# United States of America

# Food and Drug Administration

# 2014 Next Generation Standards Conference

**Bethesda, Maryland**

**Thursday, September 25, 2014**

A G E N D A

Biologics Product Evaluation:

ARIFA KHAN, PhD, Session Chair
Supervisory Microbiologist, CBER
Food and Drug Administration

PHILLIP KRAUSE, MD

CHARLES CHIU, MD, PhD
Associate Professor, Director, AVDDC,
UCSF School of Medicine

ROBERT CHARLEBOIS, PhD
Senior Scientist Sanofi Pasteur

JOHN THOMPSON, PhD
Senior Investigator Merck

Clinical Biomarkers and Personalized Medicine:

ERIC DONALDSON, PhD, Session Chair

RAJA MAZUMDER, PhD

ANDREA FERREIRA-GONZALEZ, PhD
Professor, Virginia Commonwealth University
Medical Center

ANDREW GRUPE, PhD
Senior Scientific Director Pharmacogenomics,
Celera Quest Diagnostics

CHARLES SAWYERS, MD
Chair, HOPP Memorial Sloan Kettering Cancer Center

LAURA J. VAN'T VEER, PhD
Leader, BOP Associate Director, Applied Genomics
UCSF HDFC Cancer Center

PARTICIPANTS (CONT'D):

Next Generation Sequencing Devices and Clinical Applications:

ZIVANA TEZAK, PhD, Session Chair

JUSTIN ZOOK, PhD

HEIKE SICHTIG, PhD

JONAS KORLACH, PhD

JAY SHAW, PhD

KAREN GUTEKUNST, PhD

KENDAL DINSMORE

IRA LUBIN, PhD
FACMG, Team Lead, Genetics, OPHSS Centers for Disease Control and Prevention

JUSTIN ZOOK, PhD
Researcher, Genome in a Bottle National Institute of Standards and Technologies

MYA THOMAE
Vice President, Regulatory Affairs Illumina

HEIKE SICHTIG, PhD
Principle Investigator, Regulatory Scientist, CDRH Food and Drug Administration

Food Safety and Pathogen Detection:

ERROL STRAIN, PhD, Session Chair

HEIKE SICHTIG, PhD

PARTICIPANTS (CONT'D):

MARC ALLARD, PhD
Research Microbiologist, CFSAN
Food and Drug Administration

BILL KLIMKE, PhD
Staff Scientist, NCBI
National Institutes of Health

KRISTIN G. HOLT, DVM, MPH
FSIS Liaison to CDC Food Safety and Inspection Service, OPHS
U.S. Department of Agriculture

* * * * *

P R O C E E D I N G S

MS. VOSKANIAN-KORDI: Hello everyone. If everyone could get seated, we're going to go ahead and start.

MS. KHAN: Hi. Good morning everyone. Welcome to Day 2 of the conference. I hope everyone was able to stay dry, but you have a lot of, you know, coffee around, across the hall as well as in another café.

So basically, today our session on biologics product evaluation is going to cover both the use of NGOs for novel virus discoveries, as well as some development of bioinformatics tools, as well as optimization of bioinformatics pipelines.

I'm the first speaker and I'll start the first session. The format of this session, actually, will be a little bit different than the ones you had yesterday. We're going to have four talks, and then after each talk we'll have a couple of minutes for questions that are immediately relevant to the talks.

Then we are planning to have a discussion session for the last 30 minutes. That will give opportunity for more audience participation as well as -- I've created some bullet topic points for discussion of the audience can't come up with some hot topics on their own.

So anyway, for my talk, before I start I'm going to present the same disclaimer as you've seen for the other FDA speakers that my comments are an informal communication, and these comments do not bind or obligate the FDA.

This is a very general outline of my talk. I'm going to start with presenting the need for considering new virus detection methods in CBER, and then present FDA ongoing internal and external efforts for advancing new technologies for evaluation of biologicals. Then try to present the current status of the technologies and further challenges.

So the question is why is CBER interested in using NGS, and the main concern in biologicals, safety concern, is still advantageous

viruses. This class of agents includes a variety of different types of viruses such as exogenously acquired viruses which can be replicating or latent, genetically inherited viruses such as indigenous retroviruses. Viruses that can be de novo-generated such as some novel recombinant viruses.

Virus like particles which may contain co- packaged, unwanted RNA or DNA. As well as, in some cases, the presence of a reverse transcriptase activity that is known to be produced from avian as well as some insect cells. The concern there would be whether this activity is associated with retro viral particles or retro transposons, and if those particles are infectious.

So there are strategies that need to be used to mitigate the risk of advantageous viruses, and my office, the Office of Vaccines, uses a four-pronged strategy indicated here. This is to initially identify potential safety concerns do enable development of a comprehensive testing plan

and risk mitigation strategy. This includes characterization and history of the cell substrate, and cell banking, and use of qualified raw materials.

As you notice that these points are, you know, relevant to cell based products, but some of them can also be relevant to other biological products. For cell banking, it's important to develop well-characterized cell banks, and then to use certified or tested animal drive biological materials such as serum, trypsin, antibodies. I should also point out that in cell-based products these two are the main sources of potential contamination.

Additionally, the incorporating of steps during manufacture for viral clearance and purity can be done in the case of some of the inactivated products. This includes process validation for clearance of potential viral contaminants in final product, as well as reduction of residue cellular materials such as RNA, DNA, and proteins.

Testing is very important and critical,

especially for live viral vaccines where you cannot incorporate inactivation steps or viral clearance steps because the product is a virus. In such cases includes extensive testing for known and unknown agents in the starting materials, in advantageous agent testing, and different steps in manufacturing process, and at steps with the greatest potential for contamination, and using various sensitive and broad detection assays.

This slide indicates the conventional assays that are used for evaluating advantageous viruses in vaccine cell substrates as well as in certain other biological products. The general assays include in vivo and in vitro assays, transmission like turn microscopy, as well as a highly sensitive PTR based reverse transcriptase assay for detection of retro viruses.

Additionally, specific assays can be used for specie specific viruses based upon the raw materials that you're using, the cell line. As well as sources of potential introduction of viruses, and virus specific assays such as PCR and

infectivity assays. These are described in detail in the guidance for industry, the 2010 Cell Substrate guidance that is indicated at the bottom as a reference.

In certain cases, there are additional assays, which is recommended on a case by case basis. These, actually, are more relevant to some of the new cell substrates that have been introduced for vaccines. These include oncogenisity assays if a cell substrate is tumorigenic, and to look for tumor inducing viruses, and chemical induction assays, if there are concerns for the presence of latent viruses, such as indigenous retrovirus and latent DNA viruses since such viruses may become activated due to the manufacturing conditions.

However, the list of novel cell substrates for vaccines continues to length. Here there are some examples of cell lines from different species that are currently being used. As you can see, there are a variety of species. Some of them, actually, are quite challenging in

terms of -- for regulatory review, and especially these cell lines A549 hela CM, which are human tumor derived cell lines.

So therefore, the use of such cell lines, and the variety of different cell types and species poses a advantageous agent challenge for the use of novel cell substrates that includes detection of known unknown as well as unexpected viruses. These can be tumor inducing viruses, as I mentioned, the case of tumor cell lines, latent viruses, occult viruses, and novel viruses. So the bar is quite high now, especially for viral vaccines in terms of using technologies that can detect all possible agents.

Therefore, the need for considering new technologies for broad novel virus detection. I've listed three here, by no means are these the only ones, but these are the ones that are established and, I guess, available for use currently.

Next generation sequencing technologies have demonstrated success in detection of novel

viruses, and can be used for both known and novel viruses. There are many different platforms, and the challenges of using these are the topic of this meeting. Additionally, there are micro arrays that are useful for the detection of known and related virus sequences. There's also broad range PCL with mass spec that actually has not been mentioned here, but this uses long PCR primers that are specific for virus families. All of these have demonstrated success in novel virus discoveries as indicated in the last bullet for each one.

So some of the potential applications of these technologies for biologicals are for cell substrate investigations and characterization. As you know, if could be a whole genome sequence, if you wish. Transcriptone analysis, as well as analysis of virome on the supernatant.

Additionally, these can be potentially used for evaluation of animal derived raw materials to characterize the starting material, as well as for product stability or purity, and

other possibilities in the future.

So, this slide lists various efforts that have been done in the past, and more recently also, in addressing advantageous virus concerns for vaccines. As you can see, there were PDA meetings from 2010 as well as more recently that addressed advantageous viruses and mitigation strategies earlier. Now more recently, for example, from 2011 and 2013. These have also included discussions on the use of emerging technologies, and the challenges posed by these for use for regulatory applications.

Additionally, the IABS is active in discussions on new technologies for advantageous agents, and also, in 2013 with the WHO study group held in Beijing that also raised cell substrate issues and discussions on using DNA based technologies.

As you can see, all of these do have proceedings published as well as -- or those that are in progress. In our meeting in 2013 the proceedings are expected to be published in

November/December issue this year.

There are other group efforts. These include the formation of an advanced virus detection technology users group that was in October 2012. That was a coordinated effort by FDA industry and PDA. The mission of this was to advance the tools for next generation of viral risk evaluation by providing an informal scientific forum for discussions and scientific collaborations.

This group does comprise of scientists from various aspects, from industry, from regulatory and other government agencies, academia, and technology service developers, as well as providers. Most recently this year this group has now been converted to a PDA interest group, and the participation, actually, has even broadened. We have about 100 members now. We do have very active discussions and sharing of knowledge, as well as data. So if anyone's interested please feel free to contact me.

Additionally, it was mentioned yesterday

in Dr. Wilson's talk that in 2013 the FDA genomics working group was formed with a mission to prepare the FDA to address IT and scientific challenges to facilitate its readiness for HTS data review. This is being done through establishment of robust infrastructure as well as through collaborations.

So the result of the various meetings was to identify some of the challenges for -- common challenges, I should say, shared by regulators as well as by industry for regulatory applications of emerging technologies. This included assay sterilization including assay specificity, and sensitivity using appropriate model viruses as well as sample handling and processing.

It included bioinformatics, that was much discussed yesterday, in terms of data analysis, as well as data storage and transfer. In the terms of data analysis, it was recognized that there was a need for guidelines for acceptable quality of Reed's parameters for short

reed assembly, approaches to identify a novel virus that has minimal nucleic acid sequencing model to known viruses.

As well as a need for a publicly available complete, curated reference virus database, and we've heard about databases yesterday as well. Of course, as was discussed, the need for deciding the format and security measures for data storage and transfer. Additionally, one thing that was not mentioned so far is the need for a follow-up strategy to evaluate a positive signal to determine its biological relevance and significance of the signal.

Some of these challenges have been undertaken by the advanced virus detection technologies group. A pilot spiking study for assay generalization was initiated in 2013. The viruses were selected based upon their commercial availability and relevant viruses were selected to encompass different physical and chemical properties, and a human cell line was chosen for

background nucleic acid.

Three labs are participating in this study, two industry, and my lab, and the FDA, and using at least two different NGS platforms. There are different sample types were created independently based upon shared information regarding virus titer particle count and viral genome copies.

There were different types of samples, such as viral nucleic acid spiked into background of cell nucleic acid, and viral particles spiked into cell lysates. The results are currently being obtained and will be shared after internal review and analysis, and publications are planned for broader sharing of these results.

The results from these studies, we hope, will provide important information for further development of virus reference materials for NGS technologies. This was also an important point in the questions that were raised yesterday.

Another effort that was initiated by the group that is ongoing in my lab is creating a

complete virus reference database. The approach that we have undertaken is to download all the Genbank entries, except phage bacteria and contigs. Virus related entries were selected based upon positive key words, and then tagged with non-viral negative words for manual review and decision making.

We had a Version 2 database that was internally tested and revised to improve format and data presentation. There's a Version 2.5 that was distributed through the NCBI for initial evaluation by the Advanced Virus Technologies Group. Comments were received, and additional revision is ongoing. Further efforts are directed towards clustering to reduce redundancy, to add viral related sequences from other databases, and planning a web portal for broader distribution and for public use.

However, community efforts will be needed for accurate annotation, curation, and continued maintenance of the virus database. That's where we seek all of your help.

One thing I just wanted to point out that would be noted is that there's some viruses that are misannotated in the Genbank has host because one of the questions that you might pose is why do we need another database? We have, you know, the NT database, and we have our own private databases.

Well, we are trying -- the effort is to have a complete virus database. One of the main, I guess, efforts is to include an accurate and annotate indigenous retroviruses. Based on my expertise in this subject matter I'm able to contribute toward that effort.

So some examples are indicated here that are currently taxonomically misannotated. It's highlighted. Like indigenous retroviral DNA from cat is indicated as the host. So if one were to look for viruses under taxonomy, based on taxonomy you will not pick this up.

Similarly, for the RK indigenous retrovirus LTR you would not pick that up or a mouse memory tumor virus or a retro transposing

like sequence from the ZMAs. So these are just some examples as to why we need to continue efforts to have a much more accurately annotated database for viruses.

Additionally, we had posed a follow-up strategy of a positive signal to determine its biological relevance. I won't go through the points here, but basically you can read them. I will just mention we have been successfully applying this proposed strategy in our own lab when we are evaluating cell substrates.

I would just like to show this application of this strategy successfully in our recent discovery of a novel rhabdovirus in assay lines cell lines. This was based upon MPS efforts or data from the transcriptome, from nucleic acids from the filtered supernatant, from total nucleic acids that were concentrated from the filtered supernatant. As well as from sized fractionation filtered supernatants.

This was done on the SF9 cell line which is from (inaudible) insect, and the fall army

worm. This cell line is broadly used for Bactra virus produced products, vaccines as well as therapeutics.

The strategy that we use, we were trying to encompass detection of all possible, even distantly related sequences, so the data generated from the MPS was analyzed using blast initially, by using blast N. We used it against the NRNT database. We used blast X against the protein database, and then we found that we have to use T blast X, actually, I'll indicate shortly why, against the NRNT as well as the virus database.

Then further then the results evaluated by further binomephomatic analysis. We used the CLC genomics workbench. Then we were able to successfully identify the novel rhabdovirus.

This just gives you details of the data obtained from the different samples indicated here. This is from the transcriptome, from the soup, and from the concentrated TFF sample. The contigs from each resulted in a large fragment indicated here, 5 to 6 KBs.

However, we were only able to detect this particular fragment. Then had to manually, actually, walk up the genome to assemble the other sequences here which resulted in the assembly of a full length, almost full length, rhabdovirus. That's indicated here.

This indicates the struggle in trying to assemble the sequence. Because, basically, the extent of hemology or the extent of identity of this whole 13 KB sequence to any known sequence was only within a 295 base pure region in the L protein gene.

Even there the identity was only 66 percent. The most identity 66 percent in the nucleotide level was found to this task strip virus which is a new member of the mononegroviralus. The 4 percent coverage indicates the 295 base pure region that showed this identity, so the rest of the genome did not have any identity to any known virus.

But using Blast P, indicated in this column here, we were able to get 53 percent

coverage, but 27 percent identity to the lettuce yellow model virus. You said well, what happened to the taster. Well, it was not in this database because that virus did not have a complete L protein available, so it was not included in the protein database.

Our T Blast X results were the most useful because they were able to identify the tastra virus from the NT database with more coverage. The identity was 31 to 47 percent. Therefore, we did recognize that there is a need to use all possible tools if you're looking for novel virus detection.

This is just a phylogenetic tree indicating the novel viruses up here. With the tastra virus it is distinct. It's more closely related to plant rhabdoviruses, than to animal rhabdoviruses. This result has been published this year.

So this just comes back to our original scheme for virus detection. As I mentioned, T Blast X was useful. However, this was extremely

tedious. Literally it takes about two weeks to get the data. This is where we think that having a reference viral database would be very useful because it would weed out all of the cellular background that's currently in the NT database. So we hope that in the future this will make using T Blast X or any other format engines much faster and more specific for virus detection.

I just wanted to show a picture by cryoTN that yes, we did see a virus particle, as it is one of the checkpoints for our strategy. Infectivity analysis was done under conditions of mammalian cell culture as well as insect cell culture. We were not able to show replication in an mammalian cell lines. However, there was limited infection in the insect cell lines.

These are some additional examples of efforts in the Office of Vaccines. As I mentioned, the effort on the rhabdovirus has been published. We also have looked at other technologies. This is also published on the broad range PCR electro spray ionization mass

spectrometry system as well as virus arrays for virus detection.

Additionally, there's been efforts in Phil Krause's lab, who's my co-chair for this session, he'll come up after me to introduce the other speakers. They have recently published optimization of virus detection in cells using MPS.

Also, in Costa Chumacoff's lab they're using deep sequencing approaches for genetic stability evaluation. They've published on influenza A viruses, and I think Dr. Wilson's talk also referred to their earlier work, and looking at polio virus vaccine stability. In 2010 they published a PNA paper on that.

Okay. So in conclusion I just want to mention that the new technologies may be useful for supplementing the conventional assays for novel virus detection. I should say these are not my personal conclusions. There are, sort of, conclusions of the field, and the field meaning from the various meetings we've had. These are

the generally conclusions.

Then the results from ongoing studies can provide scientific direction for additional work that may be needed for advancing new technologies for regulatory application, particularly the development virus reference materials, and that is my own addition here. Due to the complex nature of the new methods and associated bioinformatics, collaborative efforts among industry and regulators should continue to identify and address the technical challenges and efforts toward harmonization. These should be initiated among regulator health agencies sooner than later.

I think that this meeting is one of the formats, you know, towards that as well. So I should just acknowledge the people in my lab, Helen Mont, Terry Galvon, and Saya Chiziabod who were involved in the rhabdovirus study. Helen actually spearheaded that study. Aisha Aljahana, he has led the efforts on the virus database development, and Laryn Talifaria on the spiking

studies that I described.

Rodney Bristor from NCBI has really been extremely helpful in having ongoing discussions, and whatever needs to be done in terms of addressing any comments or concerns about the database. Yesterday he mentioned he'll try to do it as soon as possible if he can't do it the next day. I should say he generally gets it done the same day, so please, you know, if you have any comments do share it with him.

Also, the members of the Advanced Virus Detection Technologies group. There are sub-leaders for each of the different topics that we're addressing. Siemon Ng Sanofi leads the efforts on the sample preparation and processing. John Pol Casart on the virus selections and standardization. Kavetha Bekari on the virus database, and Adam Palermo on the optimization of the pipelines.

Again, these are just some of the -- these are just the sub-group leaders, and, of course, each of them have a lot of other excellent

experts involved, and some of them are participants in this conference. Thank you.

I'm going to turn the session over to my co-chair Phil Krause now. Oh, questions. We do have time for a couple of questions or maybe not, but we can take them.

QUESTIONER: Very nice presentation. Thank you very much for that. I leave with uncertain when you getting to the slides and the conclusion. Basically, in the last slides you mentioned the new technology will be helpful to generate a hypothesis and to suggest new assays, but it's not ready for the regulatory applications. Do you want to elaborate a little bit more on the last slides?

MS. KHAN: I think we were trying to do this toward regulatory applications. Is there a mistake on the slide? These efforts are ongoing so that we can, you know, move ahead with these technologies towards regulatory applications since there doesn't seem that they can be, you know, very useful in certain situations. We can

definitely discuss it more in the discussion session if I haven't been clear about that.

QUESTIONER: So the question is if they are building a gold standard reference database for biogenomes is Dr. Khan's group working closely with the virus pathogen database and analysis resource group? Seems like there's a lot of potential for redundancy if not. That's cited as the IPRBRC.org.

MS. KHAN: Basically, I should mention that in terms of calling it a gold standard, I think it will always be work in progress. I think that's for any database. We hope that it will be useful, you know, for the applications that this is intended, which is for biologicals.

So basically I am aware that there are many database efforts, and more are being brought to our attention as we discuss them publicly. But I think what I've also heard is that the different databases are maybe useful for specific applications and being created for specific needs. Therefore, it may not be as global as what we

want, which is to look for any virus, whether it's, you know, regardless whether it's a pathogen or not. You know, because we want to know if it's infectious, in the end.

So even if it's a sequence of a virus we're still interested in that because then the follow-up will be whether if there's a related virus can that virus be infectious. So I think ours is a little bit more global. But definitely we want to coordinate efforts and reduce redundancy if we can. But I'm hoping that once the different database efforts, you know, come nearer to finish then we can be at a point of sharing, you know, the databases, and maybe come up with a more global.

Definitely the goal of our database is to work with NCBI, at least in their current -- you know, to incorporate our database and theirs together towards, you know, a virus reference database. So hopefully, you know, as long as we're aware and we keep talking about what the databases are and what the goal are, hopefully,

you know, eventually we can come up with, you know, one, if it's possible, that can be broadly used.

MR. KRAUSE: I think Arifa's talk nicely illustrated the complexity of using these next generations sequencing techniques in a regulatory setting where you're looking at the potential for contamination of a product. Of course, if you find something that is a sequence it doesn't necessarily mean that -- I'm not the speaker, I'm just the --

SPEAKER: (inaudible), you know?

MR. KRAUSE: This doesn't work?

SPEAKER: This is for (inaudible).

MR. KRAUSE: I see, okay. It nicely illustrates the complexity of doing this in a regulatory setting where a positive result can mean a huge amount of extra work to figure out what that means, and a negative result may have a very clear consequence to somebody who's using a regulated product.

In particular, that's difficult because

when these techniques are being applied, at least initially, we're finding more of these potential positive results that request a huge amount of follow-up, and, perhaps, not as many of the kinds of results that would necessarily be concerning from a safety perspective.

So the other speakers in the session are going to address this issue from their own perspectives. I'm very pleased that the next speaker is Charles Chiu from University of California at San Francisco, who's going to talk about next generation sequencing assays for pathogen discovery in the infectious disease diagnostics.

He, in particular, has experience using these techniques in the context of a CLIA certified lab where the stakes are very high as well, and so his perceptive is going to be very valuable here. Thanks, Charles.

MR. CHIU: I just want to start by thanking the organizers, especially Arifa and Phillip for inviting me to speak with you today.

I'm going to talk about, continue, sort of, what was brought up by Dr. Khan, and discuss, sort of, our experience with using next generation sequencing.

My laboratory spans a wide variety of different, kind of, areas of focus. Part of it is in development of NGS or next generation sequencing based diagnostics. We also do pathogen discovery. As well as, looking for, kind of, advantageous agents. I'll actually give you examples of all of those during this talk.

I want to start by just disclosing that I do get research funding from Abbott Diagnostics, from Amazon.com, to establish NGS based pipelines, and from Rubicon Genomics.

So the general idea is that we want to use NGS as a way to cast a wide net to detect pathogens. This is opposed to traditional methods like culture and PCR where typically you're only targeting one agent or a limited number of agents. We're just using NGS to really pick up the full spectrum of agents that can cause disease, whether

they're viruses, bacteria, fungi, or parasites.

You might think that there probably are many different applications or potential use applications by which NGS would be very useful. In clinical diagnostics, and pathogen discovery, and in advantageous agent detection. Our approach is to use, what I'm referring to, as unbiased medogenomic next generation sequencing.

The general idea is that this is a method that's -- we're, essentially, just sequencing everything in the clinical sample. We're not using specific primers and probes. The idea is that by sequencing everything in the sample, in theory, you can reconstruct the sequence and assemble the sequence of really any and all pathogens, including viruses, bacteria, fungi, and parasites.

Now, part of what's made this method/approach attractive is the fact that now that there are these portable instruments that can be deployed into clinical laboratories, and are actually being deployed. As well as the fact that

now you have enormous sequencing capacity such as, you know, 30 to 40 million reads on the miseek, aluminum miseek or more than a billion reads in the hiseek as well as other technologies.

In addition, there are not methods to bar code primers so you can multiplex clinical samples to reduce cost per run. However, a big challenge with this whole field is that it's really required the development of both customize, specialized protocols as well as novel bioinformatics analysis tools. I'll be focusing primarily on the bioinformatics analysis here.

Our general approach is once you have your raw sequencing data, as you might expect, the vast majority of reads correspond to the hosts. So if you're looking at human samples, clinical samples, probably 99.9 percent of your reads are going to be from human.

This is really a needle in a haystack problem. We're not interested in the human fraction. Although, you know, geneticists and oncologists will probably be interested in doing

that. We're interested in the fraction that corresponds to potential pathogenic microorganisms.

So the general approach is called computational subtraction. We're identifying for reads that corresponds to human. Then we aligned those remaining reads to all sequences in -- pathogen specific databases are all sequences in reference databases. This tends to be a huge computational bottleneck. As, kind of Arifa just spoke, after the last talk, this is something that can take, easily, days to weeks using traditional blast based approaches.

In addition, the magnitude of the bottleneck actually tends not be highly dependent on the clinical sample type. So if you -- it's a little bit ironic that if you think of tissue, like, say, human tissue where the vast majority of our sequences are human. It's actually quite efficient to just subtract all of the human reads from it. Then you might end up with a small fraction that, hopefully, correspond to the

pathogens of interest that you're looking for.

On the other hand, take the other extreme, which is a stool sample, which is, essentially, an environment medogenomic sample. Where even if, and this is actual data, where even if you eliminate all human reads, all viral reads, and all bacterial reads you still have more than 50 percent of your reads that are presumably something else.

They could be unidentified, unsequenced bacteria. They might be environmental sequences. They might correspond to insects or animals, or some components of someone's diet. So this is a big problem because the magnitude of the bottleneck really depends on the clinical sample you're interested in looking at.

So to give you some idea of the bioinformatics challenge, this is actually a graph showing, kind of, how long or how expensive it is to analyze 100 million sequences. So if you take, for instance, a single machine, like a, you know, a 16 core server, it takes about three months

which is not very conducive to, kind of, short turnaround times for diagnostics.

Now, you can, kind of, move up the chain where you can go from a network to a cluster, and then finally to a cloud. On the Amazon cloud, for instance, you can actually recruit up to 5,000 cores of data. You can actually get that amount of sequence analyzed by using a blast-based pipeline in about six hours.

But the problem with using a cloud then is that the cost rapidly mount out. So this gives you an example. It actually costs almost $400 just to do the analysis in 6 hours by recruiting 5,000 cores.

So we quickly realized that we needed to find a better way to do this. So we developed a pipeline called SURPI, sequence based ultra-rapid pathogen identification. What it does is actually harnesses two algorithms. One of which was developed from a former post-hoc in my lab who collaborated with UC Berkeley and Microsoft to develop and algorithm called SNAP, which is

essentially a rapid or ultra-rapid nucleotide aligner.

We also do de novo assembly. We use, kind of a translated nucleotide or protein search to look for, kind of, you know, highly divergent viruses for pathogen discover.

I want to talk a little bit about SURPI. Specifically, Dr. Brister actually gave a talk yesterday that was a very good talk about the construction of viral reference database. We actually, for SUPRI, we actually use all of Genbank NT. We're not using all of Genbank NT because we're able to do it. They're actually some good reasons why we're choosing to use Genbank NT instead of, say, a curated viral reference datable.

Part of it is that by using that for pathogen discovery it gives us immediate access to all of the newly described partial sequences or gene-only viral sequences that are added to the database. The second thing is that this, actually, provides more efficient viral discovery.

What do I mean by that?

What I mean by that is by aligning your sequences to everything in Genbank NT you can then quickly identify all sequences that are non-viral. Whether they're human, bacterial, fungal, insect, algae, animals. The idea behind doing that is that then you can take the remaining sequences and very efficiently use a small subset of sequences to do de novo assembly of viruses, and potentially to do faster, kind of, nucleotide or translate nucleotide or protein searches which usually is, kind of, the computational taxing part of the algorithm.

Another big issue is that with curated databases such as Refseek there's a loss of specificity. The reason that exists is because viruses are incredibly diverse. The problem with doing that is in order to capture the diversity of viruses with using, I'll say a viral reference database, you need to use a very low stringency cut off.

By doing so you then end up with a

larger number of false positive. So this is why our solution was just, simply, to use everything in Genbank NT. That has its own problems, and I'll go into that in a little bit.

But the advantage of doing that is that we are able to do it because we have an algorithm called SURPI which is incredibly fast. So this gives you an idea. This is SURPI which uses this algorithm called SNAP. Aligning nucleotide alignment to the human database, which is about 3 gigabases in size.

You can see here it takes SNAP about two minutes to align 25 million reads, and about 7 minutes and 30 seconds to align 125 million reads. The comparable times for blast would be 16 and 80 days. For a billion sequences it's actually fairly reasonable. It can align a billion sequences in about an hour and a half, and it would take 2.2 years on exact same machine for blast to do that alignment.

Because it's so fast we're actually able to also do alignment to NT. I can tell you about

10 million reads by aligning to NT which is -- at that time it was about 42 gigabases. It's actually now 70 gigabases, NT. You can actually align 10 million reads to all of NT in about 10 minutes. Basically with, kind of, the use of SNAP.

So we felt that we've dealt with, kind of, the issue of -- at least partly dealt with the issue of this computational challenge of this bottleneck by implementing something like the SURPI pipeline where we can analyze data sets from 10 minutes to 15 hours.

Now, 15 hours was kind of a worst-case scenario where we took 100 million reads from a highly complex stool sample. Basically, we had to kind of go through everything. Couldn't really do a computational subtraction with it. You can see here that the costs are now reasonable. So I can get this on the cloud, for instance, I can actually get the same data analyzed, about $4 versus more than nearly $700 previously.

I also want to mention that we've

actually published SURPI. We published SURPI about four months ago. But in the meantime, we've added some bioinformatics enhancements to SURPI. One is that, you know, one of the challenges of using all of NT is there are a lot of misannotated sequences in NT.

As you might expect, there are sequences that have been misannotated as virtual. So we actually have a rapid filtering algorithm that can take the output, the raw output, and then rapidly, kind of, re annotate, you know, sequences that had been misannotated.

We've also included a method that you can do, kind of meta data tagging. You can basically provide additional information in the SAM file that will give you some metadata information. Such as, whether or not this is a clinical pathogen, where is this from, the host, etcetera.

You know, I think that Dr. Brister talked a little bit about that as well, about intents to incorporate that with the databases.

We've developed, also, a front-end visualization tool for SURPI, a web-based interface. Then we've also used SNAP in a different application. We've actually used it not only for aligning to reference databases, but also for doing taxonomic classification.

That actually becomes really, really important because, as you might imagine, viruses have conserve regions and divergent regions. The problem with the conserve regions is that conserve regions basically can represent a wide number of viruses at various taxonomic levels.

So, for instance, you can have sapid species, more than 100 different species of viruses. They may share the exact same read or short region, and by doing accurate taxonomic classification you can then really be able to identify exactly what virus you actually have in your sample.

In addition, we've also been able to do some enhancements to the memory usage of SNAP, so that now you can, for instance, put all of Genbank

NT in memory. We're able to do that. You do need a 1 terabyte memory server, but that's -- actually now, actually it's not -- I just purchase a server. It's about $8,000. It's a reasonable price, actually, to get a 1 terabyte memory server. That enables you, for instance, to align -- 10,000 reads, everything in NT in about 5 minutes versus 40 minutes before in the published paper.

This gives you some idea of the front-end visualization tool. So this is actually automated output coming out of the pipeline. So I'm not doing any, sort of, manual, kind of curating, making it look nice. So what it'll do is it will not only identify, kind of, the closest hit. This is actually a clinical sample from a patient with Ebola virus infection. Not from the West African outbreak, from a previous outbreak in the Democratic Republic of the Congo.

You can see here that the it basically found that the closest hit was Ebola Virus Zaire. It actually gave you the actual GI corresponding

to it. It gives you the pure wise identity. It will take the reads that were generated and plot them across the genome of the virus. It will also give you, what I call a pure wise identity plot.

It will tell you, kind of, at a glance, kind of, how difference is this virus from the viral genome in the database. That's a quick way to identify recombination events, etcetera, and also map that against the genome of it's -- genomic structure of the virus as from Genbank.

The nice thing about this is it will do that for all of the viruses that were detected in your raw data, and it will do it in automated fashion. We're hoping to develop some additional tools that will allow you to edit, kind of, this output that comes out of the pipeline.

