely removes the plasmid sequences, the result is a perfect replacement of the wild-type allele with the watermarked allele.

A.4. Summary of steps for integration of barcode or watermarks by homologous recombination:

1) Create plasmid vector that contains a selectable marker, and a 1 kb fragment of the genome surrounding the desired site of barcode or watermark insertion, with the barcode cassette or watermark base change in the middle. This vector should lack an origin of replication for the microbe to be transformed so that the selectable marker can only be maintained by integration in the genome.

2) Transform bacteria with plasmid and select for integration (loop-in) of the plasmid marker.

3) Confirm genome structure of transformants by diagnostic PCR.

4) Grow strain non-selectively, plate colonies, replica-plate to identify those that no longer bear the plasmid marker.

5) Confirm genome structure of recombinants, sequencing barcode cassette or watermark, and surrounding region.

B. Site-specific recombination: group II intron retrohoming

The homologous recombination method described above has the advantage of allowing perfect replacement of a genomic locus with an engineered version of the locus, but has the disadvantage of requiring efficient homologous recombination between plasmid and chromosome in the host microbe. Since this is likely to be problematic in some bacterial species, we propose a site-specific recombination mechanism that relies on engineered group II intron ribozymes. Group II introns are mobile catalytic RNAs found in bacteria, and in organellar genomes of eukaryotes. These introns transpose to new DNA sequences by a mechanism termed "retrohoming" (Belfort et al., 2002). The introns contain a gene for an intron-encoded protein (IEP) which is a reverse transcriptase required for the integration event. The IEP forms a ribonucleoprotein complex with the intron RNA that catalyzes the event. The site specificity of the retrohoming recombination event is specified by base-pairing of sequences in the intron with the target site.

Lambowitz and colleagues have studied the *Lactococcus lactis* Ll.LtrB group II intron, identifying the sequences that are responsible for targeting (Mohr et al., 2000; Singh and Lambowitz, 2001), and manipulating those sequences to target the intron to new sites (Karberg et al., 2001; Perutka et al., 2004). They refer to these retargeted group II introns "targetrons", and have used them to target insertions in both Gram-negative and Gram-positive bacterial species.

<u>B.1 Engineering a group II intron for a defined target:</u>

The targetron is usually expressed from a donor plasmid that contains a deletion derivative of the Ll.LtrB group II intron (Karberg et al., 2001). The deletion removes most of the gene for the intron-encoded reverse transcriptase, leaving a short sequence flanked by the 5' and 3' exons. Since the reverse transcriptase is required for the insertion event, it is expressed from the donor plasmid at a site separate from the intron. Exogenous sequences, such as the barcode cassette, can be inserted in the site of the deleted reverse transcriptase, and will be integrated in the genome as part of the intron (Frazier et al., 2003).

Site-specific targeting occurs by base-pairing interaction of the intron's exon-binding sites 1 and 2 (EBS1 and EBS2) and δ with the target's intron-binding sites 1 and 2 (IBS1 and IBS2) and δ'. Modification of the target specificity of the intron is achieved by manipulating the EBS, IBS and δ sequences in the donor plasmid such that they match new IBS sequences in the desired target. The Lambowitz group has recently developed an algorithm that both predicts potential intron insertion sites, and designs PCR primers to modify the intron such that it will insert into those sites (Perutka et al., 2004). To target the Ll.LtrB group II intron to the desired barcode cassette insertion site, this algorithm would be used to identify the optimal target within a "window of opportunity" of approximately 100 bp. This is necessary because some sequences are incompatible with intron insertion due to requirement for recognition of intron sequences by the

IEP. The intron donor plasmid would then be modified such that it contains the barcode cassette and has all of the required sequence changes for retargeting. Transformation of the donor plasmid bearing the engineered intron into the desired bacterial strain would result in insertion of the intron into the host genome.

