

(Tape 4)

Panel II: Predicting Pathogenesis and Adaptation to Humans

Moderator: Ann Reid, The U.S. National Academy of Sciences

Ladies and gentlemen, would you please go back your seats. This session will soon be resumed.

Good afternoon, everyone. In the interest of finishing on time, we will try to start on time.

Welcome to the second panel of the workshop. My name is Ann Reid and I am on the Board on Life Sciences at the National Academy of Sciences in Washington, DC. I spent my scientific career studying the 1918 influenza virus, so the subject of genomics and infectious disease is very dear to my heart. I've already learned a lot and I'm looking forward to learning more today.

We heard this morning about the challenges of detecting infectious disease and monitoring or infectious disease agents in the environment and in patients. But, of course, we live in a microbial world and there are far more microorganisms in the environment and even in our own bodies that we could ever hope to characterize quickly or completely. So, part of the challenge is understanding which of these microorganisms might cause disease, what is the relationship between their genetic sequence and their ability to cause harm to people.

So, we know, as Dr. Ahlquist said this morning, it is not just a matter of the microorganism. The same microorganism might cause very different symptoms in two different people. It may make one person very sick and the other person have no symptoms at all. So, the potential of genomics is to be able to understand what are the genetic characteristics of an organism that allow it to cause disease, and what are the genetic characteristics of the host that allow it to be infected.

So, our speakers in this panel are going to address this issue of how you go about determining which microorganisms are threat, which microorganisms might be able to adapt to cause disease, and I'm really looking forward to these presentations.

I have to report that Dr. Yuan is going to be late. He's the first speaker listed in your programs. He is going to give his speech later in the day. So, we're going to start with Dr. Eric Eisenstadt, who is with the Institute for Genomic Research in Rockville, MD.

Eric Eisenstadt, The Institute of Genomic Research (TIGR)

Genome Sequence Analysis as a Tool for Understanding and Controlling Infectious Diseases

Good afternoon. If we were really good, we'd be able to give each other's talks -- if we were really cooperating with one another. Dr. Yuan would have just sent his material here and I could have talked about it, and then when he showed up, he could have given my talk. But, maybe that will come in the future.

I'm delighted to be here. I just want to share with you some news. I've had two great accomplishments already in the very short time that I've been in your country. The first of those was when I registered for the meeting, I discovered that I had a middle name. My parents never gave me a middle name, so I actually thought that was quite nice. And just this morning, I learned from one of the hosts here that in some minds, I've been credited with discovering the cause of mad cow disease. So, I figure at this rate, who knows what great discoveries will be attributed to me by the time I finish. Maybe we will have some cures for the flu and the common cold.

So, I'm going to give you an overview of some work and a couple of examples of activities at TIGR and I've pitched this at a fairly high -- I hope not too superficial level. It is the 10,000-foot view of some research activities that I think relate to the topic that we're beginning to discuss now in this afternoon's session.

Based on the posters that I've seen and talking with several of the scientists here and listening to the discussion this morning, I feel a bit like we have an expression in English called "bringing coals to New Castle" – I'll be telling an audience about genomics and its importance in infectious disease, and you're among the first to realize that. But, please bear with me.

On this slide, I'm highlighting what I think is well-known to this group – namely the very short history and meteoric rise of genomics and its prominence in biomedical science and, in particular, in infectious disease.

Just around the time that several of the young people in this audience were born, this Sanger sequencing was developed and in 1986 the human genome project in the U.S. was launched, a few years later TIGR was founded with an initial focus on EST analysis. But, within a few years, they began to use this industrial-scale sequencing capability to begin a shotgun sequence analysis of . . . flu. That was finished in about a year's time. Now, here we are in 2006, hundreds of genomes have been sequenced, many of them to completion, and most of them microbes and viruses and parasites that, of course, the human genome and other mammalian cells is among them.

Genbank is reporting on the order of 100 gigabases of DNA sequence is now deposited worldwide in the databases. In the spirit of the opening remarks this morning, exhorting us to embrace the genomics revolution, it has happened – we have embraced it. There are multiple industrial-scale genomic sequencing centers throughout the world, a prominent one here in Beijing. Those centers are motivated by the power and promise of genomic analysis. We've begun to hear a little bit about that. And largely enabled by the improving and increasingly more affordable technologies that permit us to sequence organisms, sometimes even just overnight or in a matter of days.

So, we have one of these industrial-scale genome sequencing operations. The name J. Craig Venter will be known to all of you, so we have a template production lab and a sequencing lab, and I won't bore you with the details. We are looking forward to the day when one won't require this size footprint to do the sequencing. Perhaps we will have a sequencing capability on

our bench top and maybe you will be telling us about that later on. You just have to make it more affordable.

Of course, accompanying those high through-put centers are the computational grids and TIGR, just like all of the other big sequencing centers, has a major computational grid. I've grayed out all of the nuts and bolts of it because I don't think there is anything really unique about this grid to distinguish it from all the other computational grids that I'm aware of. We do it as well as others – no better/no worse. But, I do want to make the point about computational grids, looking forward into the future, that Larry Smar, among others at the University of California at San Diego, super computer center, has been pointing out in recent presentations that the world we live in now is not a flat world. It is a common phrase that is being used in the U.S. to describe how relatively connected we all are and there are relatively few impediments to movement back and forth across this flat landscape.

Larry points out that in this day of super computing and super computing power and grids, we are really all reduced to a point. We can be in Beijing and talking to a community of scientists halfway around the world with no delay in the communication if we use, say, fiber optic communication capabilities. So, there is increasingly spectacular computer power. Now, all we have to do is figure out how to use it intelligently.