In addition, we've had the pipeline -- also, it can be used as input into what we call Kroma plots. These are taxonomic plots that, essentially, show you, kind of, the distribution of the viruses that were detected at different taxonomic levels.

This gives you an example. This is actually a stool sample where you can see that, you know, about 90 percent of the reads were from norovirus. This is a norovirus infection as well as a par echovirus which is kind of a secondary infection in the same sample.

In addition, we've been really very interested in developing what I call a clinical work flow in a CLIA laboratory. So we have, actually, a mechanism for actually developing a process for actually generating this data and going from sample to sequence in about 12 hours to 2 days. We can take, pretty much any sort of sample type, serum, whole blood, respiratory secretions, cervical, spinal fluid, etcetera, and be able to, kind of, put it through, kind of, a rapid automated pipeline.

We've currently been working on some spiking experiments where we're trying to calculate using probative analysis what the limits of detection are. You can see here that for limited detection of HIV as an example of an RNA

virus we're getting about less than 100 copies per mil form. This is spiked into plasma.

For CMV, which is a DNA virus, it's about 1,000. Although, we are improving that. Actually, the latest data suggests that we are also able to get, kind of, a similar level of sensitivity.

This gives you an idea of how we're using the approach to actually diagnose a large variety febrile illnesses. So these are respiratory secretions and serum from patients in the hospital with various diseases. You can see here that we can detect plasmodium falciparum, haemophilus, influenza, rhinovirus, menuavirus, adenovirus, adenovirus senna type 7, salmonella, typhee, the cause of typhoid fever. These are all clinical samples that we did metagenomic next generation sequencing from.

I also want to mention that there also was a part about sample prep. We've been working, also, on ways to actually improve the sensitivity of metagenomic next gen sequencing. We're doing

this by combining random application with what we call probase enrichment.

On the left you actually see a serum sample from a patient with Hantavirus pulmonary infection from the recent Yosemite outbreak. You can see here that this was actually a case that was PCR negative from the serum. But, if you actually do the probase enrichment you can now detect sequences, specific sequences to Hantavirus.

Similarity, on the right is basically a sample from a patient with Zaire Ebola virus infection. You can see that it was unenriched. It's at the limits of detection of PCR with a cycle threshold of 35. You detect very few reads, but then once you enrich it you can actually get, kind of, full genome coverage.

So we think this is a viable technique that could be used similar to what's being used in cancer now with cancer gene panels, you know, NGS coupled with pro base enrichment. We think that this is a viable technique to use for this viruses

as well.

I just want to briefly go over -- end up with like three examples of how we've actually used this technique for various applications. So the first is, basically, identification of contaminants or advantageous agents. We basically reported last year that we found a virus that we called powerful like hybrid virus.

We were screening hepatitis serum samples by NGS, and this happens to be a very divergent virus. It only has about, you know, 31 percent identity on the amino acid level. We then looked at three different data sets, and we were able to, basically, assemble and identify this virus from three different data sets. If you actually look at -- one of the data sets is actually water, okay?

That was concerning because, you know, we were able to assemble the genome of a virus from water. That's a problem. So we quickly looked at that. We actually looked at the phylogenetic analysis of this virus. It happens

to be some virus that happens to be between the powerful virada and the circle virada. It happens to be a powerful virus, circle virus hybrid.

If you actually look at the data we then went through all of our NGS data, and we were finding this virus everywhere. We found it in, you know, MRSA cultures. We found it in human samples. We found it in animal samples. We found it in stool. We found it in blood. So this was very suspicious.

Basically, it was around the same time that it was actually reported that they had found a hybrid DNA virus by a group at the NIH. They actually also reported this in PNAS. They had found a hybrid DNA virus with seronet from patients also with seronegative hepatitis.

We did further work and we actually traced that this virus actually could be traced to extraction spin column from a certain manufacturer. We actually found this. I one batch we found the virus in every single spin column from that manufacturer. In fact, we're

still actually seeing this virus now in spin columns, so it hasn't been a problem that's been. But if you're doing par viral discovery you probably should take a look for this virus.

This actually happens to be the same virus, actually, as the one that was reported in PNES by the NIH which was called NIH-CQV which we now think, if you actually screen environment metagenomics, you screen through this. So I ended up screening through 78 public datasets, about 213 million reads. Out of all of those reads, the only reads that had hematology to those viruses were found, actually, in coastal sea water off of California and off of the Pacific Coast.

Actually, we have a mechanism, we actually think that this actually happens to be a diatom related virus. It happens to be the silica is contaminated. We think the silica that are typically used as components of most spin columns by manufacturers, okay? So we think that this is, ultimately, the source of this virus which we believe to be an environmental contaminant.

I want to go over another quick example. We identified a new hemorrhagic fever virus called the Congo virus. It appeared to be the cause of a very small outbreak cluster of hemorrhagic fever in Africa. I also want to point out, you know, similar to Arifa who had discovered a very diverse rhabdovirus, this was a very diverse rhabdovirus. In fact, it only had about 30 percent amino acid identity.

But if you have enough coverage, so basically we generate about 140 million alumina reads. You can actually de novo assemble the whole virus. Even though there is no reference sequence, or at the time, there was no reference sequence in the database.

So we were able to assemble about 98 percent of the viral genome solely by looking for overlaps and doing de novo assembly. This gives you an idea of the difficulty. So what you see here is in grey, this is the map of the virus, the coverage map of the virus. In grey are parts of the virus that have absolutely no hemology to

anything else in the database at the time.

If you look here in pink, pink were basically regions of the virus that were only identified on the basis of translate nucleotide or protein alignment. It just shows you, kind of, the challenge with viruses because they have so much diversity. You can't rely, necessarily, on nucleotide alignment. You actually have to go to translated nucleotide alignment in order to identify these viruses.

I want to end by just one last case which is a case that's actually in press now at CID. This was a 55 year old male who presented -- who had a bone marrow transplant that presented with a rapidly developing encephalitis that initially began as hearing loss. He had, kind of, a negative work up, negative cervical spinal fluid, negative MRI scan. He had developed nausea, fatigue over several weeks, and had a negative MRI, but then repeat MRI several months later was positive.

He had developed at that time,

basically, abnormal signal in his thalamus and his mid-brain, as you can see here by the white arrows. So for this we actually took his brain biopsy, generated 381 million sequence reads. I'm not sure if I overdid it, but I was really worried because it was tissue. Then we ran it through the SURPI pipeline.

To give you an idea of, like, SURPI it took about hours to analyze the 381 million reads. But what we found, basically, we were able to assemble directly from that the full sequence of a novel astrovirus. Now, it actually turned out that this was not really a novel astrovirus. It had actually been reported by two different groups before. It actually is an astrovirus which is called astrovirus VA-1 or HMOC. For those of you that are in astroviruses, found in children with diarrhea.

We were also able to show that this is by subsequent RNA, DNA insitu hybridization. You can see that, shown there in the red, basically that this is using a probe for the astrovirus. It

appeared to directly infect, kind of, the neurons. So this patient had an invasive astrovirus encephalitis.

It's actually a little bit concerning because it had been reported before by Ann Lipkin's group that there was a boy who was immunocompromised just like our patient who has astrovirus encephalitis, and he actually died from it. Our patient is the first adult who is actually characterized with this neurovasive astrovirus infection.

He was started on ribavirin and intravenous immunoglobulin, unfortunately, we don't know where he got the virus. Presumably it was community acquired. Unfortunately, despite treatment he did deteriorate, but he passed away about four months after the diagnosis. But at least the fact that we were able to identify it, an agent, suggests that we did have four months to intervene if there had been a therapy. Unfortunately, there is no established therapy for astroviruses.

So I just want to end by talking -- by acknowledging members of my laboratory. Erik Samoya is the CLIA technician who's developing the NGS assay. Samina Cash is a biologist who, and microbiologist, who worked on the SURPI pipeline. Scott is a bioinformaticist who development the SURPI pipeline. We collaborate with Eric Delword on the NIHCQV study.

I'd like to also acknowledge funding from the NIH and Abbott file discovery ward, and amazon.com grant, and a UC discovery grant encephalitis. Thank you very much.

QUESTIONER: Thank you for the great talk. I have two questions. First is, there are some cases where viral load or titer will drop and you'll still have symptoms, so there must be some consideration in when you take the samples when you're doing the sequencing for anything that's present too?

Then the second question has to do with dark matter. Are you seeing anything that looks like virus that you can't identify at all? I know

from previous studies we've seen long stretches of nucleotides that encode proteins that look viral in nature, but no identity in the databases. Thanks.

MR. CHIU: So the answer to your second question for us, we definitely see it. We do see these long contigs that have been formed that have no resemblance to anything else in the database, and we don't know whether they're virus or something else. For those we've been archiving them and, you know, in the hope that eventually we -- you know, somebody or -- if we would not be able to definitively identify what they are. So we are seeing dark matter.

The first question is a really good one in that you're right that it's really dependent on when you get a clinical sample. Because this is a direct detection method just like PCR. It's really dependent on there being the virus or the pathogen in your clinical sample.

Unfortunately, we've been focused mainly on what I call Hail Mary samples. Basically, you

know, samples of last resort where they've been extensively tested for other agents. As you might expect, we're getting a very low percentage of positives.

But I think that where this assay really belongs is a, kind of, early second line assay. So you would do, kind of, the frontline tests upfront, like bacterial culture, respiratory virus testing, and then as -- if it's still negative and you're considering sending out tests, more esoteric testing then you probably would want to have a test like this in place, a second line testing. But not as, kind of, a test of last resort.

MR. KRAUSE: Just to make a quick comment. We're running a little over time, so I'd like the questions to be quick and the answers to be quick. Maybe we take two more questions.

MR. CHIU: Okay.

QUESTIONER: While you're generally looking for viruses, you're also, basically, containing the full human genome of these

patients. How do you protect patient privacy, and how do you store your data so that it would be protected?

MR. CHIU: Yeah, that's a great question. So essentially, I mean, these are all protected under, kind of a VPN. That's for the hospital VPN, so it's considered clinical data. It's considered, kind of, HIPPA protected data. That's how we're keeping it.

Now, with respect to actually providing this information publicly, we've been removing the human sequences and then depositing it into SRA. Otherwise it has to go into something like DD gap.

MR. KRAUSE: We'll go to the last question over here.

QUESTIONER: Great talk.

MR. CHIU: Thank you.

QUESTIONER: How do you see this changing with advances in long read technology, and will that have an impact on some of the things you're doing?

MR. CHIU: I think long read technology

is very promising, especially for doing genomic assembly, and for identifying bacterial genomes. There is still an issue with having short reeds because, as you might expect, you might have chimeric assemblies and things that may make it difficult to interpret the data. So I think having long read technology is going to be really, really important in the future, especially for larger genomes.

For viruses, probably less useful. Because, I think, the reads of about 100 bases are probably adequate for viral identification. But certainly for genome assembly. Thank you.

MR. KRAUSE: Thanks, Charles. That was really excellent. I hope those who wanted to ask questions will catch up with you during the break. Having heard from government, and academia in a clinical setting, we're now fortunate to have two speakers from industry who will be talking about using these kinds of techniques in production flow.

The first of these speakers is Robert

Charlebois, who's a senior scientist Sanofi Pasteur. He's going to be talking about standardizing a polymorphic next generation sequencing pipeline. So, Robert.

DR. CHARLEBOIS: The right arrow to move? Okay. Thank you for the opportunity to speak to you today and also, thanks to previous speakers for providing a lot of the motivation and the background for this kind of work.

So Arifa, this morning, talked about the importance of exploring NextGen sequencing assay for example, for advantageous agent detection but also other applications in industry to get, in our case, vaccine products out the door. But it's not easy.

The main problem for industry is that it's a moving target. So the technology is evolving which is very good and it has a lot of promise which is very exciting but in terms of developing assays, we have to worry about not stag -- like not getting stuck in a particular version

of the technology when next year it's going to be so much better. And we also don't completely understand all of its strengths and weaknesses. It is still new.

It's not the same way of looking at the problems as we are used to looking. And so, we need to learn about it and as we learn and as it evolves, we need to still control and standardize this technology so that we can actually apply it in a quality environment.

Now, some of the assays that we do are research grade. You can have a lot more freedom. Some of them are, you know, characterization studies, some of them are release tests. And so, we have different levels of quality that we need to match. And we don't want to not be able to take advantage of the technology.

And so, yesterday we heard a couple of talks that talked about object-oriented programming. At heart, I'm a C++, a hardcore C++ programmer and so, I understand these concepts pretty well. And so, the way that we're

approaching the standardization of an assay is essentially to make each step of that assay a standalone part of the overall pipeline.

And so, we're essentially making an abstract functional module where you have some -- where that particular step doesn't care what came before and doesn't care about what comes after. And so, by doing that, by let's say validating the interfaces, the input and the output, then we can have the independence of the individual steps. And so, that makes it, I think a lot easier to validate the individual steps.

So eliminating dependency but we still want flexibility. We have a lot of different kinds of assays we want to run. We want to be able to reuse the steps for different assays. We want to be able to improve the steps and so, we'd have different instances of each of these particular steps. I'll provide more concrete examples in a few minutes.

And so, this is essentially what a pipeline looks like to us. The bottom part has

several streams because some of the, I mean, it isn't -- some of those steps are not in common. Like, if you're looking for genetic stability, it's quite different than looking for adventitious agent detection or sequencing your cell line for characterization.

But some of the steps at the top are in common to all of them. And so, you would have, you know, your nucleic acid extraction and your conversion to a double-stranded DNA. And I'll talk about these steps and the challenges of each of them shortly. And the sequencing, and then, you have your domain specific sequence analysis. And so, those are the big automated bioinformatics pipelines, the things that we've been talking about as big data.

And then, the next step after that, the domain specific data analysis, we're no longer in big data mode. And so, there we're in the more traditional kinds of analyses where we have, you know, lists or trees or variant profiles. We've collapsed the huge data into something that we can

use our ordinary tools for like molecular modeling or what have you.

And then, there can certainly very well be some domain specific follow-up work depending on what we find. Most of the time, of course, in industry the assays end up being pretty boring. I mean, you get what you expect. But occasionally, you have this surprise and then, we mobilize and do an investigation and try to figure out what might have gone wrong.

It's not quite the Pandora's box as people feared initially. The false-positives that we see, for example, in add agents detection are very limited in number and very manageable, at least in our experience. And so, it's not so scary. Initially, people were terrified but it's not so scary today after we've learned a bit.

And so, I'm going to talk about the interfaces too. So step one sounds sort of straightforward and for some of the applications like RNA seq or looking at bacterial DNA, it's pretty straightforward. But some of the other

applications like nucleic acid extraction for add agents detection where you have single-stranded RNA and DNA in packaged in viral capsids or in bacterial cells or fungal spores or what have you, it's not so obvious how to do this without bias.

We don't want to throw the baby out with the bath water. Through enrichment techniques, for example, and there have been some talks in earlier conferences where people have tried like nuclease and they've seen signals disappearing. And so, we have to really worry about that and so, that's one of those boxes that is evolving. We have a working version that we like that is working well for us but we know that we can improve, that we can improve the sensitivity because most of the background in our case won't be human but it'll be the cell line that we use for manufacturing like a monkey or a hamster or what have you.

But from the point of view of the pipeline, the only thing that the next step needs to worry about is that it got nucleic acid. And

so, it might need to what kind of nucleic acid it is. And so, your test requisition would say, you know, you're looking at RNA. But the point is that it's all standardized to nucleic acid.

And then step number two, there's not much to talk about. That is very straightforward. DNA, if you have double-stranded DNA to start, then that's a no-op. If you have RNA to start you'll turn that into double-stranded DNA and so on. If you have a complete unknown, then you go through all the steps of reverse transcription, second strand synthesis, maybe even some amplification.

Step three is, so it's shown here as a black box, library prep and sequencing but there's opportunities for modularization with that too. So if you have a particular instrument in-house, there might be different kits for library preparation. And so, you can sort of separate out library preparation from the sequencing.

If you have different instruments, you might have different kits and different

procedures, different SOPs and so, there's opportunity to modularize with that. And I'm not showing that here because then, the talk would be too long. But in any case, we can decide, and so, this -- we may not decide as a large community but we can decide in- house that our output of this step will always be FASTQ, for example.

And then, given that then it doesn't -- this step doesn't care about what comes after. And so, we produce FASTQ and the next step has to worry about dealing with FASTQ files.

Step four, so we've heard and we're going to hear some more about pipelines. And again, this can be modularized. So I'm not presenting our bioinformatics pipeline which is sort of the focus of this session. It's been presented at last year's PDA Conference and will be published in the PDA journal, fingers crossed. But here again, you can modularize.

So for example, our particular pipeline has two steps. We have as our first step that takes FASTQ data from whatever kind of instrument.

It will do the quality filtering. It will do the adapter trimming. It will do assembly. We do de novo assembly on the primary data. It works for us. And then, we do some annotation of the sequences.

And then, that produces a FASTA file. And then, the second part of our bioinformatics pipeline, there's also an automated self-contained module that takes a FASTA file that's properly annotated and produces lists of potential contaminants. It actually identifies everything that's in the sample. You know, whether it's hamster or bacterial or fungal or what have you.

And right now, it just produces lists but our goal for standardization will be to produce something like.xml or at least, like.xml as a technology but the goal is to have like an ontological or defined vocabulary kind of output so that we can standardize. The reason I put.xml there as a proposal is because 50 years from now, if our product's still on the market, anybody can still read an.xml file because it's human

readable. They can write a parser for it.

If we produce some kind of binary or proprietary output, good luck reading that 50 years from now. I mean, we'd have to preserve the software and the hardware and that was discussed yesterday. So that can be a challenge. We can actually, you know, print out on paper an.xml because I'm thinking, you know, 50 years ago, 1964 there were punch cards with machine language instructions for hardware that is sort of interesting but not really relevant today.

So let's try to keep this because, you know, as a manufacturer, we have to store these records for a long time because we may need to revisit them. And so, something that it is, I mean, it is a heavy format at it was discussed yesterday. But I think it's something that we could preserve for the long, long term. And then, the language ontologies tend to use, you know, well-defined vocabulary so it's easy to understand in the future.

The data analysis, so this can be a lot

of different kinds of things and you probably will need adapters or you will need adapters from our defined vocabulary to feed into these new, these various tools like molecular modeling or alignment and so on. So, sorry, in the previous step what we're specifically not doing is things like aligning to the reference genomes, looking at coverage, look at percent identity and all that. Because that can be -- we don't want to lock that into this particular module because if we want to improve that part, we don't want to have to revalidate that entire step.

So if all we do is say this sequence is from this taxon, then the next step, then we can go further and do the additional bioinformatics analysis on it. So we still do that work but we just don't do it in the same step. And so, there's lots of things we can do here. We'll need adapters coming in and adapters going out. Again, producing some kind of defined vocabulary output.

So I don't know if we can all agree on such kind of vocabulary or ontologies but within a

particular company, we certainly could. And that would, at least, make it easy for let's say a regulatory agency so they'd know, you know, information coming from Sanofi always looks like this and information coming from Merck always looks like that or what have you.

The follow-up is a lot more open-ended because it really depends on what you find. You may be able to use some assays or some techniques that you've already developed before. You may have to develop some new techniques. It's essentially problem-solving and investigation and a lot of the time, there's just nothing to do. And so, if there's nothing to do you just simply write your report, your C of A. If there's something to do, then you have to follow it up. And so, we will hope to make those kinds of reports, semi-automated and standardized for at least, you know, for our own safe if not for our customer's sake.

So now I have three simple enough examples of just showing what this would look

like. And so, suppose we want to look -- to characterize the genetic stability of a single-stranded RNA virus. So we produce those products and so, we have -- we would have our standard validated RNA extraction procedure, our TPCR. We would sequence, let's say with an Illumina or some other instrument that would give us the sufficient resolution for looking at genetic stability.

We could have our statistical variant calling pipeline that compares various samples against each other. So there could be a database in there of what we found before. Looking for, let's say, lot to lot variability or any kind of trends. Whatever we find, then, could feed into -- could be mapped onto a 3D structure so that the biologist in the risk assessment can look at it and say, there is a problem or there's no problem and go from there and write the report.

Example number two, adventitious agent detection, so let's say we're working with supernatant from a master cell bank. Our current

approach is to do total nucleic acid extraction with no enrichment. It is the needle in a haystack problem. We have a lot of host nucleic acid. And we just sequence it at very high depth so that we get the resolution that we need.

We're still working, as Arifa mentioned, there's some spiking studies ongoing to see what kind of sensitivity we get, the LOD that we can get from no enrichment and see if we have to enrich to get -- to match the existing methods. So we have our taxonomic classifier sequences. We get candidates. We then characterize them, alignments and so on and then, follow-up work.

And then finally, and just to show you some reuse, suppose we would want to look at the genetic identify of bacterial endocriduction cells. So we have our E. coli seed and we're producing our protein product and we want to make sure after the fermenters have grown up that it's all still all the same thing. And so, here it's very simple. We have a DNA extraction. This -- it's already DNA and we're reusing this myseek

module for example. We don't have a myseek but I'm just using that as an example because it would be appropriate technology here.

Again, we have all of the standard interfaces. We could use our classifier again. We don't have to reinvent something. Because we're trying to find out what this bacterium is. And so, in our database we have this bacterium and we have other strains of this bacterium. We can resolve them with our classifier. We decide that it's okay and we go on.

So this sounds very repetitious. So my talk so far has been very repetitious and that's the point is the reuse and the standardization in doing things the same way over and over again so that we don't have to reinvent the wheel every time we want to do something new. And so, just to summarize in two slides, the module is responsible for accepting the format that it gets. And it's responsible for producing the format that the next step expects.

And so, if you can prove invalidation

that it deals with input A and that it produces output B, then you don't have to so much worry about all of the combinations of all of the modules. You will, of course, in a validation have to an end-to-end pipeline run to make sure that the whole system works well. But then, that's not so onerous as mixing and matching all of the different components and all of their different parameters. So that's our proposal and the advantage is it provides a lot of flexibility in testing. And so, your test requisition may be asking for lots of different versions of things and we can provide that with this kind of an approach assuming that the different modules are validated at the right level of quality.

Then we can run the assay that we're asked to do. It allows reuse and so, that accelerates new tests. And it allows evolution so we can improve an individual test without impacting the rest.

And then, finally, I want to acknowledge people that I work with so Lucy and Lauren and

Simon and I are on the advanced virus detection technologies user interest group that was mentioned earlier. Greg and Sam have recently joined and they're working on the bioinformatics validation and sample extraction optimization respectively. Thank you.

DR. KRAUSE: Questions?

SPEAKER: (inaudible) or rather a comment. Thank you exactly this -- this is exactly the kind of input which we need to proceed with our strategization in bioinformatics, harmonization efforts. So please do join the group and we need --

DR. CHARLEBOIS: I signed up.

SPEAKER: Exactly, perfect. That's the comment. This is the exactly the reverberance we needed.

SPEAKER: So we heard yesterday that for variant detection that they often use three different methods to validate or to come up with some reasonable assessment of true variation. What do you intend to do with variant detection?

I noticed that part of your pipeline. How will you validate that? Which method will you select? Those kinds of things, how are you dealing with those issues?

DR. CHARLEBOIS: So the variant detection, so we're primarily have the problem of viral quasi-species and so, the viral seed that we start with is already quite variable. And we -- what we want to do is to ensure that manufacturing, that it's consistent and so, lot to lot, variability needs to be minimized or if there is any, we need to understand what the impact of that is.

And so, it's more of a like we do Chi-square tests, for example. That's what we're exploring to essentially look at the profiles for each nucleotide in the genome to see if there's any actual statistically significant difference among them. So it's a bit different than, I think, the problems that were described yesterday and in terms of validating that were we've traditionally been using Sanger sequencing

approaches which is -- which has been accepted in the community.

And so, we've done, you know, head-to-head comparisons of that. For other kinds of variants, I think it was mentioned on the very first day the neurovirulence, monkey neurovirulence in polio, and so, that requires a very sensitive measurement of the variability at a very particular nucleotide. And so, there's a specific test called MAPREC that does that.

And so, we're essentially doing our comparisons head-to-head against that assay which is the way that we've been doing it and submitting. To make sure that they correlate very well, that they give exactly the same result. So we're always comparing our existing methods with what we've been doing.

In some cases like with this add agents detection, we can do things that we couldn't do before. But we have -- for the things that we could do before, we have to do at least as well and it has to be consistent. So that's pretty

much all I can say to that.

DR. KRAUSE: So our final talk of the session is also an industry talk from John Thompson at Merck who will be talking about viral metagenomics analysis, BLAT-based approach and adventitious virus detection interest group activities. So he'll be providing us both an industry perspective and an idea of what the interest group has been doing and doubtless will encourage you all to join at the end.

So John? And so, depending on the timing after John's talk, we may have a short panel discussion which may give you an opportunity to revisit some of these issues and ask questions that may be interesting to hear all of the speakers in this group discuss together. But we'll see what the clock looks like.

DR. THOMPSON: Thank you. Thanks to Arifa and Philip for inviting me to this session. So I'm going to describe a pipeline that we've set up and been working with in a research mode locally. In the process, we curated the public

data and developed an automatic way to curate the public data that I'll describe.

And we've used an older algorithm, BLAT-based and I'll go into the reason we've done that. And some of the methods we use to sort of calibrate the sensitivity of that and shake it down. And then, I'll spend about a third or the last half of my talk talking about some of the activities of the adventitious virus detection group that Arifa introduced.

So like other people who have spoken today, we've centralized on the NCBI virus resources and the virus genome databases is the RefSeq collection of genomes. It's a representative sequence approach generally with one record per virus. And then, we supplement that with data from the NCBI non-redundant nucleotide, the NT collection which collects data from all the public resources.

And it includes many additional virus genomes that represent additional virus diversity that we want to pick up. And it also contains a

large number of partial viral sequences, fragments of viruses that have been deposited in the resource. And so, we've developed a system for combining record from both sources but in prior work in this area, we found that the greatest source of false positives are the partial viral sequences that are in the NT collection.

They're often deposited with vector and host sequences attached to them. And we found that many times we're getting matches to the vector and host sequences and not the viral portion of the sequence and they represent a huge false positive burden that the analysts need to wade through. So we developed this system to sort of capture the diversity represented in the partial sequences but avoid as much as possible the false positive issues associated with those.

And so, at the top right we've got the RefSeq collection that's going -- of whole genomes that are going into our virus database and then we take the -- on the left, take the GenBank NT collection and we parse it into whole viral

genomes, partial viral sequences and then, we keep the non-viral NTs separate for use in our pipeline as well.

And so, we take the whole viral genomes, remove exact duplicates and combine them with the RefSeq collection of whole genomes. Then what we do is take the partial sequences and align them to the whole genomes and we discard any partial sequences that align to the whole genomes as being already represented by a whole genome. And in that way, we're only keeping about three percent of the partial sequences. And so, we've eliminated most of the source of our false positives but we're keeping those partial sequences that add diversity.

And so, in our algorithm choice and this is now almost two years ago, which is a long time in NextGen sequencing technology, we initially rejected the short read aligners, BWA, Bowtie in this and some 70 similar algorithms are out there now that are all optimized for matching with near identical sequences to a reference sequence. And

we decided that those were too stringent to capture the viral diversity for our purposes and we fell back on a decision between BLAST and BLAT. I think almost everybody who is in the room should have known, know about and probably used BLAT -- BLAST in their past sometime.

BLAT is a BLAST-like alignment tool that if you've ever aligned a sequence to the genome using the UCSC genome browser, BLAT is the underlying algorithm. It's designed to be faster for high homology alignments. But it is tunable into the range that we needed for lower identity matching as well. And so, that's what we optimized or what we centered on for a number of technical reasons that I won't completely cover today.

But this is showing how we started to define the sensitivity and specificity parameters that we needed to use with BLAT and in this example, we're using retroviral sequences, HIV. We took the two HIV records from the reference -- from the RefSeq collection and one for HIV-1 and

one for HIV-2 as our database in this trial. And then, we picked 10,000 random HIV sequences from the NT collection to have additional diversity to see what kind of identity threshold we needed to adopt to pick up the diverse sequences as represented in NT.

And we selected that 10,000 sequences to be less than 90 percent identical to the reference sequences to make it a challenging test. And we also picked up a false positive dataset of 10,000 random non-viral sequences from the non-viral section of the NT collection. And we found that at 80 percent identity, we were picking up about 82 percent of those 10,000 reads. We picked up another hundred reads by dropping to 70 percent and another 10 reads by dropping to 60 percent. So we decided that 70 percent was sort of our threshold of diminishing returns and adopted that for further work.

Then some additional sensitivity tuning, we've done using the whole RefSeq collection of viral sequences as a database and using the entire

NT viral collection of sequences as a query test to see what proportion of the diversity represented in the NT collection would be detected by the reference database. And at 90 percent threshold that was only 61 percent of the NT collection viral sequences were detected but at 70 percent we were picking up over 90 percent of those sequences. So we think the sensitivity of BLAT at 70 percent is pretty appropriate.

And this describes the pipeline we established using BLAT. We took a two-step approach rather than compare -- because it's engineeringly (sic) difficult to compare to a 50 gigabyte or larger database. So we take the clean, raw reads and align to the smaller virus database first. And then, we take the smaller set of virus-like reads and counter screen them against the NT viral database, both using the BLAT algorithm.

And out of that we get a set of reads that match best to the viral sequences including we keep ties as well. And then, we go through a

process that some similar to what others have described of consolidating those reads and reporting on a target basis what the depth and breadth of coverage is to help interpret the data for the biologist downstream.

So as I said, even running a large Illumina dataset against just the viral database in one fell swoop is difficult. And we end up splitting both of those sequence sets into multiple smaller pieces and doing all the pairwise combinations and then, on the back end of that merging and sorting the data to get the best hits. And so, there's a great deal of splitting and merging that goes on in that process and we have a whole job control system that manages that and resubmits jobs that fail for hardware reasons and so forth.

The processing times we've been achieving are not as good as what Charles is getting with the SURPI pipeline and that's part of the -- part of why this pipeline is really showing some signs of age at this point, I think, and may

need some further development but with Roche, we're processing with Roche datasets of about 600,000 reads. We were getting things through in three hours. With Illumina reads at 26 million reads, it's now taking three to five days to process and that depends largely on the amount of host reads that need to be aligned or that might be in the database as well. The automated viral database curation process that I described runs in basically an overnight run.

And one last test I'll describe is an in-silico spike-in analysis that we did to sort of model a RNA seq experiment. We simulated 230 megabytes of human transcriptome reads and then, spiked-in a simulated set of 1X to 100X coverage of the three viruses listed in the table here. And we were primarily concerned that the counter screening against the NT collection could, in theory, remove some real virus reads if the NT collection was contaminated with some poorly annotated or mis-annotated virus sequences and so forth.

But what we found was that with the versions of the databases we're using right now, we didn't lose any of our virus reads going through this. We recovered 100 percent of our reads at every level. And so, we think the bioinformatics is actually working fairly well in this process.

The other thing that comes out of that spike-in analysis is a question of how do you interpret when you have a whole virus. And the important question is what depth of sequence do you need before you expect to see full coverage. And from the simulated data, that answer is about 10 percent. That's probably overly optimistic compared to real data that will have less even coverage.

But the point is that around -- with this dataset, around 10X coverage gets you to 100 percent coverage. So if you're looking at a fragment of a virus, alignments to a fragment of a virus that are very deep, the hundreds or thousands of reads deep and you don't see coverage

across the entire virus, you can logically conclude that that virus is not really present.

So some of the future considerations for this pipeline are we initially avoided host filtering because of the potential to remove virus-like sequences based on endogenous retroviruses in the host genomes. But I think scalability issues are forcing that -- forcing us to reconsider that option. The suggested constraints that we would apply for host filtering would be to use really high-stringency alignment methods and possibly consider masking the endogenous, the known endogenous, retroviral elements so that those are -- types of sequences are not being removed by the host filtering.