B.2 Identifying insertion events, and eliminating extra sequence;

Because of the high efficiency of targeting, it is possible to carry out the intron insertion without selection for the event. In this case, strains bearing the intron in the proper location would be identified among colonies selected for bearing the donor plasmid. These colonies would be tested by PCR to identify those with predicted genome structure at the site of integration. In cases where intron is not efficient, and it is not possible to identify insertion events without selection, a selectable drug-resistance marker could be included in the intron sequence so that only cells in which the intron is inserted in the genome would become stably resistant.

If a selectable marker is included in the intron, it is desirable, as for the homologous recombination method, to remove that marker after completing the manipulations. This could be accomplished by flanking the selectable marker with direct repeats of an innocuous sequence; recombination between these repeats would result in loss of the intervening marker (Alani et al., 1987). This method has been used extensively in yeast, and allows for sequential rounds of manipulation, each using the same marker. As described above, loss of the marker by homologous recombination would be a relatively rare event, and could be identified by replica-plating of colonies that had been growing for several generations under non-selective conditions.

B.3. Summary of steps for integration of barcode by group II intron retrohoming:
1) Identify appropriate intron insertion sequences using Lambowitz algorithm.

2) Create donor plasmid vector that contains the group II intron with changes in the EBS, IBS and δ required for retargeting to the genome sequence of choice, and contains the barcode cassette in place of the IEP open-reading frame.

2) Transform bacteria with plasmid and select for the donor plasmid marker.

3) Identify strains in which intron has inserted by diagnostic PCR.

4) Grow strain non-selectively, plate colonies, replica-plate to identify those that no longer bear the donor plasmid.

5) If a selectable marker was used to identify intron-bearing strains in 3), remove marker by homologous recombination between flanking repeats.

6) Confirm genome structure of recombinants, sequencing barcode cassette or watermark, and surrounding region.

Note that the group II intron method is not appropriate for inserting watermarks, because it leaves intron sequences in the genome. Only the homologous recombination method can precisely alter single base pairs in the genome.

References for Appendix B

Alani, E., Cao, L., and Kleckner, N. (1987). A method for gene disruption that allows repeated use of URA3 selection in the construction of multiply disrupted yeast strains. Genetics *116*, 541-545.

Belfort, M., Derbyshire, V., Parker, M. M., Cousineau, B., and Lambowitz, A. M. (2002). Mobile introns: pathways and proteins. In Mobile DNA II, N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds. (Washington, D.C., ASM Press), pp. 761-783.

Campbell, A. M. (1961). Episomes. Adv Genet *11*, 101-146.

De Milito, A., Catucci, M., Iannelli, F., Romano, L., Zazzi, M., and Valensin, P. E. (1995). Increased reliability of selective PCR by using additionally mutated primers and a commercial Taq DNA polymerase enhancer. Mol Biotechnol *3*, 166-169.

Frazier, C. L., San Filippo, J., Lambowitz, A. M., and Mills, D. A. (2003). Genetic manipulation of Lactococcus lactis by using targeted group II introns: generation of stable insertions without selection. Appl Environ Microbiol *69*, 1121-1128.

Gay, P., Le Coq, D., Steinmetz, M., Berkelman, T., and Kado, C. I. (1985). Positive selection procedure for entrapment of insertion sequence elements in gram-negative bacteria. J Bacteriol *164*, 918-921.

Hamilton, C. M., Aldea, M., Washburn, B. K., Babitzke, P., and Kushner, S. R. (1989). New method for generating deletions and gene replacements in Escherichia coli. J Bacteriol *171*, 4617-4622.

Karberg, M., Guo, H., Zhong, J., Coon, R., Perutka, J., and Lambowitz, A. M. (2001). Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. Nat Biotechnol *19*, 1162-1167.

Leloup, L., Ehrlich, S. D., Zagorec, M., and Morel-Deville, F. (1997). Single-crossover integration in the Lactobacillus sake chromosome and insertional inactivation of the ptsI and lacL genes. Appl Environ Microbiol *63*, 2117-2123.

Mohr, G., Smith, D., Belfort, M., and Lambowitz, A. M. (2000). Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. Genes Dev *14*, 559-573.

Orr-Weaver, T. L., and Szostak, J. W. (1983). Yeast recombination: the association between double-strand gap repair and crossing-over. Proc Natl Acad Sci U S A *80*, 4417-4421.

