

JUDITH BLAKE

CENTER FOR BIODIVERSITY AND CONSERVATION

SPRING SYMPOSIUM

CONSERVATION GENETICS
IN THE AGE OF GENOMICS

AMERICAN MUSEUM OF NATURAL HISTORY
APRIL 5-6, 2001

DAY TWO

PART IV

THE ROLE OF EXPANDING TECHNOLOGY
IN CONSERVING BIODIVERSITY

ROB DeSALLE (MODERATOR), Curator,
Division of Invertebrate Zoology,
American Museum of Natural History

RD: ... One of the charges that we gave ourselves as organizers of this conference was articulated quite nicely yesterday by George Amato. If I remember it correctly, we

charged ourselves to assess the past legacy of genetics in conservation and also to look forward to the future of conservation genetics. Both sessions today will attempt to do that looking forward into the future.

At first glance, this first session might seem like a hodgepodge. But in Dr. Conway's wonderful talk yesterday, he articulated the genomics and new genetic technology to bring us vast genetic and reproductive knowledge of organisms that will be the tools of the 21st-century conservation biologists.

This session is about some of those tools, including genomic databases and what their impact will be on our way of dealing with conservation. Another set of subjects we will touch upon in this session concerns genetic modification and cloning. For those of you who did make it in this morning from the poster session last night, you will have noticed that there were a lot of posters on genetic resource banking. And I believe Ollie may touch on that this afternoon.

These are all new things for us, in this field of conservation genetics. And with these new things come new questions—new social, ethical, and legal questions. And so we've also included in this session a talk by some lawyers. (Laughter) And I'll refrain from my lawyer jokes until I introduce them, okay?

Without further delay, I'd like to introduce our first speaker. This is Judy Blake. I first met Judy last year at an NSF workshop, and I was totally impressed with how calm she is. She said she was nervous just a second ago. But with someone who has to deal with so much data—so many data—and so big of a problem as annotating genomes, I was just amazed at how calm and organized she was. And I think Judy has made some really great contributions to genomic science, and I think she's one of the few people who actually thinks across genomics to conservation biology.

Judy has written many, many papers—in single papers she has citation indices that outdo total citation indices of many scientists—by an order of magnitude, with a single paper. This is how influential her work is, and has been.

Judy is now at The Jackson Laboratory in Bar Harbor, Maine. She runs, or directs, the Mouse Genome Informatics Group there. And her talk today is on “Comparative Genomics and the Conservation of Biodiversity.” Judy?

(Applause)

COMPARATIVE GENOMICS AND THE CONSERVATION OF BIODIVERSITY

Judith A. Blake, Research Scientist,
Mouse Genome Informatics, The Jackson Laboratory

JB: I want to thank the organizers for inviting me to this symposium, because I really enjoy talking across groups. I am involved with genomics right now, very deeply, but I have a long-standing interest in biodiversity and conservation genetics. This kind of talk does make me nervous, because I’m forced to think across these disciplines and to try to synthesize what it is that we have in common and where we might go together.

The charge I’ve been given today is “Comparative Genomics and the Conservation of Biodiversity.” I am dividing this talk into two parts. The first part looks at some whole genome comparisons. And this means whole genome comparisons. Because we have whole genomes now, and that’s been such a major advance in our thinking. We have a complete set of sequence from an organism. And secondly, I’m going to look at integrated information systems for data

exploration. This is where, I think, there is a lot happening that will impact on the field of conservation genetics.

I want to start with the microbial genomes. This is where the first sequencing was done—with *Haemophilus influenzae*, when I was at the Institute of Genomic Research. Microbial genomes have been a fascinating study in the impact of a new technology on understanding biodiversity. The first sequences were done more as demonstrations of the technology. But since then, the sequencing continued, and we have over 40 completed genomes now. Microbial life forms are known to occur in all parts of the earth—from high-temperature archaea to methanogens. There are many, many interesting genomes, and it's thought, quite amazing I think, that fewer than 1% of them have been described in any way.

