UNITED STATES FOOD AND DRUG ADMINISTRATION

CENTER FOR BIOLOGICS EVALUATION AND RESEARCH

CBER SCIENCE SYMPOSIUM

DAY 2

Washington, D.C.

Tuesday, September 17, 2024

PARTICIPANTS:

CHARLES CHIU, MD, PHD
Professor of Laboratory Medicine and Medicine, Division of Infectious Disease Diagnosis, University of California, San Francisco

SHIRIT EINAV, MD
Professor of Medicine (Infectious Disease) and of Microbiology and Immunology, Stanford University

VISWANATH RAGUPATHY, PHD
Staff Scientist, Office of Blood Research and Review, Centers for Biologics Evaluation and Research

GABRIEL PARRA, PHD
Principal Investigator, Office of Vaccines Research and Review, Center for Biologics Evaluation and Research

ROMMIE AMARO, PHD
Professor, Molecular Biology, University of California, San Diego

FRANK DELAGLIO, PHD
Principal Investigator, University of Maryland/National Institute of Standards and Technology

NOBUKO KATAGIRI, PHD
Staff Scientist, Office of Plasma Protein Therapeutics, Center for Biologics Evaluation and Research

LISA PARSONS, PHD
Staff Scientist, Office of Biostatistics and Pharmocovigilance, Center for Biologics Evaluation and Research

PARTICIPANTS (CONT'D):

HUSSEIN EZZELDIN, PHD
Senior Staff Scientist, Office of Biostatistics and Pharmacovigilance, Center for Biologics Evaluation and Research

WINSTON COLON-MORAN, MSC

ALAN BAER, PHD

ROBERT DOROSKY, PHD

BYUNG WOO KIM, PHD

STEPAN SUROV, PHD

TAPAN KANAI, PHD

SABARINATH NEERUKONDA, PHD

ATUL RAWAL, PHD

ERICA SILBERSTEIN, PHD

BRYAN VISSER, PHD

DAVID MCGIVERN, PHD

RANADHIR DEY, PHD

DARON FREEBERG, PHD

MARISABEL RODRIGUEZ, PHD

* * * * *

P R O C E E D I N G S

(9:30 a.m.)

DR. MCGIVERN: It is 9, it's after 9:30. Good morning and welcome to today's session on emerging and reemerging diseases. I'm David McGivern from FDA's Office of Blood Research and Review, and I'll be chairing this session with my colleague, Dr. Ranadhir Dey.

Our first speaker is Dr. Charles Chiu. Dr. Chiu is Professor of Laboratory Medicine, and Professor of Medicine in the division of Infectious Diseases at University of California, San Francisco, and director of the UCSF Clinical Microbiology Laboratory.

He leads a translational research laboratory focused on the development and clinical validation of metagenomic next generation sequencing and host profiling assays for the diagnosis of infections, outbreak, investigation, and pathogen discovery. Please join me in welcoming Dr. Charles Chiu.

DR. CHIU: Thank you. And I'd also like

to begin by thanking the FDA for inviting me to speak with you today. The title of my talk actually has been a little abbreviated. It should be actually Metagenomics and Host Profiling for Diagnosis of Infectious Diseases. And so the motivation of doing so is that we face many challenges in infectious disease diagnosis, especially for emerging infections.

One is that we know that conventional microbiologic tests have poor diagnostic yield. In fact, approximately 50 percent of the time, we're unable to identify the cause of meningoencephalitis in patients admitted to intensive care units despite extensive testing.

For pneumonia, it's anywhere from 15 to 62 percent depending on the study that you look at. And even for sepsis, approximately 20 percent of these cases remain undiagnosed.

In addition, the typical infectious disease patient undergoes multiple tests without a diagnosis as a result. And available tests may be limited in both sensitivity and/or scope. This

contributes to a long diagnostic odyssey for many patients, especially patients who are affected with emerging infections, which contributes to prolong hospitalization, progressively invasive tests, and unnecessary treatments.