Perutka, J., Wang, W., Goerlitz, D., and Lambowitz, A. M. (2004). Use of computer-designed group II introns to disrupt Escherichia coli DExH/D-box protein and DNA helicase genes. J Mol Biol *336*, 421-439.

Posfai, G., Kolisnychenko, V., Bereczki, Z., and Blattner, F. R. (1999). Markerless gene replacement in Escherichia coli stimulated by a double-strand break in the chromosome. Nucleic Acids Res *27*, 4409-4415.

Singh, N. N., and Lambowitz, A. M. (2001). Interaction of a group II intron ribonucleoprotein endonuclease with its DNA target site investigated by DNA footprinting and modification interference. J Mol Biol *309*, 361-386.

DISTRIBUTION LIST

Director of Space and SDI Programs
SAF/AQSC
1060 Air Force Pentagon
Washington, DC 20330-1060

CMDR & Program Executive Officer
U S Army/CSSD-ZA
Strategic Defense Command
PO Box 15280
Arlington, VA 22215-0150

DARPA Library
3701 North Fairfax Drive
Arlington, VA 22203-1714

Department of Homeland Security
Attn: Dr. Maureen McCarthy
Science and Technology Directorate
Washington, DC 20528

Assistant Secretary of the Navy
(Research, Development & Acquisition)
1000 Navy Pentagon
Washington, DC 20350-1000

Principal Deputy for Military Application [10]
Defense Programs, DP-12
National Nuclear Security Administration
U.S. Department of Energy
1000 Independence Avenue, SW
Washington, DC 20585

Superintendent
Code 1424
Attn: Documents Librarian
Naval Postgraduate School
Monterey, CA 93943

DTIC [2]
8725 John Jay Kingman Road
Suite 0944
Fort Belvoir, VA 22060-6218

Strategic Systems Program
Nebraska Avenue Complex
287 Somers Court
Suite 10041
Washington, DC 20393-5446

Headquarters Air Force XON
4A870 1480 Air Force Pentagon
Washington, DC 20330-1480

Defense Threat Reduction Agency [6]
Attn: Dr. Arthur T. Hopkins
8725 John J. Kingman Rd
Mail Stop 6201
Fort Belvoir, VA 22060-6201

IC JASON Program [2]
Chief Technical Officer, IC/ITIC
2P0104 NHB
Central Intelligence Agency
Washington, DC 20505-0001

JASON Library [5]
The MITRE Corporation
3550 General Atomics Court
Building 29
San Diego, California 92121-1122

U. S. Department of Energy
Chicago Operations Office Acquisition and
Assistance Group
9800 South Cass Avenue
Argonne, IL 60439

Dr. Jane Alexander
Homeland Security: Advanced Research
Projects Agency, Room 4318-23
7th & D Streets, SW
Washington, DC 20407

Dr. William O. Berry
Director, Basic Research ODUSD(ST/BR)
4015 Wilson Blvd
Suite 209
Arlington, VA 22203

Dr. Albert Brandenstein
Chief Scientist
Office of Nat'l Drug Control Policy Executive
Office of the President
Washington, DC 20500

Ambassador Linton F. Brooks
Under Secretary for Nuclear Security/
Administrator for Nuclear Security
1000 Independence Avenue, SW
NA-1, Room 7A-049
Washington, DC 20585

Dr. Darrell W. Collier
Chief Scientist
U. S. Army Space & Missile Defense Command
PO Box 15280
Arlington, VA 22215-0280

Dr. James F. Decker
Principal Deputy Director
Office of the Director, SC-1
Room 7B-084
U.S. Department of Energy
1000 Independence Avenue, SW
Washington, DC 20585

Dr. Patricia M. Dehmer [5]
Associate Director of Science for Basic Energy
Sciences, SC-10/Germantown Building
U.S. Department of Energy
1000 Independence Ave., SW
Washington, DC 20585-1290