So, of course, to make again an obvious point to this community, genomics is much more than sequencing. Sequencing, per se, is by far the easiest part of the genomics operation. It is highly automated and given DNA, you can generate sequence with no problem. It is reading the sequence and figuring out what it means – that is the truly hard part. So, all of the value-added at genomics institutes like ours is provided by all the subsequent computational and experimental tools. For example, the assembly of sequence into larger structures, annotation of sequence to predict open reading frames, genes, and gene products and so on, and of course, database tools for bioinformatics analysis to help predict antigens or pathways or networks. Finally, the suite of functional genomic reagents and assays that enable experimental biology to take place. For example, making clones of open reading frames available or purified proteins or microarrays. So, the simple and again, obvious point I'm making here is that it takes a rather large community

of scientists to really practice genomic analysis and extract the biological meaning from sequence.

The drivers for genomics research in the United States and undoubtedly in this country as well, are up here and they are prominently the rational development of better diagnostics, of vaccines, and therapeutics, and increasingly as we learn more about the diversity of organisms in our world, an additional genomic driver is the development of phylogenetic tools and epidemiological tools.

I've used the terms "predictive phylogene" and "predictive epidemiology". That is not a discipline that exists right now. We certainly don't do phylogene in any predictive way. We treat organisms as discovered objects, and only after they are discovered do we then classify them and try to relate them to other organisms.

I look forward to the day when, based in large part on genomic analysis and genomic science, that we'll have some truly predictive tools, and so even before we see an organism, we will be able to say something about its existence on first principles. If we had that capability, that might begin to enable the development of a predictive epidemiology. Rather than waiting for an outbreak to occur, we could begin to infer based on the space of organisms that we are predicting is out there, we may be able to anticipate new kinds of infectious disease outbreaks.

The other point I want to make here, and excuse me for these capital letters, but I think it is an important point – translating genomics research into products like diagnostics and vaccines and therapeutics is a two-way street. It requires constant and iterative pulling that is back and forth between the basic research community on the one hand, and the applied research community on the other. The basic research community cannot expect that just by doing genomics and throwing the information over the proverbial wall that the applied research community will be ready to catch it and then turn it into useful products. It really has to be communication at every step along the way between those two communities. That is really easy to say and it is very, very hard to do.

Let me tell you a bit about a major supporter of research activities at TIGR. TIGR is a not-for-profit institution, which means that we are dependent on funding primarily from the U.S. government, primarily in the infectious disease area from the National Institutes of Health. The National Institute of Allergy and Infectious Disease is supporting three major resource activities at TIGR that I'm going to tell you briefly about, where there is one of two microbial sequencing centers, the other is at the Brode Institute in Boston, Massachusetts. They are supporting a pathogen functional genomics resource center and they are supporting TIGR as one of six bioinformatics resource centers.

The Microbial Sequencing Center – and by the way, much of the information I'm going to describe to you is readily available on the internet on the web. The Microbial Sequencing Center at TIGR and also at Brode was established to generate and then deposit sequence information, including annotated information, for a wide variety of pathogens and vectors and the information about the various genome projects that are funded by the microbial sequencing center can be found by visiting this website.

A hallmark of the Sequencing Center is that all data generated at TIGR and the Brode is rapidly deposited in public databases so that it is available worldwide to the research community. That is a condition of working with the NIAID. The Microbial Sequencing Centers were set up to enable not only the sequencing of infectious disease organisms, but the propagation of that information into the public sector just as soon as it was available.

One of the prominent Sequencing Center activities underway at TIGR is the influenza virus genome project. I'll say a little bit more about that later. I'll just point out to you here that as of last week, we sequenced our 1,000th flu genome.

The second of the resource centers is this Pathogen Functional Genomics Center at TIGR and here we develop resources such as the ones I've already mentioned – DNA microarrays, gateway cloning, reagents, protein expression, proteal mix, and comparative genomics tools, and generating with these tools data and databases. All of this is distributed to the international pathogen research community. Any of you who might be interested in finding out what

resources are available for use to help you study the pathogens of interest to you, I welcome you to feel free to contact me personally, but a good entry point would be to visit the website here for a description of how you can obtain access to those resources.

Lastly, we are one of the six bioinformatics resource centers where we develop and support bioinformatic analysis for the research community.

So, let me know talk about the first of the two specific examples that I wanted to share with you – examples of some ongoing research at TIGR. The first of these will permit me to talk to you about what we are calling the pan genome idea, and what its relevance and application is to vaccine development.

So, the idea is simple, and we have heard discussion of it already. This is just another term for it, if you will. The question we are asking here is what is the genomics space occupied by a bacterial species? We know for the 40-50 years that we have been studying bacterial genetics, the bacterial genomes are dynamic objects – they engage in very active information exchange with other organisms via mechanisms that I think you all understand, via the movement of plasmids and affage, and other bacteria and so on. So, it is reasonable to ask if you sequenced an organism, what does that really represent about the space occupied by that type species? It turns out that of all sequenced bacteria, only 8% of them (a very small number) have had more than two different isolates ever sequenced. Far and away, most of the organisms listed here have only have one representative sequenced.

What do you learn if you look at one organism and begin to sequence multiple isolates from the organism? You know the answer because we have heard a little bit about this already from the morning session. Here is one organism that I'll use to illustrate – streptococcus – a gallic . . . It is easier to call it group B strep. I've abbreviated it here GPS. It is of medical interest because it is responsible for most meningitis in newborn infants.

When Herve Tetlan, who is one of our investigators at TIGR, working in collaboration with Kiron, began to look and fully sequence multiple isolates of the group B strep, what he observed

was that as he sequenced different genomes, the different isolates, that the number of genes that were common to the isolates plateaued at about 1,800 genes or so. That is, for this sample size, it looked as if there was some poor sequence that all isolates shared and the number of genes that appeared to be identical was on the order of 1,800. At the same time, what they noticed was that as they added an increasing number of strains in their analysis, although the number of new genes added began to decline, it didn't reach zero. It looked as if, again in this small sample size, began to look as if it was plateauing, suggesting that no matter how many additional strains of group B strep one sequenced, you would discover on the order of 30 genes that you hadn't seen before. This has to be extended to far greater numbers of isolates, but we have looked at other bacterial species and we find that for some, this is the pattern that obtains, while for others you only need to sequence a few -- bacillus anthracis is a prominent example of that and you discovered that you've really captured what appears to be all the genomic space of that organism.