In addition, the need for doing this testing and the lack of assays contributes to very high costs or high cost of burden of patients who have infections or potential infections. For instance, meningoencephalitis patients in intensive care units typically incur around 50 to 100K. And then pneumonia and sepsis can incur up to $8.1 billion in costs per year.

So my interest has been in developing kind of clinical metagenomic sequencing for diagnosis of meningoencephalitis, and meningitis, and encephalitis. And the general idea is the use of essentially rapid pathogen detection. So what we've done here is the goal is to take a research based assay to do metagenomic sequencing and be able to, in a single assay, be able to diagnose the full range of infections that may affect

patients.

So this has been clinically validated at UC San Francisco since 2016. We've been running spinal fluid metagenomic assays in our diagnostic laboratory here. And we've also have developed respiratory, and plasma based metagenomic assays. And the principle of this has been demonstrated in several papers that have come from my laboratory. One includes a paper in the New England Journal of Medicine titled Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis.

The general principle of clinical metagenomic sequencing is it's essentially shotgun sequencing of DNA and RNA, because bacteria, viruses, fungi, and parasites, they're inherently DNA RNA based, there's no reason that with a single sample, we can in principle be able to diagnose the full spectrum of infectious diseases.

In addition, we've been able to show that this technique is universal in the sense that you can actually apply this across a wide variety of different body fluids. So what makes us quite

attractive for the purposes of emerging infections is that you don't need to add targets as new targets arise.

So to give you a good example, many of the respiratory virus panels that were available at the time when the COVID pandemic happened, really, it was very challenging for manufacturers because they had to ensure that by adding a new target, in this case SARS- CoV-2, would not affect the performance of their assay. And so what we really want to do is we want to develop methods where we don't have to rely on constantly updating the composition of our panels, depending on the current epidemiology of the disease at the time.

And metagenomic next generation sequencing, because it doesn't apply any primers or probes to target any individual pathogens, is very attractive for broad based diagnosis of infections.

I want to move now into kind of where I see though the next generation. I mean, traditional microbiology has always been based on

detecting the pathogen. So we're always interested in either culturing the organism, detecting antigens from the origin. Example would be the spike protein on the SARS-CoV-2 viral particle.

PCR testing, or nucleic acid testing, or perhaps the most broad based testing, nucleic acid testing, which would be DNA and RNA sequencing of pathogen genome. These are all essentially direct detection methods, meaning that you're looking for the pathogen.

Now we do have in infectious diseases other ways to diagnose infection. So one would be serology, which had been used for now more than a century looking for antibodies that may be specific to a certain type of infection. In infectious diseases, we can also identify like cell profiles, such as from spinal fluid, cell counts and differentials, to be able to distinguish, potentially distinguish between, for instance, bacterial infections, which cause bacterial meningitis, versus other organisms that

cause meningitis, such as aseptic meningitis, including viruses, and parasites, and fungi.

And then we've also developed protein biomarkers. A good example would be procalcitonin, as a way to distinguish potentially between bacterial and viral infection. Now, however, one thing that you should note is that really we lack assays, abundant assays, on the host side. And my interest has been in potentially leveraging metagenomic sequencing as a way to not only look for the pathogen, but also look for host response signatures.

But we do know that disease itself, that infectious diseases, is really highly dependent on an interplay between the pathogen and the host. A great example of COVID-19 where you can have the spectrum of disease ranges from asymptomatic infection to critical illness.

Okay. I want to give you an example of potentially how we can do host based diagnosis of emerging infections. So a good example are diseases where there's both an acute and a chronic

component. So a great example was Lyme disease, where you have, now this is a disease that has various stages. In early Lyme disease, or acute Lyme disease, for instance, you potentially can identify the organism by directly looking for Borrelia burgdorferi in the blood.

Or you can look for antibody testing. In fact, the gold standard testing is actually 2 tiered antibody testing. However, there is also kind of a later phase of Lyme disease. So either chronic complications of Lyme disease, including arthritis. You can have a condition called post treatment Lyme Disease syndrome, which is where it's a poorly characterized syndrome, where patients continue to have symptoms, neurologic symptoms, and signs despite successive, presumably successful treatment for acute Lyme.