Ms. Shirley A. Derflinger [15]
Technical Program Specialist
Office of Biological & Environmental Research
SC-70/Germantown Building
U.S. Department of Energy
1000 Independence Ave., SW
Washington, D.C. 20585-1290

Dr. Martin C. Faga
President and Chief Exec Officer
The MITRE Corporation
Mail Stop N640
7515 Colshire Drive
McLean, VA 22102

Mr. Dan Flynn [5]
Program Manager
DI/OTI/SAG
5S49 OHB
Washington, DC 20505

Ms. Nancy Forbes
Senior Analyst
DI/OTI/SAG 5S49 OHB
Washington, DC 20505

Dr. Paris Genalis
Deputy Director
OUSD(A&T)/S&TS/NW
The Pentagon, Room 3D1048
Washington, DC 20301

Mr. Bradley E. Gernand
Institute for Defense Analyses
Technical Information Services, Room 8701
4850 Mark Center Drive
Alexandria, VA 22311-1882

Dr. Lawrence K. Gershwin
NIC/NIO/S&T
2E42, OHB
Washington, DC 20505

Brigadier General Ronald Haeckel
U.S. Dept of Energy
National Nuclear Security Administration
1000 Independence Avenue, SW
NA-10 FORS Bldg
Washington, DC 20585

Dr. Theodore Hardebeck
STRATCOM/J5B
Offutt AFB, NE 68113

Dr. Robert G. Henderson
Director, JASON Program Office
The MITRE Corporation
7515 Colshire Drive
Mail Stop T130
McLean, VA 22102

Dr. Charles J. Holland
Deputy Under Secretary of Defense Science
& Technology
3040 Defense Pentagon
Washington, DC 20301-3040

Dr. Bobby R. Junker
Office of Naval Research
Code 31
800 North Quincy Street
Arlington, VA 22217-5660

Dr. Andrew F. Kirby
DO/IOC/FO
6Q32 NHB
Central Intelligence Agency
Washington, DC 20505-0001

Dr. Anne Matsuura
Army Research Office
4015 Wilson Blvd
Tower 3, Suite 216
Arlington, VA 22203-21939

Mr. Gordon Middleton
Deputy Director
National Security Space Architect
PO Box 222310
Chantilly, VA 20153-2310

Dr. Julian C. Nall
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

Dr. C. Edward Oliver [5]
Associate Director of Science for Advanced
Scientific Computing Research
SC-30/Germantown Building
U.S. Department of Energy
1000 Independence Avenue, SW
Washington, DC 20585-1290

Mr. Raymond L. Orbach
Director, Office of Science
U.S. Department of Energy
1000 Independence Avenue, SW
Route Symbol: SC-1
Washington, DC 20585

Dr. Ari Patrinos [5]
Associate Director of Science for Biological
and Environmental Research
SC-70/Germantown Building
US Department of Energy
1000 Independence Avenue, SW
Washington, DC 20585-1290

Dr. John R. Phillips
Chief Scientist, DST/CS
2P0104 NHB
Central Intelligence Agency
Washington, DC 20505-0001

Records Resource
The MITRE Corporation
Mail Stop D460
202 Burlington Road, Rte 62
Bedford, MA 01730-1420

Dr. John Schuster
Submarine Warfare Division
Submarine, Security & Tech Head (N775)
2000 Navy Pentagon, Room 4D534
Washington, DC 20350-2000

Dr. Ronald M. Sega
DDR&E
3030 Defense Pentagon, Room 3E101
Washington, DC 20301-3030

Dr. Alan R. Shaffer
Office of the Defense Research and Engineering
Director, Plans and Program
3040 Defense Pentagon, Room 3D108
Washington, DC 20301-3040

Dr. Frank Spagnolo
Advanced Systems & Technology
National Reconnaissance Office
14675 Lee Road
Chantilly, Virginia 20151

Mr. Anthony J. Tether
DIRO/DARPA
3701 N. Fairfax Drive
Arlington, VA 22203-1714

Dr. Bruce J. West
FAPS - Senior Research Scientist
Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

Dr. Linda Zall
Central Intelligence Agency
DS&T/OTS
3Q14, NHB
Washington, DC 20505-00