The estimate is that we'll have 115 or more completed microbial genomes in the next two or three years. Some of them are in private hands, being analyzed by private companies, but many of them are in the public domain. Many of these genomes are from uncultivated microorganisms—I mean to say, these are organisms that can't be grown in the laboratory—so this sequencing technology is enabling us to discover new things about our world. There seems to be hot spots of microbial diversity at depths below 100 feet in the ocean that are only just being described and understood.

Also, a lot of these microbial genomes are very special animals—for instance, you may have heard of *Deinococcus radiodurans*. This organism survives 1.5 million rads of radiation—which is 3,000 times the lethal dose for humans. Its genome has been sequenced, and it is now being genetically engineered to transform divalent mercury into less toxic forms. And, it also might come to be used in radioactive waste cleanup.

This slide I'm showing now is looking from the Department of Energy Joint Genome Institute—JGI—a relatively new consolidation of Department of Energy labs. Last October they had a microbial marathon, and they sequenced 15 microbial genomes in a month.

So, of course, the first thing you want to do when you have all this information is compare things. And I have to say that comparative genomics is really an exciting endeavor right now—just looking at systems and saying: What's happening here? What's happening there? For example, these are two of the first genomes that were sequenced—*Haemophilus influenzae*, as I mentioned; and *Mycoplasma pneumoniae*. And people are starting to look at and compare functional systems between these two organisms. So this slide is a comparison of transport-system characteristics from two distantly related organisms, and we want some statements about where different aspects of transport systems occur in different groups of organisms.

Of course, all of this started with the Human Genome Project. And one of the great things that happened was the recognition that we weren't going to start right off and sequence the human genome. And so the concept of model organisms was developed. These "model organisms" were considered key organisms that would give us insights—they are typically useful in the laboratory, and well-known biologically.

Initially, the first genome that was planned to be sequenced was *E. coli*. It actually ended up, I think, being fifth or sixth. This slide shows a set of species whose genomes have been sequenced now. And, again—very elemental comparative work being done. For instance, among the organisms whose genomes are sequenced, genome size does not correlate with the numbers of genes. This statement itself illustrates how basic our knowledge is in this world of whole-genome comparisons.

I'd like to mention particularly that upwards of 50% of all the genomes that have been sequenced, up to 50% of the sequence elements that we think are coding for gene products, represent sequences that are dissimilar to anything we've seen before. We can identify new gene families of sequences, but we don't have any idea what the gene products of these families might do. So we have a lot of discovery work to do, to understand those genes and their role in these organisms.

Here, again, we're looking at comparative gene classes. And when we look at conservation genetics, or taxonomic studies, often we're looking at genes, and we're doing comparative analysis between genes in different species. Here sequences are being functionally annotated and described based on certain sequence characteristics. In this analysis, an overall similarity of sequence is not the defining factor. One of the great genome informatics projects going on right now is at the Swiss-Prot protein database, a resource based in the U.K. and in Switzerland. They are doing a very careful analysis of the actual architecture of a protein, the particular elements that must be present in a given configuration, in order for the protein to have a certain function. So it's not the overall similarity—it's that there's a certain critical amino acid; a certain structure to the molecule that gives it its function.

Again, though, you're seeing here the comparative gene classes in this slide. And here are the model organisms. When we start looking at, say, acetylcholine receptors, we see that there are 17 of these elements in gene products in human; 12 in fly; 56 in worm. And we're just trying to sort out the numbers. I often think of this process as comparable to the early explorations of Humboldt, going down the Amazon or wherever, and just picking up beetles and other natural objects and throwing them into bins. And that's really what we are

doing. We're just binning genes, at this point, because we have no idea how it's all going to sort out when we're comparing them.

We go a little further, now, and here's an example of gene-specific similarities—and now we're working into the gene sets that we know something about. Recently the fly genome was finished. And one of the first things that we observed was that we could identify in the fly genome genes that had high degrees of similarity to genes recognized as being important in human diseases. We heard the other night about recessive disease. Of course, only a small portion of diseases and syndromes of human interest that have a genetic component are understood to be due to the nature of a single variant. Most diseases occur with a complex interaction of genetics and environment and are mostly beyond our full understanding at this time.