So, the implications for bacterial taxonomy and practical implication in this case for vaccine development are important. Classical taxonomic approaches for classifying isolates usually rely on the invariable core genes that we associate with organisms in a core genomic feature such as 16S ribo. . . RNA typing. But, it turns out that the features that we most care about, from a practical perspective, if we are interested in developing antibiotics or vaccines, those features are determined in large part, at least for the group B strep case and we suspect for others, by the highly variable parts of the genome. So, it is going to be important to be looking again at the full genomic space occupied by the organism rather than just deducing from having sampled a small set what a good vaccine approach might be, or a good therapeutic approach.

One example – I won't dwell on the detail here – of the power of that is that it turns out in a study supported by Kiron, that combinations of antigens where the antigens come from both the core as well as the pan genome, the highly variable portion of the genomes together protected against 92% of all tested GBS strains, one of the antigens being from the core genome and three of the antigens being from the pan genome. So, that is example number one.

Let me move to the influenza work that is currently underway at TIGR, and I know you'll be interested in hearing about this. This part will introduce you to the whole genome sequencing

operation that is underway at TIGR. I suspect you already know about it, but I'll introduce it to those of you who aren't aware of it.

We have a pipeline, as we call it, which starts with the collection and amplification of flue isolates, the extraction of the RNA and shipping of that RNA to TIGR. It is to say these first two steps don't obtain, in our facilities, we don't have the ability to work with pathogenic organisms. We work with an international community of investigators who do the isolation and the RNA extraction. Once its nucleic acid, we are at home with that and we take it from there. We use technology developed in large part by . . . Geddon at TIGR to amplify the RNA segments by RTPCR tiling. We sequence those amplicons, trim and assemble the sequence and with a little bit of editing, review for quality and the data immediately gets submitted to the database at NIH – the NCBI databases.

We have estimated that if push came to shove and we were asked to rapidly run this operation on RNA samples that were delivered to us, we could turn this around in the order of 48 hours.

The results of the genome project are deposited in the database via the agreement we have with the Microbial Sequencing Center. I've told you that we've already reached over 1,000 genomes. We do on the order of 50 genomes a week at TIGR. The data, again, is released and all the information and access to it is available via the web on sites like that.

So, one already published finding based on the whole genome sequence analysis of the influenza is highlighted here. That is, an examination of samples from outbreaks in New York in the 2001/2003 timeframe and also the 1999 or so timeframe, the phylogenetic analysis of the sequence data that we obtained, revealed that there were two populations – one was the major population depicted here, and one was a relatively minor population depicted here. The whole genome analysis revealed that a subsequent outbreak or variant that arose that was resistant to the vaccine that had been prescribed for use that year, arose by, as it were, a donation of the segment 4 – the HA segment from this relatively minor population to the major population. That accounted for its relative resistance to the vaccine and its epidemiological prominence in New York in that flu season.

The conclusions from that analysis are listed here. The variation, it is clear, occurs in all of the gene segments. A lot of focus, I know, is placed on characterizing the HA and the NA segments, but the whole genome analysis permits to look globally across the virus and one discovers – and it is not a surprise – that there is variation everywhere. The whole genome analysis reveals that in any given flu season, there are multiple lineages, multiple populations of flu that are co-circulating at the same time. Re-assortment of variance of even the same subtypes can lead to the emergence of epidemiologically relevant and antigenically novel flu.

So, this leads us to believe that whole genome sequence analysis, based on random sampling, can by revealing the presence of these co-circulating strains, before antigenic novelty has emerged, can become a very powerful weapon, a very powerful surveillance tool for any country that is interested in understanding what is out there and what the potential is for a major outbreak or pandemic.

I mentioned that we work with many collaborators. They are listed here. They are numerous and they are, indeed, widely distributed.

... be here for the entire week. I'm also making some visits to other institutes in the Beijing area and possibly throughout my schedule is wide open, and I'd love to have the opportunity to talk in more detail if you like about some of the avian flu whole genome work that is now getting underway at TIGR. We also have some other RNA virus sequencing projects, rhino virus and corona virus, and we're beginning to engage in a discovery effort to characterize the diversity of novel viruses in the human and animal population and also in the environment.

A point of contact for you is David Spero – one of our scientists at TIGR. Again, I'd be happy to help make the introductions to David if you would like. Maybe we can even send him over here.

Let me just end with some brief remarks about some major challenges that I see for controlling infectious disease and some immediate opportunities that I think lie ahead. We're getting very good at generating data in this genomic world that we live in, but I think all of you would agree

with me that although on the one hand we appear to be drowning in genomic data, we nevertheless feel that the data we have is limiting. If we have a lot of it, it isn't necessarily the right kind of data. We constantly need to be developing better tools and improved technologies to gain access to data that would be of value for infectious disease work.

We're fundamentally limited by our meager understanding of biology. Biology is much more complex than physics. The physicists have had rules like $F = MA$ for several hundred years, and we have some rules, but we don't have any formulas, we have very little predictive capability in biology, and we're going to need to improve our fundamental understanding of biology before I fear we will have some really powerful weapons for controlling infectious disease. And, we are challenged, as I mentioned at the outset, in moving from basic research to products and applications.

A couple of opportunities that I see. One is to improve the way that we now educate and train our young scientists. I'm a believer in the importance of interdisciplinary education and training and, in particular, bringing ideas and concepts from the worlds of mathematics and physics and engineering into the world of biology to enable the development of new technologies and fundamentally to enable the development of what I'm calling new mathematical and computational frameworks for thinking about biology and biological data.

Finally, I've mentioned this before – we need better communication between the applied and basic research communities. One really needs to be sure there is a two-way street between those two communities.