And we've also considered going to de novo assembly. Early on we avoided that because we can align individual reads but when you do a de novo assembly you need a certain depth of sequence to form a contig. So we thought there's a potential for some loss of sensitivity there. And so, we didn't go that route initially but I think

that's really attractive because the longer contig sizes greatly improve your ability to evaluate the presence of about a -- coverage of a whole virus as well as identify potentially novel viruses.

So just to summarize it that at this point then, we've presented what's a practical automated viral database curation approach that relies on the existing sequence sources and balances capturing virus diversity with the need to avoid a source of large -- large source of false positives. And we've shown that a BLAT-based viral detection pipeline can work fairly well but there are definitely scalability issues as the size of datasets continue to increase on us.

And then, I've presented an approach to sort of calibrating and testing the sensitivity of that -- the pipeline. So for the rest of the talk I'm going to summarize some of the activities, some of the bioinformatic activities of the virus detection interest group that Arifa introduced. And so, I can pretty much skip this slide. Arifa

introduced this already.

This group was formed in October 2010 and I think it's a great example of several companies, multiple companies, in the industry recognizing that this is sort of a precompetitive space. We're all going to have to go through this and come to some sort of agreement on a standardized process eventually. And we can accelerate that process by sharing information that these companies, I think, have logically considered a precompetitive space.

So the virus detection group has been broken into four subgroups and I'm going to focus on the C and the D group. But this, as Arifa mentioned, the sample A group is involved in evaluating sample preparation methodology and the B group has been setting up a spike-in, excuse me, a spike-in experiment. And those two groups together are sort of setting up the lab side of the process of how do we -- how do you optimize isolating viral nucleic acid.

And then, the C and the D groups have

been working on the bioinformatics aspect. C has been focused on standardizing viral sequence databases. And D has been focused on comparing and contrasting the bioinformatics pipelines.

So the focus on subgroup C which is chaired by Kavitha Bekkari of Merck, the group did an initial survey of data sources and used and there's a great commonality, I think. Most people are using the NCBI RefSeq collection and NTB sources. And I've described those a little bit already.

The NT collection, again, this slide is old. Charles told us this is up to 70 gigabytes now. And I described the distribution of whole genomes and partial sequences so I'm not going to dwell on that here.

I want to speak for a minute about the genomes neighbors that Rodney Brister set up because this is what we use to identify the whole, the punitive whole viral genomes from the NT collection. The problem is that the sequences are coming in so fast that the curation tends to lag

behind the sequence collection and that there are many potential, full-length viral genomes in the NT collection that do not yet have that complete genome tag on them.

So the approach that Rodney and his group took were to compare coding regions between the RefSeq entries and the NT collections. And if all the coding regions seem to be present, then they tag it as a genome neighbor or as an otherwise known or otherwise thought of as a punitive whole genome. And we capture -- we use that genome neighbor approach to capture the whole genomes from the NT collection. And it's quite a bit better in terms of the numbers of genomes we're picking up than just relying on the complete genome tag.

So Arifa's lab de -- Arifa described this so I don't have to spend a lot of time on this either but this is a slide from Arifa's group and where she and Aisha Aljanahi have been at the PDA have been working on a keyword approach with a series of positive and negative keyword selections

so that you, up until now or in our pipeline, we've been relying on the taxonomy tags of GenBank to pick out viral sequences. And so, this is an attempt to, again, avoid relying -- overly relying on the annotation and pick up viral sequences more comprehensively and that's being evaluated in the context of the interest group.

So subgroup D is chaired by Adam Palermo at Genzyme Sanofi and has been setting up and well, initially surveying the different methods that people use and working toward setting up a comparison of different methods. And so, Adam put together this slide after trying to sort of generalize the input we got from different companies and different groups that are working on this that pretty much everybody is starting with reads, doing some kind of quality filtering.

Some people filter against the host. Some people skip that step and then, some people take the high quality reads and preassemble them with de novo or other methods. Some people take the raw reads into the database alignment step.

Some people are aligning not just to a viral database but to a more comprehensive pathogen database.

And then, you have a set of punitive contaminating, potentially contaminating reads, that are indicative of some kind of contamination that are being put through some kind of a post-processing analysis to determine how reads are mapping to each target and what extent of coverage of those targets is taking place so that the biologists can begin to sort through and interpret those results.

But as you can see, there's considerable variation and choices made at each step by different companies and it's really sort of impossible to a priori define what the best pipeline strategy is. And that's sort of the basis for the interest group in the first place.

So this group has turned focus on setting up a pipeline evaluation and of between spike-in simulations and public datasets, we've actually settled on some public datasets as a

first trial and several groups are participating in this comparison. So we've picked several sub -- public datasets. One is virus particles from a human fecal sample that represents a complex metagenomic sample and should be likely to have many things that may not be in the databases yet. So it should be a very interesting dataset to analyze in these pipelines.

The second one is an HIV infection experiment that includes a human cell line, plus and minus HIV infections. So with the HIV infection, there'll be a massive predominance of HIV sequences in that dataset that hopefully everybody will pick up. And the uninfected control serves as sort of a negative control that hopefully doesn't contain much in it and will be a test of how well each pipeline deals with false positives.

The third dataset is also an uninfected human cell line but it's a particularly large dataset. This may be 5 to 10 times larger than what most people are currently using. This is

being collected on the new Illumina X Ten machine, 120 gigabytes of data from this cell line.

So there's several companies and organizations that are currently participating on that. We've come to some agreement on how to summarize the data at the hit or target level. And then, in particular cases where there are discrepancies between pipelines, we're likely to delve down into the read level on selected samples or targets.

So the different groups have downloaded these different datasets and are in the process of running those now. In the next few weeks or so, we'll start comparing results.

So the last slide I'm going to present is kind of, I think, a consensus from discussions in the interest group of some rules or some thoughts about interpreting the data and that's -- I posed it as a series of three questions. And the first question is is the complete virus present? And for that, we, I think, generally agree that evaluating the depth and breadth of

coverage is critical for determining whether a complete virus is present but it's hard to establish precise thresholds and to a large degree, that's going to depend on the different library preps that are being used and how even their coverage is.

And so, that's going to be something that evolves with the methods that are being applied. The second question is is the viral titer increasing during the culture and I think this is even more important than the depth and breadth of coverage. Because we've seen in Arifa's nice example that you can get partial coverage from a known virus from a novel virus. So even if we had partial coverage of a virus but we say that titer increasing during a culture run, during a time course from a culture run for a biologic or a vaccine preparation, that would be a great cause for concern and for -- generate follow-up activity certainly.

And the second -- the third question is really is the detected virus infectious to humans?

So if we found a full length virus or we found a partial virus that was increasing during a culture, we would be relying on the information from the public databases and collective knowledge about where the human infectivity risk lies.

So that's it. I want to thank the collaborators, my collaborators at Merck, Kavitha Bekkari did a lot of the bioinformatics for this pipeline and Paul Duncan provided a lot of biological guidance. Joe gave us engineering support for a lot of this splitting and merging we're doing. And I've also listed the interest group organizers and subgroup chairs and certainly, many other people in the interest group deserve credit as well. Thank you.

SPEAKER: So I actually have two questions from the online audience. The first is a little broad. It may be difficult to answer concisely but what is your opinion on the pros and cons of k-mer based viral detection?

DR. THOMPSON: So k-mer is something that's been -- I've been introduced to but we

haven't had time to play with yet. It's -- from what little I understand of it so far, I think it'll be a greatly different way of looking at the data but it's potentially much more scalable. So I'm anxious to see somebody, if not us, go that direction.

DR. KHAN: Okay, and the second is what is the FDA's perspective? Also not directly to you, I guess, on applying NGS first sequence variant detection in cell banks and production cells used for the production of other biologics, i.e. antibodies?

DR. THOMPSON: Well, I'm not from the FDA so I don't know that I can speak to that.

DR. KRAUSE: Neither -- do I need this microphone? I'll speak into your tie. So neither Arifa nor I are from the part of the FDA that deals with that issue as well but I think we're all faced with the same question of how it is that these methods can be standardized in a way that they provide reliable results. And I do know that there's great interest, in particular, in the

monoclonal antibody production field in favor of using these kinds of techniques to rapidly detect potential fermenter contamination events. And so, I think the discussion we have here today is directly relevant to that issue.

MS. VAN 'T VEER: I'm Laura van 't Veer. I had a que -- my background is more in cancer, molecular testing. And I was wondering, in the cancer next generation sequencing there are a number of these dream challenges organized and one is ongoing actually on pipelines for next generation sequencing REM files to FASTQ files which I think they're probably very specific elements for viral testing but some might be actually very similar to what you're doing.

And since, Merck is also working in the space of oncology and actually in next generation sequencing, as I'm aware, so do you actually exchange part of the pipeline would actually would fit nicely with the previous speaker who actually showed that there are modules you could sort of validate across maybe different organisms,

different testing capabilities?

DR. THOMPSON: Well, I think that's kind of what we're doing within the interest group. We have this trial going of different bioinformatic pipelines. We have the spike-in test going where different labs are applying different methods for isolating nucleic acid. So I think that's sort of essentially what we're doing within the interest group.

Ah, I see, look for commonality between the cancer work and the virus work. That's a good idea.

DR. KRAUSE: Well, great, I'd like to thank all the speakers but I'd also like to ask them to come up to the front. We started a little bit late and there were two hours assigned to the session and so, we have about 10 minutes for a brief panel discussion.

And so, the two questions that I'd like to put to the panel, one of them is, which the speakers can be thinking about as they get up, one of them is where do they see the most immediate

promise for these techniques in biological product evaluation? Where do they see the most immediate applications where these things are going to be used first based on everything that they've seen?

And the other question is a more technical one which is that we heard from several different speakers about different methods of analyzing next generation sequencing data. We heard about the value of casting a very wide net using programs like Tblastx and we heard from, I think, Charles was the winner in terms of his SNAP program which was able to analyze data very, very quickly. And of course, we all know that the faster you analyze your data, there may be some tradeoff in what it is you're looking at versus not.

And so, I guess the other question that is -- and then, we had the BLAT discussion which is somewhere in-between, I think. And so, the other question that I would like to sort of hear some broader discussion on is where is the right location in this intensity of analysis tradeoff

for this particular application, what are the parts of the analysis that we can perhaps do without that may be present in some of the more complicated and time intensive programs? And how does one arrive at the optimal algorithm for analyzing these kinds of data in a virus detection setting?

So maybe we can start with that second question since it's fresh in everybody's mind and then, as we close then people can give their vision for the future.

DR. THOMPSON: Okay, all right.

DR. KHAN: Your microphone is not on.

DR. THOMPSON: There we go, ah hah. So the question was?

DR. KRAUSE: The question is is that there are many different programs that do the -- and algorithms that do the sequence analysis. How do you decide among them in terms of the time versus intensity of sequence analysis tradeoff and arrive at the right degree of intensity of the algorithm for looking at the data?

And what tradeoffs, in particular, can easily be made with this kind of data that maybe you couldn't make with other kinds of sequence analysis?

DR. THOMPSON: So we felt that working at the read level was going to give -- offer us the most sensitivity and that's where we started. And then we had, you know, long- term plans to move into the de novo space. And I think, you know, some people may be beating us to that and I think we're going to be looking -- we're going to be agnostic about the algorithms and look at what other people are doing with de novo and see if the advantages really outweigh the potential detriments in that.

So, you know, I think we do have to pit these different approaches against each other and figure out what the pros and cons of each are.

DR. CHIU: Yeah, I mean, I think that what's really needed with respect to centerization is actually more precise benchmarking of algorithms. And this is a big problem because

everyone is using their own algorithm. They're even cobbling the algorithms together into pipelines and yet, there's very little published data on actually benchmarking, you know, algorithms even ones that are commonly used. And I'm actually quite surprised by how little I actually see about comparisons.

And this is important because it makes it impossible to determine which algorithm is better and which one's worse or for different characteristics. And the truth is that there's no one algorithm which is the best algorithm. It really depends on the application.

And so, the second point I want to make really quickly is that I think it's really important that this is not a yes or no answer. That there's one algorithm that's the best for every possible application and this is precisely why I think it's really important to discuss how we're going to construct pipelines and how we're actually going to analyze the data.

Because I actually think that a tiered

approach would probably be the best at least for adventitious agents where you can use a very precise and sensitive algorithm to get rid of, say, host and then use kind of a more broad encompassing algorithm maybe translate nucleotide alignment to identify novel agents.

DR. KRAUSE: So can you, Charles, just add a little bit also about -- because I was struck by how fast your SNAP algorithm is. And of course, that may have to do with your amazing computational power. But it sounds like probably you're making some tradeoffs in that algorithm. Can you say a little bit about how that works? What is it doing differently from more time-intensive algorithms?

DR. CHIU: I mean that's a good question. Basically, it's -- you would -- the best way actually would be to speak with Bill Bolosky from Microsoft who actually developed the algorithm. He actually gave a talk where he talked about the details and if anyone's interested in specific details, there are some

changes to the algorithm beyond what's typically done with other kind of seed-based algorithms.

It is a seed-based algorithm. So it is a k-mer based algorithm and I can talk about that in more detail. But I think the tradeoffs are that it is still a nucleotide aligner meaning that you still need to have very good alignment. It can tolerate some degree of mismatch but not a huge amount of mismatch.

So you can set that as the editus. And so, it's important to realize that that's not the single algorithm that I think is going to work in all cases. I mean, clearly it doesn't do translated nucleotide alignment, for instance. Which is -- or, like, it doesn't do anything, like, similar to BLASTX or Tblastx and there needs to be. And that's why we really use it to identify known sequences upfront and then, we use a different aligner such as BLASTX for novel sequences.

DR. KRAUSE: Okay, Robert?

DR. CHARLEBOIS: So we were very

concerned about using anything, any aligner that was too stringent like BLAT or anything else. And so, we wanted to be very conservative initially. We can change our minds later. And so, we're using BLAST with a very low, like, 1E minus 5 threshold.

And we compensate for that later on by doing some statistical analysis on the hits to clean it up and we get quite good results. But the thing is, like, for example, I had done an alignment of all of the members, all of the current members of a viral family. And the longest contiguous stretch that they had in common was two nucleotides.

And that is alarming if you're thinking about potentially finding a new member of that particular viral family. So I mean, individuals will have, you know, stretches that are more common. So we wanted to pick something very relaxed to start and then, clean up the signal after in a subsequent step.

So we do have very different approaches

to things and they may all end up at the same place in the end. And that's going to be very interesting to find out. So our different approaches -- so for clinical samples, time is really critical. For a manufacturing thing where you want to release your product in three months from now, it might not be so critical. Of course, we still want to do it faster and we're going to explore faster ways but we wanted to start ourselves with a base that was non-controversial using a, you know, really standard tools like BLASTN and BLASTX that nobody would argue with.

They might say, oh, it's really slow and we'd say, yes, it is. But then, and then, learning about these new techniques like I read the paper on serpa and it's quite exciting. But I'm a little scared because I want to see it head to head, you know, to see if it misses things.

DR. KRAUSE: Is there something you'd like to add, Arifa? You can say no.

DR. KHAN: I agree with everything that's been said. I think it is challenging. I'm

really excited, you know, to hear, you know, the different, you know, people sharing their information. I think this is really what's important is to openly discuss, you know, the different approaches people have.

So as Robert said, you know, so at the end of the day we can hopefully make a science-based decision whether all or any one is or which one may be applicable for certain situations. That's all.

DR. KRAUSE: Well, great, so I guess we're almost at the end of our time here.

DR. KHAN: I think there's a question from the --

DR. KRAUSE: Oh, there is a couple of questions from the audience though, so why don't we just take those. We'll start there and then we'll go there and then, we will finish up.

SPEAKER: I'll be loud, it's okay. Okay, thank you. So NT is doubling very, very quickly almost. So the BLAT and the BLAST-based approaches, I mean, yes, you know, they have been

the standard for many, many years. So do you see within the next year, for example, having huge challenges to continue? So what is the timeline by when you think things needs to be resolved that which pipeline is maybe one of the better ones and is it scalable?

DR. KRAUSE: Answers? I think obviously we're faced with a field that is moving very rapidly and what we say today almost surely will not be true even a year from now. And it's also the case that there are many sequencing projects that have important sequencing that aren't even present in NT.

And so, if you cast a wide enough net even now things get very complicated. And I think all we can do is as people in the field is work together to figure out how this can be managed. And that's the importance of meetings like this. To the degree that viral databases can be culled from the genomic databases upfront and so, work can be done that doesn't need to be repeated each time, that obviously can be helpful also.

But I think we all recognize that and as with all of big data, as the data continues to increase potentially exponentially, the complexity of these issues keeps increasing.

DR. KHAN: I also wanted to mention, I mean, that's a very interesting -- it's a very important point but I think it does indicate the need for others to get involved. And so, these challenges are not only for us but I think perhaps we'll get others to develop further tools that would help us facilitate the analysis and meet the challenges. So I'm hopeful that the discussions here will maybe go out broader to others who are non-biologists maybe more, you know, computer experts or bioinformaticians or who can help develop tools that might address these current and future challenges.

DR. KRAUSE: So Robert had a comment also but we need to pass this down to him.

DR. CHARLEBOIS: So we can never play catchup with the sequence data and we may need, in the future, to adopt a heuristic approach. And

so, Rodney Brister yesterday talked about, you know, the -- and also, others have talked and even yourself have talked about having a representative genome. And so, if we, I mean you could potentially miss something but if you have a representative genome based on sequence space definitions, you could first see, you know, can it potentially match something in that area. And then, you focus in on the next level to get more detail.

And so, heuristic approaches always have costs and tradeoffs but it might be a way to just organize the biological sequence space so that we can deal with this explosion of data in the future. I think there's -- we're going to have to pay that price, I think.

DR. KRAUSE: So that's a very thoughtful answer. And I see there's one more question. I also see somebody standing by the stage with a hook. So we'll go with this question and hopefully just one or two very quick answers to it.

SPEAKER: Do you see any platform-based biases that might be introduced particularly in relation to the choices of parameters that you use for each of those steps in your processing?

DR. KRAUSE: Short answer? Yes. Anybody else?

DR. THOMPSON: Well, I think that's exactly what we're trying to sort out by beginning to compare different pipelines and different choices that people are making.

DR. KRAUSE: Okay, well, thank you very much to the audience. Thank you very much to all of the speakers. This has been a great session and it's part of a great meeting. So thank you.

DR. KHAN: So our next group is the clinical biomarkers and personalized medicine that is charged by Dr. Eric Donaldson and co-chaired by Dr. Raja Mazumder. Our first speaker will be Eric Donaldson, so the stage is yours.

DR. DONALDSON: All right. Good morning, and welcome to the Clinical Biomarkers and Personalized Medicine session.

I'm going to give a brief introduction to the session and give a little bit of background on two new drug applications that were submitted last year, reviewed, and approved that use next-generation sequencing for their resistance analysis. These came through the Division of Antiviral Products and it's sort of a launching point for our discussion.

First, a disclaimer that myself and the members of the panel do not represent official FDA policy. This is an ongoing discussion and we're here to learn. The outline for the talk this morning is I'll give about a 10-minute introduction. I've asked each of our distinguished speakers to give about 15 minutes, and then the idea being that if there's time at the end we'll have 20 minutes for discussion. I know that we're already running late in time so if it turns out that we don't have time, then I have a question that I'd like to pose to all speakers and everybody in the audience at the end.

But I wanted to start this session with

just a few definitions so that we're all on the same page. Now, the official FDA definitions are listed in the guidance, which is bullet point number four on this slide. But just for the sake of conversation this morning, personalized medicine is customization of health care based upon analyses of individual biomarkers to identify the most relevant therapies for an individual. We heard this described yesterday as precision medicine, which is kind of along the same lines. The biomarker is a measurable indicator of a biological state or conditioned response, and a companion diagnostic is a medical device or assay that identifies patients for a specific treatment and/or dose based on the discovery of those biomarkers that are predicted for the drug response, and companion diagnostics are paired with a specific drug and often approved together in collaboration with CDRH and CDER.

In clinical practice, personalized medicine is generally used to select subgroups of patients that are most likely to respond to a

specific treatment, and there are some examples of these, particularly in the field of oncology, where a number of important biomarkers determine which drug is used for a specific type of cancer. For example, EGFR mutations to treat lung cancer. Some of our speakers will go into details about these particular ones in more detail in their talks so I won't spend a lot of time there.

Another example is for cystic fibrosis with (inaudible), which is indicated for gating alleles -- one or more of the following gating alleles that are listed there.

Now, I've included two antivirals at the bottom. These are not specifically examples of personalized medicine in a sense that they don't have a companion diagnostic to go along with them. However, both of these have viral determinants that determine which drug or which patients or which patients that are infected with which viruses should receive those drugs. For example, Maraviroc is for treating HIV-1 infected patients with the CCR5 tropic virus only, and Simeprevir,

which was approved last year, is for treating HCV patients, Hepatitis C virus patients with genotype 1, but with the genotype 1A, subjects with Q80K present at baseline in the protease do not respond to Simeprevir, so it's actually recommended in the label that if you're infected with genotype 1A, you be screened prior to receiving Simeprevir to see if that polymorphism is present.

These are examples of viral determinants that are biomarkers, but viral resistance is also a biomarker that we're very much interested in in the Division of Antiviral Products. And there are several viruses -- I've just listed a few here -- that are important. I mentioned the Q80K polymorphism, but in addition, R155K is a common resistance- associated substitution to the Hepatitis C virus protease, and if a subject has that or a patient has that at baseline, they've actually lost the entire class of HCV proteases that are approved. So that's an important biomarker for understanding whether or not a patient will respond to a particular therapy.

With Sofosbuvir, which is a polymerase inhibitor that was recently approved last year, L159F and C316N may be important markers present at baseline that predict failure with Sofosbuvir. That's still yet to be determined.

Human immunodeficiency virus, baseline resistance profiles are often taken for treatment, experienced patients who have failed heart, to establish which regimens they will respond to currently. And then of course, with influenza A virus, the H275Y substitution and the H1N1 neuraminidase is predictive of resistance to also Tamivir.

So identifying resistance-associated substitutions and baseline resistance polymorphisms is essential for determining the safety and efficacy of a drug and for identifying subgroups of patients that will benefit the most from using that drug.

So in our division, we take that very seriously, and in general, the sponsor will provide all of the information on resistance

analysis from the cross clinical trials in the format of the summary. Our division also requests that they send the raw sequence data, which to this point has been Sanger data, Sanger population sequencing, and generally what we ask for is a baseline sample and a time of failure or multiple times of failure after failing a treatment or a regimen in clinical trial so that we can compare. And we're asking for sequences of the target. So for a polymerase inhibitor, we're asking for the polymerase gene.

And then our viewers, the virology viewers, actually do an independent analysis of the resistance data because we feel it's extremely important to understand exactly what's going on and we want an independent assessment of resistance so that we can describe that accurately in the label.

So as you can imagine, next-generation sequencing adds a whole level of complexity to this resistance analysis. And we're seeing within our division, it started with these two

submissions. We now know of at least 25 submissions that are coming in the next year or so that will include next-generation sequencing, driven much by cost as Carolyn Wilson pointed out yesterday.

But what we love about next-generation sequencing is it allows you to take a much deeper look into the data. So with population sequencing, you get sort of an average consensus of what's in the sample. With next-generation sequencing, you get an opportunity to look at the minor viral variance in the population, and so you might be able to identify baseline factors that are present before treatment that will emerge and later become associated with resistance.

But, of course, with next-generation sequencing comes the technology challenge, which really a lot of times comes down to interpretation. And so with Sanger you have a nice trace that is arguably one interpretation for one sample. With next-generation sequencing, you have millions of viral sequences in the tube,

hundreds of alignment algorithms, variable filtering criteria and assemblies, no standard analysis pipeline, which means many different possible interpretations.

What the crux is for us is with variant detection, and we heard yesterday that somebody is using three different variant detection methods. And the idea is that we want to try to identify real variants versus those that may be sequence error or introduced by PCR or any number of different ways. The problem is that without a standardized analysis pipeline, different sponsors are using different variant calling techniques, and these all were designed for different purposes in some cases. There's very little standardization of cross methods for calling these variants with little overlap among the different callers. In fact, Ara Attal did a study recently comparing five different variant detection systems and they were sequenced by alumina 15XMs from four different families and then compared the five different methods and found that there was 57.4

percent congruence across those concordance. It was even worse when it came to insertion solutions; it dropped down into the mid 20 percent concordance. And each one of these methods actually identified 0.5 to 5.1 percent of unique variants of their own.

And so, as you can imagine, if you're comparing the results from one detection system to the other, there may be major discrepancies. And when it comes to identifying those key variants that are important for resistance, we want to make sure we know what we have and what is the truth. And Oscar Wilde puts it best, "The truth is rarely pure and never simple," particularly when it comes to next-generation sequencing.

So when these submissions were coming, we had several questions that we asked, and we have attempted to find ways to work through these solutions. And for experimental design, we wanted to know what exactly is being sequenced. Is the target amplified? Is it multiplexed? Are there minor variants preserved? Are there experimental

artifacts? Is there something in the way that they're preparing the samples that might leave behind important information? Or is there something in the sample that shouldn't be there? And in fact, with those two submissions, we saw examples of PCR contamination across contamination of multiplex samples, both of which could have an impact on whether or not you call a minor variant real or not.

In addition, for the data analysis, how is the data being analyzed? Is it reproducible? What programs or parameters are being used? What we learned from this first early experience is that both sponsors used entirely different pipelines and many of the components of the pipeline were proprietary so that we did not have access to those.

And then when it comes down to regulatory review, how do we make a decision based on the data? What data are needed? Are the summary data from one analysis pipeline sufficient to do this? And we felt early on that this was

not. Knowing of the issues that I've described earlier, we felt that it would be important to have something to compare it to. And so our strategy was to come up with an internal analysis pipeline that we could then ask the sponsor for the fast queue sequences, do a quick analysis of the data to see if our results match those of the sponsors, and to identify any discrepancies in that way.

That then introduced another problem -- data transfer and storage. These are large files, a terabyte or more. So how are we going to store these as an agency? How are these data stored long term? Is encryption required? How is the data integrity managed? And we have to think about this in the context of regulatory data which is stored for 30 years in some cases, so how do you maintain data integrity over a 30-year period? And then, of course, data confidentiality. How are we protecting the proprietary and patient information associated with this? And so this was really a major collaborative work to put together

this analysis pipeline. And I'm just going to quickly show that it was a collaboration between CDER, CBER, CDRH, with support from the Office of the Chief Scientist.

So now that was our experience as a division with two applications that came in last year. We know that this field is exploding and expanding, and so we've gathered a distinguished panel of speakers who we've asked to discuss from their perspectives what the different issues are with next-generation sequencing and clinical data or drugs.

And so I want to just introduce the panel now. Dr. Andrea Ferreira-Gonzalez is Professor of Pathology and Director of the Molecular Diagnostic Laboratory at the Virginia Commonwealth University Medical Center. We have Dr. Andrew Grupe, Senior Scientific Director from Quest Diagnostics. Dr. Charles Sawyers will be joining us, hopefully, remotely. His airplane was canceled. His flights out of New York were canceled, and so his slides are here. He's going

to try to call in and present the presentation. But he's the Chair of Human Oncology and Pathogenesis Program, Marie-Josee and Henry R. Kravis chair of the Memorial Sloan-Kettering Cancer Center. And then Dr. Laura van't Veer, leader of the Breast Oncology Program and Associate Director of Applied Genomics from UCSF Helen Diller Family Comprehensive Cancer Center.

So we're very happy to have this panel of distinguished guests. Our first topic will be presented by Andrea Ferreira-Gonzalez, and her topic is going to be "Next-Generation Issues Related to the Clinical Lab."

MS. VOSKANIAN-KORDI: Just a small note, due to logistics issues since Dr. Sawyers is doing it remotely, we're going to do him last.

DR. FERREIRA-GONZALEZ: Thank you very much. I would like to thank the organizing committee, but also the Association of Molecular Pathology for inviting me here today to share some of the issues that we currently have in the clinical laboratory as we apply next-generation

sequencing technology.

My laboratory is currently using this technology for the identification of mutations or genetic changes, specifically for cancer-actionable jeans. So that's how I'm going to do the context of my presentation.

I can pass the slides. So let me introduce first why we have started incorporating next-generation sequencing technology in the clinical laboratory. And again, specific for the cancer application. Mutation screening has been an integral part of the analysis of oncology patients for many, many years, and in the past, the molecular diagnostics laboratory has had some interventions in helping the diagnosis, identifying the different markers or prognostics, or even detections or determining what is the best therapy for patients, but it was very limited. And the reason it was very limited is because cancer is a very complex disorder with many, many different genetic alterations and many different genetic genes. And it was very difficult for the

laboratory to have a comprehensive way to look at all these different changes at one time.

With the advent of next-generation sequencing, now we have a tool that we can apply to really look comprehensively at what is the genetic status of the particular cancer. And what we're seeing is that more and more every day we have at least a new gene that has been associated with different cancers. And in the past, we have actually associated specific gene mutations to specific cancers, and the cancer is in a specific location. And as we do more and more next-generation sequencing, what we're starting to identify is that actually there is a huge amount of commonality between some of the genetic changes that occur in a different multiplicity of cancers and we are starting to see a movement towards the new description of cancer in a more molecular way. Very similar to what is happening in the hematologic oncology processing where the WHO is starting to classify them using these molecular targets.

So I put here a list of the different mutated genes that have clinical utility, and I'm not just talking about clinical validity, but actually clinical utility. So finding the status of that particular gene in a particular cancer or condition is very important to not only helping the diagnosis, prognosis, but also in the therapeutic determination of where these patients are. And many of these markers of each tumor has been identified, that play a very specific role in very specific cancers, but now we're starting to see that there actually might be a role for some of these other different cancers.

As we continue to the application of these technologies, we are also faced with the decrease in the invasiveness of the procedures that we are putting our patients through. Fifty percent of the specimens can detect cancer or laboratory today for doing cancer genotyping analyses or fine needle aspirations, and due to the fact that we use fine needle aspirations, the tissue that we're getting to the laboratory are

extremely small, are extremely limited, that then we have much DNA that we can extract, allowing us to do a large number of different tests with different platforms. We have to start having to consolidate and get the most information that we got.

But with that, it's also very important that when we have single gene testing today, we have a large number of different platforms that can be used that require a large amount of DNA, in some instances we are not available in the laboratory. And we are seeing an increase in the test volume for these different clinical applications due to the fact that we have this amount of information now with clinical utility.

So let me show you what has happened in my laboratory in the last year. In the past, we used to have three different tests for the detection of KRAS, EGFR, or BRAF, and these are laboratory-developed procedures, the first one and the second one, and this is a cryogen procedure for the detection of EGFR mutations that require

if you needed to ascertain these three particular genes in one patient sample, a large amount of DNA. And in some instances, we didn't have enough to be able to do the three tests, and we had to choose which one we would do first, and that was the only information we can get from patients.

Earlier this year we developed, implemented, validated a test that is commercially available using the Ion PGM Cancer Panel v2 that allows us to ascertain the mutational status of 50 different genes that could be applied for different types of cancers. As we implemented and validated these types of assets in the laboratory, we went through a very rigorous procedure. To date, even though we ascertained 50 different genes, we only have reported seven for specific cancers being colon, lung, or melanoma due to the clinical utility of these genes has been demonstrated for a limited number of these genes.

As we brought the clinical testing into the laboratory, we followed their rigorous procedure for not only test development but also

test validation. And as we went through the test validation, we were able to obtain a significant amount of information that was necessary to develop the quality management and that was then implemented to monitor the quality of the testing as we move forward, but also maintaining the accuracy of the testing that we do. During the test development -- we were able to do the test development, and where you do test validation, sometimes simultaneous or not, but we had to validate not only the wet part of the laboratory testing but also the dry part of the laboratory testing which I consider the pipeline.

So we did a validation of the pipeline by itself and then we did validation of the wet part of the laboratory and incorporated the validation of the pipeline and then validated the entire process together from the wet part and the pipeline together, and then we had to do modifications as we went along until we got to the right type of conditions necessary to meet the specifications that we require from the design of

the testing that we're bringing to the clinical laboratory.