But here is a slide of some of those single-gene diseases. And here we can see—for instance, in the upper block—there's polycystic kidney, PKD-1 gene, and it has a very closely related gene in fly. And so, in fact, the fly, as a genetic model, is being used to study human diseases. And that only reinforces for us how closely related we human beings are to all other organisms.

Of course, the human-mouse comparison is the one we hear the most about. Mouse is the primary model system for understanding human biology and disease. That's because it has much of the same physiology; it has a short generation time; we are genetically able to manipulate this organism to specifically study human disease processes.

Here we see some of the work of Lisa Stubbs. She set out to analyze the sequence of human chromosome 19. And there are 15 relevant conserved regions in mouse, and she sequenced all those, as well (she has a paper coming out in *Genomics* really soon, and I've been talking with her and hearing her talk). And so she's able to look at these conserved regions between mouse and human, and

begin to come to some new understandings about genomic level similarities and differences between mice and humans. Human genes, for example, appear to be on average much bigger—the introns are much bigger than the same gene in the mouse. Mouse genes are much smaller. Segments in the mouse genes duplicate more, so there's more members of a gene family in mouse. And so we're starting to get these bits of information from this kind of comparative work.

Of course, in our own work in the Mouse Genome Informatics group, we've paid very close attention to the orthologous gene groups. We focus on comparative analysis of human, mouse, and rat. We do, actually, collect information for 18 different mammals, a representative set of mammals that are important in various ways. And so that's yet another element of comparative genetics that is mostly yet to be realized.

I'd like to end this section of the talk by saying that one of the impacts that this genetic technology revolution is having is that we're moving genes onto different genetic backgrounds. So in trying to understand the function of genes, it's at a point where we choose the experimental system that works for us: Is it fly? Is it mouse? And so, in many cases, we have instances where we're moving human genes onto specific mouse backgrounds, in order to study the function of the human gene. So the interest is in moving from understanding the similarities between the sequences to understanding the function of the genes—and moving beyond, into understanding how the organism generates a certain phenotype.

At the Jackson Laboratory, we have a very large induced-mutant resource. It's a national repository of specifically developed strains of mice. Scientists create particular mouse models, typically moving a particularly constructed gene onto a particular genetic background. And then, when their work is done, we at the Jackson Laboratory will preserve that genetic construction and make it available to the scientific community. And here in this slide we see a listing of a

particular group of those induced mutant mouse strains—a web page showing available neural-2 defect mouse model resources. So in comparative genomics, the data generation has been phenomenal. There is lots of data, and a fair amount of preliminary analysis. We are amassing large catalogues of genes, and putative genes—about half of which have no prediction of function at all, because we've never seen anything like them before. And in fact, we're moving, then—the first flush of genome projects is over, and we're rapidly—there's whole sets of people moving into describing what we call the “transcriptome,” which is that set of all the gene products produced from a given genome.

As many of you are probably aware, a gene can produce multiple gene products. I think the biggest number I've heard so far is one gene with 138 exons. And it's predicted it could produce 38,000 gene products—and that's before they're modified by methylation or anything else that might influence their function as mature proteins. Ultimately, we want to get to understanding, then, the proteome. So what does this get us to? And what is happening in the development of information and analysis systems to manage all this data. So this is the role of expanding technology. We've had all the generation of sequence data, and now we're looking at visualization and annotation. And, most important from my perspective—and, I think, for this community, as well—is data integration.

Here's an example of visualization of genomes. This slide presents a linearization of a circular microbial genome—*Neisseria meningitidis*—and a specific strain. This is available publicly, this visualization. And actually, what happens here is, you can click on any of these, and you keep drilling down further and further. This is about a 2-million-base-pair genome. And you drill down further and further—you can actually get to the actual DNA sequence of a given segment of this gene.

This can be compared to a kind of GIS system. It's just a linear model. One can envision having all kinds of information available to someone on a transect in this way. Of course, there's a lot of data collection and standardization that goes on in order to be able to have this kind of presentation.