And for these, for chronic manifestations of Lyme disease, we really have no tests. We sort of lack accurate tests for either diagnosing the disease in the chronic stage, or for monitoring the disease. And ideally, we would

want to have targeted tests for different stages of a disease.

Another good example is long COVID and acute COVID, where we now have new methods of treating, new treatments, or vaccines for addressing kind of acute COVID. But we really have nothing for long COVID. And part of the reason is that we don't really understand the pathogenesis, which almost certainly involves the host response.

Okay. So my interest now that has been moving towards, can we expand potentially for addressing acute infections, can we expand beyond simply looking at for diagnosing acute cases of disease to where we can better understand not only the infection itself, but also the host response to the infection.

And so the general idea here is we want to employ a technique which is part of metagenomic sequencing called RNA sequencing or host profiling. And the general idea is, with metagenomic sequencing, you sequence all of the

DNA and RNA in the cell or in clinical samples.

But however, most of what you end up sequencing is actually not from the pathogen. The goal of metagenomic sequencing is the needle in a haystack approach. I'm interested in looking for specific fragments of DNA or RNA that may correspond to pathogens.

On the other hand, most of what you end up sequencing is actually the human host. So what you're actually sequencing is, with RNA sequencing, for instance, you're actually doing an RNA sequencing experiment. You're actually sequencing the host transcriptome.

Similarly, by doing DNA sequencing, you're, I mean, technically you're sequencing the host genome and in principle depending on your coverage, you may be able to even assemble the genome or identify genomic markers. I'm going to focus in this talk on the RNA side because my feeling is that by looking at the transcriptome, the idea is can we actually simultaneously compare not only looking for the pathogen with sequencing,

but also characterizing the host response profile.

And we've been able to do so from spinal fluids. So this is actually an LDA plot, linear discriminate analysis plot, which is analogous to a PCA plot, for those that are about principal component analysis plot. And you can see what I've done here is I've actually taken a wide variety of different infectious and noninfectious conditions and essentially plotting them to look for what are the similarities in the host response, and the RNA, patterns of RNA gene expression among these different conditions.

And in principle you might be able to look at this and say, well, if I had an unknown sample of a patient that had, say, an unknown meningitis encephalitis, could I essentially place it on this plot to identify, potentially be able to identify does the patient have a bacterial viral fungal parasitic infection, or even a noninfectious etiology, such as malignancy, or vasculitis, or autoantibody syndromes.

And so that's the general concept that

we want to do with host profiling with metagenomics. Now, we've been able to show that this actually does work. So for certain types of emerging infections, so one is actually enterovirus associate acute flaccid myelitis.

So my group was actually initially involved in some, essentially a worldwide effort to try to identify the cause of acute flaccid myelitis in children. And it seemed that the we identified enterovirus, and others as well identified enterovirus D strains as sort of implicated in acute flaccid myelitis.

But one of the things that's made analysis of this very challenging is that you have children that present with the acute flaccid myelitis syndrome, and yet, we almost never can detect the actual virus in spinal fluid. I think the yield from the CDC was roughly two out of 500 patients actually ended up having detectable enteroviral RNA from the virus.

Now they were able to make the diagnosis because we were able to look at other sample types

like respiratory secretions for instance. Many of these patients presented with essentially an acute febrile illness, a self limited viral illness that came back several weeks later, or a few weeks later with acute flaccid myelitis.

And so one idea is can we leverage, for instance, the host response to better develop diagnostics for this kind of condition. So the general idea here is, well, why don't we leverage the tools of precision medicine, and the idea here is we want to leverage the diagnostics that have transformed, for instance, cancer or rare disease genetics. But we want to use this for infection.

And so essentially the idea here is that we want to identify RNA gene biomarkers and using AI and machine learning actually develop models that can help us predict responses to infection and be able to use to either diagnose or to predict infections. And so this is a technique. It's a very standard technique in the research world, which is called RNA SEQ or RNA sequencing.