We were lucky to start to see a merge in number of different guidelines for different areas of testing, starting with inherited disorders, but we're starting to see some evolving for cancer applications. But these are very important. The CDC started working early on, and we're going to hear more from Dr. Ira Lubin, who is present here today. The Division of Laboratory Sciences and Standards got involved very early in developing an extraordinary successful program called the Gate RM Program, but also convene a group of experts to start looking at some of the issues necessary for the validation of these next-generation sequencing applications, and what are some of the recommendations or some of the practices that are starting to emerge?

MM9 is a great document from the Clinical Laboratory Standards Institute that allows to provide some information; also, guidance on how to bring next-generation sequencing. And

there are other ones, too. The Association for Molecular Pathology, which I'm a member of the Whole Genome Analysis Workgroup, has been working very diligently in trying to deal with these issues, and we have produced several white papers and are working right now on another white paper on how to validate or best move forward with this type of testing in the clinical laboratory.

But as we go through the testing, I'm going to show one of the cases that we did in the laboratory just to show you why it is so needed this type of testing. And as you can see here, we got a final aspiration specimen from a 79-year-old Caucasian female, who had a history of right lower lobe nodule, and she was actually followed for a year and a half. She had a negative PET scan that then continued to evolve and change in size, and after the final aspiration procedure, there was a diagnosis of invasive lung carcinoma which was modularly differentiated, but the point to put here is the size of the actual cells that are actually needed to be analyzed within the final

aspiration specimen. Just to give you an idea, this is a glass line and this is the marked area by an anatomic pathologist letting us know that this is the area that we need to analyze within the tissue, that then we need to extract the DNA and apply it to the next-generation sequencing.

And this is important because with that amount of tissue, we could do EGFR testing, determine if these individuals could be eligible for some of the molecular type drugs that have occurred in living proof. But there is a large number of other genes that have the mutation, and some of them are exclusively -- one you have a mutation and one you don't it in the other one. But if we have to apply this test for the EGFR that we had last year, from this specimen we couldn't get enough DNA to actually be able to analyze it. Some people recommend using KRAS first as a triage and then go to the EGFR, but I think it's better just to see if you actually have EGFR.

So we did apply the next-generation

sequencing assay to these particular patients, and we actually found a number of different mutations. And as you can see here, even though we do test for 50 different genes, we only report seven of those. And you can see that the patient had an AKT1 KRAS mutation. There were pathogenic, and it had a BRAF mutation that was likely pathogenic.

So this individual still could receive some of the molecular-targeted drugs for EGFR, but at the same time, if this didn't work, we already know that there are other drugs that are going through clinical trials that might have relevance for the AKT1 mutations and so forth. And it gives you a complete picture of the status of that particular cancer sample.

As we go through the data analysis, and this was kindly provided by Dr. Lubin, again, he presented these at the annual meeting of the Association of Molecular Pathology last year, we go through a different number of steps, as we do the data analyses, and through the validation of these particular tests, we have to validate each

of the different steps of this pipeline alone, and then we have it with the wet all together. We actually multiplex or barcode our specimens. We can run up to seven specimens per sequencing reaction which allows a higher throughput, but it gives us enough depth in the reading of the electronic file that gives us a sensitivity of 4 percent. We have to have 4 percent of the DNA that we analyze mutated to be able to be detected.

So you have to really determine or validate that your algorithms to be able to demultiplex the data -- see, all the data coming out, you have to demultiplex what reads correspond to each of the different patients and make sure there is no crosstalk. That needs to be validated. And then the secondary alignment, variant calling, and then from there, the tertiary analysis and where you have the variant alleles and you have to annotate these and prioritize them. And then the final clinical interpretation which is the most labor-intensive, to be honest with you.

And then again, I'm not going to go into detail on this, but you can see here the different steps that are needed to go through and all the information of the genes that have to be annotated to remove those genes not associated with the presentation at that time. You have to have patient information, not only the phenotype but also some of the variants, and you start having to annotate. That allows you to remove variants that are unlikely to alter in function, and then you have to combine all this information to develop at least the prioritized list, and then compare it to the clinical grade assessments and clinical result reporting. It takes a longer time to do the data analysis sometimes actually than generating the wet laboratory data.

There are annotation databases that we use continuously. And as you can see here, I put a very limited number of different databases that we can use to start trying to get information. So, the ultimate information, is this change clinically relevant for the current presentation

of my patient?

We use a large number of different databases -- dbSNP, Clean Bar, HGMD, Cause Big, My Cancer Genome, but we have to use all of them. We cannot use a single one because all of them have caveats. Some of them are better to ascertain different changes and the information might not be updated and so forth. But it takes a large number of these annotation databases to be able to build confidence in some of the data that we're trying to produce.

So we do have some interpretation issues when we get to that point. In some instances, there is very limited evidence on the clinical utility of specific mutation for that particular cancer that we're seeing it. There are few mutations that are listed in consensus management guidelines. Institutional and national efforts to gather and cure it are just starting. For example, some of the efforts that were mentioned earlier today at the NIH label, My Cancer Genome. Sometimes we don't know what the clinical

significance or well study mutations in different tumor types. And as we start doing genotyping with a certain number of genes, we're starting to see these mutations coming up new, and we don't know what the significance is in that particular organ. Actually, the clinical significance in other mutations in targetable genes is very important. For example, unreported mutations in EGFR, they want to confer susceptibility to the different molecular targeted drugs that we see. So this is the interpretation of the information that we need to start trying to get a handle on.

As we move clinically, there are a number of different missing pieces in the different pipelines that we use and even the laboratory information systems. We don't have good tools to actually very rapidly and easily determine the adequacy of the amplicons. How the amplicons are working and actually how your PCR reactions, for example, in our case, are working without having to really start digging into the info informatically. There's not an easy way to

ascertain these. There's no easy way to visualize the results as we go through the data analysis.

Conversion of variants into standardized nomenclature is highly -- I think is highly desirable. We are supposed to be reporting all the gene changes used in standardized nomenclature, so I think it's very important that as we look at these pipelines, that we start using the common language that we have to use clinically, which is standardized nomenclature. We have to be able to monitor frequency of specific mutations in our own samples. We have to develop our own databases. There is not a commonplace where we can go start depositing some of our information to monitor over time what happens with these changes and then how they compare with other sites. Their link into genomic databases is also very important.

Inclusions or exclusions of variants for the reporting interface with the laboratory information system. Today, we have no way to represent this information within the laboratory

information system or the electronic medical record, so we are actually putting PDF or paper as an image into the electronic medical record that cannot be leveraged. We cannot develop clinical decision support tools and so forth from this information.

And then generation of clinical reports replaces in different forms of these reports and there is not clear guidance or good guidance to be able to determine what should we put into the different clinical reports.

Additional problems, as we see that as we come up with the current technologies, it's very difficult to start aligning with confidence some of these reads. So the long range actually is going to help as we move forward. But it's very difficult to identify variants, so we don't have very good tools to be able to identify variants.

The different data formats is a limitation today. How do you describe the quality of data generated, changes from the different

platforms from the different vendors, so sometimes it's hard to know when you're talking to individuals using different platforms, when we talk about the quality, just something basic as the quality of your data, but how do you actually talk apples to apples?

I think some of the opportunities that we have is actually looking at interoperability of prodigals conference, where we can call get together and look at the different prodigals that we're using for the different steps and then come up with maybe common prodigals or recommendations to which we can actually use and then start comparing.

I think also the development of benchmark tools for variant call might be really relevant or important to move forward, where we can actually all work together to develop this tool. I know the NIST is having a significant effort in trying to work towards that goal in developing this type of benchmark tool that we can use.

As you can see here in the little graph on the left, you can see the different software that you can actually use to create or generate a bond file. And as you can see from previous speakers, there are even more than this. And they produce different formats, different information that it's very difficult to actually then start comparing between the laboratories.

So one of the challenges I think will be standardization of all these different topics that I just mentioned. But I think also it will be important to look at difference reference materials, because as we develop the tools and we try to come up with a common language, we have to be able to prove that these common languages or tools can talk to each other or actually produce accuracy. So developing of reference materials, and not just only materials as a form of DNA or form of a specimen, but also electronic files that can be used as reference material to calibrate all our different aspects. So the publication of the Genome in a Bottle that NIST is working very

diligently that I think now you're going to have a speaker that is going to talk specifically to this will be very well received. These would allow us to calibrate the pipelines in the electronic file, but also calibrate the test because we can use materials with the electronic file that allow us to (inaudible) through the development, validation, quality control, and so forth.

The reference materials for it I think will be important that we can all get together and talk about what are our needs and what materials can be developed that can be used for validation and compliance for quality control, and I think maybe doing something similar to what I heard earlier in the other group, having different laboratories working together with exchange specimens, exchange electronic files at different levels and see what you obtain at the end, is starting to come up with common censuses or some commonalities that we can all come up with a tool that can be used for the different areas of the testing.

And with that I would like to end my presentation. Thank you.

(Applause)

SPEAKER: I have a quick question in terms of the bioinformatics. How do you version control the annotation databases, like CFNC and dbSNP that change every few months?

DR. FERREIRA-GONZALEZ: Yes, it's very difficult to keep that. Not only the version of the software that the manufacture is giving us for the instrumentation, it's a huge issue. The version of the human genome reference. You know, we were 19 to 20 and we had to remap all our algorithms to that. So it's very hard to maintain that and it's very challenging. So we are trying to keep some of the reference on our own local server, but it, again, becomes very challenging. So having to remap to the current version all the algorithms that need to be revalidated, so there is a huge issue there on that.

SPEAKER: Do you have a criterion for when you upgrade to a new version of dbSNP, for

instance, or do you just automatically take the latest one?

DR. FERREIRA-GONZALEZ: We automatically take -- that's what we are doing all the time.

MS. VOSKANIAN-KORDI: I have a few questions from the online audience. The first is, how are you evaluating the abundance of false-positive sematic variant calls during test validation?

DR. FERREIRA-GONZALEZ: So what we do is we have specimens that we've already characterized. We did for colon, for lung, and for melanoma, so specimens that had already been characterized for very specific hotspots. We knew the sequences around those areas. So as we went through the validation, we knew what we needed to obtain, and then as we started getting information or data out of the instrumentation we started to -- that they were called by the software, we started identifying artifacts. But then we started veiling up masks to block or just ignore to be honest with you as we go through that

analysis.

MS. VOSKANIAN-KORDI: And, sorry, let me find the second question.

What is the decision-making process on a negative finding from a clinical sample? Are validations done or required for a negative sample?

DR. FERREIRA-GONZALEZ: So we have samples that are already being characterized for specific genes that were negative. So as we run them through the instrumentation and they come in negative, they're fine. So those samples that we knew were positive that might come up negative, they will have to be confirmed with a third method or another method to see the discrepancy between the two different values.

SPEAKER: One of the IT challenges or bits that need to be fixed or filled out that you mentioned was converting a variant into standard nomenclature. It seems a lot of people are using standard nomenclatures, like HgbS or even the variant call format, but the problem isn't that

standards don't exist but when you create such a thought, you can create a different nomenclature from the same variant so that two labs producing the same variant can produce different HgbS or different variant call format, and those things are difficult to compare. So the challenge to me isn't really a standard nomenclature but a directly comparable nomenclature that I can say that these two things are identical or not.

Can you talk about that a little bit?

DR. FERREIRA-GONZALEZ: I think you're bringing up an important issue. One of the things that we see that sometimes will have discrepancy (inaudible) because sometimes you can go by the difference of a base where it has to do with which reference you're using of the sequence. So maybe the nomenclature needs to go back to the reference of the sequence and from the reference version that you have, then how do you compare to one versus the other one. But you have to provide the reference, the specific reference version that you're using.

That brings also the point from the previous individual that asked, version control.

SPEAKER: Thank you for the talk. In terms of reference materials, what would you consider the optimal reference material? Is it a plasma-based cellular disease mimic or is it something else that hasn't emerged yet?

DR. FERREIRA-GONZALEZ: Well, you know, ideally with a patient specimen. Plasma has a certain advantage. You can manage what you actually have specifically in different locations, but it doesn't have all the structure, secondary tertiary structure that you might have from the whole human genome. I think the closest you can actually get to a real patient specimen is the best way to come out to some of this reference material. Sometimes it's not going to be possible, but the closest that we can get to that will be very useful.

Using cell lines, again, it's not the best because isolated cell lines are not in the context of the patient specimen, but at least you

have the whole genome structure there that you have to deal with on the wet part and even the electronic part of the sequence that you're obtaining.

DR. DONALDSON: Due to time, we're going to move on. So thank you.

(Applause)

DR. DONALDSON: Our next speaker for this panel is Andrew Grupe. He'll be talking about "Sequencing and Truth: Utility for Common Reference Standards."

DR. GRUPE: Thanks, Eric, for the introduction. And thanks to the organizers for the invitation.

I wanted to take the opportunity today and combine actually two different things. One, talk about our experience at Quest Diagnostics, but also I've been chairing the Reference Material Selection and Design Working Group of the Genome in a Bottle Consortium, so I wanted to combine those two aspects and really talk about reference materials and the utility of reference materials

in this section.

And maybe a show of hands here. Before Andrea's presentation or before you came to the meeting, who has heard about Genome in a Bottle?

All right. So that's maybe 20 percent. And who of you have attended a work session of Genome in a Bottle in the past?

So that's maybe five or six of you. All right. Very good.

So we all know that sequence information is becoming more and more important in order to make clinical diagnoses or also to make treatment decisions. And in the past, just like Andrea already iterated, that was kind of a linear hierarchical process that was being used where initially, sanger sequencing was used or genotyping, and you had your sample and you had your most likely gene that you wanted to analyze, so you looked at that gene and you worked your way down, looking at gene number two, number 3, number 4. And that could entail a long time until you found a result, so the patient had to wait a long

time before they were informed.

Do I need to do something different?

MS. VOSKANIAN-KORDI: It's this one here. We're having an issue here.

Is this better? I'm also closer.

DR. GRUPE: I can try to get closer to the mic. That's a situation where I'd like to be shorter.

So next-generation sequencing technology avoids some of that conflict so you don't deplete the sample, you have a much faster turnaround time to get information back to the clinician.

We've seen that the times to find relevant clinical markers and translate them into practice has become shorter and shorter over time. In the beginning, we were talking about decades. Now we're talking about years in order to translate that information. So therefore, I think it behooves us to think about the whole process in terms of analytical validation, clinical validation, and also the regulatory review process, how we can address that timeframe and

provide still optimal and timely care for patients.

So next-generation sequencing, as you all know, we look at a really large universe. We can generate so much more sequence compared to Sanger sequencing. So validation of that technology becomes much more complex than for a sanger-type of assay or a simple genotyping assay. And so therefore, I think it is really critical to have well characterized reference materials available for this process, and those reference materials, they probably will differ depending on the application that you would like to use. So there will be different reference materials for inherited diseases or oncology if you think about noninvasive prenatal testing or infectious diseases.

So I want to spend some time on the Genome in a Bottle Consortium and the reference materials that are being discussed there and being developed. I want to talk about synthetic constructs and some ideas about their utility.

And then also simulated data.

When you look at the National Comprehensive Cancer Network Guidelines for recommendations on how to treat cancer patients, you see an evolution there in terms of the types of mutations that are listed. When you go back and look into the recent past, you might see a single gene listed for certain cancer types, or you might actually see specific code-ons or specific mutations listed that are relevant. If you look today, that information is really increasing in terms of the number of genes, but now we're not only looking at specific mutations that are in the guidelines but we're looking at whole axons or whole genes, so that adds another complexity level also in terms of the analysis and finding those mutations because you're on longer just limited to having a specific sample with a specific mutation, but you need to analyze the sequence as a whole.

So the increase in terms of clinically-relevant genes and mutations, that is

also expressed in the number of tests that are being offered. Andrea gave an example. Here is an example of what we do at Quest Diagnostics. Traditionally, and still we're offering specific panels that look at specific cancer types and have specific genes for when we do Sanger sequencing. And then recently, we've also released a panel that looks at 34 genes and analyzes those genes using Next Gen sequencing technology. We all know that Next Gen sequencing technology comes along also with potential errors. We've seen one slide from Eric. Here's another example where groups have looked at the error rate using different platforms and actually also using different genomes to analyze that. So again, stressing the point that if you do sequencing on such a large scale, you really need well very well characterized, high confidence reference samples that are available.

Now, when you think about errors within the process, we're always talking about the wet lab process and we're talking about the dry lab or

bioinformatics pipeline process. Where do these errors really occur, or what processes can contribute to these errors? In the case of oncology, the tissue is usually formalin fixed and embedded in paraffin, so that can lead to deamination of the sequence, so there is a potential error that gets introduced or changed in the sequence. When you share your DNA, there have been reports that that can introduce oxidative damage.

We all know about PCR-based errors. If your sequence, you're variant that you're trying to find is close to a homopolymer or in a high GC-rich region, that's going to cause problems. The error rate will be associated with the REDPs that you have, with the quality of the alignment, the trimming that you do, and then actually, pseudo genes, they also will mess up your data and make your interpretation more difficult. So when you have reference samples, you can strategically place those reference samples and analyze your process by looking at the reference samples and

see how they perform. And also, looking downstream in the process when you actually don't have samples for specific mutations, you can think about using a synthetic dataset for your analysis to test your bioinformatics pipeline.

So we've been talking about reference samples, but what might be the different requirements for reference samples that you might want to think about? So when I think about DNA sequencing, I want to look at single nucleotide variants. I want to look at small INDELs, probably copy numbers and rearrangements.

So when I look at inherited disorders, there is a specific -- well, there are fixed variables that you identify. So if you look at SNPs or single nucleotide variants, they are most likely to be homozygous or heterozygous only. If you look at copy numbers, it's going to be a homozygous deletion or heterozygous deletion, or you're going to have two, three, four copies of that chromosome or chromosomal region that you analyze.

And so in the Genome in a Bottle Consortium, which was initiated by NIST and stakeholders from academia, from the government and industry, we're discussing, and we are developing reference materials, reference methods, mostly reference data to really use that information and asses the confidence in human genome variant calls and to enable the performance assessment of clinical genome sequencing.

I'm sure Justin, who is here, who will talk in the afternoon, will go into more detail, but the characterization of those reference materials is based on multiple platforms, on Illumina, on Life Technology Platforms, PacBio, Complete Genomics, and others. So there is really a very comprehensive dataset that is being assembled and analyzed to come up with those very high confidence genotype calls.

Here is a list of the reference materials that we have selected so far. NA12878, that's probably something that is familiar to at least some of the old-timers who were involved in

the Human Genome Sequencing project. There has been a lot of data been generated using Sanger sequencing, but there has been new data been generated on the next- generation sequencing platforms, and there are high- confidence genotype calls available that Justin actually has put up on the server and that are available to be used right now. The sample itself is available from Coreal right now, but it will also be released by NIST shortly, probably towards the end of the year.

And then there are additional samples that are in different stages of the process. For the European Ashkenazi trio samples, there will be initial sequences available later this year, and currently, these samples are also available through Coreal, but in the future they will also be available as reference materials through NIST.

The reason why we've selected a number of different reference materials, different ethnicities, was that we didn't want to focus on only a single sample because you can over-optimize your algorithms when you use a single sample, and

we wanted to have a variety of genetic information context but which we can stress the system and analyze the systems.

Coming back to this table, and if you think about not only inherited diseases, but if you consider cancer, sematic mutations, noninvasive prenatal tests or infectious diseases, now you're no longer left with homozygous or heterozygous genotype calls but you're left with fractions, allele fractions where you want to report on allele frequencies that are maybe 4 percent, 5 percent, depending on where the cutoff is for your test. Copy numbers are also going to take the whole range, between zero and any number, not just whole numbers. So what your test needs to do is your test really has to have a much higher performance, has to have higher sensitivity and has to maintain the specificity as a test for the inherited diseases.

So that obviously also has implications for the reference materials that you might want to consider. I've walked you, on a previous slide,

through some of the errors that can get introduced when processing your samples. For inherited diseases, many of those errors are probably not that drastic, that they will not have a big impact on your end result, but you have to keep in mind they will be additive also. So when you look at lower allele frequencies, your test has a much higher sensitivity, so small mistakes that are being made throughout the process, they accumulate and you will see them in the end, and they will impact your results.

So reference materials that we discussed at the Genome in a Bottle workshop were to simply mix some of the reference material that have been selected to get a range of allele frequencies that you can analyze, although those allele frequency differences would be based on polymorphisms that are present in those reference materials.

In the case of cancer, we thought it would be very useful to have specific mutations because some specific mutations play such an important role to have these available, and there

is utility for cell lines that contain human-relevant mutations. There are some vendors who offer these. And then we've been starting to discuss synthetic constructs where mutations have been engineered in plasmates, and these can be used to test your process. I'll come back to those in a minute.

Here are some examples of how we've used some of this at Quest Diagnostics early on in the development of this next-generation sequencing test, OncoVantage. So on the right hand, the table shows you that we've used those cell line references with engineered mutations to determine can our test detect those mutations, and then how well do we match up the frequencies for these mutations that have been in the reference material established using droplet PCR.

The other way we've used these cell lines was to analyze and to optimize our bioinformatics pipeline. If you think back a year or two years, there has been a real need to be able to identify short deletions or insertions

using the different alignment processes, and some platforms simply were not able to do that. And so we've looked into these platforms and tried to optimize the parameters so that we were able to detect those deletions because we thought those were really critical for a test like this, and we would not be able to make a test available without being able to detect those deletions.

FFPE treatment, I've already mentioned it. It impacts your DNA sequence. So within the working group of Genome in a Bottle, we've discussed experiments and how to test this. Actually, ABRF, headed or represented by Don Baldwin at our workshop meetings -- ABRF stands for Association of Biomolecular Resource Facilities-- is planning to take cell lines and compare the results of FFPE- treated cell lines and untreated cell lines to determine what is really the impact of FFPE treatment as the end result.

In terms of synthetic controls, in my mind, and I think in a lot of people's mind, there

is the question, well, are synthetic controls -- how can we use them? What are the limitations? And are they really useful in terms of substitution or supplementing patient samples within the process?

So one hypothesis that we want to test is are synthetic controls good surrogates for DNA isolated from a clinical sample? And in order to do that, I think we have to acquire evidence to either support or refute that hypothesis. And so what we want to do is we want to come up with a study design that will enable us to answer that question, come up with answers to that question. And we're actually looking also for input from multiple stakeholders, because if there is information like that, we would hope that a lot of labs and companies would be using that information and translating that information into practice. And we would anticipate that a study like this would not only be conducted by individual groups but by multiple different groups in order to put that on solid footing.

Simulated sequences, the value really is that very often we don't have samples available, patient samples, that carry specific mutations, or we have specific questions towards our bioinformatics pipeline. So we can use a dataset and manipulate it in a way that we engineer mutations into that dataset. And that can be specific to a platform. The platform that you're using in your lab can be specific for the variant type that you want to analyze, and it can be specific to the context of the sequence. So you can put it next to homopolymer sequences, et cetera. And you can determine whether your pipeline will detect that specific variant, what are the frequencies that are attainable with these types of variants, and which adjustments do you have to make to your pipeline.

Here is one of the examples for two mutations that we've generated such simulated data for. What we wanted to find out -- well, what is really the average frequency that we will detect within our pipeline, but also what is the range of

frequency that we will detect versus the expected frequency to give us a good idea about the sensitivity of the pipeline.

In summary, one point that I haven't made but I think is obvious to everybody is when you look at sequence and compare research applications and clinical applications, there's an important difference between research and clinical applications where you have to have a much higher performance level on the clinical side versus the research side. That information is being used to make critical decisions for patients, so you have to be sure about what you're reporting. Next-generation sequencing will really benefit from a well-characterized group of reference standards, common reference standards that are suitable for analytical assay validation.

One point that I want to make about the Genome in a Bottle reference materials is that one of our selection criteria also was that there is a fairly broad-based consent available for these samples. So if somebody -- another group wanted

to come in and produce derivatives of that material, put other mutations in there, or you can think about different things, that the consent would allow for that. Synthetic constructs and simulated sequence data, they all can be very useful to assess your pipeline and assess different parts within your pipeline, and I think the next steps really are to discuss and determine what the limitations or what the use of synthetic standards can be specifically also in the oncology area, specifically also in the oncology area for analytical validation, but also to run them every time when you do a sequencing test to include that within your test.

And then for simulated datasets, I think there also had to be agreement. How do you assemble the simulated datasets? Which information can you draw from these simulated datasets and come to that acceptable conclusion within the community?

With that I want to conclude, and thank you for your attention.

(Applause)

SPEAKER: Hi, Andrew, that was a fantastic overview.

On your last point, in your conclusion, the simulated data, I mean, obviously simulated data is fantastically powerful. You can generate thousands of fake samples, millions of variants, and use them for validating pipeline updates. We've, in our diagnostics lab, done that extensively and it's very useful. The challenge, of course, is it's difficult to simulate all the things that are going on biochemically in the raw data coming in. So, for example, if base modifications, because of the FFP treatment, are local sequence-context specific, or pull-down efficiency, which of course varies by mutation type, have you looked at all of that issue, and do you have any thoughts about how to generate simulated data that might better reflect the underlying biochemistry?

DR. GRUPE: Steve, I think you make a very good point. There is a lot of complexity

that goes into the whole process and so it requires really consideration of many of those points. I think it will be difficult to have a standard simulated dataset for that purpose, because that simulated dataset probably has to be consistent with the workflow that you really want to test. So that's probably something that you have to take into consideration when generating those simulated datasets.

SPEAKER: So we've been trying to work on such a thing. As you point out, it's a challenging problem, but we'd love to see if a consortium around that might be of interest.

SPEAKER: Another consideration for cancer samples, you can have at the same time gross genomic changes, a mixture of (inaudible). And what I've seen in the clinic is that you should -- you usually did cytogenetics and then the molecular oncology part. Now, everything is done at the same time, so that might affect the results, too, and something that has to be taken into account, I think.

DR. GRUPE: Yeah. I agree with you. Specific cell lines probably have endogenous rearrangements. Specifically, if you take cancer cell lines, they are a mess, but if you look at specific mutations within those, I think there is value for those cancer cell lines. The reference materials that come from normal donors, they might be different because they will be EBB-transformed B cells most of the time and they probably will have less of those limitations that you refer to.

MS. VOSKANIAN-KORDI: So a few more from the online audience.

One was, you spoke to this a little bit, but a question on whether or not it is actually suitable or appropriate to use cell lines versus patient lines as a reference, and the comment from the audience was that it is not clear what the different institutions or government agencies' position is on this or whether there is policy on this yet, and my understanding is that you're actually working to study this to kind of help form guidance for the policy. So the policies are

not yet in place, but if you could just clarify an opinion on that, that would be --

DR. GRUPE: Yeah. I think I can't make any FDA policy.

MS. VOSKANIAN-KORDI: Sure.

DR. GRUPE: I'm the wrong person to answer that question. And I think we just need some studies in order to substantiate our decisions one way or the other.

MS. VOSKANIAN-KORDI: Well, and then the other question is what is Quest's position on how to handle potential false positives in patient samples that show deleterious impact?

DR. GRUPE: That's a good question. That's a piece of discussion also. For example, for the OncoVantage test, right now we're only analyzing tumor tissue and not blood samples, so we don't necessarily get direct information on the germ line DNA other than through the tumor. It's also limited to those 34 genes, so you're not looking at a whole genome scan or whole genome sequencing for these samples, so therefore,

incidental findings would also be limited.

(Applause)

DR. DONALDSON: Our next speaker will be Dr. Laura van't Veer, and her talk will be "Impact of Next-Generation Sequencing on Drug and Diagnostic Co-development."

MS. VAN'T VEER: Thank you very much for the opportunity to have me speak here on this interesting topic and I will particularly focus on the diagnostic code developments of next generation sequencing and other molecular tests because in that aspect they're all the same. I think we are in an exciting era, the era of here are my genes and what is my drug. And people are pouncing on our doors and want to know the answer and sooner than later is actually what is required. So how do we get there?

So companion diagnostics and in particular companion therapeutics there is an exciting time ongoing in oncology because there are over 1000 targeted drugs in development. We're approaching 50 that approach full approval

in oncology. There are numerous biomarker technologies for screening of diagnostics that would apply to those targeted drugs. Some are really specific, others are more broad. But we're discussing today of course standards, are these tests all reliable and how do we make them reliable? And the clinical trials that are currently ongoing some immediately test drug biomarker combination but sometimes that's actually done after the fact. And then there is the issue of efficacy of off label drug use. So once you find a mutation in one tumor type where you may have a PMA approved diagnostics for a targeted drug you may find that mutation in another cancer type and then what do you do. And so off label drug use of course is always possible but it may I think particularly in this setting of next generation sequencing we may want to have some further guidance.

So how do we manage all this so we make fast progress to actually deliver on the promise of precision medicine? The biomarkers for agent

selection has been extensively documented in a report from the Institute of Medicine under leadership of Gil Omenn which actually nicely, besides only this figure, lays out a landscape for diagnostic tests and how do they come from being identified to actually really clinical utility. And also in the end as you can see here in the bottom that they end up in practice guidelines and reimbursements. And although we're focusing here at this meeting and what are the requirements for FDA clearance and FDA approval there is actually a whole other part that we're not addressing today which is in the bottom to get this in these practice guidelines and being reimbursed. And having experience myself that actually the requirements for each of these steps are not the same which is confusing. So once we have actually a biomarker that can be used for agent assignment there are two ways how that can be tested in prospective clinic trials. One option where the test does not direct patient treatment and you can just test it in the background, or when you really

test this to direct patient management, a so called FDA investigational device exempt status is needed. And so those are being tested for some of the targeted drugs currently in practice. And sometimes maybe a particular test has already been cleared and you can use it as part of the IDE as well for a new indication. So the targeted drugs, they're over a 1000 that under development or in use or actually targeting what we call the hallmarks of cancer which has been nicely described over the years and consists of processes like sustaining proliferation, evading the immune surveyance, and many new drugs have seen the lights that are very effective in model systems and some start to show their efficacy when used in cancer patients.

So accompanying those tests are a number of prognostics and today's topic is predictive tests that are either marketed as LDTs, 510Ks, or PMAs and that's for all of these different indications. And so you can see here for which cancer types are actually already tests being

used. But there is sort of a flip side to this because many of the companion diagnostics are based on maybe a single clinical trial or retrospective prospective assessment. Then there comes the issue of me-too test because if you can identify a mutation with one technology and it's easy to do with another then sometimes this is being used as a laboratory developed test which is of interest to the FDA with their latest announcements to actually also oversee these type of LTDs. Sometimes one trial per diagnostic test is not sustainable for the future, and one test per approved drug might also not be sustainable. The regulatory oversight is variable so there is lots to improve to actually make us all go to a more structured and efficient process.

I'd like to take you to one example in breast cancer where we tried to develop such a structured process also in communication with the FDA on several steps. In breast cancer there are also a number of targeted drugs in development that are guided for their use by the biology of

the tumor. And for this particular example we've been working with the FDA guidance to use pathological complete remission as an early end point for accelerated approval. So oncology that means you can give a treatment to a patient when the tumor is still in the body and you're observing in a six month period usually where at surgical resection at the end of the periods there's still tumor left in the patients or whether that's all gone, being identified as pathological complete remission. There's an end point that is only accepted for these high risk breast cancer patients.