So, let me just end here by acknowledging the people who really do the work, and I won't read them here, but the pan genome effort that I described for you is an effort involving investigators with Kiron, Harvard Medical School, and my colleagues at TIGR. On the influenza flu project, our investigators in viral genomics, our sequencing closure group, informatics group, and continued collaborating with Steve Saltzburg at the University of Maryland, flu experts at the Wadsworth Center, Jeff Talgrenburger at the Armed Forces Institute for Pathology, Ed Holmes at Penn State and bioinformaticians at NCBI. Thanks very much.

Moderator – Thank you, Dr. Eisenstadt, for that very interesting talk. I'd like now to introduce Professor Huanming Yang. He is the Director of the Beijing Genomics Institute and he is going to speak to us about genomics: a new tool for combating infectious disease.

**Yang Huanming, Institute of Genomics,
Chinese Academy of Sciences**

Genomics: A new tool for combating infectious disease

Thank you, chairperson. I also thank the organizers for inviting me and providing me the opportunity to meet friends from the United States and in China, both old and new, to communicate with colleagues in the fields of both genomics and infectious diseases.

We all know that . . . is inspiring and influential to the public as genomics . . . human genome project and the human cancer genome project I will use a lot of slides as the background for my talk. But, I'm notorious for not keeping time. I will try to behave.

I would like to organize my talk into five parts: (1) a single . . . word with common . . . ; (2) . . . two pillars of genomics; (3) . . . as well as in China; (4) pathogens . . . but we are still learners.

Suppose we were taking the U.S. space shuttle or China spacecraft running into the sky and looking back to our homeland through the windows. That is really a wonderful globe. It is so beautiful. It is so lovely. Nothing would be more beautiful in life if we compared what is said about an American lady who was the first one to walk into space, and the Chinese ancestors in the . . . we can see . . . same beautiful language to express the same passion for life. But, if we have another closer look at our homeland, diseased world, exhausted world, wounded world, broken world, . . . poor world, a world with many children who are still hungry when they go to

bed, just as the idea an aging world, an unhealthy world, a sudden outbreak of SARS in that year, and the widespread of flu in China and other countries. It is possible worldwide spreading for . . . our world is not that secure. We have realized that a single and . . . world we are facing the same common challenges. It also reminds me of a great book by the U.S.A. National Academy of Sciences in 1968 entitled “Biology and the Human Future of Man”. The Chinese version was published a few years later, and I found myself actually taking dozens of pages of notes on that book and found there were only 19 chapters in the Chinese version, but . . . in English version. That is a great part of the book. . . . Then also we see opportunities which would . . .

That is the world that is . . . and the biotechnology . . . with that book and biotechnology requires knowledge in bio life sciences and the genes to play with. This is the reason we need genomics. Genomics is . . . to provide genes and the knowledge about the genes and the genomes. Genomics is the upper . . .

What are the genomics is? That is . . . secret of life. Thanks to the contribution by an American and then a British . . . in America, now we know, if not all of the secret of life, is hidden in

(remainder not transcribed)

Moderator –Professor Yang, let me be the first one to say thank you for a very interesting and very inspiring speech. I’m sure many others will say that to you today.

Finally, I’d like to introduce Dr. Gary Anderson from the Lawrence Berkeley National Laboratory. He is going to speak to us on environmental influences on emerging pathogens: the development of a high density microarray to measure microbial community dynamics.

**Gary Anderson,
Lawrence Berkeley National Laboratory**

***Environmental Influences on Emerging Pathogens, Development of a High-density
Microarray to Measure Microbial Community Dynamics***

(Tape 5)

What I'd like to talk about today is something that hasn't been talked about too much so far, and that is the role of the microbial community on both pathogens and on emerging pathogens.

I would like to start with a little bit of history in that for about the last 100 years or so, culture enrichment has been the primary source of identification and characterization of pathogens. Most of what we know about pathogens we first grow them in culture in selective media, isolate them, and then study their properties. But, in fact, there is organisms that can be cultured are a very small part of what is out there in nature. Typically, less than 1% of the organisms can be cultured, and even by using extraordinary methods, usually no more than 20% of organisms can actually be cultured in any particular sample.

You've already heard some talk today about the 16S ribosomal gene. This is one of several different stable biomarkers and there are other ones and I won't go into the whole history of why it is used. But, it has been accepted among many evolutionary biologists for describing differences between organisms, at least at a gross level.

One of the advantages is that because it is part of a protein assembly machinery, it is less susceptible to homologous recombinations. So, usually the 16s ribosomal gene of a particular organism has always been a part of that organism. It is a structural molecule as represented here. It actually functions as a structural RNA molecule. So, there are certain particular regions that are very well conserved. This has the additional advantage of being able to use for PCR amplification. But, the main reason why we use it is just because since it has been established so

much by so many labs, there are currently well over 200,000 sequences of this particular gene in the database.

So, there have been some novel or new pathogens which have been characterized specifically by 16s sequence. One that I'll just briefly mention here is *Trofarima whipli* is the causal agent of whipples disease. This is a wasting disease which, for about 100 years, was known by microscopy and then later by electron microscopy, that there were bodies which looked very bacterial-like in the epithelial cells that no one was ever able to culture until just a year or two ago. This organism was then – with the advent of 16s ribosomal identification, it was identified by that method and found to be a member of the actual micif phyla.

So, what we use is a particular microarray called a phylo chip to study microbial communities. We use the affymetrix platform so this allows for massive parallelization. We can have many probes which is what can be considered in essence a very comprehensive look at microorganisms in an environment. So, what we do is this microarray is able to identify multiple species in a mixed population. There are two characteristics about this microarray to allow it to do that. First, we use multiple probes for each identifying organisms that we identify, which are targeted to specific regions of the either amplified DNA or the 16s ribosomal gene. The other characteristic is that we use something which is common to the expression array analysis world which is mismatch control oligonucleotide probes. So, in other words, for every probe that we put on there with an exact 25-base match to a 16-s target, we put a mis-match probe which in the central position has a mis-match nucleotide. For interactions which are sequence-specific in theory then this should have a lower hybridization intensity as compared to its perfect match partner. It is hard to see here, but if the red bars indicate the perfect match probes, the green showing the mis-match, are typically much lower for a sequence specific interaction.