But the general idea is since we're

already doing metagenomic sequencing to identify RNA viruses, why not leverage the sequencing data to now look at host responses simultaneously. And we can do so in using machine learning, using a machine learning based analysis workflow, we can actually identify, for instance, features that can help us kind of predict types of infection.

So this is how Google Image, for instance, works in that I could, for instance, feed the algorithm, say thousands of cars, for instance, in this case. And the general idea is by identifying cardinal features, you can then train the model to identify not only the cars that were used in the model, but also potentially new, essentially new brands, new models of cars.

So this is the idea is that with machine learning, you traditionally take your dataset divided into 80/20 percent. 80 percent is used to train the model, and 20 percent is used to actually test the model. And so this has been done in the machine learning analysis workflow. And essentially the idea is can we develop now a

classifier for actually differentially diagnosing neurologic illness.

Now you might expect this would be very important, for instance, for emerging infections, because emerging infections, as I said before, all of our tests are really highly targeted. Right. So if you have a new, say, a new coronavirus, there's no guarantee that existing assays are going to pick up that coronavirus.

So if you have a new Bunya virus that's causing an emerging infection, there's no guarantee again that your assays will be able to pick it up, because there's no genome representation in the database.

And the same is true as well for fungi and even fungi and bacterial strains. And so the general idea here is can we leverage host response to be able to diagnose disease. So instead of taking just the non-human reads, looking for aligning those to identify paths, why not take the human RNA reads and generate the host response.

And so we've developed, in fact, the

classifier. This is a very complicated slide, but in short, the general idea is we're using machine learning and cross validation to be able to identify essentially factors or models, classification models that can predict disease.

So I'm going to kind of go quickly through this. But the general idea is, let me kind of skip through this. The general idea is now we've actually developed models, and this is based on spinal fluid RNA sequencing, where we can distinguish, for instance, potentially between autoimmune disease, bacterial, viral, and fungal infection.

And indeed, many of the genes that we identify that appear to be associated with these makes sense. So if you look, for instance, here like with bacterial infection versus everything else, one of the genes is PI3, which is a well known antimicrobial peptide. For viral infection, the top gene is actually, as you might expect, an interferon associated gene.

And autoimmune noninfectious diseases

like auto AINI, which includes, for instance, cancer, malignancy, vasculitis, autoantibody syndromes. There are a number of neurodegenerative autoinflammatory genes that are represented kind of in that set.

And so what we can essentially do here is we can actually develop essentially a use the training set, 80 percent of your data, generate a fairly accurate model, and then apply that to the test sets to be able to diagnose infections.

What's also quite important is, and what's interesting here, is that if you actually classify a priori, let's say you take, in this case I took 1,000 infections, and I took spinal fluid data. We did clinical metadata in a blinded fashion. Look to see what was the clinical score. In other words, what was our clinical confidence in the diagnosis.

Did I know for sure that this was a viral infection, did I know for sure this is a bacterial infection. What you can do here is you can actually show based on the clinical confidence

score that the accuracy of the model depends on a priori, how confident you were.

And in fact, if you actually decrease the confidence level, the score here varies from 1 to 4. You can see that the accuracy actually decreases. And what that suggests to me is it's not that the model is being less accurate, it's more likely that the clinical diagnosis was actually not accurate at the time.

And so this highlights potentially the utility of this as an adjunct to aid in clinical decision making. Okay. I also want to state that this is not only for general categories. So traditionally, host response has been like bacterial versus viral infection. Can we distinguish between that.

Now, that's important for certain applications. For instance, like if you want to do a workup of sepsis. And there's a lot of interest, several companies have developed tests that can distinguish between viral and bacterial infections, for instance.

I'm actually a little more ambitious. I'm interested here in potentially being able to use host response to identify specific types of infections. Now, for many of these infections, though, they happen to be rare or uncommon, you may not have a good many examples of them. So we're using essentially what we call a leave one out algorithm.

