

Host: This is the podcast from *Clinical Chemistry*. I am Bob Barrett. First, we all became acquainted with the genome, then the proteome, and now a whole host of “-nomics” have dominated research in clinical chemistry and laboratory medicine. These have helped to break down some barriers among laboratory disciplines, and now there are many projects involved in characterizing the microorganisms that are associated with human health and disease.

In the fall of 2008, the National Institutes of Health announced awards for its Human Microbiome Project. Dr. James Versalovic is the Head of the Department of Pathology and Director of the Division of Molecular Pathology at Texas Children's Hospital and an author of a paper in the May issue of *Clinical Chemistry* on metagenomic pyrosequencing and microbial identification.

Dr. Versalovic, can you tell us something about the Human Microbiome Project? What exactly is the human microbiome?

Dr. James Versalovic: The human microbiome is, technically, a set or collection of genomes belonging to microbes, and these microbes live in and on the human body. There are vast microbial communities that essentially are part of us as human beings as *Homo sapiens*, and these complex microbial communities remain largely uncharacterized and not well understood scientifically.

So the Human Microbiome Project is seeking to understand the genes and the microorganism containing all of these genes and how they impact human health and human disease.

Host: Now, what is the difference between microbial genomics and metagenomics?

Dr. James Versalovic: Microbial genomics is a relatively old science and by that I mean it has been around for 15 to 20 years. The individual bacteria or fungi that are cultured in the laboratory can be isolated as pure microorganisms and these single microbes can then be characterized in terms of their entire gene set or genomes.

The term “metagenomics” refers to the collection of genomes of complex microbial communities, and

these communities include many different types of microbes.

Only with recent developments and DNA sequencing technology in the past several years have we been able to address the issue of metagenomics and specifically in this case the microbiome. So the key difference is that in metagenomics we are seeking to understand the genomes of entire microbial communities in contrast to individual microbes with microbial genomics.

Host: In your opinion what are the trends in DNA sequencing technologies and the implications for biology?

Dr. James Versalovic: DNA sequencing technologies have evolved rapidly during the past decade, and in this century already, as brief as it has been, nine years, we have seen the proliferation of entirely new sequencing chemistries and the informatics and computational biology applied to the sequencing technologies.

Throughout the 1990s there were gradual improvement in DNA sequencing that culminated with the completion of the Human Genome Project in the years 2000 to 2001.

Since then, we have seen the advent of next-generation sequencing technologies, and there are a variety of next-generation DNA sequencing technologies that have greatly increased the power and throughput of DNA sequencing. These developments in the sequencing technologies have enabled us now to confront entirely new problems in biology that were simply difficult to comprehend in the past.

The Human Microbiome Project is one example. The field of metagenomics, as a whole, is now feasible as a science in biology, and so we are simply able to do much more in the genomics sciences now due to these developments in DNA sequencing.

Host: What are the challenges with the databases and organism identification?

Dr. James Versalovic: There are a variety of challenges with respect to informatics and the ability of store and manage very large information banks. The databases, as they have been structured to date, have been geared to accommodate a large amount of DNA sequence information and other biological information. But

quite frankly, we are dealing, now, with differences in orders of magnitude and huge scale up challenges in terms of the information.

So in addition to the developments in DNA sequencing, there has been a parallel enhancement and the abilities of scientists to deal with information, but quite frankly scientists right now are quite challenged by the complexity and the magnitude of the information that is being generated by a variety of DNA sequencing projects that include the human microbiome and metagenomics.

So as many new microorganisms are being identified via these metagenomics projects, we are now identifying and also detecting many organisms that are not well understood and many organisms, many microbes that have simply not been listed in the databases.

So we are also are refining our tools for searching and aligning DNA sequences, so that we can effectively match sequences to sequences already in the databases or in many cases we don't have an exact match, and so we have to do basically a nearest neighbor type of approach, where the DNA sequences basically can be grouped with their cousins or relatives and again because of the tremendous amount of new information and new microbes that are been identified via metagenomics, has posed challenges and fortunately scientists are now pooling resources and working collaboratively together in the various genome centers to tackle these issues and refine our search and identification tools.

This challenge in informatics will remain a major challenge in microbiology and metagenomics for years to come.

Host: Within this expanding field how are genome centers adapting to the metagenomics era?

Dr. James Versalovic: Genome centers are having to retool because of the increased complexity, the refinement of the DNA sequencing technologies, the massive amounts of data sets and information, now, that are being gathered in the centers are certainly challenging the centers to rethink how they carry out the work.