High risk breast cancer patients, there are number of new strategies that either target the immune environment, target the (inaudible) gene, target stems cells, the PI 3- kinase pathway, are several ways to target the (inaudible) to a receptor. The trial that all describe where we are using this is the I SPY 2 trial, investigation of serial studies to predictor therapeutic response with imaging and

molecular analysis where the trial PIs are Laura Esserman and Dawn Berry. I'm chairing the biomarker work of this particular trial. Investigational arms, you can see here the different processes in the middle column that were also shown in the hallmarks of cancer that are all options for high risk breast cancer patients to be tested in this Phase II trial. Under several drug arms, investigational arms in this trial in parallel up to five in comparison to the standard of care. The trial design is novel because it's uses as I said PCR as an endpoint, an early endpoint. Final assessment will be by long term survival, but the first endpoint is PCR to assess the response rate of investigational agents in combination with standard therapy as compared to standard therapy alone. Randomization is done adaptively and I will not show the details of that, that can be a whole different talk, but suffice it to say that it uses a pre-specified and automated algorithm where the randomization probability to each of the arms of these different

targeted drugs as the study proceeds is based on the molecular signature of the tumor, and actually based on MRI ongoing through the trial, and finally the assessment of the pathologist in the surgical specimen of PCR. This same algorithm triggers the decision to graduate when approximately 60 to 120 patients are enrolled because if at that time the adaptive randomization doesn't show superiority over standard of care for this particular trial it is decided that investigation was being dropped because the incremental increase of success is too low for it to go forward. It doesn't mean there's no effect, but not in this trial process as chosen how to use it.

But in addition, and that's more important for today's talk that we implemented the framework to qualify promising biomarkers for companion diagnostics to make it as efficient as possible. The trial design in a simplified version here, there are two investigational arms, the middle one and the bottom one as compared to

the top which is standard of care for breast cancer patients before surgery, but we have up to five -- and actually currently five investigational arms in addition to the standard of care. So here you can see that at different time points MRI assessment is done as shown in the bottom as well as biopsies. And for eligibility the biopsy before randomization is used to actually define eligibility and is used within the adaptive randomization. There are tests that are being used for eligibility and randomization are either FDA cleared or are available as standard of care like some immune is to chemistry tests. And if the patient shows to be high risk by either one of these features as indicated here the patient is on the study and this one in particular combination of these factors is used in the adaptive randomization leading to defining eight subtypes that include ones like triple negative for breast cancer, hereto two positive, (inaudible) very high in those. The evaluation of the investigation for the standard is being done.

So there's a long story in between these two slides which is to show you that last year the first drug connection, this year the second drug were reported at meetings to have graduated successfully from the trial arm. Here I'm showing you a PARP inhibitor Veliparib which was given in combination with Carboplatin where you can see that three different signatures all were to negative patients, hormone receptor positive for two negatives and hormone or triple negative were being evaluated ongoing through the trial and going the trial it was identified that the triple negative group had a more favorable response than either the full group or the hormone receptor positive for two negative. So there was an enrichment ongoing for this particular subtype which meant that the drug graduated only after 115 patients had been randomized to this treatment arm or to their concurrent controls. And so that's a real efficiency increase because the number of patients needed to come to this conclusion of graduation because of the enrichment has been

decreased enormously. In the right two columns you can see the probability that in a phase three trial this drug will be better than standard of care.

What I would like to point out to you because it's important for the qualifying biomarkers that although we were successful in identifying the tripe negative group as a graduating signature where 52 percent of the patients achieved a pathological complete remission as compared to 26, that still means that only half of the patients benefit from the drug so there is room for improvement. So one of the programs and very active programs we're running in parallel is how can we use the biology to identify companion diagnostics and we have a process to qualify biomarkers to be tested on a qualifying level based on earlier evidence. And it's all to see if we can increase this 52 percent in an investigational way to a higher proportion by applying those biomarkers. So the way how we set up the system is that we're using established

biomarkers or ones under an IDE so it's an investigational device exempt status for stratification and randomization as I explained. But the qualifying biomarkers need to be run under CLIA and for some of them we actually have an IDE platform that we're using to qualify these biomarkers. So these are hypothesis testing not yet used for treatment assignments, but they could be subsequently in a Phase III trial be validated and then become companion diagnostics.

Of course in the background there is a lot of exploratory analysis which is more hypothesis generating and once the biomarkers are used as an IDE this facilitates as part of the IND the process towards the FDA PMA approvals. So the companion diagnostics that we tested in the I Spy trial for the PARP inhibitor Veliparib in combination with carboplatin are listed here and you can see that the biology speaks because PARP inhibition is effective in the DNA repair or blocks DNA repair processes and is considered to be very effective if that is deficient. So here

are a number of tests that we have been qualifying and I cannot show you the results yet. They will be presented this year in December, but just for the purpose of this talk assume that some of these qualified what do you actually need to do next. So this is like an overall scheme as was published in clinical cancer research earlier this year showing that if you go from an adequate test performance in the laboratory setting and if you then validate it in a clinical validation you need to go all the way to clinical utility. And on top of that the regulatory oversight of course can be several. In Europe it's mainly ISO certification, in the U.S. CLIA. There can be a European CE marking or here in the United States it's a 510K or a PMA process. For treatment assignment it is the PMA. And of course as I said before guidelines and reimbursements in the end will actually define the clinical utility.

So where are we with I Spy 2? I think for those that I just showed you, although not all of them I can say, but some of them actually made

it into the final phase of clinical validation. So that means those that have been successfully qualified were run under CLIA and some actually were from an IDE platform, they can now be used in a Phase III trial to really show the clinical utility of the tests in a predefined way even though you can still choose if you want to have treatment assignment based on the diagnostics or not. If that is successful then of course as I started my talk and it's very much of interest and for our next generation sequencing there can be a lot of off label use which could eventually lead to additional indications if sufficient evidence has been accumulated. But if you use a particular test and particular technology and that's particular for next generation sequencing, once you have either 510K or PMA in case of treatment assignments, if there are improvements in technology one always needs to keep in mind that you need to update the PMA if that's the technology you're using. So that's a separate discussion.

Just to end to show you one slide which was inspired because of the whole database session yesterday because besides that you can do all of these analyses in isolation and here you can see the I Spy clinical trial data and the forced database we're using as well as molecular systems and hospital systems. We also actually are thinking, and this part is actually functional, we're also thinking how to actually be better in analyzing data by using semantics. And this is a slide that I got from Sue Dubman who's actually overseeing the I Spy IT systems. So I think clinic trials very much also needs to be much more documented in an automated way so that analysis can be done more efficiently.

I'd like to end with to actually acknowledge the I Spy 2 trial study team which has a large group of working group tiers but also many individual molecular oncologists are active as agent chaperones in the trial, the project management office, and the oversight, the sponsor and the funding and the oversights where you can

see from the FDA, I see there Janet Woodcock and Richard Pastor, but I'd also like to acknowledge actually CDRH previously Steven Gutman, but currently Bob Becker, and also the work of Liz Mansfield in guiding how to use next generation sequencing which is currently being employed for one of the drugs in the trial. And I would like to end to say that also I think lots of the thinking how to implement all of these diagnostics is based on the work with Jon Retzlaff and Rasika Kalamegham from AACR where we have a very active working group between academia and industry, particularly pharma and diagnostic companies. So when you think about what are the efficiencies and what are the requirements that we need to actually come to companion diagnostics.

So I hope I didn't leave you too much in a sort of twist here. I'm happy to take any questions. (Applause)

MS. VOSKANIAN-KORDI: Due to time limitations we're only going to take one or two questions so if anyone has any.

Okay. Our next speaker is going to tune in remotely so I'm going to set him up if -- introduce him maybe.

MR. DONALDSON: Okay. Thank you very much, Laura.

(Applause) Our next speaker will be Dr. Charles Sawyers and he'll be sharing in the age -- the topic of his talk will be Sharing in the Age of Clinical Genetic Data.

MS. VOSKANIAN-KORDI: I want to see if he can hear us first. Dr. Sawyers? Is the volume on?

DR. SAWYERS: This is Charles Sawyers in New York. I apologize for not being able to be there in person. I don't see my slides appearing on the screen yet. And I cannot hear my voice over the background so I don't know if I'm live. Okay, my slides are up. Thank you. I apologize for not being able to appear in person. I've been listening to the session, enjoying the conversation. I'm going to cover this topic from

a slightly different perspective and that is leveraging existing clinical data to help us discover new biomarkers in a validated way.

So I'd like to begin by reminding you that many of the companion diagnostics that we've heard about, particularly in lung cancer, came about from clinical observations that set sets of patients were responding to targeted agents and that led to investigations of the molecular characteristics of those responders versus non responders which elucidated the predictive biomarkers. As many of you are well aware that prior to clearance of information sequencing instruments many cancer centers have invested heavily in sequencing large numbers of patients either with very focused panels of genes or of larger, you know, exomes. I can tell you that at Memorial Sloan Kettering here in New York we've committed to sequencing 10,000 patients over the next 12 months just at our center alone and I know of many similar efforts at other large cancer centers. So I mention this because it presents I

think an unprecedented opportunity to try to leverage this existing data to figure or discern more clinical correlations between different mutation patterns.

So on the next slide which I don't know that I can control -- can someone advance the slide for me please? Got it. I just want to refer you to this paper that was published earlier this year in Nature by Gaddy Getz and colleagues to introduce you to this, you know, concept called saturation analysis. So while this was done for a different purpose it led to a thought experiment that I'd like to share with you what I've done recently with Gaddy. So what's showing here on this slide is from his paper, an attempt to leverage the 5000 tumors on which whole exomes had been completed and are in publicly accessible data bases estimate how many tumors would you need to sequence to discover, you know, it includes all cancer mutations. And of course it depends on the frequency of the mutation and the tumor type. What you can see on the red line here is that

mutations present in 20 percent of patients with that particular tumor type can -- all those mutations can be discovered, you know, with several hundred patients. But most mutations are not present at 20 percent so we have a significant number of tumors left to go to define the universe of mutations. And what this paper shows you is that it detects mutations that are present at a two percent frequency and you need to do about 2000 samples of each tumor type. If you assume there are 50 different types of tumors it's only 100,000 tumors. Well 100,000 is a large number, but we all know with the cost of sequencing today that's an attainable goal and one that many consortia I think will achieve over the next several years. But the real reason I show you think is to ask the question can we apply this thinking to biomarker discovery. How many patients would we need to discover a biomarker? We're watching through a well known example in oncology which is drugs that target mutant BRAF present here in less than 50 percent of patients

with melanoma. So this is a famous waterfall plot from a New England Journal paper that described the response to the BRAF mutant patients treated with Vemurafenib. And you can see that most -- you know, a huge number of patients are responding which is why this drug was such a game changer when it was first studied. You can also see by the red line that only about half of the patients who the response biomarker have tumor shrinkage greater than 40 percent which would meet the criteria for objective response that's established among clinical trialists. Now we all recognize that there's a smaller fraction of patients which I've depicted here as 20 percent which are having the best responses. And the question we'd all like to know is there a second biomarker, what I would presume biologically is a modifier of response that might predict for those patients who have the best response. And if you knew the answer to that maybe you'd treat them with Vemurafenib alone and you'd treat the other 80 percent with some kind of combination. So how

many patients would it take to discover such a biomarker? That's the question. So using saturation analysis Gaddy ran the same mathematical algorithms and what you can see here is that -- of course it depends on the frequency of the biomarker in the population, but if it's present in one out of five patients remarkably you need to only 200 patients to discover that biomarker. If it's present in 5 percent of the population you need 1200 patients and if you scale this all the way you hit a saturation at 1 percent of biomarker frequency, you need 6000 patients. Now depending on your point of view that's a large number, but I can guarantee you there's 6000 patients who'll think (inaudible) lower number receiving Vemurafenib in the U.S. today. So if we can just leverage the existing clinical practice that's going on we're in a very powerful position to refine ability to give these drugs more precision.

So there's a medicine area that I'd also like to walk you through which is the phenomenon

that you're well aware of is that most of these drugs are effective for short periods of time and the patients develop resistance. So again using the same drug in the same disease the patients tend to respond for about a year before relapse. So in another paper published earlier this year by Levi Garraway and colleagues in Boston there reported the results of whole exome sequencing from a small cohort of patients who had a biopsy done, pretreatment, and then one had relapsed. Now we know that melanomas have a large number of mutations at baseline due to UV exposure most of which are passenger mutations. So the biological evolution of this disease is depicted here with the heterogeneous nature of the tumor at presentation, each of these shades representing different sub-clones which have been produced through different driver mutations that accumulate over time in a background of passenger mutation. So what Gaddy and I did was look at this published data and noted that the background mutation rate in tumors at diagnosis was roughly 12 mutations

per megabase, but now post treatment patients, you know, roughly in a year now have 13 mutations per megabase which across the genome the enormous number of mutations to attribute causality for resistance. However, you know, it's fairly straightforward to realize that what is most important here are the new ones or the additional mutations that have appeared since treatment started and that's a very small number and therefore allows you to make a calculation as to how many patients would it take to discover causes of resistance and it's only fair in existing patients to discover all of the resistance genes for Vemurafenib that would be present at at least a five percent frequency in patients with a power of greater than 90 percent.

The last though I'd like to leave you with is another clinical scenario which gathered a lot of momentum about two years ago when two of my colleagues at Sloan Kettering, Barry Taylor and David Solit, reported a whole exome sequencing analysis of a single patient who was treated on a

clinical trial of a drug called everolimus which inhibits mTOR in a trial of patients with bladder cancer. This was the only patient on the trial that had a complete remission, that had a remarkable complete remission. And what they found when they sequenced that patient were mutations in two genes, TSC 1 and NF 2 for which biology, you know basic cancer biology conducted in laboratories had clearly predicted dependence upon t he mTOR pathway. So this is the classic (inaudible) story, an anecdote, but by knowing this and going back and looking at the remainder of the patients in the trial they were able to identify four other patients who had TSC 1 mutations who had partial responses. So this had led to now a prospective study of everolimus in patients who have TSC 1 or TSC 2 or NF 2 mutations. Since that report many additional similar stories have been published or are on their way to publication with different drugs and the National Cancer Institute is launching an extraordinary responder initiative which has

already led to queries from several hundred investigators who had several hundred patients who would fit the criteria for such a response.

So, you know, the summation of all of this is that there are remarkable clinical experiments being done. The data is sort of out there on the table, but it's been difficult to harness. And we are very involved as a community in efforts to build the infrastructure to allow such data to be shared and leveraged more easily. So I want to call your attention to an organization called the Global Alliance for Genomics and Health on which I serve the executive committee which was initiated just in the past year based on a series of meetings held about 18 months ago amongst leaders not just in cancer but in mendelian diseases, infectious diseases and other communities. You can go to the website for details. The first in person meeting was held earlier this year in London. The next one will be held in October at the Human Genetics meeting in San Diego. This slide is a bit out of date but

there are well over 150 partner organizations now spanning 20 countries and 6 continents which have a number of working groups which you can query on the web which are coming up with data sharing standards, consent standards, and infrastructure for data sharing. I'd also like to let you know that very soon the American Association of Cancer Research for which I just completed my term as President is structuring a trial project amongst several cancer centers in the U.S. and Europe to share sequencing data of the type that I've described that's already happening at our centers to try to leverage the power of this approach.

I'm going to stop there and you can tell me what to do next. And if there are questions I'll try to take them. Thank you. (Applause)

MR. SIMONYAN: Questions? Maybe if there are questions we can ask them to read them for him because he cannot hear.

MS. VOSKANIAN-KORDI: That microphone is not on.

SPEAKER: Very different in that my talk

was sort of edificating we need to have structure at trials to actually validate these biomarkers whereas his presentation shows us that we need sort of the world to actually understand what are the right biological markers to understand response and resistance. So my question to him would be how does he think about validating the findings from the global alliance? So do we need trials or do we just need our common sense to actually start using the -- it would be like an off label testing that you start to promote. The sort of complexity of the mutations is so huge that to make sense out of it you need to combine the whole testing from the whole world. And so how do we validate it?

(Audio Interruption)

DR. SAWYERS: -- from the extraordinary responders program would make their way into the clinic. I don't know if there is a formal position, but the approach that, you know, in the academic community we're thinking is the extraordinary responder insight is an insight that

leads them to a more formal, you know, test of the drug in the companion diagnostic mode for approval. That at least would be the way the everolimus example would be proceeding in which a set of biomarkers for the TOR pathway, TSC 1 and 2 in particular in a clinical trial would move forward.

(Audio Interruption)

MR. DONALDSON: Okay. In the interest of time we're going to not have the panel but I would invite you all to continue this discussion with the working groups that will be forming and if you have any ideas over lunch please let me know.

Thank you very much to the panel and thank you for your patience and for part of your lunch hour. (Applause)

MS. VOSKANIAN-KORDI: We will try to start again at 1:00 o'clock. I know it's a little late; I understand some of the speakers had a little more to say than the time allotted. If you guys could gather back in here at 1:00 it would be

really appreciated. If a lot of people aren't in here we'll try to hold it back a little bit but just would like to say on the time frame. Thank you.

(Recess)

DR. TEZAK: Okay. We're going to start the session. I want to thank everybody for having a really short lunch and coming back. We're just a little bit late. So next session is Next Generation Sequencing Devices and Clinical Applications. And actually the session that just ended was like a really, really intro to our session so I want to thank Eric and all the speakers from the previous session for giving really good intro both to the NIST standards and Genome in a Bottle Consortium efforts and also to CDC led efforts for a standardization of next gen sequencing, data formats, and testing.

So I'm Zivana Tezak; I'm at FDA. I'm in CDRH Office of In Vitro Diagnostics and my co-chairs are Justin Zook, who's going to speak a little later from NIST, and Heike Sichtig who is

from FDA and she's going to also speak a little later. What we're going to do is -- so this is the agenda and probably we're going to have to shorten maybe Heike's presentation and the panel maybe to 30 minutes so that we're still done around 3:00. So the first talk will be -- I'm going to introduce all the speakers first, the same as what Eric did. So the first talk will be from Dr. Ira Lubin from CDC, and he's been actually as previous speakers mentioned very involved in leading and forming various consortia on standardizing the human genome sequence representation and the formats. Then the next talk will be Dr. Justin Zook from NIST who has been leading Genome in a Bottle Consortium. So we have the data format standardization, the first talk and the second is actually about reference materials. And then the third talk is going to be from Mya Thomae who is VP for Regulatory Affairs in Illumina and she's going describe the first four clearances for next gen sequencing devices that Illumina paved the way for others. And then

the last talk will be Dr. Heike Sichtig from microbiology division and she's going to talk about the infectious diseases efforts and thinking that FDA has in some research efforts. And that's going to be followed by the panel discussion. And we are very happy that we have four current micrograde developers, all the representatives on the panel. And Justin agreed to moderate the panel and our panel is going to be Dr. Korlach from PacBio, Dr. Shaw from Ion Torrent, Mya Thomae again from Illumina, and Kendal Dinsmore from Complete Genomics.

So without further ado we're going to start the session.

DR. LUBIN: Okay. Back from lunch and reenergized. So I want to thank the organizers for the opportunity to present on one of the efforts that we have underway at CDC in which we have established a national work group to explore the development of a clinical grade variant file and the data standards therein that will be useful -- that will support interoperability for

applications relevant for clinical and public health settings. So the intention is to take a look at variant files and by variant files we're thinking about the VCF and similar files that are currently used during the course of next generation sequencing and how we can evolve these to really present unambiguous descriptions for the sequence elements that are representing therein, preferably with a haplotype background and also to provide the capability to communicate these sequence elements, genomic representations among clinical laboratories and within healthcare infrastructure that conforms to adoptive standards such as HL7 that will support systems interoperability.

So the applications that we are seeking to address, to develop use cases today at best are challenging to put into practice and often require custom solutions for those in which we can actually address today. And this includes exchange of sequence information among laboratories for quality assurance. Proficiency

testing is probably the best example that folks are knowledgeable about here. To be able to take sequence generated from the laboratory either at the stage when a variant file is being generated or those results that are found to be clinically relevant and deposit them into medical data bases or disease registries to outsource variant file data in a uniform manner for downstream analysis and interpretation and ultimately to be able to message that data to patient records making it amenable for clinical decision support systems or reanalysis at a future time when there are new testing indications or as new knowledge is developed that would then trigger the need to reanalyze that data.

Looking at the broader sphere of how our health IT system is evolving we're now at a time when we are seeing health information exchanges come on line. In some settings more successful than others but this is the direction we're heading. The continued integration of electronic health records, data warehouses where clinical

information is being deposited and available for analysis to provide a more timely improvement of patient care and assess quality metrics and how that care is being delivered. Also we have patient portals coming on line and the capability of patients being able to see their data in more real time and mechanism in which we translate what's coming off of the test and being made available to patients.

So a little over a year ago we established what we're calling the Clinical Grade Variant File Specification Work Group and this was established in collaboration with other Federal agencies based on previous work groups we had that saw a real need in trying to develop a more standardized way for genomic representation both in the variant file and how it can be communicated to other settings. We're also aligned with the HL7 Clinical Genomics Work Group and we have a number of others throughout the community, clinical laboratory professionals, informaticians, and folks who are active with a number of

professional organizations as well as translational researchers who are involved with this group. Now we also realize that the outcomes, the opinions of any work group is the consensus of that work group. So we just released a website VCFclin.org in which we will begin publishing the outcomes of our discussions on that site eliciting feedback from the broader community that can inform our final recommendations because as all of us has probably learned the opinion of even a small group of people is probably not shared by the larger community, so we really need this consensus building process. Our intent is to take the principles and recommendations that are developed and use those to inform use cases designed around the applications I noted earlier. And this will form the framework for pilot studies in an effort to move the recommendations and principles into practices. This is not something that can happen quickly but in taking sort of this structured approach to identifying what needs to be done and then figuring out how to do it the

hope is that we can move forward in an orderly fashion that will be least disruptive but useful to the broader community.

So in thinking about the variant file as it's currently integrated into next generation sequencing we were thinking about this in three stages, three parts of the process for performing next generation sequencing. The first is to consider the data that populates the variant file and what needs to be considered in preparing that data before the variant file consumes it. The second is the variant file itself and to what extent we need to further constrain data within that variant file. And for some applications such as laboratory exchange of data that variant file may be the end point assuming that we can get the variant file into a format that can be readily understood by other laboratories for the comparison of data. And third is to take the data in the variant file and to translate it into formats amenable for messaging through an IT system using adopted standards such as HL7. So

for example we would want after downstream analysis and interpretation on the laboratory side we would want to certainly translate the outcomes of that in HGVS which is recognized by HL7 along with other established descriptors. And that would be the format that those particular test results would be messaged.

So this is a basic representation of the work flow when performing next generation sequencing and in thinking about what we need to do standardize the way sequences are presented we actually have to look back to the stage of alignment and then work through this and see what else we need to consider. So taking a look at the first step which addresses the preparation of the data for input into the variant file the work group discussed primarily two issues, and that is the assignment of genomic coordinates and variant callers and how they need to be evaluated and set up. And so here was recognition and broad consensus within our work group that the alignment or mapping of the sequence reads needs to be done

against a version reference assembly. And the reality today is that a lot of labs are doing panels even though there's a transition to exome analysis today and those doing panels will often align to a laboratory reference sequence. And this sequence or may not be deposited in a standard database such as RefSeq, HLA also has their own database in which there are a uniform set of methods used to cross-map that sequence against the reference assembly. And so if the sequence that you're aligning against is in RefSeq for example you're pretty well set. You can go back and NCBI has done that cross-mapping so you can get your genomic coordinates. If it's a sequence that is not deposited in these standard databases then we would strongly recommend that it get deposited and NCBI allowed to take that through their methods to assign genomic coordinates. Laboratories are discouraged from doing that alignment themselves because aligners vary, the way aligners are set up vary, so there is the probability that if you were to do that

cross-mapping the genomic coordinates that a laboratory would assign may have the potential for higher error and those coordinates being wrong.

In terms of the variant callers, the output of the variant callers are what populates the variant files and the work group is recommending that the data that be collected include just not the variance but the reference calls, the no calls, as well as phasing information to allow the greatest possibility for applying metrics so you can really understand the quality of the sequence that is being generated, as well as permit you the greatest opportunity for interpreting that sequence and developing the data set that you'll analyze to derive what is clinically relevant. So one of the advantages of taking the phasing information is when you have variants close by to know whether they're on the same allele or not in the absence of family studies which you may or may not have, generally not. You really need to use the phasing information to show the allelic representations.

And in this figure we actually have under alternate a single haplotype representation, but you'll see that depending on how the variant caller is set up you can have different representations of this. In the absence of phasing you would not know which output is associated with another in terms of the allelic association. This also addresses the question of the size of the event that you should output when you have control over that. And when possible it's better to output large events where you can see all the allelic associations in the same segment, but when you don't have that opportunity then you really do need phasing information that you're putting out such as a base or just a few bases at a time.

In thinking about the variant file now the work group wanted to focus on content and not format. And the rationale there is that there are translators that will allow you to translate from one format to another. But if the content is sufficiently constrained then it should appear as

the same content from one file to another which then it's amenable to being understood in different settings and supports interoperability. Today variant files in clinical settings are sufficiently different that it's anecdotally it's challenging to take a variant file from one clinical lab and have it understood by another. And this is the message that we have received particularly from NCBI with respect to another project in our group, the GeT-RM project that collected data from various laboratories and almost in all cases the laboratory needed to be contacted to make sense of what was in the file that was communicated. And it's a matter of the implementation of existing specs, not so much the specs themselves in that laboratories will adjust the way they set up their variant file to accommodate the platform they're using in their downstream analytics, and because this is all contained in the lab they can deal with the issues that come up. When you talk about taking the data within a variant file and messaging to an external

setting then you really need to think about the content and have agreement on what it should it look like and how it will be interpreted. The good news though is that for several of the features within the variant file, the designation of the chromosome, the genomic coordinate, a simple variance, these are fairly well established and should be understood even in the current state when there's files moved from lab to lab or lab to another setting. There are limitations though. In the new reference assembly I believe we have about 170 or so alternate assemblies such as HLA where there are multiple alleles that are highly divergent compared to the primary assembly and there has not yet been a consistent means to represent these in a variant file, although there's discussions under way with other groups and how that might be undertaken. Also there are issues with complex variants and haplotypes and how they're represented. You can have different representations and I showed you a basic one earlier that they look different but it actually

turns out to be the same variant. So work has to be done to assure that when dealing with these complexities that there is some level of common structure that these are represented and can be communicated or otherwise tagged as something that may be more difficult and needs to have a closer look.

The other issue which we'll come back to later deals with the metadata. And the metadata supports the sequence, provides context to the sequence. And the metadata that typically currently exists in these files as text type data, not structured data. So there is a lot of interest, and some laboratories are doing this, in trying to put that data in that XML format that permits it to be more structured in mind and more useful for computational processes.

One of the questions that comes up, particularly from the clinical realm is how do you assess the quality of this data. And this came up earlier in Dr. Gonzalez's talk is that there are no common metrics that transcend platforms,

pipelines or assays. So the work group could only recommend that sufficient description be put into these variant files or otherwise made available that describes the filters or metrics in a uniform manner. And what that uniformity is has yet to be established to provide the ranges, the cut offs, and what's being thought of as good versus poor quality. And to put enough context in there that someone who is reasonably knowledgeable, who's outside the laboratory can look at the data and make some determination on the quality of that data.

So the variant file that -- a VCF for example that constrains to the 4.1 specs and where you're dealing with for the most part relatively simple SNPs, INDELs, what have you, should be exchangeable among laboratories. The challenge is that there's enough variation at this point that it's not. So there needs to be additional efforts to encourage more uniformity to provide a means to take whatever the lab has optimized for their pipeline into a format that can be readily

messaged to other laboratories or other settings. In terms of messaging that through a healthcare infrastructure it's possible to put an HL7 wrapper around that and move it to where it needs to be. For the results that are to be reported as clinically actionable you do want to assign the HGVS and in time assign LOINC codes and put this in a message that is more amenable to the standards that have already been established in an implementation guide by HL7. And so I explain that here and the sort of things that you want to at least assign to your clinically relevant variants are the ID numbers H, G, and C. Human Genome Nomenclature Committee assigns ID numbers, names, and symbols to each gene. The caveat is the names and symbols sometimes change. So while it's not standard practice it seems reasonable that ID numbers should go along with all of those with names or symbols perhaps. Very variable. Version HGVS nomenclature, it's easy to say but the reality is that HGVS nomenclature is not always applied in a uniform manner and some of the

specs for HGVS are not well established for certain types of variants, particularly the more complex ones. So in order to have more uniformity of HGVS this is another issue that needs to be addressed in terms of thinking about how we can promote use of that. And it is the adopted standard for how variants are described. Other identifiers, and we mentioned LOINC codes in time; we're not quite ready for the full set yet I think but as we move more into the electronic messaging era we'll need to consider those. And I just listed a number of documents there that contain a number of the standards that have been currently adopted by HL7. These are constantly getting re-discussed and new documents are being developed. And NAACCR has standards applicable to cancer registries but they're weak in the genomics area right now, but that should be changing within the next two years as genomics becomes more a part of the discussions of that and other groups.

The metadata as we mentioned is important for providing context to the sequence

data that's included. When all of this is within a single lab the metadata supports the analysis that's done in the lab and typically there's minimal metadata to support what needs to be performed in terms of the downstream analytics. When you're talking about taking the data in a variant file and messaging it to an external source where it can be analyzed and interpreted at a later time by someone else the work group considered four general categories that need to be considered to provide more robust metadata within that file and those are listed here. And I just want to emphasize the idea that metadata together with sequence data is what is needed.

And finally I'll just end by mentioning the machine versus human readable data. At the level of variant file software is deciding what is deposited in that file. The data within the file then undergoes further analytics with additional software analysis afterwards. So that data needs to be machine readable to a great extent. For messaging, there you're talking about a different

format at least for what's clinically relevant in terms of the chromosome, the HGVS nomenclature. And when reporting to physicians you absolutely need a format that physicians understand. So for example when you're talking about cancer you need to use the formal name of the biomarker so they know what to base their decisions upon. Providing genomic coordinates or refs or even RefSeq, HGVS is not necessarily -- would be useful to the physicians at all times, particularly those who are less familiar with how all of this is evolving. So in summary these kinds of data standards are required to support systems interoperability we have a work group that's addressing this, a website to get community feedback, and it's just not about the standards, it's what you do in the laboratories as well. We talked about the aligning and its importance in the mapping in terms of identifying -- assigning the right genomic coordinate, and the importance of the metadata in being able to understand sequence context.

So that's my world wind tour of what our work group is doing and thank you for listening.

(Applause)

DR. TEZAK: Thank you, Ira, so much. So I think this was a really great talk to show into the weeds what kind of problems people are dealing with here. Unfortunately we ran over so we have no time for questions right now but Ira will be around.

Next up Justin Zook from NIST.

DR. ZOOK: Okay. Thanks, Vahan, for organizing this and I'm happy to tell you some of what we've been doing in the Genome in a Bottle Consortium. Andrew did a great job this morning of giving an introduction to some of the activities that we have and so I may skip quickly over a couple part of the talk, but I'll talk generally about Genome in a Bottle.

And so the question that we're trying to address is that so you sequenced a genome, how well did you. And these reference materials that we're developing are designed to help answer that

question. So we formed this Genome in a Bottle Consortium to develop the infrastructure for performance assessment of next gen sequencing and part of the purpose of this is to support science based regulatory oversight. And so the FDA is actually funding a fair amount of our work at NIST in this area I know is interested in using the reference materials that we're developing. When we started this there weren't any widely accepted set of metrics that you can use to know how well you're doing in variant calling and so we're working towards that but still aren't quite there yet. And we're developing standards to address this. So this includes human genomes that we're making into reference materials, so these are really well characterized whole human genome cell line, extracted DNA from them. And these are being characterized and will be disseminated by NIST. And we're also developing tools and methods that you need to use these reference materials. So if you sequence this DNA and run it through your pipeline how do you compare your variant

calls to our in a standard way.

A lot of the motivation for doing this was because if you compare different sequencing technologies or different bioformatics methods you end up with different answers for variant calls, and so this is just a comparison of three different platforms, bioinformatic pipelines, and you end up with hundreds of thousands of variant calls that are unique to an individual pipeline. And similarly, this figure was shown this morning also for exome sequencing pipelines, even with those exact same sequencing data if you run different bioinformatic software you can get largely different answers. And so we wanted to move beyond just doing these sort of venn diagrams where you say oh, these disagree with each other, to actually understand why they disagree with each other and how do you know which one is right when they disagree with each other.