And the idea behind it is imagine I want to validate a test for, say, mycobacterium tuberculosis infection. Well, if I only have 10 cases, what I do is I take one of the cases out, and the remaining 9 cases in all the other samples generate a model, and then see how well does that model predict that one case, and then you go on, so on and so forth, for the remaining cases.

So that's called the leave one out method of validation. And essentially we did this for several categories. And you can see here that, in fact, we're able to develop essentially host response classifiers for a variety of different types of infections, parasitic

infections, herpesvirus infections, coxy and histo, which are emerging, dimorphic fungal infections, enterovirus, associated acute flaccid myelitis, D68 and A71, MTB, and cryptococcus.

And this gives you an idea of how specific you can, these are what we call ROC curves or receiver operate characteristic curves. And you can see that the performance actually is quite good for cases of acute flaccid myelitis, but not as good overall, interesting enough, for like more general cases of enterovirus meningitis.

And my interpretation of this data is that the AFM or acute flaccid myelitis happens to be a very specific type of syndrome, so it's more easily distinguishable than more of a generic enteroviral meningitis.

But what's interesting is also that the more distinct syndromes are actually the syndromes that we're interested in developing better diagnostics for. I want to give you another example. This is an example of a very specific classifier that can be used potentially to

identify neuroborreliosis from spinal fluid or Lyme disease.

And if you look here, these are essentially what the model predicts are the 10 top genes associated with Lyme disease, at least in the neuroborreliosis, neurologic Lyme disease. And you can see here what's interesting is also post priori, we actually found out that what I have highlighted there in asterisks are actually the, those are actually genes that have been reported in the literature.

So you can see the vast majority of genes have actually been already reported, which helps to kind of validate the model in the first place, even though the model itself was developed in an entirely unbiased fashion.

Okay. I want to end by actually talking a little bit about how we can actually apply this in practice. So what we're going to see, and this is if you happen to have, say, an unknown outbreak. The idea here is that we not only want to be able to identify the organism, potentially

using metagenomic sequencing, but can we simultaneously look at host response.

And these are actually an example. These are 5 cases, again, spinal fluid from 5 cases of meningitis and encephalitis that were essentially unknown. So like the first example is the case of MTB. That was negative by sequencing. It was culture negative as well.

You can see here clearly a bacterial infection. A case of GABA receptor (inaudible) encephalitis. This is an auto antibody syndrome that was read as a noninfectious condition. If you look on the bottom, a case of Zika virus that was read as a clear viral meningitis.

And that's actually quite interesting given that Zika virus was actually not represented on the original model. So it gives you, perhaps, the hope that should we identify, say, a novel viral pathogen in the next, say, 10 to 15 years, which we may indeed do, this type of approach would have ability to detect it without actually having a specific diagnostic test targeting the

genome or the pathogen itself.

And this is currently what we're envisioning would be the results report, is not only would you get a metagenomic result. So this is a case of Lyme disease, again neuroborreliosis. So not only can you detect in principle the Borrelia burgdorferi bacterium, but the idea is could we couple that with the host response to be able to characterize not only the infection itself but also the patients host response to the infection.

Okay. I wanted to give one last example on terms of how we've actually been able to apply this to an outbreak. So in October of 2021, four recipients of organs from a common deceased donor in the United States, developed of febrile illness and encephalitis 2 to 6 weeks after transplantation.

Essentially, the blood donor provided blood to the tissue organ donor and the deceased donor then donated organs to various recipients. And unfortunately, two of the recipients ended up

dying from this because they developed essentially post transplant encephalitis.

And so the key question was, you know, what is this. And we actually identified this as a case of yellow fever virus. And actually, the way it was identified was a detection of a single read in a patient who, in an infected patient, and it was this detection of a single read that precipitated a CDC investigation, wherein they found out that the blood donor had actually received the live yellow fever virus vaccine 3 weeks prior to donation.