So, in designing this array, for the probe design, it was critical to distinguish the bacterial groups. So, we took advantage of all of the sequence information that was out there and through a series of algorithms, we clustered very similar 16-s sequences into like groups. We removed basically what we considered junk from there and just remained with high quality sequences. We created what we called these probe sets which are just groups of oligonucleotide probe pairs which are

specific to an organism. So, we have typically a minimum of 11 probes for each probe set. A probe set can be considered to identify one particular type of organism.

So, a lot of this – we have a website where we put a lot of our data information which we call green genes. You can just access it at greengenes.lbl.gov (for Lawrence Berkeley Lab, where I'm from). This has not only all the probe sets and the databases that we use, but the various algorithms for making your own probes.

So, each of the probe sets that we use in this array is placed in a file genetic context. So, what I mean by that is based on 16-s sequence, organisms are related in various levels of phylogeny to each other. We use that information in the probe sets. So, as I said, we have a minimum of 11 probe pairs for each probe set. The probe sets are then placed in a hierarchical order for typical, classical microbiology from family to orders to classes to phylo to domain. Domains, being of course, bacteria and arkaia. So, basically we have 455 different families which represent all the different families out there. We further subdivided these into sub-families and into the approximately 9,000 probe sets. Each one of these has, at the 16-s level, about 1% variation. So, this can be considered somewhat like a species level identification.

So, this is an example of what a probe set would look like. This is for an environmental organism, *disulfalvibrialgaris*. In this particular case, there were three different sequences in the database of high quality and full length. We found a number of regions which were specific to this particular grouping of sequences, as well as regions which were not unique. So, we designed these 25-based probes based on these unique regions as part of this whole 1,500-base 16-s gene sequence.

So, the actual sample preparation and how we use these arrays is quite simple. As I said, we designed an array that was then made by affimetrix, so it uses pretty much just their standard technology. What we start with is we extract the genomic DNA. If I have time at the end here, maybe I'll talk a little bit about our RNA work too. But, regardless if it is DNA or RNA, after amplifying the fragment, we fragment into small, say 50 to 200-base fragments, and label each

one of these, do the hybridization, and then come back with a fluorescent probe to this biotinylated target sequence.

As I mentioned, we had 500,000 spots. This is done by a photo-lithography process and so each spot is about 18 by 18 microns.

This is an example then of a sample of what we would see on a computer screen – just different spots of the 500,000 spots would light up. This is our microarray. This is compared to the size of a U.S. quarter. So, for the half that doesn't know the quarter, it is about that size. So, it is just basically about a microscope slide in length. Just a blow-up of one little part of this microarray, the computer software then determines X/Y coordinates of each probe and it is typically very accurate. So, we don't have any cross-over of one probe into the next cell to give a signal there.

In the remaining time, what I'd like to do here, and since I don't have much time, just talk very briefly on three examples of using this array for looking at microbial community dynamics of human health. For the first example, I'll talk about what are the typical pathogens that we find in the air, and do they change over time? Then I will very briefly talk about the use of antibiotics in microbial communities in the lung, and also of the diversity of bacteria of the GI, gastrointestinal tract.

So, first in talking about aerosols, I heard from a number of people here that there were sandstorms in Beijing a couple days ago and sandstorms sadly are typical in other places too, for instance, sub-Saharan Africa, because of global climate change, desert regions are getting bigger and the sandstorms are getting more violent and stronger each year. What has been shown to happen is that these sandstorms, in addition to kicking up hundreds of millions of tons of sand and dust, also brings over pathogens from Africa to the Caribbean, Florida, and Central America. It is correlated with coral bleaching. It is correlated with respiratory infections, allergies, and a number of other things. But, the air can be thought of as a very extreme environment for bacteria. It is very high in ultraviolet radiation. It is also very low in relative humidity. It is a very harsh and extreme environment. So, for organisms to survive in these long transports, quite often they can either cluster together or are somehow shielded by other particles.

We looked at the microbial diversity of a pathogen surveillance program. This was a major U.S. program that is over \$100 million a year to basically be an early warning monitor system for biological attacks, biological warfare. What was seen is that each year there had been numerous newspaper articles about there have been a number of documented false-positive detections. So, what was critical to know is what are the pathogens that potentially could be detected. Some of the pathogens that they are interested in are endemic in the United States anyway. So, how often do they just naturally happen, or is it just a close relative.

One of the ways we went about this was to look at two cities which had pathogen collection systems and this is the state of Texas in the United States and two cities of equal size – a little over a million people each and about 100 kilometers away from each other in Texas. So, we looked at 17 weekly samples for these two cities to start to scope what the variability of microorganisms in the air are. So, we wanted to assess the diversity by time, by these weekly samples, and by location to see what kind of variation we would get.

So, one of the first things we did in using our microarray is we wanted to compare it against a more costly but proven technology of clone library sequencing. So, we sequenced 16-s clone library from one of the weekly air samples and we compared it with the microarray results. What we saw in our clone library, we had about 420 clones of high quality, 1,500-base sequence. We found that there were two dominant organisms which together were over a quarter of the entire sequences found which were *Bacillus megaterium* and *Bacillus pseudomegaterium*. So, it was dominated by just a couple species. But, what was interesting is that then this was anorganism closest sequence match. After that, it went to one or two sequences per bacterial type – in other words, a very diverse sample – as diverse as most samples that had been seen.