The various aspects of these challenges include the strains, of course, on personnel, the ability to handle the sequencing technologies and the information

generated with current personnel, the current structures of the laboratories, and the computational resources available in these centers.

So consequently the centers are now examining various options and developing new tools. Automation is now being ramped up — automation in terms of DNA sequencing, but automation also in terms of computational biology.

New software tools and algorithm are being developed that affectively enable the genome centers to automate much of the conversion of data to meaningful scientific information by automating search and identification algorithms and the ability to bin or categorize sequences into various categories and to then deal with metadata.

The Genome Centers are now also thinking about different metadata sets, metagenomics, and metadata in terms of clinical data, phenotypic data, RNA data, protein data.

So the Genome Centers are adapting to these challenges by increasing automation by incorporating new technologies and coupling these technologies with new software tools, and also rethinking how these work groups are organized. How people work together increasing effectively the collaboration between the centers to pool resources more effectively and take advantage of distributed computing and networks and so we are seeing a variety of adaptations and certainly many more to come.

Host:

From your perspective, what about the main challenges and informatics in computational biology?

Dr. James Versalovic:

Well, I have alluded to several of these points in the prior answer and I guess I could amplify further and maybe highlight a few points.

Informatics in computational biology are certainly being challenged simply by the scale and the magnitude of data sets that are now being generated by these next-generation DNA sequencing technologies.

So the simple issues or seemingly simple issues of data storage and management of large sets of data is certainly a primary challenge in informatics and how to effectively manage these very large databases will be an ongoing issue for many years.

Additionally, the area of computational biology is being challenged because there are new kinds of data sets being generated. So it's not just a matter of quantity but also quality. We are now being immersed in these metadata sets, which include genomes of vast microbial communities, many microbes that are simply unknown, and in fact many genes that are entirely unknown.

Genes that may code protein sequences that represent entirely new protein families. A good example is the recent project headed by Craig Venter in the Oceans of the World, in which marine microbial ecology was tackled, and many new sequences have been identified, gene sequences that frankly have no match, or even near-match in the databases.

So our tools in computational biology will have to be refined to incorporate these new kinds of data, sequencing data, as well as data evolving from protein biochemistry and structural biology that will enable us to link these sequences with reputative functions.

So these are few of the key challenges, and clearly we are going to have to evolve the informatics to go hand-in-hand with the developments in DNA sequencing.

Host: What about the implications of next-generation sequencing from the diagnostic laboratory in the 21st century, and the implications for personalized medicine.

Dr. James Versalovic: This is an interesting question, because medicine as a whole and laboratory medicine in particular are fields that are just beginning to confront the challenges of metascience, metagenomics, and next-generation sequencing.

The diagnostic laboratory has been applying DNA sequencing, now, for more than a decade in the clinical laboratory and the hospital, and medical centers, and reference laboratories throughout the United States, and internationally as well.

However, next-generation sequencing introduces several possibilities and the challenges that go along with these possibilities. Clearly there is much excitement about personalized medicine and the potential ability to sequence entire human genomes of individuals to predict susceptibility to human

disease and the potential to use specific pharmacologic agents based on the human genome.

The human microbiome introduces another dimension to this whole aspect of personalized medicine. By being able to sequence and understand the human microbiome in the future and coupling that information with human genome sequence data we may have a more complete picture of the human individual and that more comprehensive or holistic view of personalized medicine may introduce new opportunities for selection of particular drugs and prevention of human disease the customization of particular dietarian or nutrition strategies for human individuals or particular subpopulations of human beings.

These are a few of the possibilities that next-generation sequencing is now introducing in to diagnostics because it enables us to tackle the whole problem of metagenomics.

So, again, I just want to highlight the fact that next generation sequencing is now enabling us to think about the human microbiome in addition to the human genome, and it makes this evaluation practical potentially in the future as sequencing costs come down, as new medical informatics tools are developed to rapidly translate the wealth of DNA sequencing data in to practical clinical data. We do have the opportunity to transform the diagnostic laboratory in this century.

It will probably take another decade, at least, to translate and port these technologies in to the mainstream diagnostic laboratory, but clearly a number of physicians, pathologists, clinical laboratory directors are now beginning to discuss this issue at national meetings, and we as a medical and scientific community are already considering the implications of these advances for medicine and personalized medicine in particular.

Host:

Our guest has been Dr. James Versalovic, Head of the Department of Pathology and Director of the Division of Molecular Pathology at Texas Children's Hospital.

Total Duration: 14:48 minutes