And it had turned out that the blood donor had then donated to the deceased organ donor, and then had 2 of the 4 recipients developed a fatal encephalitis. And eventually, this was actually proven by antibody testing, and unfortunately, autopsy testing of the brain.

But one thing to note, though, is that we detected a single read out of 15 million reads. Now we could have easily missed this by metagenomic sequencing. But what are we going to

do in the future? Well, what we can do is we can, again, now do kind of the host response analysis. And this data is now published where you can see that all of the affected recipients, and these are the recipients that receive organs other than the cornea, the cornea recipient had no symptoms. But the other recipients who were affected and were symptomatic all had a very strong viral signature.

And what's quite interesting as well is in the viral signature, again, I've highlighted with the asterisk there. Those are actually genes that have been linked to association with either yellow fever, vaccination or infection.

So in hindsight, let's say that we had not detected with metagenomic sequencing. Well, we actually have another shot at potentially being able to identify the cause of this outbreak by host response analysis.

So I'd like to end by thanking the members of my lab who were involved in this work, funding from the NIH, BARDA, the CDC, the Delve Bio, and the Steven and Alexander Cohen Foundation

for their Lyme disease support. My website is there and thank you for listening.

DR. MCGIVERN: Thank you, Dr. Chiu, for your informative presentation. We're open for questions, but while people are thinking of questions, I'd like to ask a question myself. What's the timing of the results and the analysis, and are they fast enough for clinicians to make decisions based on the result?

DR. CHIU: That's a great question, and I didn't get into that. So metagenomic sequencing, at least the metagenomic sequencing that we run on lumina instruments in the clinical lab, has a turn around time that varies from 24 to 72 hours. And that's an interlaboratory turn around time.

The actual, and we actually have a paper which is shortly to be published in Nature of Medicine that goes over our seven-year experience. But in short, the actual turn around time ends up being somewhere on the order of like 4, 3 to 4 days. And that's primarily because there's also

time incurred upon having the sample sent to the lab, the need to batch samples.

So unfortunately, it becomes there is a delay in kind of the turn around time. And as you know for infectious diseases, you really want to minimize that turn around time. We are indeed trying to develop methods using, for instance, nanopore sequencing that will help us get same day turn around time.

But with respect to the host response though, or metagenomic profiling, we're also very interested in developing sort of leveraging this information. For instance, I'm interested in transferring these models over to a very rapid testing model. So we're currently working with Thermo Fisher on TaqMan array card, which is essentially a multiplex RTPCR type of platform, where we can get the same sort of results within 2 hours.

But currently with sequencing itself, we have been able to do same day sequencing using a nanopore sequencing platform. Thanks.

MR. KUMAR: Yeah. Hi. This is Sanjay Kumar from Office of Blood at CBER. Charles, excellent as always. So my question here, I think more extension to something you answered. Do you have a pattern emerging from pathogen specific immune response, then you can associate with acute disease, to help (inaudible) management, and also long-term disease, chronic infection.

You know the turn around time, the early most is very rapidly 2 or 3 days. I mean patient is dead really. So can you pinpoint the immune responses that can really aid in clinical diagnosis? And I would listen.

DR. CHIU: Yeah. I think you have like two excellent points there. So the first question is, you know, is it useful to have a test that's like just a few days. Now, actually, you'll actually see in our paper, it was useful.

So part of the reason it may also be the clinical syndrome. So for instance, encephalitis, I mean, these patients are in the hospital for weeks and in some cases months. So there is, and

many of them as I said, more than 50 percent of them actually end up being discharged from the hospital without a diagnosis.

I mean, they're empirically treated with antimicrobial agents or autoimmune, and immunomodulatory agents. And yet we still don't have a definitive diagnosis. And so my feeling is that metagenomic sequencing by itself even having a turn around time of say 2 days, or 1 to 2 days, is still going to be clinically useful.