So the reference materials that we're characterizing like this are designed to understand the performance for a particular part

of the measurement process. And that part is depicted here in yellow on the right hand side. So there are pre-analytical steps where you might extract the DNA from the sample, and our samples are already extracted DNA so they do not test that part of the process at all. But they do test the library preparation, the sequencing, alignment, mapping steps, variant calling, and whatever confidence estimates you associate with the variant calls. And so it will test that sort of black box there that's in yellow. If you get the right answer that doesn't mean that you're perfect it just means that you're doing well for this sample in this particular situation and so you can proceed with caution and so have this red light over here that is sort of -- it gives you a yellow light or a red light. So if you get the complete wrong answer then you know that there's something going wrong. These also are not designed to test the downstream steps, like the clinical interpretation or functional annotation at this point. And these are just normal genomes that

we're characterizing right now.

So the reference materials that we have in the pipeline right now are all 10 microgram samples of DNA that are isolated from a large batch of cells. So we have between 2000 and 10,000 tubes of DNA from each reference material and each of them is from a single large batch. And so this helps to control for any types of mutations that might happen during the growth of the cell lines because we've evenly distributed those mutations over the entire batch. We have a pilot genome, NA12878, that we're hoping to release as a NIST reference material later this year and we already have characterization information that I'll be talking about in a bit. Then we've also selected a couple of trios from the Personal Genome Project for the reasons of consent that Andrew talked about earlier. And so one of these is of Ashkenazim Jewish origin and the other is of Asian original. And these are mother-father-son trios that we're developing because you can do some error analysis by looking

at the trios and also do phasing of the variant calls that you get.

So our goals for the data to accompany these reference materials is to have approximately zero false positive and false negative calls in our high confidence regions. And we want to include as much of the genome as possible in our confident regions. So we don't want to just take the inner section of those venn diagrams that I showed earlier, we want to -- if they disagree with each other we want to understand as much as possible why they disagree and decide which one is correct when we can. We also want to avoid bias towards any particular platform and take advantages of the strengths of each of the platforms so we're using data from multiple technologies in this process, and avoid bias towards any particular bioinformatics algorithms as much as possible.

So for our pilot genome we have 14 and actually now even more data sets from 5 different sequencing technologies. And so these are from

Illumina, Complete Genomics, SOLid, 454, and Ion Torrent. And then we also will be incorporating biodata as it's generated and I'll talk about that a little bit more later. The process that we've used to do this integration was described in a recent paper in Nature Biotechnology and I'd encourage you to look at that paper if you're interested in learning more about the details of the methods that we used, but in general we find a list of candidate variant sites, so we go through with a really low threshold and look for any site that might possibly be a variant and we'll look more closely at those. Then we find the concordant variants and this sort of helps us to see so if it's likely to be a true variant what are the characteristics of these variants. And so we use those characteristics of where if you have a new sit that has sort of abnormal characteristics where it doesn't look like the true sites then we will trust that data set less at that location than another data set that would have sort of normal characteristics. So for

example if one of the data sets has strand bias at a location then says it's a variant and another set does not have strand bias and says it's not a variant we would trust the one that does not have strand bias. And then we assign a confidence level and part of that process is also excluding sites where all of the technologies have evidence of bias. So if all the technologies have low mapping quality then that would not fall in our high confidence regions. Or if it's a segmental duplication, or if it's in a potential structural variant or in a repetitive region, we exclude those regions. And so part of this last process now more recently is also taking the work that the Illumina Platinum Genomes Project is doing as well as Real Time Genomics where they've analyzed the pedigree that our pilot genome is part of so it's a 17 member family pedigree, and they phased everyone and looked to see which variants are inherited properly as you'd expect from the phasing, and that has helped to move some of our less confident sites into the more confident sites

by ones that were consistent.

So after we have these calls questioned a lot of people ask us how you compare it to these Genome in a Bottle calls. And if you're not asking that question you probably should be because it's not actually a trivial process to compare to the calls. So this is something that the Genome in a Bottle Performance Metrics Working Group has been working on and we've developed the specification for what a tool that compares variant call to ours should do in terms of what the input should be, the outputs, what type of licensing it should have, and the definitions of the different performance metrics, like sensitivity and specificity, and false positive and false negative. And if you search on our blog page you should be able to find this specification.

More recently we've been working with the Global Alliance for Genomic Health Benchmarking Working Group which is a new working group that was just formed in the past couple of

months that's sort of taking this work that the performance metrics was doing and actually trying to develop a tool that will do these comparisons. And so we're further developing sort of the standard definitions for these performance metrics and developing some standard methods for benchmarking that can be used so that if you compare your calls to our you know it's done in the same way as if someone compares their calls to ours. We're also developing some standard data sets that people can use if you just want to test about informatics pipeline. And also looking at how you stratify performance. So how you can stratify in terms of how well you're doing for SNPs versus INDELs versus complex variants and then also for different regions if the genomes. So exome versus genome, or repetitive regions versus normal regions. And there are a variety including some doing that.

So I just wanted to give a few examples of some preliminary uses of our high confidence genotypes. These are available from an FTP side

that's at NCBI and you can go to the genomeinabottle.org website to find a link to this FTP site. This FTP site includes both our high confidence calls as well as a lot of data that people have generated on these reference materials. So if you just want to download FastQ files and run them through a bioinformatics pipeline you can do that. And we have exome data and genome data and multiple technologies, and I think potentially some targeted data also. And so these data are already being used. And the data that most people are comparing to are a set of SNPs and INDELs that we have which cover about 77 percent of the genome. And so we have a BED file that specifies the high confidence regions and you should only compare your calls to ours in those regions. There's a GCAT website that's been developed where you can do interactive exome comparisons and you can upload your data to that site and compare you calls to ours and get a variety of different types of performance metrics from it and you can compare to other

bioinformatics pipelines that people have run as well.

The short article, Perspective Peace, about the first FDA -- of the authorization of next gen sequencer by Francis Collins and Margaret Hamburg mentioned that the reference materials that we're developing played a role in this first FDA authorization and we're hoping they'll continue to play a role in the future. Mount Sinai School of Medicine published a paper in BMC Genomics earlier this year where they used our calls to benchmark clinical exome sequencing. Qiagen published a paper in BMC Genomics where they looked at somatic mutation calling methods and they mixed the reads from our reference material in with reads from another genome to see what levels of somatic mutations different pipelines could detect.

So this looks like it got distorted a little bit in the conversion but this lays out what our development plans are for the reference materials over the next year or so. So the first

line is related to our pilot reference material. We plan to release the NIST reference material which will be number 8398 by around the end of this year. And we also are developing a set of preliminary large deletions calls so we're starting to move into looking at structural variants and will be continuing to develop those over the next year and the methods that we're using to develop those so people can do performance assessment with structural variance as well. The next line is about the Ashkenazim Jewish Trio that we're developing into reference materials. We've already collected some data on that that's on our FTP site and I'm actually about to post a blog about the data that's there. We have Illumina and Complete Genomics' data so far but we're also collecting Ion data, bionano genomics, optical mapping data, and solid data that will all be publicly available. And we'll be similarly generating SNPs and INDELs for those and over the next few months we're excited to be collaborating with Mount Sinai School of Medicine

to do 100x PacBio sequencing over this trio, so this will be whole genome sequencing and we think this will really help to get at some of the larger types of structural variants as well as some of the more difficult parts of the genome and help us to characterize them. And that data will be made publicly available in stages as it's produced so that anyone who's interested in helping us to analyze it can help. We're also getting Illumina assembled long reads which is often called molecular data and we will be continuing to develop those and then we'll follow a similar process for the Asian trio in terms of getting all of this data together and characterizing them. And hopefully we'll have those reference materials released around the end of the year, end of next year.

The future directions include for germline mutations we want to look at additional ancestry groups, African Americans and Hispanics in particular we'd like to have represented as well as a large family because we've found that

for NA12878 it's been really useful to have this pedigree. And then we're also looking at somatic mutations and Andrew talked about those earlier this morning. I just wanted to highlight that there are some commercial members of Genome in a Bottle Consortium that I've already either already developed FFPE cell lines based on the cell lines that we're using as reference materials or they're thinking about doing it. And potentially might introduce mutations into them or modify them in other ways.

So if you'd like to get involved there are a number of ways you can do that. You can use our integrated calls and give us feedback. For NA12878 these calls are available on our FTP site like I mentioned. You can help with sequencing and analyzing the new Genome in a Bottle samples, especially with structural variant calling or with analyzing the data from the long read technologies that we're getting. We also hold biannual workshops in January on the Stanford University Campus in California and August here on the NIST

campus in Gaithersburg, Maryland. And you can get involved in the Global Alliance for Genomic Health Working Group where we're developing methods to get performance metrics. And you could also sign up for genomic newsletters at genomeinabottle.org and feel free to email me with any questions that you have.

Thanks. (Applause)

DR. TEZAK: Thank you very much. So we have time for a couple of questions if anybody has any.

SPEAKER: No? Yes? There we go. Sorry. A few questions again from the on line audience. One, any recommendation of reference materials for targeted mutation panels?

DR. ZOOK: So it depends sort of on how targeted the panel is. So I think our reference materials will likely have some mutations in whatever targeted panel you're using, but I also recognize that depending on how small the panel is there might not be very many that are your panel. And so I think for those panels there are some

spikens that Mickey Williams group that has developed that Andrew had mentioned and I think sometimes probably just using cell lines that have mutations in those regions are what you need to do. It ends up being a combination of different methods of doing validation when you have a targeted panel.

SPEAKER: Okay. Next, what interactions occur between NIST and the FDA with regard to any regulation of reference materials?

DR. ZOOK: We are non regulatory so we don't set any of the guidelines at all for what's done, but we hope that our reference materials are useful in the regulatory process. So we definitely talk with the FDA as part of the process and they've funded us to do a lot of our work, but we don't set the regulations.

SPEAKER: And if no one else has any questions -- one more -- any plans to also integrate human phasing data from Illumina?

DR. ZOOK: So our most recent set of variant calls for NA12878 does include phasing

information though it's based on the pedigree instead of using that. I'm assuming they're referring to the molecular phasing data. One of our next steps is to think about how we integrate different forms of phasing information so you can do phasing in a variety of different ways and so we want to think about how to integrate that data together but I've not done that yet.

DR. TEZAK: So our next speaker is Mya Thomae from Illumina.

MS. THOMAE: Good afternoon, and thank you so much for inviting me to speak today. What I thought I'd do today is just go over the clearances that were granted to Illumina at a fairly high level. I haven't included a lot of the data, but for folks that don't tune into regulatory submissions all the time, there's actual a fair amount of interesting work that's been just done on the intended uses and the various products that have come out.

As everybody knows, last November there were four clearances that were released all in the

same day: Two for CF assays, one for an instrument, and then one for a set of reagents. FDA was kind enough to do a press release on this, so we ended up in the FDA news. There was also, as Justin said, an article in the New England Journal of Medicine, and then Jeff Sharon also did a blog on this. It was a pretty good day back in November for Illumina.

The first thing that I thought I'd talk about is the instrument platform, that's the MiSeqDx. For the regulatory geeks in the crowd, I've put in the product codes and the device types. What's really interesting here is this was done as a de novo 510(k). De novo 510(k)'s are utilized when the risk of a device is similar to other class II devices but there isn't necessarily a specific, predicate device.

The instrument was cleared as a de novo, and the very exciting thing that happened is FDA gave it a class II instrument, which is similar to what was done for previous multiplex instruments like the arrays, but they gave it a status of

exempt from premarket notification. That means -- and we saw this this week with Life Technologies -- is that folks that come in after Illumina and have a similar intended use are able to get their instruments on the regulated market without having to do a 510(k).

This was an incredibly innovative approach that FDA took to make sure that sequencing instruments can get out there and a fairly expeditious manner to be used in the diagnostic marketplace, though the intended use of the instrument reflects the data that was put into FDA. It's very specifically for use with the MiSeqDx universal reagent kit, which I'll cover in just a minute, but it's also very specifically for use with DNA and also with peripheral whole blood samples. Similar to other intended uses, it is very reflective of the data that was submitted to FDA.

Along with this there's the universal reagent kit 1.0. Again, a new regulation for this: Reagents for molecular diagnostic test

systems. The reagents were given a class I status, also exempt from premarket notification. Again, this allows folks that have instruments, next-generation sequencing instruments, and related reagents to be able to get both the instruments and the reagents registered with FDA in a very least-burdensome manner. The intended use of the universal reagent kit, pretty straightforward: Library prep and sequencing.

Along with these, there were also two CF assays that were cleared. The first CF assay is pretty similar to other CF assays that have been cleared in the past, and this one threw as a traditional 510(k). For those that want more data, you can find the CF assays in the traditional 510(k) database on the FDA website, but just so you know, I looked up the other day, and I was like, 'Huh. There's only two Illuminas there.' The de novos are actually now in a separate database, so if you need to look up all of the data that is in here you'll see it on the right-hand side; there's one for 510(k)s and then

there's one for de novos. They're just in two separate databases now.

The 139-Variant Assay went through as a class II. This is how other CF assays have gone through, but the interesting part about this assay is that it is an extended panel. It's a much larger panel the most of the other 510(k)s that have been cleared, and that's reflective of the next-generation sequencing technology. I won't read that to you, but it's very similar to other CF-intended uses.

Then there was also a second CF assay that was cleared, this one for a slightly different intended use, and I think that's probably, again, class II, but I think the intended use is actually the most interesting part of this. Again, it's talking about specifically one can be sequenced with this, but the intended uses a bit different than the other CF panels in that the test is intended to be used as an aid in diagnosis of individuals with suspected cystic fibrosis -- again, a little bit different from the

intended CF assay -- and the test is most appropriate on the patient has an atypical or nonclassical presentation of CF or when other mutations have failed. This is a bit more of a specific intended use that was put out there to cover this additional content since it's a bit different from how the normal panels are used.

That's actually all I have today. I'm happy to answer any questions, but it was a really big day. There was a lot of work that was done to get four assays into FDA and get them all cleared, and I know that was true on the FDA side as well. I know we're trying to make up a little time, so I'm happy to stop here or we can take some questions or I can answer things as part of the panel.

(Applause)

DR. TEZAK: We have time for several questions.

SPEAKER: Mya, that was great. Can you just comment a little more on the process by which that particular intended use in your last slide

was decided and in particular the restriction to patients with atypical presentation or for whom the classic mutation panels were negative?

MS. THOMAE: I can't talk about everything that happened, but my understanding is -- and this happened actually before I was at Illumina -- this could be done as an all-in-one panel, but the entire content of the panel was very difficult to cover in a single-intended use. Zivana, I don't know if you want to comment further on that, but it was going to be difficult to find a single, intended use that could cover both of them. I think that the discussion was about making sure that each of the portions of the panel was appropriate.

DR. TEZAK: Your question was why to CF. The 139 actually has clinical utility because the clinical utility comes from the CFTR2 database, so it was really the regular regulation for CF as the previous tests were cleared. The one which is for the whole gene was a novel approach, and we didn't look at every single variation that can come out

of it because it's the whole gene obviously. It was the whole sequencing and not just the panel, not just clinically significant mutations, and we were looking at classes. It was differently analytically validated, and the clinical utility is not as clear.

SPEAKER: Just further clarification: The one finds 139 mutations or variants, but there's over 1000 CTFs that are known that might be related, and the other one's sequence is a whole other thing. Have we found any of them and does the first thing only look at 139 or is it probes?

DR. TEZAK: The first one does look at only 139, that's the output. The tests are very similar underneath, but the output is just 139 for the first one.

SPEAKER: They're both sequencing?

DR. TEZAK: They're both sequencing.

SPEAKER: One just doesn't show everything sort of?

DR. TEZAK: No, 139 only shows 139.

SPEAKER: All right.

SPEAKER: A few from the online audience: Is the MiSeqDx approved for tissue-derived DNA from tumors?

MS. THOMAE: (off mic)

SPEAKER: Okay, and what is the regulatory implication MiSeqDx if someone wants to get an RNA-Seq-based assay cleared?

MS. THOMAE: You need to come in and talk to Illumina Business Development and Illumina Regulatory, and we'll figure it out.

SPEAKER: Just as an added bonus, if you could find the original press release, and we could maybe make that -- you got it? Never mind, sorry.

MS. THOMAE: Got it? Okay, great. Thanks.

DR. TEZAK: Just the last question.

SPEAKER: Given the talk two talks ago about the variability in terms of calling and algorithms, etc., what would the extent of computer systems validation done to the

instruments where the intermediate steps are all validated or just the end result?

MS. THOMAE: There's a lot of different kinds of validation that goes into a system like this. There's definitely software validation that occurred, that was part of the 510(k). There's also analytical validation that's done, analytical studies, and then there's also the clinical studies. I think there's multiple levels of validation that go into it, but I think at the end the whole system is validated as part of the FDA clearance process. (Applause)

DR. TEZAK: Our next speaker, Heike Sichtig from FDA.

DR. SICHTIG: Hi, my name is Heike Sichtig. I'm a principal investigator and regulatory scientist with the FDA in the microbiology division. I'm going to be talking about how we are going to go about enabling NGS-based technologies for clinical diagnostics. This is strictly from the microbiology perspective.

First I want to just show a quick disclaimer where we know that the (inaudible) is moving very quickly and we know that new policy and regulatory issues are being brought up at the moment. These thoughts presented in this presentation are preliminary and do not represent finalized FDA policy. I just wanted to make that clear in the beginning. We are currently working on a draft guidance, but until such time that it's released, we really encourage everybody to come in and use our pre-submission process for any outstanding questions currently.

This presentation today I will outline first because NGS is really a cutting edge tool that is making it into the clinical market. I want to outline possible approaches to validation studies for next-generation sequencing systems and really highlight the difference between metagenomic versus targeted or what is also called Custom Amplicon sequencing and then also talk about the use of sequence outputs in combination with databases to potentially evaluate

performance. Then I'm going to quickly talk about the Interagency Working Group on NGS feasibility that we are currently working on and then really focus this presentation on the MicroDB reference database that we are creating with funds from FDA, and then also touch on these microbial reference materials.

First, for the potential validation strategies, this is a 10,000 foot view of how we're going to go about that. For industry folks in the audience, they already know that we really look at everything from a systems perspective. We really start from specimen collection and then go all the way to the clinical (inaudible). If you look at the green box, that would be something that we look at that would come under regulatory purview, at least from the microbiology perspective.

At the FDA, for IBD devices, we really do apply risk-based regulation. As you just heard, the MiSeqDx clearance got classified as a class II device, and for microbiology devices that

will most likely be the same case. There are some class III devices where you would submit something that is called a premarket approval, or PMA application, and that would be for something like the human papilloma virus or hepatitis viruses, but most of the bacteria that we're looking out would be class II.

Here is the FDA's general concept of how we evaluate diagnostic devices. On the left-hand side, if you focus there, there's a picture of the samples which could be, for example, bacteria, virus, or fungi. We would apply a simple comparative performance, and on one side you have the new device which could be an NGS device, and on the other side you have the reference method which could be a PC or amino acid or composite thereof. Then you apply a 2 x 2 comparative analysis where you want to show clinical sensitivity and specificity. But really the bottom (inaudible) currently is each possible organism needs to be confirmed by this reference method, and that becomes exceedingly burdensome

for metagenomic samples because showing every possible organism in there is almost impossible.

Here again, as I mentioned previously, we really look at evaluating the whole system from sampling or through pre- analytical steps, amplifications, sequence detection, assembly, and all the way down to the clinical (inaudible). We would also look for validation studies for each of these steps, but these would be different for what is called metagenomic sequencing versus targeted sequencing.

For targeted sequencing, we already have an approach called Multiplex, and there's a guidance out if you want to look that up. That would be a good first step to look at.

For these metagenomic sequencing validation studies, we would most likely look at the classical, two- armed approach that we applied before where on one side we have in-house analytical studies and on the other side we look at clinical evaluations. For the in-house analytical studies, we most likely will have to

take liberties -- we understand that -- for the limit of detection and the inclusivity studies because of the exceedingly burdensome showing of metagenomic samples that they cannot detect all of the organisms that are in the sample. We would also look at interference studies, contamination studies, and carryover studies.

On the other side, for the clinical performance, we will stick with multisite clinical users to look at the device, how it performs at different sites, and this should really be designed to test the data analysis pipeline all the way down to the final core determination. The performance could also be established possibly through comparison to a database, but again, we really encourage you to use the pre-submission process to discuss this with us.

For precision and reproducibility, we could suggest using a panel of representative microorganisms. FDA is funding this to create a microbial reference standard to be used as a tool to show precision and reproducibility, but I will

touch on a later slide.

All of these studies were also discussed at a workshop that we had on April 1st, and the workshop agenda discussion paper and webcast is available online. You can simply write down the FR dock note or write me an email if you'd like to get this information or just simply Google it. There we also discussed how we are going to go about working on this draft guidance, and this will be coming out soon.

Now I'm going to change gears and just briefly touch on the feasibility group that I'm a part of where FDA, NIH, NIHID, DTRA, (inaudible), Livermore, and CDC came together. We really wanted to focus in on looking at NGS feasibility, and then we looked at bacteria only. We want to conduct a small pilot study -- which is still ongoing -- to generate information to evaluate quality of existing sequences in the public domain. This is still in progress and we're still churning through the data, but publication should be coming out soon.

The overarching goal was to identify the preexisting high quality deposits in the public and then really use that information and build from there. Then after we have established these quality metrics, we would like to use that information and set our sequence outputs for our ongoing sequencing efforts that we're currently doing that will be the reference database. Then utilizing existing standards, if they're available, for technical and (inaudible) of metadata because why should be reinvent the wheel when some working groups or other people have already looked at these things. We really wanted to pay attention on connecting antimicrobial resistance phenotype to genomic deposits at the clinical collection site.

A very preliminary view of what these feasibility studies showed or what the suggestions that we will be coming out with will look like are that multiple levels of reference databases are likely where you have, sort of, high-quality genomes on one side where these can be used for

validation and clinical use, and then on the other side you would add other available genomes for the testing and development process only. Extensive screening will also probably be required of human and other hosts: Chimeras and artificial constructs. The other suggestion would be to really use separate bacterial, viral, and fungal reference databases. This should all be publicly available.

This really highlights the current need for having a robust, standardized, and high-quality microbial sequence database in the public sector containing representative samples, metadata, high-quality raw sequences, assemblies, annotations all available for the public.

This is a slide given to us courtesy of NCBI, and it really shows the exponential growth of the number of genomes in GenBank, and at the same time not the same growth for the number of new species, but for diagnostic processes. The representation of real species really needs to be increased, and that really highlights our efforts

that we are spearheading with the targeted sequence effort that is funded with limited funds by FDA at the moment to identify and fill gaps. All these (inaudible), assemblies, annotations, and metadata that get generated are sent to NCBI to be made accessible to the public. All these results that are generated should be traceable so they could be reevaluated as necessary.

We identified a very small set of 600 clinically relevant (inaudible) microorganisms initially in collaboration with NCBI, and we are applying a highly controlled and documented approach, and we found a very, very good solution with the Institute of Genome Sciences at the University of Maryland where we are applying a hybrid approach of using Pacific Biosciences sequencing and Illumina Sequencing. We are also using metadata and standards that were already established by Sifsen and also a template that was made by Dr. Lynn Bright for antimicrobial resistance data that we also adapted for our data. All this, again, is pushed out to NCBI. On the

left-hand side, you see a list of our collaborations with clinical labs and repositories where we currently receive our samples from.

Here is the sequencing strategy that we are using for creating the reference data. One is Illumina (inaudible) sequencing, and then on the other hand we also perform Pacific Biosciences long-insert sequencing for all these samples. Then for assembly, we are doing various assembly algorithms, and then there will be a bioinformatics expert that will be sitting there and deciding on which one gets pushed into NCBI. But we will have all these assemblies available later on. Also what they're doing is they're applying (inaudible) and GCQ4QC, and then they're using an automated pipeline for annotation, and they also give us base modification detection data that will be most likely uploaded as well to a database.

Here are just some very preliminary sequencing and assembly steps for the first batch. Here on one side we had samples from Rockefeller

University that was, sort of, a uniform set of (inaudible) samples. Here you can see that already 23 out of 50 on single-contact status, but they're still churning for the data and more might qualify later on. Then on the other side you see a more diverse sample set from Children's National Hospital and Sifsen where you have now 18 genera represented. Here currently we have 12 of the 41 in single-contact status, but again, they're still working on it and as time goes on more might qualify to be in single-contact status. That would be our goal for these.

Here's a quick snapshot on how this data will be represented at NCBI, thanks to Bud Klinke for putting this data really together. The ID for this is 231221 if you want to check it out at NCBI. On the left-hand side, you have BioSample database sort of, like, an example of how this data will be represented. There are two samples here that are shown: One is from the Children's National Hospital and one is from Rockefeller University. On the right-hand side you see the

sequencing archive database, and this is where (inaudible) will be available. For BioSample, for example, if you go into the Children's National Hospital sample, there you see all the fields for the metadata that would be required in order to qualify for our database. Also, if we have the antimicrobial resistance data available, they'll be put in by a template that we adopted from Dr. Lynn Bright.

Then all the assemblies that get generated get put into the assembly database at NCBI, and here's an example. For example, IGS looked at one assembly, and they used the asymmetry in GCQ which then suggested a misassembly, and then various assemblies were selected previously. This really showed that they should use the HGAP 2 assembly to push forward to NCBI.

Lastly, I just want to briefly touch on the microbial reference databases. Justin had talked about human reference samples, but they're also, at the same time, working on microbial

reference materials that got funded by FDA. There are four (inaudible) rated bacterial standards that they're working on, and they're growing large batches of these as you can see from the table.

Then on the right-hand side you see where you can these extracted DNA for testing, library preparation, sequencing, mapping, (inaudible) calling, and confidence estimates. Then these were also generated with multiple platforms. The de novo was made with Pacific Biosciences, and then also platforms like Illumina, INTRON, and Ingen technologies were used for validation.

So, just wanted to go very briefly (inaudible) that in order to realize NGS as a regulated device for microbial diagnostics. It is necessary to ensure that the (inaudible) information used for interpretation of results is of suitable quality and that it should include appropriate metadata. Efforts are currently underway. There will be a difference in validation strategy for targeted versus

metagenomic as I outlined previously, and we will enable a streamlined approach for regulatory evaluation as was highlighted by many, many commercial developers and clinical end-users.

The Division of Microbiology devises that FDA will continue with the current efforts to augment existing sequence information in the public domain. There are many people to thank for it. As you can see on the slide, I want to specially thank the FDA Micro team and the collaborators and everybody involved. I'll take questions.

DR. TEZAK: I think you have time for maybe a couple of short questions. I want to ask the panelists if they can come up here while we're answering questions. If there are no questions we can start the panel.

SPEAKER: If the red light is on that means the microphone is working, and if you guys could pass around the other phone so people can hear us. Whenever you need the slides to be changed let me know.

SPEAKER: We're excited to have a few representatives from a variety of the sequencing companies here to talk some about where they see their technologies and sequencing technology in general going in the future. I think I'll let each of them introduce themselves, and maybe we'll go from Mya down that way. First if you could introduce yourself, where you're from, and then just describe in a couple of minutes your technology, your company, and where you see your company going next with either the current technology that you have or with new technologies that you might develop.

MS. THOMAE: I'm Mya Thomae. I'm Vice President of Regulatory for Illumina, also was founder of MyRAQA which was purchased by Illumina back in July. I think we've talked a lot about Illumina Technology today, so think everybody understands that. I think Illumina's been pretty public that we're really going through a clinical transformation process at Illumina and really trying to bring NGS into the clinic.

I think we're just getting to start with what we've done so far. The CF Assays are just what we're starting with. I think there's massive potential in all parts of the clinic, and we've already announced some collaborations with PhRMA to try and do some cancer-panel- type of work in competing diagnostics as well.

I think the sky's the limit, which is why I'm here, so think there's going to be a lot of interesting things in the near future.

MR. ROSENFELD: Hi, I'm David Rosenfeld. I work for Complete Genomics where I manage their data analysis pipeline. Complete Genomics is different; we're not an instrument vendor. We're a genome-sequencing service, and we only do whole human genomes. That allows us to focus our efforts in various ways. We have our own proprietary and, sort of, novel strategy for doing that. We built our own instruments and we use our own sequencing chemistry and our own assembly algorithms.

About 2 years ago we were purchased by

BGI, so for the last 2 years we've actually been developing technology for them. They have a lot of our sequencing instruments in China now, and they've actually used it to develop a noninvasive (inaudible) assay that's been cleared by the CFDA in China.

I can't speak to a lot of future technology that we're working on. The one thing that I can discuss is things that we've announced already, and that involves our LFR technology. That's something we call Long Fragment Read. It's a way of basically separating long fragments of DNA early in the process, tagging the shorter fragments that come from that, and using that tag information through the assembly process to get much higher accuracy in our short variant calls. We can basically count how many similar tags show up in the assembly and filter using that. That process also allows us to phase the genome, and it allows us to assemble from a very, very small number of cells, so it's a pretty exciting technology. It has difficulties in

commercialization that we're working on, but anyway that's what I have to say.

DR. SHAW: Thanks. I'm Jay Shaw. I'm a Senior Director at Thermo Fisher. It was discussed earlier; we just announced registering the PGM Dx platform with the FDA. Similar to what Illumina has done, we've gone with a genotyping claim for isolating old genomic DNA from whole blood. Plan to move into content. We've also announced a collaboration with PhRMA for cancer diagnostics, which will bring us into the arena of detecting somatic mutation.

In terms of where we want to move in the future, it's still a fairly complex technology. We view automation as being really important; simplifying the process for the end user and just making it so that we can push it out further and further into less-complex labs so that it can be used there.

DR. KORLACH: Thank you. I'm Jonas Korlach. I'm the Chief Scientific Officer at Pacific Biosciences, and I was told that we would

be able to show a few slides, and I took advantage of that offer.

At Pacific Biosciences we've taken a little different approach in terms of the sequencing performance in that, as you know, there has been a fairly strong focus on the throughput of the sequencing systems. At Pacific Biosciences, we wanted to build a technology first and foremost that provides the highest quality of data in all the four areas that are relevant when evaluating really any sequencing technology, and you see them up there.

As you may know, we have, by far, the longest reads and the least bias, which gives you the opportunity to sequence the entire genome of microorganisms or also the human genome. Then as was mentioned in some other talks, we have the unique capability of looking at the chemical changes in the DNA, the epigenetic methylation, and other changes as part of the sequencing project.

In the next slide you see the

development by us and others that now make it feasible in an automated fashion to generate finished, microbial genomes and (inaudible) outlined that with the University of Maryland for comprehensively describing and characterizing pathogens. This has now been used by the community.

The next slide just shows a partial list of publications that have been described that use that capability to now obtain in a very efficient, automated workflow and also cost efficient to get comprehensive information about the pathogen of interest. I just want to show one recent example in the last slide which was a paper that came out last week in Science Translational Medicine by Julie Segre from across the road here at the NIH Clinical Center using this technique to understand the horizontal gene transfer, the plasma trafficking, that happens in the multidrug-resistant enterobacteriaceae like klebsiella that caused these multi-drug resistant hospital outbreaks.

This figure is from a commentary in science showing that you really -- because these events are complex and it's now realized that in microbiology, hospital- acquired infections, food safety, and in the human genome the variation is not just limited to snips. There's large structure variations. The acquisition of new plasmids or rearrangements of plasmids play crucial roles in understanding whether a particular salmonella strain will become an outbreak strain or not. With the technology, it's now possible to understand that more comprehensively, and that's being utilized not only in terms of hospital-acquired infections.

I talked to Michael Art; they're using their two Pacific Biosciences instruments to generate finished genomes to have reference strains for the major foodborne outbreaks, and that has two advantages. First, you may actually understand better what's going on and what makes that particular strain an outbreak strain. Secondly, having this complete genetic and

epigenetic information may give you more sensitive and specific markers now to track the evolution of the outbreak and to track the source in conjunction with the other technologies.