We also looked by our array and by taking the exact same sample and with that split sample, we found that we actually detected 170 taxa – so over double the number of organisms by the array than what we did by the clone library. What was nice was that we found a majority of the taxa that we did detect by the array were confirmed by the 16-s clone sequence library. There were an additional 8 taxa which were not seen. A couple of these were novel and a couple of these we

don't know why we didn't see them and several of these were just seen by the clone library as single isolate hits. So, statistically it is possible that it was just below the detection limit.

We noticed that we had a great many more taxa which were not seen by the clone method. We actually looked and we found nine of these groups that we just tested right off just to see if they were there, and were indeed there. I'll show you how we did that. In this case of *Nitrospira*, we found that we had these 30 probe pairs of this probe set which was for this particular organism here – *Nitrospira moscoviensis*. It looked very much the perfect match is stronger than the mismatch all the way across, and very much like a sequence-specific interaction and plenty of signal there, but it was not detected by the array. So, what we did was we went back to our original sample and we designed primers specific for this particular group of organisms, or for this *Nitrospira*, and we found an amplified product of the predicted size. So, we sequenced that and confirmed that yes, that *Nitrospira* organism was present.

We have gone to much more verification since both by the specific PCR method as well as by quantitative PCR/QPCR. But, by having 17 consecutive weekly samples for two different cities gives us a great opportunity, as opposed to a single snapshot in time of a 16-s clone library, see not what is there but what is changing. So, that is what we wanted to do.

So, one of the things we looked at, we looked at pathogens and we tried to compare it against various environmental parameters. For instance, for micro bacterium tuberculosis, which was detected in these samples, we saw that as the average temperature for the week increased for either of the two cities, there was generally an increase in the signal of organism detected on the chip or the abundance of the organism. It is a loose correlation here, but it gives some idea of the power of what you can do.

We took this to a more sophisticated level and did multivariate regression tree analysis and this is a little more sophisticated type of comparison analysis in which we looked at all the environmental factors that we could discern from these 17 weeks, as well as all the changes in organism abundance for presence as determined by their microarray results. By combining both of these methods together, we could see what were the predominant environmental factors which

shaped the differences from array sample to array sample or from week and city samples. We could also see what are the organisms that varied from sample to sample, week to week and so on. We were actually able to find both particular organisms and I just have – the phyla that they are in are indicated here. These would be their patterns for the different types of clusters that were present, and their environmental factors that were important. What we found for environmental factors was that in this type of analysis, it goes from the most significant environmental factor on down as you go down the tree. The week in which the sample was taken from was the most important environmental parameter, not the city which was very interesting. So, more important than what city it came from was what week it came from. The first three weeks from either city were different from all subsequent 14 weeks, for instance.

The next most important factor was temperature, and then a number of factors down below that – the mean particle size of 2.5 microns, the mean sea level pressure, and so on down, were indicated by this. We saw the number of different types of organisms that responded.

Since we're looking at a comprehensive of everything that is there, these are all environmental organisms which are going up and down which are the most significant. Pathogens are included here too, but they weren't the most significant organisms going up and down by week by week.

Very briefly here I'll talk about just two other snapshots of what we're doing with University of California San Francisco, we are looking at intubated patients and looking at lung micro flora from individuals that get pneumonia. In this example here, looking at this array, a patient comes in and for whatever reason is intubated and develops pneumonia, in this case by antibiotic therapy – developed pneumonia – *Pseudomonas aeruginosa*. Then later started to recover. What was seen was that by this type of analysis, this is a heat plot here, looking at this on the y-axis here are the different microbial groups that are changing – red being more present or strong signal, green being either absent or a weaker signal. We see that the before and after samples were both more diverse and clustered together as opposed to while active pneumonia was happening.

But, the other interesting thing was that the majority of the organisms present here have also been seen in oral flora, oral cavities. So it looks like in these intubated patients, that is one of the significant routes of bacteria into the lungs.

Also, we have looked at high density array of gut microflora. Other people have too. There has been a study both on mouse and on men using extensive clone libraries done by other people. We have started a microarray study and we found some additional phyla present that are present in human guts. In looking at Crone's disease, we looked at someone who relapsed into a Cron's infection and compared to a number of healthy people, again by one of these heat maps. Basically what we saw is that we couldn't see anything here – the bottom line – the difference in gut micro flora varied more from individual to individual than from disease to individual.

The last thing I'll talk about here is a little work we are doing on RNA analysis. Basically, instead of using – in all the work that I've talked about so far we've amplified 16-s ribosomal gene and placed on the array. We also looked directly at ribosomal RNA. RNA is usually made of about 20,000 copies per cell. So, with no amplification, we can directly label and using a specific fluorescent tag, put it on an array. What we can determine not only is its presence without amplification, but also relative, metabolic influences because as you would assume, the more copies of ribosomal RNA you have, the more protein assembly machinery is going on in a cell, indirectly saying more metabolism is going on. This is unfortunately from an environmental aspect, but in looking at two different treatments here, looking just at changes from one level of organic carbon versus another, we actually saw -- I want to highlight this because of all the 9,000 different groups that potentially you could detect, there were just a few that significantly had either a greater amount of fluorescent signal from the RNA in one treatment or the other. What was very interesting is that these delta proteo bacteria and these arkeil and acetobacteria have been implicated in going into syntrophic relationship. That is what we determined what was going on in this particular sample.

This could be done in human studies as well to see not only what organisms are there, but what are the most metabolically active which could be important in certain clinical cases.

So, in conclusion, our microarray data, what I hope you get out of this – it complements cloning sequence libraries in the assessment of microbial diversity. We have actually done quite a bit of work to confirm that our array results are, indeed, real by both specific PCR and QPCR. What we find is that the clone library consistently underestimates microbial richness. Although at the 16-s level with our probe arrays we don't get a detailed look at what is there down to the sometimes specific pathogen species or strain, it gives us an initial target of what to follow-up with more precise sampling techniques.