We have many, many sequence-annotation tools. Here we have curated sets and computational sets of genes. And when we annotate BACs (Bacterial Artificial Chromosomes) we use these kinds of tools. This slide shows an interesting tool, first developed at Berkeley. The grad students who developed it were brought into the Celera fly-genome project, and there they developed a proprietary application. Now the annotations and visualization software is being redeveloped—by Berkeley, again—in conjunction with the Sanger Institute. And now this is all freeware, and available to anyone who wants to use it. And we in our group are using it in various ways in our project of integrating mouse sequence with mouse biology.

In the Mouse Genome Informatics Databases we represent the biology of the laboratory mouse. Right now, as might be guessed, we're consumed with sequence and the integration of sequence data with other biological attributes of the mouse. We've just recently accessioned 21,000 full-length cDNAs, and we have done the analysis and determined that 12,000 of them represent new genes in the mouse. On the other side of the equation, there are major mutagenesis centers now that are starting to generate and characterize mouse mutants based on phenotype analysis, at the rate of about 4 to 8,000 a month. There are several major international mutagenesis programs. Some focus on heart, lung, and blood disorders; others focus on neurobiology phenotypes. And all of this, the genome sequence and the phenotypic analysis, comes together—at some point, we hope—in generating the understanding of the whole phenomena of the laboratory mouse as an organism.

Three main projects in our group are the mouse genome sequencing project, headed by Carol Bult, which is integrating the high throughput mouse-sequence data with the mouse data; the gene-expression project under the leadership of Martin Ringwald is looking at expression data and integrating and understanding the micro-array and expression profiles; and the mouse genome database overseen by Janan Eppig.

I want to talk now about how we generate unique designations for genomic features. This is a very important and difficult task. We do a lot of indexing and co-curation with the sequencing centers—and with the different sequence repositories, such as NCBI and Swiss Prot—in order to have a representative set of information about a gene, or any other genomic feature.

So what that allows us to do, then, is to integrate data including, for example, multiple publications, various sets of mapping data, and information about phenotypic alleles. We have a representative consensus map representation; a little minimap of the chromosome, showing the location of a gene on a chromosome. And I'm going to spend some time now talking about the classification systems we are using, which are standard systems, and how we then carefully curate links to collect a set of sequence associations for this gene.

Ultimately, we want to move beyond the mouse. We want to be able to speak the same language when we're talking to people describing other organisms. I was just at a meeting recently where we had a big discussion: What is the definition of development? What do you mean when *Arabidopsis*, which is a plant—what do you mean when you use the word “development”? When does the process of “development” start? Different research communities will give you different answers to this question. When does the anatomical element called the heart—where does that start in the developmental process? What are the components of the “heart”? What is outside the concept “heart”? What is inside

the term “heart”? Does it include the circulatory elements? What about the pericardium? Typically, the pericardium is not considered a part of the heart, yet when researchers search for information about pericarditis, they might logical query for “heart disease.”

So we’re moving towards developing a common language for biology, at least for molecular biology. And this work has come, again, out of the model organism community, where we want to use the power of comparative genomics to look at and compare information about shared genes. We’re really looking at a set of genes that happen to be on different genotypes. You know, one way to approach this problem is to say it’s the same gene, whether you have it on a fly background, or on a yeast background, or on a mouse background. But we were using different languages to describe our knowledge about the function of the gene. And so when we would use a term—like cell-division cycle number 42 gene—is this what the orthologous gene is called in all of the model organisms? And what is our collective understanding of this gene’s function?

So we all got together, that is to say, folks working on the informatics of the model organisms. The project I will describe to you now actually started with mouse, yeast, and fly bioinformatics communities. We figured that if we could have a common language of describing the molecular biology of our three organisms, then it would help the whole biological community to communicate more effectively. And that’s what we were having a problem doing, as we were doing our comparative analysis.

So we formed the Gene Ontology Consortium—now expanded from fly, yeast, and mouse to include *Arabidopsis*, *E. coli.*, and others. The microbial genomes are talking with us, and want to use our standard vocabularies. And other groups are now involved as well, including *C. elegans* (the worm).