With regards to where we're going, we are for researchers only right now. As you may know, in October of last year we announced a strategic partnership with Roche for leveraging the power of this technology into the diagnostics market. Now, as you know, that take a few years, but we are working with them very closely and ultimately getting this technology into the diagnostics base as well.

DR. ZOOK: I have one or two more questions here that I'll just let whoever wants to talk about the questions answer. If you in the audience could be thinking about any questions that you have also, we'll open it up for questions for any of you also.

The next question that I had was what types of sequencing you think will be done clinically, say, 5 years from now or so? Do you

think it will be whole genome versus targeted sequencing? Do you think it will be just human or will there be a lot of other microbial sequencing? Will it be single molecule or amplified or in highly specialized labs or in more (inaudible) pipelines or even maybe bedside sequencing? Just interested in your thoughts as to where it might be moving in the future.

DR. SHAW: I think from a technology standpoint 5 years is a long way away, but 5 years for a regulated device is a short period. As a person in development, I always appreciate how long it takes to do things. I don't think we'll see this moving, in my opinion, that fast to getting to the bedside in terms of 5 years.

The problem is you start the development process, and you have to lock down. As the company continues to make improvements, they're always asking, 'Can you roll that in? Can you roll that in?' The problem is you have to start the process all over again. You have to make the commitment early on, and it takes a while to get

this in. Within the next 5 years, my opinion is I don't think we'll see bedside sequencing going on.

MR. ROSENFELD: That was, kind of, a complicated question maybe. From our perspective, we're all about scale and doing things at a big scale. We certainly expect the clinical markets to be much bigger than the research market was. We think whole genome sequencing has a place there. It's not the only thing that's going to be there, and certainly not in 5 years. It won't be the only thing that's going to be there, but we think it's got a place there, especially around cancer.

But a lot of it is going to be what brilliant breakthroughs come through in the assays that get developed that really drive a market because that's where the investments going to be put into on our part and our competitors' parts. That's where you'll see the really rapid progress.

MS. THOMAE: I would agree with Jay that I think bedside sequencing in the next 5 years -- we haven't even gotten PCR to the bedside yet.

But I don't know, Zivana, if you're going to cut us a break on this one. I'm just kidding. (Laughter) I think that's probably ambitious in the next 5 years, but I do hope it's before I retire.

I do think there's a lot of work in human, but I think the microbiology applications to this are very interesting as well. I really appreciate the previous presentation on it, and I think there's a lot to be learned.

DR. KORLACH: Yeah, I agree with what was said, but I want to maybe add that I think there will be a gap in the research arena in what we will be able to do.

I'm going to make a prediction -- it may not be true -- but I think I would expect that 5 years from now we will be able to generate de novo genomes for not only microbes but also the host, the human genomes. I think we'll eventually free us from the reference and from a reference-based research --I'm talking about the research setting -- and would be able to de novo assemble human

genomes in an efficient and fast manner 5 years from now.

That relates to then a challenge both with respect to the regulatory environment -- how do you deal with that? How do you deal with the structural variation with everyone being a true individual with respect to their genetic blueprint? Then also bioinformatics challenge because right now we don't really have the infrastructure and the methods to interpret that data and take that information from literally hundreds of thousands of individuals for which this information, both in terms of the host and then the pathogen in a metagenomic setting, will be created.

I think it's going to be very exciting. There's a lot of work to do in all of those areas. All the sequencing technologies are advancing extremely rapidly, so I think we should be thinking ahead and stimulating all the young bioinformaticians to be aware of that and get excited about that.

MR. ROSENFELD: If I could add to that, that LFR process that we have actually does allow for de novo sequencing of the human genome, and it creates all kinds of interesting problems for data formats and things like that where all these resequencing standards that we have work a certain way, but when you're outside of that and your first step isn't mapped to the reference, all that stuff doesn't make so much sense anymore.

DR. ZOOK: The last question I had was much shorter. What challenges do you see with getting the new technologies into the clinic in the near future?

MS. THOMAE: I'm a regulatory person, so I usually see the regulatory challenges first. I think there's the issue of how do you find the right validation strategies? Are you trying to look at everything you might find or are you trying to look at representative sorts of things? What are your reference standards? This can become a pretty complicated conversation.

Then I think there's a lot of

interesting challenges, and I know FDA is thinking about this too. When you start to pour it over content, for example in competing diagnostics, that's already been approved for other PMAs, and you want to maybe put that content on your cancer panel, sort of, how do you do that? What's the best way to do that? I think the regulatory challenges are pretty interesting, but I think there's a lot of good thinking being done about it right now.

Then, of course, we can't always be U.S.-centric. There's a lot of interesting challenges around the world with the different kinds of testing that's done. Some of it, like the prenatal testing, has various different cultural impacts. Not everybody maybe wants to find out everything that Americans do. I think there's a lot of very interesting questions to be handled as this moves to the clinic.

MR. ROSENFELD: I think almost everything is a challenge to get into the clinic. It starts with quality, that's a challenge. I'm

not sure any of us are good enough yet. Then the regulatory stuff, and then there's just a huge challenge with acceptance by doctors and having them be educated to understand what all this stuff does.

We're really way ahead of a lot of things in terms of technology and trying to push it into the clinic when the clinic isn't asking for it, so it's pretty hard sometimes.

DR. SHAW: I also think as we move forward with these large, complex panels, it's hard enough to develop a single analyte test. Now we have to factor in for all of the analytes that we have to deal with, so sample availability, performance across all of these analytes is going to be, I think, the real challenge in terms of generating the body of data you need to prove the performance of these systems.

DR. KORLACH: I fully agree. I'm a method developer; I'm a little more removed, but Eric Schadt, who is at Mount Sinai and just doing clinical sequencing there with the different

technologies, he had a Mendelspod podcast this week, and he was asked the same question. He highlighted the need for better education and openness of the clinicians. We need to train the existing ones and bring up the young ones who are having a much more open mind with respect to these kinds of datasets and engaging in the opportunity to use those for making clinical decisions.

I'm personally interested in maybe, for this forum, two philosophical questions in how this concept of personalized medicine and knowing something about someone in a very detailed fashion is going to work out with, on the other hand, having lots of cohorts and trials and medicine that works for lots of people and to save lots of people and so forth. To me, that's an interesting paradigm consideration on how we make that work where eventually we will have the information that we know for a certain person that drug works, but how do you prove that? How do you validate that, and how do you get that into an environment that makes the public feel safe and feel good about the

process?

DR. ZOOK: Thanks, everyone. Are there any questions from the audience? There are microphones around.

SPEAKER: They're focused on the bacteria, so are any of these methods applicable, sort of, the soup of blood? You don't have to isolate first? I want a follow-up question if they're -- it wasn't clear because that would be great, and there are some tests that now with DNA testing.

MR. ROSENFELD: Like with our LFR test that works from a small number of cells, you do have to do a cell selection out of the blood, but you're --

SPEAKER: You have to isolate the bacteria?

MR. ROSENFELD: You're isolating either a circulating tumor cell or (inaudible) DNA (inaudible) way it works.

SPEAKER: I was talking about the bacteria groups here and whether they can attack

just a raw sample and figure out what it is.

DR. KORLACH: Right now with Pacific Biosciences I think typically it's done with isolating the bacteria. I know of a few studies where they have taken just the whole sample, and then you sequence everything for identification (inaudible) what's there, and I know that's also been true for the other technologies. Joe Direesi gave a very nice talk in February using the Illumina system to sequence a patient, taking just the blood and sequence everything and then find that there was a certain pathogen in there.

I think that will likely become more and more common that you don't need to do these manipulations, and you can jus take the DNA of the entire mixed sample, sequence everything, and then bioinformatically filter out what you're interested in.

SPEAKER: The follow-up question is there's a lot of tests currently for probes mostly, probe-based or probe PCR, for identifying any of a sequence of bacteria or whether they have

the various resistance genes. Will these methods replace those? Are there advantages or disadvantages assuming you isolate?

DR. KORLACH: Yeah, from my perspective, my prediction is that these methods will be replaced eventually as it becomes more standard and cheaper and a, sort of, automated workflow because in addition to what you think you already know is there with the PCR method, you need to know what those -- and the primary sites have to be there and the genome and so forth. The sequencing gives you de novo information and gives you information about something that you didn't know was there. It's a hypothesis-free approach, which I think inherently is more powerful.

Right now, I believe there still is a hybrid approach. When you have a de novo reference you can design specific primers to a certain outbreak that makes the tracking of that more efficient. Ultimately, I think it'll just go to sequencing.

SPEAKER: Do you see a role for, let's

say, paleobiology where you're looking at, for instance, mummies or other human tissues for human diseases or pathogens? You start out with small, short sequences, so you're halfway there, but they're from complex genomes.

DR. KORLACH: Maybe I'll start. We've had a couple papers where that's been used. Now obviously you don't need the long (inaudible) from Pacific Biosciences, but the fact that you don't have to amplify the DNA helps there. There have been some studies there also looking uniquely at the base modifications in these ancient samples. From our point of view, that's been demonstrated in the research setting and is feasible.

MR. ROSENFELD: I'm pretty sure no one's used our technology for that. (Laughter)

DR. ZOOK: As far as you know.

MR. ROSENFELD: I think I would have noticed. I do the QC on the genomes. I think I would have noticed.

SPEAKER: Thanks to the efforts of all of your companies, we have technology now that can

sequence willy- nilly. We all can talk about ways to validate and do it medically responsibly. Indeed, we can sequence far better than we can clinically interpret. Most variants that aren't obvious, common polymorphisms that you find in a patient are VUSs.

Even worse, of the stuff we can clinically interpret, we can diagnose far better than we can treat. We can definitively say, we know what's wrong with you. I'm so sorry, there's nothing we know to do about it today.

My question is the fact that information about clinical utility at variations that are discovered during sequencing, both sematic and constitutional, is changing very, very rapidly, what's the best way to incorporate that into clinical practice but to make sure it's done responsibly? Let me start with Mya just because I like to put you on the spot.

MS. THOMAE: Great question. I hadn't prepared for that one. I think some of it's about being willing to go back to the data that has been

generated for a patient as we understand more. That gets tricky in terms of saving large gigabytes of data, but I think you have to be willing to figure out how to report things and potentially be bucketing them appropriately at the time the test is done but then be willing to go back and relook at it.

For example, there's a number of consortiums working on this in cancer and trying to bucket things by known clinical relevance, maybe clinical relevance, and we can find this, but we don't know what it means. There is the Actionable Genome Consortium, something Rick Klausner's been working on, that's really trying to do that for cancer, and I think that's probably needed all over the place; in HLA and everything else.

But I think it is a process of revisiting it periodically. How that happens in the clinic, how our payer system works to be able to support something like that, how we make sure physicians get this kind of information, and not

just at the large institutions but across the board; those are questions that are pretty tricky. They're definitely not FDA or regulatory driven. I think they're more in how healthcare is delivered.

I'm working on world peace, and I'm working on fixing our healthcare system in the United States at the same time. I've got a couple PowerPoint presentations that I'd love if you could work with me on those. It's pretty big issues.

SPEAKER: Which is harder: World peace or fixing healthcare? (Laughter)

MS. THOMAE: Fixing healthcare.

MR. ROSENFELD: I'm not an expert on this, but it seems like different technologies are going to move at different rates all the time. If we want the understanding of what all the polymorphisms mean, to speed up, it means you've got to do bigger studies and you've got to, sort of, free all that data that's already been generated.

Right now it's all in little islands of individual studies, and there's no way for people to share it. That was discussed yesterday as part of the cancer discussion and NCI discussion. It's that way for almost every dataset that's out there. I know we've produced almost 20,000 genomes for our customers, and there's not way that can all be put together to do a study. That's going to continue to be the case unless somehow there's some kind of framework that's put together to fix that. That would be, to my mind, a big first step in trying to move that analysis along.

DR. KORLACH: I fully agree, and just a small note to add. What I'm impressed by and what I take heart in is the work that's been done at places like Mount Sinai, at the Brode where people are looking at this in new ways, a very integrative approach, lots and lots of data, developing new math and new statistics to treat those kinds of data. I think that's the first step.

We don't really know about the genomic variation at all-size scales in the human population, and we need to find that out first by generating a bunch of data. Then it's going to take a couple of years to really be able to understand and treat it appropriately to then make informed decisions for the clinic.

DR. ZOOK: We could maybe take one last question depending on how long it takes.

SPEAKER: This question is somewhat related to the last one. Hopefully it's a simpler answer. We've talked about personal genomes; there's a 100,000 genome project in the U.K. Given regulatory issues and reimbursement issues and diagnostic issues, we can envision the future where everyone is walking around with their personal genome. Instead of doing a lab test, a physician orders a bioinformatics test. I get consent to look at a particular region and then get the answer instead. Of doing all the informatics up front, I do it as I need it.

In that future where everyone has their

own personal genome in their medical record, how long do you think it'll take to get that? Will it happen in my lifetime? Will I have it at some point? Will it happen in my kid's lifetime? When do we start planning our organizations around not doing lab tests but doing bioinformatics tests on a genome?

DR. KORLACH: My prediction is that you're going to have that in your lifetime, and you're going to have to do that --

SPEAKER: Ten years, five years?

DR. KORLACH: Something like that; between 5 and years. I think you're going to have to do it multiple times because when you develop a tumor then you want to sequence that tumor too. Then in terms of gene expression with the different tissues, there's going to be changes there. But 10 years, I'd be surprised if it'd take longer than that.

It's going to be expensive and maybe not reimbursed initially, but I know in Japan some of our users there are already approaching a certain

customer base that are perfectly happy to spend a few $10,000 to get that kind of (inaudible).

SPEAKER: I guess I'm looking for when is it becoming the norm that you assume that everyone has their personal genome?

DR. KORLACH: I can't answer that. (Laughter)

SPEAKER: Yeah.

DR. KORLACH: (inaudible) because that goes beyond the scientific feasibility.

SPEAKER: Right, right.

DR. KORLACH: There's a whole bunch of other things that needs to happen for that.

MR. ROSENFELD: For sure less than 10 years scientifically; there's no question. You can almost do it now if you wanted to if you're willing to pay for it. I'm not sure if it'll ever be the norm in the United States. If you look at market research studies of that kind of thing, most people don't actually want to know those things. There's actually only a small population of people who are interested in getting their

genome sequenced, so I don't think it will be the norm.

SPEAKER: We get the data so that we can do the test later cheaper, so we're not actually getting any information other than Gs and Ts, but then that's something to mine and interrogate when you want to know it. That would be the model.

MR. ROSENFELD: I understand. People are nervous about that. It's a question of trust of what's going to be done with those Gs and Ts.

DR. KORLACH: The culture could change over the next 10 years possibly.

DR. SHAW: I think the point you just made about the difference, the somatic mutation, detecting it later when there's a disease state means multiple times in your life potentially as opposed to an early sequence that'll carry you through your lifetime and the need for that.

DR. ZOOK: Okay, let's close by thanking all the speakers and the panelists in the second -- (Applause)

MS. VOSKANIAN-KORDI: At this point,

we're going to take about a 15-minute break, and we'll reconvene at 3:15 for the food safety and pathogen protection section.

(Recess)

MR. PETTENGILL: So I guess we'll get started. Thanks everyone for coming. I guess this is the final session of the two day meeting. This one's on food safety and pathogen detection. I'm James Pettengill and then Heike Sichtig will also, will be coordinating this. I think we have three talks that will focus on food safety and pathogen detection, from, one from the FDA, Center for Food Safety and Applied Nutrition, one from NCBI, dealing with some of the database issues and performing whole genome sequencing, and then a final talk from somebody from FSIS.

I think it's sort of safe to say, at least from the FDA SIPSANS perspective that next generation sequencing has sort of really revolutionized the food safety and pathogen detection. I mean, I think we'll be hearing a lot about that today from the first speaker, and then

I think another big part of it is to bring the whole genome sequencing and food safety into practice and in production, it's been a sort of multi-agency collaboration. I think we'll hear a little bit about that today also. So with that I guess we'll get started.

Our first speaker today is Mark Allard from the FDA Center for Food Safety and Applied Nutrition. He'll be talking about FDA Genome Tracker surveillance and source tracking for Salmonella Listeria. So we're just zooming down into a bacterial genome here. We'll get to the base level here soon enough. (laughter)

SPEAKER: I just want to say as an introduction to my talk, I hope all of the food and genetics folks (inaudible), I guess what I want to say is what we're doing to bacteria, will be the future for (inaudible) and for food the future is here and now. And so what we'll be talking about microbial trace backs (inaudible)

MR. ALLARD: Thank you and I'll proceed. Essentially, to bring next generation into the

regulatory lab, and I represent the research end of the Center for Food and Safety and Applied Nutrition, which essentially supports our regulator state and federal laboratories, we did timing testing to say, will next generation lab response be at least as fast as traditional typing procedures, which is PFGE, and of course the answer is yes. The current technology is relatively rapid; in three or four days to a week you can have an answer. This may not be the clinical level, at fast enough, but in food safety and trace back of outbreaks, this is a much great improvement over the traditional methods. Essentially what we're leveraging with whole genome sequencing of bacteria, is old ideas in evolution that dated all the way back to Darwin and that is that the principle of inheritance will betray the original birth place of organisms. And we've known this from the beaks of finches to the traditional museum curation and taxonomy and the genome is just additional characteristics that will trace organisms right back to the source.

And so for FDA, it's critically important to understand where an outbreak occurred, where the contamination is arising in the food supply and to be able to remove it from the food supply. And so in 2012, we had sort of a watershed event for FDA in an example of a sushi outbreak. And this is the traditional view, is you see a large cluster of clinicals all of identical genotype. This would be done based on PFGE, but you don't know where the food is from and so the trouble with -- the difficulty with the PFGE is for some bacteria like salmonella, they are so clonal that they all share the same PFG type. So it doesn't really help you very much with source tracking, because your last ten years of isolates could all have the same pattern.

But a genomics provides additional resolution and in essentially actionable information in helping us identify where the food came from. And so, and then when you actually find the food and sequence it, in this case it was a single point source outbreak, the bacteria from

the food is directly imbedded within the clinical samples. And then when you actually do an inspection, all the environmental swabs will also sit in the middle of that.

But what's particularly important with this example was we wanted to say, well, are there any investigational leads that can be provided by just going into our freezers, diving into them and sequencing all the Salmonella Bareilly that we had in the freezer, or Bareilly that was of the same PFG type. And we got about a dozen isolates. Six of them were PFG identical, and if you look at the link just above the outbreak isolates, there's a sample from India from shrimp from an inspection five years earlier. And genetically, it's about 25 snips away, so it's not the source. The outbreak cluster's only two to five snips away. But it is geographically a lead. And the PFGE supported that it's somewhere in Southeast Asia but it could be the gulf or even potentially the Pacific. Okay, but genomically, it says we're on the west coast of India, and in fact, if you do

overlay this on a geographic google earth map, so you're laying the tree on top of google earth. This is what an outbreak looks like along the eastern seaboard and this is a picture of where the contaminant is coming from. And in this case, the Salmonella Bareilly from five years previous had occurred from eight kilometers away from the actual produce packing house, where the contamination occurred. The reason we did such a good job in geographically source tracking this isolate was it was a repeat offender.

Now we don't know if the isolate is reoccurring contaminants in the wild from that region and the fish that come in occasionally carry that, or whether this is a contamination within the fish processing packing house that's in the equipment that is sporadically coming out. But essentially this gave us the understanding of how powerful the source tracking could be with doing whole genomes and comparative clustering. And so the idea was, well we should sequence everything in our freezer or at least the last

five years of outbreaks, put them in a database and use them essentially to provide leads of where to send inspectors. And the technology is driven by the drop in cost of equipment and so now it's feasible to sequence hundreds if not thousands of bacterial isolates to draft sequence quality levels.

To build at the national database, we need three components. The first are labs to generate the whole genome sequence, and we're interested in the distributed network because that's where the isolates are coming in, a local, state department of health, a local clinic or a federal inspection facility. We need network management. And then we need sequencing storage data provider and analysis and essentially, we've off, put off our data storage and partnered deeply with NCBI and so they're taking over all those costs, data storage. And we've been following them in recommendations on how to move data efficiently up to the NCBI. And Bill Klinke (phonetic 010:12.4) will be talking about those --

that partnership and the part that the NCBI does.

But essentially, by shifting the cost over to the MIH NCBI, we have more funds to provide to more state and federal laboratories, to actually put the sequencers and reagents into a distributed network. Though the basic model is to have national and international distributed networks of sequence uploaded to a common place, this is the NDC which is the NCBI, EMBL and DDBJ database, and we call this database Genome Tracker. And then all the sequence, all the draft sequences that I'm talking to you about are publicly available in real time, uploaded, and that means that our federal partners can get access to it, our state and foreign public health laboratories and of course academics and industry. The data's transparent. All of the data's available.

And so our initial network -- the network as it exists today is six state labs and eleven federal laboratories. But his is rapidly expanding as state networks; state departments of

health purchase their sequencers and are ready to start uploading to the database. And we're in actually the second year of a pilot study. And the pilot is to build a reference database to source track unknown clinicals against the known food and environmentals that are sitting in the database.

And you can go to this genome tracker database. It's just a bio project at NCBI. I think by the end of the year, right now, we have 10,000 preregistered strains and 6,000 genomes. This slide's already out of date. I think it's 7,500 Salmonella and almost another 1,000 Listeria. So we're going to top out of 1,000 isolates of mostly unpublished draft genomes, uploaded in real time, available to anyone in the world.

We are also providing, and this is, not we, that's the royal we -- it's Bill and his colleagues at NCBI -- they have an ftp database where they do initial draft analysis and upload a tree that could be visualized in their software

called Genome Workbench, also free. And so we have existing databases now of e-Coli, Listeria and Salmonella that are actively growing and people are collecting it, state and federal partners. And essentially, I'm just showing you an example where you can see the new tree, and I think for some of these databases we get a new tree every day that we're averaging five to six hundred new isolates every month and the trees are enormous, you know, when you're talking about thousands. It's hard to actually see the data in a coherent way. And we're working with NCBI and other third party software providers to give us new tools to rapidly say, is there a new cluster? Because what we're really interested in at FDA is, is there a growing cluster? Is it new? Is there a new association within a food and environmental sample? But frankly, the CDC has been using this and just understanding which clinicals are associated, because usually the traditional PFG would say ten are in a cluster. They're all identical PFG. But really there might be two or

three clusters there, of the independent sources of contamination arising at the same time. And so over time you can look for this. Like I say, we're working on tools. And if you dig into the tree, you can get the associated metadata, which says that it's clinical or environmental, it's associated with peanut butter, cantaloupes. And so we're using this data. So, like all of the genomics that you've been hearing about, as the cost drops and the applications increase, the rapid rate of increase of genomes is rapidly increasing and so the Salmonella and Listeria, which are what we're primarily emphasizing, are also rapidly increasing.

I mentioned the CDC, FDA, NCBI and USDA, Listeria real time sequencing project. This was a project whereby the CDC decided to get involved last year and they said for their first project, they said for all Listeria, we'll sequence all Listeria, clinical cases in the U.S. this year, and FDA and USDA agreed to do all the food in the environmental, as well as dig into our freezers

and sequence and upload them into the database. And essentially, paired with this, the CDC already has programs where they're actively doing the traditional epidemiological questionnaires of every patient, okay? So you have good epi paired, at least for the patients that answer the phone and will talk to a regulator, to add this. And so this is real time surveillance, but we've also been doing real time surveillance of some Salmonella. So for example, the states wanted to look very closely at another very homogenous Salmonella genome, the Salmonella Enteritidis, which is what's in chickens and eggs.

So in Listeria, it was a watershed year, 2014 was a watershed year, combining inspection in clinical and food isolates, the FDA closed Bruce Cheese based on genomic information. This wasn't epidemiological information that was supported by genomic information, or used at a later date or as a case study a year later. This was in real time, use of genomics to show the close relationship with the foods that were coming out of this

company and the clinical cases of the people that were getting sick. And this was significant in that at least one person died and several babies were boarded and lost. Listeria is a killer for the very old and the very young. And on this is so, actively we've done cases and now, once you've done a successful case for the government, they don't ask you to stop, they give you more and more and more cases to start doing this on a regular basis. And so we're now in a zone where we're looking at a new case about every month and continuing to essentially integrate. There's a lot of integration. This is a new tool and the epi's need to know who pulls the trigger. When do we say we should start sequencing? There's a whole sort of coordination that's directly affecting regulatory pieces of FDA.

I was going to, as long as this is a standardization meeting, I want to talk just a little bit about validation, though this is usually Errol's job, to do this part. We've done a lot of validation. Essentially, if you're going

to use this as a regulatory tool, then you're going to close down companies or put people in jail, you have to have a validated approach that it's reproducible and you get the same answer. And so we've done technical performances on standards. So we work very closely with Justin Zuck and PICA in helping build the reference material and sequencing reference material on multiple platforms. We've also done several intra-laboratory variation sequencing studies with hundreds of runs as well as typical intra-laboratory variation and then also testing the actual bioinformatics pipeline because the pipeline that brings you to the interpretation also has to be validated, reproducible and accurate, and so you also want to have good epi to back it up to say you're getting the right answer, or, if you're sequencing the same isolate, you know what the answer is.

And what this shows is that essentially there is variation from laboratory to laboratory and people can have a bad day, but in general the

sequences either work or they don't work. It's pretty obvious when a sequencing run fails. And you can see this in measurements like the N50 or the number of contigs, or actually even just the number of reads in the instrument. It's relatively straight forward. When you do cost comparisons, and you just count up the number of snips that are different in what should be identical runs, you rarely see more than one or two differences and this is in the same range as within a single point source outbreak, where we see say 1 to 5 snips in the intra-lab comparisons, and ten snips even lower.

And then, I don't have a picture but essentially when you load it into a file of genetic tree, that's the other and most important piece is, would you ever make a different regulatory decision, to say this is part of the outbreak or not part of the outbreak? And we never see these cases. And so we're working to publish that, those results, because this kind of key validation, but I like to say, all the data is

up in the Genome Tracker database, or much of it is. And so people can look at this data and carefully study it themselves or test their own validated pipelines.

So the submission test, NCBI also has to be coordinated, integrated and standardized across the network and so we're working carefully on what kind of information and metadata goes up with the sequence data, the level of coverage, and the basic QC that's done. Here's another example where a QC score, where you can see, it comes right out of the top where the number of reads is much smaller and you can instantly see that there was some problem with this particular run. The fact is is that these robots are very robust and they don't make very many mistakes. The biggest mistakes are always human error. It's mishandling, mislabeling, cross contamination, things that the bacterial community has long known are issues. And so you have to do careful bacteriology and microbiology to control that.

There has also been a large an ongoing

discussion about what's the minimum metadata. We're releasing this to the public. And so there's less data than we can -- as a regulatory agency we know much more detail about an inspection than the sampling. And that's behind the fire wall of FDA and we would share that with a legitimate regulatory body, like a state department of health or a foreign body. But it's not made publicly available, and so it's all linked to a key. And I'm not going to give you examples. Go to Genome Tracker and just randomly pull out samples. There are many needs in this. We need widely available commercial solutions. We need customized solutions. We need more automated methods for microbial, anti-microbial resistance, virulence, and we need very -- we're working closely on easy forms to interpret physicians, epidemiologists and researchers working directly with our epidemiologists at FDA and the regulatory bodies in the food safety to directly support this. Thank you.

(applause) Questions?

SPEAKER: Thanks, it's very interesting, impactful work. Thanks for the talk. Just wondering about the (inaudible) that follow genetics. What kind of techniques are advocated -- is it phasean, maximum likelihood or (inaudible)

MR. ALLARD: So we've used many methods and that's a slide I didn't leave in and so, but the current FDA procedure, if you do a search at gethub for SIPSAN you can see our whole published procedure where it's transparent. The data analysis is full transparent and described with open source software. The fact is is that what FDA is interested in is the very tip of the tree, because you were saying, is this clinical and another clinical match and does it match a food, and so we don't really care if the lower parts of the tree are correct or not. It's the very tip is where the regulatory decision is. And because these are such recent events, almost any phylogenetic method will work.

SPEAKER: Okay.

MR. ALLARD: There are very low levels of homoplasy.

SPEAKER: Okay.

MR. PETTENGILL: No questions from online.

MR. ALLARD: I just want to say with my -- oh go ahead.

SPEAKER: Sorry, quick question. Have you considered including academic partners in the tracker network?

MR. ALLARD: We already have academic partners in the network. Early on, essentially, most departments of health don't have bioinformatic excellence; don't have any genetic or genomic experience. So for example, our New York lab in Wadsworth in Albany works closely with Cornell food safety and in our site in Florida, works closely with University of Florida in Gainesville. And so, but usually we're working with, directly with state partners because we're trying to support and improve this for regulatory action and in department of health labs. But yes,

not only that, anyone in the network, we've asked them to be regional hubs and reach out, because isolates could come from a microbial ecology, or a vet path laboratory at a veterinary school, or other food safety groups that are studying what's in the -- what bacteria are in the tomato field or the lettuce field. And we're looking for industry involvement. You know, the idea is food industry could learn more from this kind of database, the same way that the hospital acquired infections folks are saying, how is the pathogen moving through the hospital? Food safety and industry should want to know the same thing. How do we get this bacteria out of our equipment and what's imbedded reoccurring? And then the last way that we think industry will use this was, it's really a litigious issue. They don't want to get stuck with the hot potato and be blamed for the recall. They'll look upstream and say, what products came in that brought contamination into their product.

MR. PETTENGILL: So our next speaker is Bill Klimke from the NCBI and he'll be talking

about lessons learned from a real time Listeria project.

MR. KLIMKE: So this is Marc's sized. Can we get this more my size? It doesn't grow? It doesn't --

SPEAKER: You can sit down if you like.

MR. KLIMKE: We have ones that you know, push buttons and they move up and down. This is a new building. They should, they should all be robotocized already. So the good news is Marc gave my talk already, so I'll be happy to take any questions.

SPEAKER: Sorry about that.

MR. KLIMKE: He's stealing my trees. So Marc gave an actually a pretty excellent introduction to our involvement with this. So I'm going to have to, I'm going to be able to skip past a lot of this. Just to note, we are really early on the technology curve. I think we're in the early adopter stage. And I know a lot of people say that they want standards, but I think standards is a good thing to discuss but if you

expect standards to drop out of the sky into your lap, it's going to take a lot of work. So I'll come to some of the issues later on.

So when FDA approached us, about this project, they mentioned that they have to do a lot of food safety. There's an enormous amount of food and the obvious solution to the food safety issue that we all stop eating is somehow politically untenable. So they're going to do a lot of food inspections and there's increasing amount of food being imported into the country. And as Marc pointed out, the basic idea of subtyping these molecular methods is, can you do a match between what is making a person sick and what is in the food source, whether it's the spinach that you bought at the supermarket yesterday or is it something that you bought or ate in the cafeteria this morning.

And current methods don't have the resolution that is necessary for all isolates. PFG works pretty well. It's not completely useless but it doesn't work in all cases. And

that's probably best distinguished in the sort of schematic of a tree where you see a postal type pattern that's sort of occurring in different parts of the tree. PFG pattern you would not be able to distinguish these isolates, but using whole gene sequencing, you can.

You saw this model where NCBI is playing a very central role in accepting data from all across the world and then the analysis can be done basically anywhere. This slide as well, and I think that you saw this slide but for the first part of the initial kick off, the first pilot year of this project, data has been flowing from the network into FDA and then into NCBI. FDA really wants to get away from this model where they are sort of brokering the data. They want to have all this distributed network where all the partners can submit to us.

All right, after FDA approached us, CBC also discussed starting a pilot project for Listeria whole time, or real time sequencing using whole genes. Marc was telling you about some of

the success of that. Just to give you an idea of why Listeria was chosen -- it's got a low incidence in the population but it has high mortality and morbidity. As you were mentioning, the people who are old. There's a small genome size that has excellent epi to do the action between the sequencing surveys with people whose current typing methods are problematic, though that's a problem with this midget desk.