We've been able to show that actually. That need for diagnostic yield. You know, that being said, it's always better to actually be able to kind of get results in a more rapid fashion, and that's why my interest has been in potentially moving these onto portable platforms, either highly multiplexed, you know, RTPCR or PCR platforms that will allow us to get kind of the 2 hour turn around time. Or at least the same day turn around time that you really need for acute infections.

On the other hand, for chronic diseases,

so you introduced another important point in that we found that the host response differs for, for instance, chronic manifestations of acute diseases versus acute disease themselves.

So for instance, a good example is the host response to COVID-19 infection is quite different than that in long COVID patients. And so what I really feel is that we would need to develop assays that target both the acute stage and the chronic stage of these diseases.

And so a related study that we're currently doing, for instance, we're identifying, again, using RNA sequencing as well as cytokine profiling, we're identifying biomarkers that can distinguish between long COVID, chronic Lyme disease, and myalgic encephalomyelitis chronic fatigue syndrome.

So we are trying to address that as well. But it seems to me that it's very likely the signatures are going to be different.

MS. ELKINS: Anybody else in the room, I have quite the list coming in from online, so I'll

try to do them justice. First, how do you avoid bias in MGS because the genome copy number of the pathogen of interest is very low.

DR. CHIU: Okay. Yeah. So yeah, it really has, to date, it's primarily been developed as a qualitative test. Meaning it's either that you detect it or not. So the idea is that you have to set certain thresholds. And so these are obtained empirically. We essentially use ROC curve analysis to identify what is the optimal threshold where we call something positive or not.

We have not been using this in a quantitative fashion, which is what the questioner might be asking about. But we are actually working on that, and it does appear that it may be potentially we might be able to obtain quantitative information. So in fact, not only be able to detect an organism by sequencing, but also be able to identify the viral load or the pathogen load.

But we currently have developed this as a qualitative tests with specific preestablished

thresholds.

MS. ELKINS: So that may relate to a follow up part of that question. The host response obviously overlaps in several pathogen infections. So how do you avoid misinterpretation or ambiguity? But if you consider it qualitative, that may be part of the answer.

DR. CHIU: Yeah, yeah. But that one is actually quite interesting, because it is true the host response does overlap. But we found out that you can identify differences. So a good example, for instance, is one thing that we found out is we tried to find a differential host response between coxy infection and histo infection, both dimorphic fungi that causes similar illness.

And we were unable to. So what we ended up doing is we sort of combined it into what we called dimorphic fungal infection. So I think that as we get more experience with this, we'll know kind of at what level resolutions to sort of classify these responses.

But another good example would be acute

flaccid myelitis, which is we know it's caused very likely by certain strains of enteroviruses, where that appears to be very specific, it's not the same as enteroviral meningitis. So I think over time we'll have better, more experience in terms of how to classify these types of infections.

MS. ELKINS: Thank you. Next, have metagenomic methods or analyses been applied to controlled human challenge models or subjects immunized with attenuated, live attenuated vaccines?

DR. CHIU: It has been. So there have been studies, for instance, done in challenge models of acute infection like the, metagenomic sequencing has been done, for instance, in college age volunteers who were basically infected with influenza, for instance.

So the answer is yes. And host response profiles have also been done in challenge models, as well as in animal models as well. I didn't talk at all about the literature, but my interest

is, of course, is being able to kind of translate that to clinical use.

MS. ELKINS: Next, have you considered host genotype, for example, in order to determine what aspects of the host response detected are functions of the baseline characteristics of the host versus acute response to infection?

DR. CHIU: Absolutely. And in fact, that also raises kind of the possibility that we could sort of do everything with a single test, because with clinical metagenomic sequencing, we sequence both DNA and RNA. I didn't talk at all about the DNA side.

We have been able to show, for instance, that we can use DNA analysis sort of a liquid biopsy concept, where we can identify, for instance, essentially a copy number of variants to be able to, potentially be able to identify leukemias or other cancers or malignancies on the DNA side.

So certainly, there's a lot of, and there's a lot of interest, for instance, in

looking at methylation profiles, cell and tissue types of origin, as well, to be able to better diagnose disease. So I think there's a lot of room to mine in that area. Our focus has really been on RNA to date, but we eventually want to incorporate DNA information as well.