We have set about creating structured, controlled vocabularies for molecular biology. Why do we want structured vocabularies? We want to standardize our annotations. We want to be able to ask complex queries across all these genomes. We want to be able to ask questions such as “What are all the genes in the model organism systems that function in the initiation of cell division?” for example. And for that we need a standard set of terms which also have to have standardized definitions. What the definition is isn’t as important as the fact that we have defined a definition for the term we’re using. Thus we would be able use terms (concepts) with a clear understanding of the definition of the concept within this system.

The Gene Ontology Consortium has built three separate vocabularies: “molecular function”—what a gene product actually does; “biological process”—a set of broader concepts such as DNA replication; carbohydrate metabolism—a larger process that this gene product is involved in. And, finally “cellular component”—where is this gene product found?

The goals of the Consortium, then, are to develop these structured vocabularies—these ontologies—each term of which has a unique ID, has a definition, and has a defined relationship to other terms. It is a hierarchical system, but it’s not a simple hierarchy. Many terms have multiple parents. And I can review that for people who are interested.

And then, we each, in our own areas of expertise— So the mouse people—mouse scientific curators—annotate our gene products to a specific term—with a term, with a reference, with a citation, and with an evidence statement. We are documenting why we think that this gene product is involved in this molecular function, for example. In the end, of course, we realized immediately that if we toss all of these annotations into a common data resource, then the biological scientific community can query that resource across all these

shared collected annotations, across the genomes of any organisms that use this annotation standard.

So here we see a representation. Again, I showed you this before—but now you understand more about these new terms we have. And we have various ways of browsing the hierarchies and looking at the definitions of each term. And we have an evidence statement. This association of the term “skeletal development” with this vitamin-D receptor gene was inferred from a mutant phenotype presented in this reference.

And throughout our annotations, in everything we do, this is a basic paradigm— if you make any data association, assign any attribute to a gene object, there must be an evidence statement and a citation for this annotation event.. This is very important...because one of the problems we've had in the first flush of genomic information is the transference of putative function based on untraceable statements or perhaps because of some general sequence similarity. And so by setting the standard of providing evidence and citation for attribute assignments we are trying to drive the data accumulation with some statement of confidence in the knowledge presented.

And so we have the Gene Ontology Consortium, and I'll give you the web url at the end of the talk.

We have a series, then, of structured vocabularies. And this, I'd say, is the major thing that we're working on now. We have the GO vocabularies. But also, there are anatomies for each of these organisms that are being standardized; nomenclature for each of these genomic features categorized in the integration process; also, now the development of phenotype vocabularies; disease models; and, of course, many other smaller ones.

The impact of genomics, then, I think, has been a whole-genome view of comparative analysis—and we're just beginning to understand the implications of

that—and the development of an integrated information system to handle all this comprehensive data, including data-generation and analysis tools, integrated information systems, and all these shared structured vocabularies.

Where do I think this takes us, in terms of comparative genomics and the conservation of biodiversity? Well, as I showed in the microbial systems, one event that I think was somewhat unexpected, was that we've had a great interest and impact on the increased discovery and analysis of biodiversity—particularly in the microbial area. And I certainly heard it yesterday, in people talking about genotypes, and understanding the genetic diversity in various groups of organisms. And so the tools that have been developed have been very important in that.

On the genomic side there's certainly been an increased recognition of the diverse organisms as sources of comparative information, and this has been very useful from both perspectives. I think when people were coming from the biomedical approach of wanting the human genome for study and understanding of the human condition, sometimes the recognition of the power of genetic diversity wasn't there. And it certainly has been eye-opening for many folks.

And finally, I think these new tools—some of which I showed you today—for data visualization and integration can be useful models—or, perhaps, even used directly in extending biodiversity representations and data exploration.

I'd like to acknowledge the many people who have contributed to this work, and I list them here. Most of them in our group, in Mouse Genome Informatics, but also bottom-left are major players in the Gene Consortium. The url for the Gene Ontology Consortium is <http://www.geneontology.org/>. And there's a shot of down the road, the coast of Maine.

Thank you very much.

(Applause)