(laughter)

So you saw this template for pathogen bio sample information, sort of the connection of what is a sample, when the sample is collected, where and who collected it? This is established template in the bio sample database at NCBI. The template exists. The current discussions between the different agencies are that contextual level of detail that can go into each one of these fields and that's an ongoing discussion between our different agencies, but this sort of minimum metadata of what to collect is actually I think well established for tens of thousands of samples

now.

The (inaudible) has this model that is sort of an analog to what we're trying to do here, is you have this weather. Weather is global. It affects everyone and it's not responsible for borders. It crosses and affects everyone, and there's all these instruments out there collecting information. And you can integrate that information to predictions. For example, we have Hurricane Sandy and its early predictions of where it will cross and what state it will impact. In the early hurricane season you don't really know where it's going to make landfall and so you wouldn't -- you can make predictions but you wouldn't actually evacuate Boston for example. But as you get the better resolution, you can see all these models in real time, both the raw data and the predictions. Then you get better resolution. And doing this publicly is important because scientific methods have not established patterns for doing this. There's not an easy button you can push and just say well which city

should we evacuate?

So how do we get involved in this or what is our role? So Marc gave you some examples of some outputs, but this is sort of our pipeline right now. The short reads arrive through the network and get deposited in SRA. And we pick up those short reads and we do a little bit of modification, trimming, and using camers we determined the nearest reference in the reference tree and this is for all bacteria, which is 26 to 27,000 genomes right now. So using camers (phonetic, same as earlier) we identified the nearest neighbor and in parallel, I have to explain this in parallel. I'm going to say it again -- in parallel, we do both reference assisted assembly and a panel of the novo assemblers. The reason to do this is the reference assisted assembly is very quick, very rapid and we can usually, for the close neighbors or clonal outbreaks we cover almost the entire genome sequence. And then the novo assemblers are mostly there to pick up the mobile elements --

plasmas and phage. This is different what most other people do. Most other people either do read mapping to a reference that's been pre-selected or they just do a de-novo assembly. And so we think our method is better and we've done extensive testing on this. We combine that final into a final assembly, do read mapping to measure the quality of the assembly and from there we can now re-cluster that, or we can on to do snip analysis. We can do annotation and deposit in a jim bank. In fact, we've actually finished the engineering for that step and so, the first Listeria genome from the project has now been deposited jim (phonetic, same as previous) bank and then very quickly we'll fill in the gap for the other almost 10,000 samples that have been submitted. So those will be in jim bank (phonetic, same as previous) shortly as well, besides just being in the short read archive.

Then Marc gave you this comparison between PFG and post net (phonetic 0:32.48.0), then we actually, comparison between whole genome

sequencing and PFG pattern typing. And we actually did a real time experiment last year with New York State where they were looking at whether they could do the PFG pattern or the whole genome sequencing typing in the same amount of time. Effectively, it was the same amount of time. The only thing I want to point out on this slide, a pointer -- people online can't see this, but you see this stack of very short parallel steps? That's basically NCBI's involvement. Everything upstream is all the laboratory response, the collecting the sample, culturing, laboratory prep, the sequencing and then submission to NCBI. So the work we can turn around within approximately on average 4 to 5 hours the result, which we haven't even optimized yet, so we think we can turn this around even more quickly in the future.

What have we done so far? So as Marc was saying, we have -- this is based on the beginning of September, so we had almost 9,000 samples in total. We have Listeria from CDC and the FDA, and even I should point out, the Food and

Environmental Research Agency in the U.K., so this is starting to expand into a global international network. And I was just at the Global Microbial Identifier Meeting in York, U.K. which is sort of an ongoing attempt to build a global system like this, although there's still far too much talking at that meeting and not enough doing in my opinion. The U.S. is a forerunner and a leader in this field. And I as a Canadian feel that we should expand this to a more global system. But we have commitments from the public health agency and England to submit all their Salmonella to (inaudible) Laboratory, so this really is starting to become a global system.

We have all the Salmonella from the FDA. The first year of the project was a great deal of retrospect and samples being submitted but we're starting to get commitments from states to sequence their clinicals to start to make those connections as well. We have some e-Coli as well.

This is just sort of the number of samples that have been submitted per day over the

last year and you can see the spikiness of the samples. And there are a couple of downtimes that are partly due to our problems, so sometimes our computer network has had issues. We had a thunderstorm that flooded our computer room, so that went down for a while. But those are things that we're going to have to be aware of in a global network system. And we also get spikes when I think people have gone on holidays and then come back and then have to submit 100 samples on a Friday.

But there are a number of issues that I see. People have presented two types of talks at this conference. One is the optimistic version and the more pessimistic version. Everybody who knows me knows I'm a pessimist so I'm going to present some of the issues that have arisen. As can be expected, when you're doing real time sequencing, you can run into some problems, and some of these are as Marc mentioned, human errors. Some of them are cases of the wrong organism being submitted and so somebody isolates what they think

is Salmonella but it turns out to be (inaudible). We've seen mixed organism cases. We have some data issues with how much data we think we can trust, so less than 5x coverage, probably not enough, duplicates, those are just human errors, making mistakes. (inaudible) is only on here, I think Jonas has already left but it's mostly because we haven't built the system to handle pac bio assemblies yet because we were assuming, if you're sequencing pac bio you're going to do the assembly yourself.

I'm going to go through a couple of these errors in detail. So incorrect organism -- this can be both in what is (inaudible) right now. Everyone knows that this is probably a problem. We have, as I said, almost 27,000 assemblies and some number are obviously going to be wrong. We're working on trying to fix that, so here's an example of a Campylobacter jejuni, which clusters with Campylobacter e- Coli and we informed the submitter of this that they probably had the wrong organism name and they agreed with us and they

made a fix. And then here's an example of Enterococcus Faecalis which clusters with Enterococcus Faecium and this was supposed to be an HNP reference genome but apparently it was not a reference genome. We informed the Brode Institute that this was a problem and they are going to correct this very soon. It's amazing what public outing of your mistakes does to your cleaning up systems properly.

We have ways of labeling this data as problematic so this is a couple of examples from our assembly database. They are listed as anomalous assembly. Two of these were misassembled and this last one is actually, was submitted as Bacillus Subtilis but it's actually, someone has cloned a Cyanobacteria genome into Bacillus Subtilis so this actually links the publication describing that process. It's not that the experiment was done incorrectly but the fact that this is Bacillus sort of, you want to give warning that there's something else going on here.

And then a quick tangent on, we have ways of labeling reference material, that you've heard lots of talks about reference material at this meeting. In the (inaudible) database we have this field called reference material, which sort of, we can give you a description of what that reference material is. You probably can't read that, but it's like a post snip standard reference material used to do pattern typing. And we also have links, for example to ETTC culture collection where if it was a culture collection, if you got it from DSMZ or ATCC, you could actually provide that information as well.

So what else have we seen as errors? We've seen mixed samples, so if you ignore the colors, if you just look on the X axis, you see the assembly size. So this is a Listeria and you see the nice clustering at around three mega bases for Listeria, and then on the left is, on the Y axis is the N50 (inaudible) length and so in the upper left you see a lot of high quality assemblies, single contig or just a few contigs

(phonetic, same as previous). The bottom left there's some bad assemblies. But the ones that are problematic are everything on the right of this X axis. And what is going here? We have some that are almost double the size you'd expect of Listeria. Well, here's an example of a contaminated sample with both Listeria and Staphylococcus. So we have a fairly decent assembly for the Listeria but we have a lot of Staphylococcuses for a lot of fragmented contigs. So this is obviously a case where this sample cannot be used for trace back because we don't know what happened to this sample.

Then we have a problem with the platform. So Illumina, most of the data that we're getting from this network is Illumina. And we have sometimes the case in Illumina where we have carry over contamination. So what we have on this slide is four samples. Histogram of the coverage and the number of vases, and the sample on the upper left, the upper right and the bottom right are decent high quality assemblies. The one

on the bottom left, I don't know if you can see this, but there is a little spike of low coverage bases in this assembly. And so this looks like this is carry over from one run to the next run in your sequencing run. And this is a case where NCBI is going to make a recommendation. It's not a standard, but it's sort of a recommendation to our partners to explore the use of combinatory barcoding to sort of separate this issue out. So this is of the case where you would, for example, sample one, you would use barcode A and B, and sample 2, you would use barcode C and D. And so next week when you're doing a sample 127, if you see barcodes A and B, you would recognize it was from a previous run. You could throw that data away. So they're exploring this and this is where the collaborative nature of our network helps.

And then the more insidious case of a mixed sample is where it's the same species. So here we have a fairly decent run. It assembles very well, three point almost two mega bases, very few contigs (phonetic, same as previous) but when

assembling map the reads back to the assembly to measure the quality, we recognized a whole bunch of alternate alleles. And you can recognize this when you see; this is in Genome Workbench or tool where you can sort of load a bound file. At the bottom you see this coverage graph and then all these little red lines are errors. If you zoom in on that, what you see is for -- you see a set of variants, a core occurring in a set of reads, and these variants complement each other very well. So that's unlike what you would see with a normal assembly where you have -- you always see variants or errors in that actual read, but they're not suggesting that the -- what we think is happening here is that we have a mixed population of Listeria in the sample.

So you could imagine, this is easy to see, right? There's a lot of nice red lines and they all look like they're in the same position and they have the same base. But imagine you were in a laboratory and you were seeing an outbreak and you had two strains that were very very very

similar to each other. Could we possible tease apart that this one is different by 12 snips and this one by 15 snips, across three mega bases? They may be very problematic. So this is something to be aware of when doing this type of trace back investigation using the whole genome sequencing.

All right, so the idea is that whole genome sequencing would give us this resolution and as Marc said, we really want to focus on the tips of the tree -- what is the epidemiological relevance. And we want to distinguish independent events, so we want to separate things that are assembly errors or repeats. And you can see this in a (inaudible) comparison of two assemblies. You see the variances of these sort of accumulate along the graph, or sample on the X axis and assemble on the Y axis. Get a uniform distribution and these two samples; they're so divergent that they wouldn't even be considered part of the same outbreak. But if you don't properly account for repeats, then you'll get a

spike like this in the graph, where in a single part of the genome you see a huge increase in the number of variants. And you want to properly filter those out, so you're only getting the independent events as a measurement. And we're working on using an algorithm developed at the Sanger Pathogen Center and we're looking at ways of how we can rapidly detect these. Because you really want to get down to this, as Mark said, the tips of the tree, and rapidly report on these cases that we think are sort of clonal or near clonal (phonetic, same as previous) indistinguishable. So we're working on outputs for both -- right now we're doing a (inaudible) tree of all the Salmonella. That's what Marc was saying we produce routinely daily, depending on whether we get samples or not, and we're working on rapid ways to do a snip tree. We've looked at both parsimony methods and rapid distance methods.

And so the idea is sort of to report to the people, something like continuing along this weather theme. You know, you can look on the

weather app on your iPhone or whatever phone you have and you can see that there is either a thunderstorm happening or there's a tornado watch, which means there's something to be aware of, but you know, don't be worried, or a tornado warning, which means, get underground as soon as possible. So this is a mockup of sort of like, you know, what's happening in Salmonella today in the U.S.? You sort of come and you can see how many samples have been updated and then get like a rapid report of a new clinical to an environmental link and you can look at the tree, just within this small part and see that, you know, there's a clonal link between two isolates for example.

So I'd just to acknowledge all the people in CBI -- it's a vast army of people and we know there's a vast army of people at both CDC, FDA, UC Davis, well okay, we'll ignore that for now, USDA, state labs, and various partners and collaborators that we worked on this project over the years. And I'd be happy to take any questions.

(applause)

SPEAKER: More a comment than a question, food processing has been very successful in increasing the, or actually in decreasing the amount of bacterial contaminants in food, whereas it's been much less successful for decreasing bacterial -- sorry, viral contaminations, so if you have the enteric viruses that are breaking through and not decreasing in terms of food borne pathogens. So do you see an increasing trend towards using your NGS or setting up libraries for viral based pathogens as opposed to bacterial based?

MR. KLIMKE: Probably a question to ask FDA and USDA, not us.

SPEAKER: No, we're planning to do norovirus and those are what's included. That's the plan.

SPEAKER: Any other questions? So briefly from the online audience -- you touched barcodes, but the question was, do you see barcode bias specifically with Illumina and if so, how do

you deal with it?

MR. KLIMKE: Do we see barcode bias?

SPEAKER: That was the question, yeah.

MR. KLIMKE: We haven't thoroughly investigated that based on the submissions that we get, because we don't always -- we're not always informed at submission time what barcodes have been used. So we would have to actually probably contact people and collaborate with them to determine that.

SPEAKER: Any other questions?

MR. PETTENGILL: Okay, our third and final speaker, is Kristin Holt from the Food Safety and Inspection Service, which is part of the USDA, and she'll probably actually not overlap, maybe much with any of the previous speakers. And she'll be talking about the use of next generation sequencing data to enhance food safety and the need for data format standardization.

SPEAKER: All right, just a quick interruption, after this talk, we will have quick

closing remarks, before we end the session. Thank you.

MS. HOLT: Okay, hopefully I've got the mikes correct. Great, okay. Well great. I appreciate this opportunity to continue the discussion of the use of next generation sequencing and food safety. I know that during the workshop, a few people have asked, kind of polled the audience for things, and I have a question for the audience, so, how many people have a relationship with food or eat food? Raise your hand. (laughter) Okay, well I actually had a kind of a hidden motive in asking you that question, because I'm the last presenter today. And I just wanted to kind of get some oxygen and endorphins going, so, because, and I'm going to move, try to move quickly there, and I do have some repeat content I think. Okay.

So the Food Safety and Inspection Services, the public health agency in the U.S. Department of Agriculture is responsible for ensuring that the nation's commercial supply of

meat, poultry and egg products is safe, wholesome and correctly labeled and packaged. So I underline meat, poultry and egg products because that's what we're about and FDA is about all the other foods. And since we're talking about next generation sequencing, I think it's obvious for me to give you an overview of the FSIS office of public health science laboratory services team. So we have three field service laboratories that are responsible for microbiological and chemical analyses of foods and environmental swabs. Again, this relates to meat, poultry and egg products. And we have a strong focus on the control of data integrity and standards. Our labs are located in Alameda, St. Louis and Athens, Georgia, and Athens we also have some specialty units, pathology branch and an outbreak section of the eastern laboratory or OSSL. Also in Athens, we have our executive associate for lab services, our lab quality assurance staff and I'm going to call out Cathy Pence whose here, to show how we're into quality. We wanted to be here at this meeting and

find out about data standards. And we also have our food emergency response network staff.

The mission of the FSIS laboratories is to perform analyses on meat, poultry and egg products. We have a very robust food sampling program, at least for food. This is very very robust. We get over 100,000 samples and perform over one million analyses a year, so our labs are very busy. Again, a strong emphasis on quality. We have a quality system and this graphic shows, you know, the types of documents that we have at all of our laboratories. We have quality manual, SOPs et cetera. And our labs are accredited to the ISO 1702.5 standard.

Now I'm going to drill down a little bit and talk to you about bacterial isolate characterizations that we do at our eastern laboratory which is located within the Richard Russell Research Center in Athens, and there's a picture taken on a beautiful spring day. So what is the current role of FSIS in bacterial isolate characterization? So we do pulse field gel

electrophoresis, or PFGE, we've done that since the beginning of PulseNet, we were an original member in that network. We upload quite a bit of patterns and meta data to go with it. I don't know why I have just numbers back to 2006, but we've got data going all the way back to the mid 1990's. We do MLVA; molecular serotyping of the Salmonella isolates using CDC's new molecular serotyping method. We do anti-microbial susceptibility testing and we have storage of lots of isolates.

Now this is really busy, so I just want to orient you to a couple things and make really a point about the work load that the folks have there in Athens, in characterizing the bacterial isolates. So the top left graphic is e-Coli 0157 and the top right is Listeria. And those programs have been pretty stable -- not a lot of changes there. But if you look at the Salmonella in the middle, which is on a different order of magnitude, you see kind of a gap in Salmonella in the early years and you say well, they do poultry,

surely they've got a lot of poultry isolates. Well actually, our sister agency, the USDA or as was doing the characterizations and we took those over recently, so then we had the big jump in the numbers. And then the bottom row is for non 0157 (inaudible) and enterococcus and those big jumps are new programs that we actually have recently implemented. So to give you kind of a sense of the jump, in numbers we had 492 isolates characterized by this crew of people in 2010. And just this first six months of this year, they did over 3,000. So this is a very busy group.

So why the increase in isolates? We just have new regulatory testing programs, we are getting isolates from baseline studies that we're going to characterize all those isolates and sometimes that can be a lot of isolates and we picked up the responsibilities that ARS had before. We're doing NARM sequel sample testing and we do special projects and outbreak investigations in the recent Salmonella Heidelberg investigation associated with chicken. We had a

lot of isolates that we did characterizations of.

So why is FSIS becoming involved in whole genome sequencing? Well it fits well with our FSIS strategic goals, our strategic plan and this is just a sampling of some that seem to relate, but the one at the bottom is, you know, we have goals to use innovative methodologies, processes and tools and this of course is a perfect fit and we have high level support within our agency to move forward.

The benefits of whole genome sequencing, now this is a really busy slide, and I actually curtailed some information. I think I could have given you four, five slides with tiny font. But let me walk you through a couple things here. So the first thing is really, I think, visionary. FSIS anticipates that whole genome sequencing will eventually replace PFGE, serotyping and other testing and we will quickly be able to quickly tell us about a strain's virulence and resistance potential, it will streamline our lab methods. And in the clinical world, we're looking forward

to the fact that we might get resistance and virulence data and information really really quickly. And of course, it's a powerful and discriminatory tool, compared to PFGE. It's fast and cheap. We're going to be more efficient with our resources. It's going to improve surveillance. You know, for epi investigations that we think we're going to continue to see that we have to do trace back and epi in investigations, not just rely on matches from whole genome sequencing between a clinical isolate or food. We got to bring that extra investigation data to the consideration, because we wouldn't use lab evidence alone to take regulatory action we have to, in a food borne illness investigation. It's going to improve the classification of ill patients, meaning that it's more likely that a case is a case and we're doing the trace back on the right person and the right food.

So how have we been involved? We've been involved in the global microbial identifier that was mentioned earlier. We've been

contributing isolates for sequencing to the Hundred K Genome Project. We are part of that Listeria real time surveillance project that you heard earlier. We're not actually doing the sequencing yet ourselves, but we're close. So to participate in the project we have contributed a 168 isolates, sent them to FDA and they've kindly sequenced them and uploading the data to NCBI. And we've also sent Salmonella isolates over.

We've received two sequencers in April 2014. We have two aluminum (inaudible), so that was very exciting for us. We may -- taken our scientists and done (inaudible) to CDC, NIH and FDA and they graciously opened their doors and showed us everything, answered all our questions and continue to collaborate and help us move our whole genome sequencing activities along. We have had successful sequencing runs with Listeria, Salmonella and STEC and there was perfect timing when we got our sequencing, we had in our experience, and FDA had a proficiency test that we participated in. So that was really great.

We're developing our knowledge and application within the agency. We're all very used to PFGE terms. So we're trying to bring people along to hold genome sequencing terms. And what's missing here is that we're not yet uploading directly to NCBI yet, but we hope that will come very soon.

And then I won't spend much time on this because this was already mentioned. We're participating in this project, and calls and conferences, et cetera where this activity is discussed.

And where do we get our Listeria isolates? We have a regulatory testing program of ready to eat meat and poultry products. We do environmental swabs with these manufacturing facilities, and if we had a food borne illness investigation, we collected samples. We may have some isolates from there. We are planning to do the real time sequencing at our lab in fiscal year 2015 and we think we'll have about a capacity to do about 10 to 20 isolates every two weeks. And

we do have banked isolates, so we will be pulling some of those out and doing those.

Just some extra considerations for shiga toxin, producing e-Coli 0157 and now we're doing testing for the 6 non 0157 serum groups. We're going to do the real time sequencing there. And if you're -- the graphic there is I guess if you're curious, what sero groups we're seeing are. We're seeing a lot of 0103. And we have banked isolates. Another consideration is for Salmonella and Campylobacter we have a lot of isolates and I think this is the same challenge on the clinical side. So we're going to definitely do that -- you know, whole genome sequencing for food borne illness investigation samples. But we're waiting to kind of see how CDC, FDA, NIH and FSIS come together on a strategy to take on Salmonella and Campylobacter because it's a little different than Listeria. There's a lot more isolates.

So what are the data considerations for us? They're basically the same as for everybody. Data storage and transmission, of course, for us,

we have the quality standards that we have to live up to. And then also, the epidemiological understanding of how we're using the new data -- that's a learning curve for us, but I think it's a learning curve for everybody. But we think it's really a great tool. Successes -- you know, commercial software is there, our IT infrastructure's going, getting better and the agency is learning more about sequencing. And I guess everybody kind of had a graphic like this. This is just basically; we've got our flow mapped out.

Meta data -- this has been pretty straight forward for us. We're going to provide and we are providing with the isolates we're sending the FDA, the product type with you know, what kind of food item it is, the year the sample was collected, the state where it was collected, any subtyping information we have and the meta data is available as soon as we pass the isolate or we're going to upload. We're not going to have any lag on there. It will go with the sample or

the raw reads.

And the assessment of sequence quality -- this is what we have. And I guess one of the draws for me to come to the meeting was also to learn, you know, what is the scientific and public health community using for quality standards? Because we're regulatory agency and for us to take an action, we really need to be doing what everybody else is doing, that there's consensus. And I guess we're learning that I guess there's not some hard numbers and hard agreement and we're really just kind of getting there and I appreciate FDA hosting this meeting.

So for the scientific and public health approach, you know, we know there are a lot of benefit from whole genome sequencing. For us, we do recognize that the Information Quality Act pertains to us and other agencies. And USDA created guidelines on the information quality. We of course, have to have that strong data quality data standardization, so that we can use that information and defend and take an action.

So I'm going to end with just kind of a little bit of a story with the focus of how the sequence data from just one person's isolate could be very very important to crack open a food borne illness investigation. SO CDC estimates that each year roughly one in six Americans get sick, 128,000 are hospitalized and 3,000 die of food borne disease. And I'm going to talk now about Salmonella causing about 1.2 million illnesses in a year. Most of the time you're thinking about poultry products, but when you have an investigation that starts, you don't really know it's really going to be the poultry. There have been Salmonella outbreaks connected to vine vegetables like tomatoes, fruits, nuts and nut butters and of course, sometimes it's not in food at all. It's backyard poultry, or somebody's pet reptile or amphibian.

And there are many benefits we see with whole genome sequencing too -- the epidemiologic and the trace back investigations. Food borne illness investigations are multifaceted, they're

multi-disciplinary and they're collaborative undertakings. We believe in the three legged stool of investigations, with the legs being laboratory, epidemiology and environmental health, and I highlighted in blue, you know I guess the clear connects with whole genome sequencing. For a laboratory, I think it's pretty obvious, but for epidemiology, we think we're going to have better cluster protection, better case classification, so we're going to rule out some of those people who probably just kind of sporadic illnesses that are not part of that outbreak. That's going to be really powerful for us. And for trace back, again, we interview people and we're going to be tracing back the right food.

So to walk through, I'll just kind of walk through quickly here. So you know, if you eat a contaminated food item, for Salmonella it takes a couple of days for you to exhibit the symptoms and be ill, and then you decide to go to the doctor. The doctor asks for a stool specimen. That stool specimen then gets tested at the

clinical lab, Salmonella is identified and then a clinical lab will send the isolate to the public health lab for further characterization such as PFG. And then the case is confirmed as part of the outbreak. So it can actually take maybe two to four weeks after you ate that contaminated food item to being confirmed as part of the outbreak. So the significance of that is, when the health department calls you up, they're going to ask you what you ate, maybe four weeks ago. And they're going to ask you the seven days, or three to seven days before you got sick. So what were you doing five weeks ago and what did you eat and where did you go? It's pretty tough, so that's one of the things I wanted to just keep that in your mind, as we move through the rest of my slides. And the other thing is again, I think whole genome sequencing can kind of cut some of this timeline down and give us some information really quickly.

So now we're using PFGE and I'm just going to use that as an example because you know, this was a new technology 20 years ago. It was

innovative. It got awards. But you know, it's been used for 20 years. And it's strong, it's powerful and maybe it's not as discriminatory but as much as whole genome sequencing, but when that isolate gets to the health department, we've got it, the labs do standard methods, you can roll the data all up together. It's not a perfect tool, but it is a great tool and it's worked really well. And we're probably going to continue to use it for a few more years now. So basically, so that person who got sick, and had that isolate, go to the health department, this will be compared either at the state, you know, looking at the central database or at the CDC and roll up the national data.

So this is kind of a hypothetical investigation that I put together. It's kind of got hallmarks of some real investigations. So you imagine there's a Salmonella Heidelberg with a PFG pattern A, it's a sub-cluster because it's maybe a common clone of Heidelberg. But we get down to those people who four or five weeks later, you

interview them and you have a good food history. For some reason, they really remember what they did, where they ate, what they shopped, where they shopped etcetera. And so we might have a big outbreak, but we're only working with a handful of people who actually have what we call good food histories or good exposure history. So every person here is really important and we don't want to lose that data because somebody said it doesn't meet a quality mark. So for this scenario, we've got to interview these four patients -- oh, sorry -- so by interviewing them, there's some kind of a poultry item we're worried about -- chicken or something else. And we're finding from the trace back going to the grocery stores, the shopper card loyalty data, that it looks like these are big chicken eaters. And believe me, this is how this really works. People eat a lot of chicken in the United States. And they eat it from all kinds of different places.

So we get okay, establishment A is showing up on three out of four, but yet

establishment B, that their chicken's showing up on three out of four and establishment C. So I'm really excited to see this fifth case come on from Minnesota to kind of triangulate and move it out of the area. What did they eat? And they had a good history, establishment A. So establishment A essentially kind of lights up. And we are then like okay, we have a lot of confidence here. And then, for case patient two, say the shopper loyalty card information was slow to come in. It comes in and voila, there establishment A lights up, and we really have it firm.

But I guess this is just to illustrate that if we lose the data quality or we don't have standards, we're not all together, we can lose one person and it can really break up our ability to then take some kind of action like a product recall, do a public health alert, go to that facility and check them out, which we probably would be doing anyway, regardless of that we'd be checking out several of these.

So anyway, just want to kind of walk you

through the, this is kind of where the rubber meets the road, and we need good data quality standards with things all coming together. But thank you.

(Applause)

MS. HOLT: We're not officially part of the Genome Tracker network. But FDA has been really just incredibly gracious. Send us your isolates, send us your isolates, send us your isolates. So we have sent lots of isolates and they have done the sequencing and I mentioned before, we visited them and they walked us through all their protocols et cetera. So we appreciated FDA's work.

SPEAKER: We appreciate your (inaudible).

MS. BOLT: All right great.

SPEAKER: I have a quick question. So with the increase in sort of Next Generation sequencing being used and given that you pointed it out, sort of the importance of you know, good exposure, history and sort of the epi and sort of

that linkage information. Do you know if there's a sort of concurrently novel method being developed, to sort of better you know track people that are sick and get information from them, that sort of would complement the NGS input?

MS. HOLT: So really that's probably a question for CDC and the state health departments. So I'll just say though, yes, enhancements to getting the questionnaires standardized, getting questionnaires collected, you know, people interviewed quicker. There's a lot of, I mean, there would be a whole 'nother presentation on that, so yeah. On the epi side, there's a lot of activities and there's a lot of close, close work between the epi's and the lab folks with this Listeria whole genome sequencing, kind of learning together as we go.

SPEAKER: You know, in that example that you gave, there's also the possibility that all of those A, B, C and D, received chicken from the same points or it's going backwards one level. And so what -- do you immediately go out and

investigate them anyway, even if you found that commonality of A? Because it could have been one layer backwards?

MS. HOLT: Yeah, and that does happen. So when we do a trace back on a person, it starts largely say, like with the grocery store. Then it goes to the distributor, then it goes to that next point. And sometimes that's a processing plant, and then it goes back another step to a slaughter plant. So yeah, it carries all the way through. And then triggers to actually go to a facility. We can look at their data just remotely, their micro profile from testing over the years. But we can, you know, really, we don't have to go do a full formal investigation. We can start, you know, going back and tracing and looking at them and talking to them even and saying, what can you tell us to help? So we have a lot of latitude. And then we have some of the more formal things that are, you know, are kind of listed at the bottom there, where we have described protocols. So we have a lot of latitude.

MR. PETTENGILL: That concludes the final session, but there's going to be a few closing remarks.

MR. SIMONYAN: I'll take like five minutes of the time that is left. We today got all of the stakeholders. Oh, sorry. Today and yesterday, we got representatives from all of the stakeholders. We got device manufacturers, we heard from them. We heard from consumers of the (inaudible) information, those being hospitals, research institutes, regulating organizations or consultants, or like national health organizations. We didn't hear actually from hardware manufacturers and we didn't hear from software manufacturers, from industry. And that is also important. We also want their input into the standardization effort and I'll make sure, I mean, I know they are in our lists with just perhaps, short sightedly didn't give them a chance to talk or maybe the time was not enough, but I'll make sure that they also have a talk when we are working on standards, and when we are working on

bioinformatics validation as well.

So what our vision is, what the procedure would be after this, we have developed the draft documents which will be introducing what we are trying to do. But after this meeting, we got a lot of input from all of you, by talking to you, we want to get a few days on editing this draft, trying to incorporate the advice, the ideas, the meaning which we have put into our brains, and then, release those drafts after we rotate (inaudible) maybe a day or two, we will release this draft by clearly stating that this is a draft. There is nothing written in stone. It is a suggestion where we can start our conversation, and how we can continue the conversation. At that point we will send an email to all of the participants in this conference and all of those people who registered for continued workshop participation and then we will outline what is the procedure, when do we meet, when is the first meeting, what is the agenda of the first meeting and ask you for agenda items on it.

Hopefully this will work out and important thing is that all of us are here for one and only goal, with different perhaps means but there is only one ultimate goal behind all of this, that's the health of the humans. And the impact we can have to health of the humans. People who use and reuse this work, but it is important to remember those humans are our children, our mothers, our fathers and us. So by having that beautiful idea, I want to close this conference by having that beautiful idea, however we do and whatever we do, we are doing it for humans. Thank you. And let's thank all of the organizers.

(applause) And Carolyn Wilson, Carolyn Wilson wanted to say a Couple of words from FDA perspective.

DR. WILSON: Okay. I will be quick. I know it's been a long two days. But I do want to also just echo the thanks to all the speakers, all the people who were here for a portion or all of the two days, and offered up their suggestions, their questions. I think what I came away from

these two days understanding is that there's a lot of concerns, challenges, and there's a lot of shared understanding of what those problems are in the field, in order to make the potential promise of this technology really give the full benefit, as Vahan said to you, the public health. And I think that at least now that we have a shared understanding of what we need to do, and I assume the next steps are really to prioritize what those needs are -- what I got from it is some will be easier to tackle than others. And that you know, some may take longer and maybe we don't even know enough yet at this point to know how to tackle them. So I think those are additional things that people need to think about going forward, but again, thank you everyone, and we look forward to continuing this dialog with the community.

(Whereupon, the PROCEEDINGS were adjourned.)

* * * * *

CERTIFICATE OF NOTARY PUBLIC

STATE OF MARYLAND

I, Mark Mahoney, notary public in and for the State of Maryland, do hereby certify that the forgoing PROCEEDING was duly recorded and thereafter reduced to print under my direction; that the witnesses were sworn to tell the truth under penalty of perjury; that said transcript is a true record of the testimony given by witnesses; that I am neither counsel for, related to, nor employed by any of the parties to the action in which this proceeding was called; and, furthermore, that I am not a relative or employee of any attorney or counsel employed by the parties hereto, nor financially or otherwise interested in the outcome of this action.

(Signature and Seal on File)

---

Notary Public, in and for the State of Maryland

My Commission Expires: November 1, 2014