Finally, as seen by our air work, I think one of the real advantages here is that by looking at a number of samples over a number of weeks, and looking at associations, you can start to select candidate organisms responsible for certain effects without any prior selection in the system.

So, with that – thank you.

Moderator – Professor Yuan has still not arrived, so we will save time for him at the end of the day. But, we do have time to take some questions now for the three speakers we've heard this afternoon.

Question and Answer Panel Discussion

Question – For these years since the beginning of TIGR, how much As a not-for-profit institute, how much from the taxpayers that TIGR has spent since its beginning.

Eisenstadt – The honest answer is I have no idea, but a partial answer might be that in the past calendar year, we have received on the order of \$50-\$60 million from the U.S. taxpayer. How's that?

Question – Another question – do you think genomics is cost effective or not, because the people in China are complaining that we are spending too much There is also one thing you know . . . acknowledge the contribution by USA again that . . percent of all of the sequence data now in the public domain are released or made by Americans. Then, the UK is the second and then China is the third. The reason is no country, if it is smart enough, would be . . . as Americans to spend the money for others . . .

Eisenstadt – For maybe the second time in my life, I'm speechless. But, you began by asking what do the American people think?

Question - concerning the production or amount of data . . .

Eisenstadt – I don't know what metric you would want to apply to evaluate the cost effectiveness of the research dollars. The number of therapeutic agents that are in the marketplace now as a result of public investment . . .

Question – My second question is how have you – how does . . propaganda to get the money from the public to spend the money on something that will be freely shared by others including . . in Iraq.

Eisenstadt – You're right. This is a very interesting discussion. I guess I've long been a believer that the sharing of this data in the public domain does far more good than the risk associated with the possible use of that information by people who intend to do bad things. In part, I believe that because there are so many inexpensive ways to do evil, to do bad, and relying on biotechnology to do something nasty is actually very, very hard.

Moderator – This is a discussion that I think will probably generate a lot of talk over dinner, but it is a rather complicated one for a big group to get into. Can we see if there are any questions on the science that was discussed this afternoon?

Question – Can I ask Dr. Anderson something about your dynamic analysis of the micro flora in the . . . you did something for the dynamic analysis – did you see any significant difference for different human beings – what sample size you have done for this dynamic analysis?

Anderson – If I understand your question, basically for the analysis what we see is we monitor then each of the 9,000 different groups of organisms that we can detect, so it is a very broad-based detection method, but only certain of them actually pass the limit of actually having enough probes to be considered to be there. So, if we look at those over the time of any type of population and actually I mentioned our few cases of clinical samples here. Primarily what we use this for is for environmental samples. In that case, it works just as well for human samples too. We can see what dominant organisms appear to be changing with the particular event that we are looking for, be it pneumonia or the amount of chromium being biodegraded in a subsurface sediment.

Question – I'd like to follow-through Professor Gao's question. Could you tell us is your sample size? What sample size is needed before you can analyze the gut flora, the data in the dynamic, to get some meaningful result in respect of its dynamics or other?

Anderson – That is a good question actually. As I mentioned before, we do both 16-s amplification and just direct RNA labeling. If we do 16-s amplification, we actually use very strict guidelines to help reduce specific amplification of individual sequences which could skew the whole population. So, we tricks like pooling together a number of what is called gradient PCR – a number of different temperatures that we do our samples. We only do 25 cycles and we use multiple different primer sets. So, that is the tricks we do. With that, what the answer would be then is that we start with anywhere from 1 to 100 nanograms of genomic material to do our very low level of amplification. What we need to put on the array is actually about 2 micrograms or between 1 and 2 of micrograms of material. For RNA work, again we just do direct RNA labeling so we need to get at least 1 microgram of 16-s RNA. So, it takes a much bigger sample to do that.

Question – I think you mistake my question. Not exactly the quantity of sample that you put on the microarray, but actually how many samples of these either RNA samples or DNA samples need to analyze before you can tell something, let's say, tell us about something about the dynamics of the gut flora.

Anderson – We use enough to do statistics. What we typically do is we like to have at least 3-4 different replicates, experimental replicates – not array replicates – different from start to finish from one sample. We do multiple samples then. So, for a particular – it is just whatever we have access to, but we do a minimum of three replications per sample type, and we do as many sample types as we can to get the most significant information. That is why, for instance, when I said for the air we did 17 samples. So, doing 17 samples for each of two weeks gave us a very powerful statistical analysis to see what would vary within that dynamic range then. So, it is a matter of finding the right statistical tool.

Question – What sequence did you use for the microbial quality of the human gut?

Anderson – We used all the sequences that are available in the data set which have been previously sequenced. What we do then is we have clustered them. So, think of these as just either species or genus level identifications of organisms from the gut. So, we are looking at the whole broad diversity. So, what we are seeing then is the diversity of what is known from in the known universe – not what is unknown.

Question – I have a question for Dr. Anderson. I think along with the other questions, I think the question that they meant was how many individuals do you need to sequence because with one particular individual the inter-individual variability might be greater than whatever variability you see within the gut flora. So, you would potentially have to look at multiple individuals or a population of individuals to get a true idea of what the overall gut flora or flora environment is like, especially if there is a unique, and it is probably likely that each individual has somewhat of a unique gut flora micro environment.

Anderson – I've been purposely trying to avoid answering that question because I don't know. It depends, I think. As I showed for our gut micro flora, we started with a very simple test just to see if one Crone's disease patient would be different from a 9 or 10 healthy people. The answer was we couldn't tell anything. So, it depends on the individuals and what you are looking at. If there is something very glaring and obvious, you could probably do it with less people, but that is a question for a statistician. It really varies on the clinical sample and to get statistically useful information, especially not just – and again this is just for presence, not really for causation, this is just saying that something is present in this particular case, it is still variable by how variable the organisms you are looking at. So, for humans, I don't have an answer. It is quite a few.

Question – . . . working in Beijing, and I have a question for both Eric Eisenstadt and Yang Huanming, and that is, how do you determine which organisms you will sequence? How do you determine the priority, given that you perhaps face different constraints or needs in terms of resources and time. There are many organisms out there and I'm just wondering how the priority setting process works?

Eisenstadt – The priorities for Microbial Sequencing Center activities are actually set via a dialogue involving the research community, TIGR and the NIH program managers. There are overarching guidelines – among them, the organism has to fall into one of several published categories. I briefly mentioned them. For example, we have a grading system in the United States – category A through C pathogens in terms of their severity and how they have to be handled and how serious a biothreat agent they are. Organisms that are near neighbors or related to those are also in the mix. Plus, there is increasing interest in general in emerging and re-emerging infectious disease. All of those categories are on the table and can be proposed as organisms that can be funded. The proposals for funding can come either from TIGR or from Brode, but also from collaborators outside of those two institutions, including I imagine even from China. So, if you really want to pull a fast one and can get the taxpayer to do even more foolish things that you've been alluding to, why not propose some organisms and let NIH pay for it.

Huanming – First, I do appreciate the . . . or the policymaking system in the States. I'm most impressed by the white paper movement which actually . . . attitudes from many communities in the States to positive attitudes. Generally speaking, our colleagues in any community would agree that sequence is not everything. But, without sequence, you can do nothing. The second . . . situation is quite different. Genomics is not only regarded as a . . . against many small science projects, it is also regarded as a rich . . . which can only be done by the rich. And then . . . genomes, of course, I have taken this opportunity to thank the governmental funding agency, especially the most. But, concerning the timing, we are always raised our own money before we get funding. It is the case for the human genome, the case for the . . . genome, and then the case for the Then for the chicken genome, we have 33 sequences from . . . We will get another 17. Then for the pig, we got the funding actually from Denmark for \$40 million U.S. dollar. So, in China, genomics has played a big role in other fields that . . . genomics project. The government funded at least \$250 million because China has already generated the sequence of the . . . genome and paved the way for other functional studies.

But, you ask what is . . . or what is the order? I want to say no order or

Eisenstadt – It may be a piece of cake to do a small organism, but your question really begs a really important one and that is just because – I'll put this as a statement – just because an organism doesn't have an advocate doesn't mean that it shouldn't be sequenced. So, we ask ourselves all the time, what should we really be doing with our resources and I don't know the answer. The advocacy system works up to a point. For near neighbors who stand up to defend the near neighbors of a group of organisms that need to be sequenced, you generally don't have a scientific community that is motivated and organized around – let's do the near neighbors. So, it is a problem.

Comment – Could I put a quick plug in for JGI here. The Joint Genome Institute in Walnut Creek, California and Eddie Ruben is on the board of this panel here, I'm a member of the Community Sequencing Panel and the Joint Genome Institute actually does a significant amount of microbial sequences every year. We meet twice a year and take proposals from all over the world and rank them based solely on peer review, on just scientific interest. As an increasing

amount of sequence related to Department of Energy missions, but basically we have sequences given to people from all over the world and the next round is in May and there is 200 gigabases of sequence are applied for. So, it is growing every year.

Atkinson – As much as I loved all three of these presentations, Huanming’s question motivates me to reverse the question on him and Eric and Gary as well. The session was talking about predicting pathogenicity – human adaptation. So, the question might be couched somewhat differently. Scientists never have difficulty asking for money for their own project. But, the question of why one should look for governmental support is the question of predicting pathogenicity relative to humans because you’re asking for societal support of one’s research. So, it is very revealing to listen to the answer’s to Heather’s question in various forms here. But, what should be done, Huanming, Eric and Gary, to communicate the urgency of your priorities into more societal support – whether it is money or students, human resources may be even more important than financial resources. So, what would you do to convey this based on the three types of presentations you made?

Huanming – I’m always ready to be questioned. I learn more from answering any question than to give a talk. . . . I think all you have said is very encouraging and very inspiring, but the purpose is not of Chinese characteristics because . . . really too small. . . scientific research, even though it has been doubled again, still not enough. There about a million scientists or researchers in the field of biology or life science related fields. Too many people – too little rice. Then it is more and more difficult to get support, not communication with other fields. Then in China, the funding system is just under reform or improvement. I think we will learn even more from you in the coming years. . . morning we have just heard . . . on the middle and long-term sense and the technology research . . . But, if you think of the large community of research . . . , nobody will get rich in the lab or institute from this . . . of projects. We hope we can make the best use of the resource . . . which can be done . . . outside America . . . really good mechanism. Of course, I also accept that nothing can be done by Americans alone in the United States alone.

Comment – My answer may surprise you because a premise of your question, George, is that we are resource-limited and that we need to be generating even more support than we have now. I have, for many years now, believe that we are not resource-limited – we are idea-limited, and if we were smarter, we could do a lot more with not only the resources we have, but even with less. I know that is pretty controversial. It is one of those things that it is easy to say and very hard to do. But, if we demonstrated results that mattered, that changed lives, that is the single-most powerful way to engender public support. I do this. This is what we learned and this is why you should care. The best way to communicate that is to deliver some products.

Comment – I think we obviously need to increase funding for all levels of microbiology research and from the talks this morning, it could be seen that basically from the time that antibiotics have been discovered, the public has a perception that we've cured infectious disease, when in fact even of the diseases we know, we really haven't. But, because there is so many ailments out there which have a microbial component suspected or unknown even yet, say asthma and other syndromes like that, I think it is imperative that more detailed studies be done on all levels to see just what is really out there for microbes and how it impacts on human health.

Moderator – I'm sure we could continue to talk about this for a long time – setting these priorities and figuring out how to train the biologists of the future to be able to deal with this massive data and extract meaning from it is a fascinating topic. We're going to stop and take a tea break now. Please try to be back in 12 minutes to put us back on time. That will be 25 minutes until 4:00.

Could you please join me in thanking our three speakers?