MS. ELKINS: So the next one is fairly open- ended, and I'll warn you that I can't keep up here with the influx. Would you elaborate on how to achieve machine learning accuracy?

DR. CHIU: Yeah. With machine learning accuracy, the best way to achieve accuracy is actually numbers, well annotated data and more of it. I can tell you when I spoke with my (inaudible), when we first tried to develop it, he wanted 100,000 cases of meningitis encephalitis.

I told him I'm not giving you without 100,000. So we're going to have to make do with fewer cases. And my feeling is especially for emerging infections, you're not going to have a lot of cases to test with. But there actually have been kind of advances in both, in not only in

diagnostic medicine, but also in AI and machine learning.

There's a technique called GANs that we're currently applying. It's called Generative Adversarial Networks. And it's essentially a method by which you can essentially use machine learning, but it applies for very small numbers of samples. So for instance, 10 to 15 samples, potentially you can generate fairly good data by virtually expanding it using this technique.

So there are some advances, so parallel to some of the advances that are being made in AI and machine learning, we're hoping to adopt that for use in diagnostic medicine.

MS. ELKINS: Got it. Does your model account for age-related changes in immune responses to different classes of pathogens?

DR. CHIU: It currently does not. And again, it's because we simply don't have as many samples. But certainly we we'd want to do that. But again, it's difficult if you don't have a lot of cases to actually work with.

But eventually, there would be the possibility that we could stratify the data as well by age. We haven't done that to date. By either age or sex.

MS. ELKINS: Next, leave out one testing is still susceptible to one of the samples being wrongly diagnosed. So have you challenged the methods using a wrongly diagnosed sample?

DR. CHIU: Yeah. So what's interesting about wrongly diagnosed samples is that that appears to be, again, that's sort of dealt with by having numbers. So for instance, in fact, I even know in some of the machine learning models that they're either wrongly diagnosed, or they might potentially be wrongly diagnosed.

But it's important that you sort of minimize those. And in fact, when we do actually do the training set, we want to make sure that that's less than 10 percent of your samples. But it turns out that if the number of, quote, wrong classifications that are made a priority, if it's less than 10 percent, it doesn't appear to affect

the performance of the final model.

It seems to be dominated by the well performing samples. But it is something you definitely have to take into account, especially because for many of these diseases there's no gold standard.

MS. ELKINS: This may be partly answered by the fungal example, but can the approach be used to distinguish more than two genetically closely related viral infections, such as consecutive flavivirus infections?

DR. CHIU: We have tried to do that. We've actually tried that. It doesn't appear to work very well, probably because the flavivirus responses are very much in common. So we tried, for instance, for distinguishing West Nile virus meningoencephalitis from Powassan virus. And we were unable to do so.

So I mean, this is clearly a limitation. If it has a similar response, similar pathogenesis, it's probably going to have a similar host response as well. So it's really

being able to identify which responses are specific and which or not. And I think generating more data will help us do so.

MS. ELKINS: Is there a risk that nucleic acid sequences from a resolved infection can result in a false positive?

DR. CHIU: No. It's more likely that host responses from a resolved infection, so one thing I should mention is host responses from a resolved infection we found are clearly not the same as from an acute untreated infection.

So the best time to actually do host responses is as early as possible when a patient is maximally symptomatic or is untreated. If you have a patient that's successfully treated, we find that the host response disappears very quickly. At least the host response of infection.

There may be a host response of treated infection or partially treated infection. We haven't explored that. But it's more likely that a resolved infection is going to yield a false negative host response, rather than a false

positive.

MS. ELKINS: Now we're starting to overlap. Let me see if I can combine a bit. Once you identify host markers by genomics, can you use them as a simple blood assay?

DR. CHIU: Yes, yes. In fact, we are doing that. Our hope is to move it to a rapid RTPSR platform. And we're working with Thermo Fisher on developing a TaqMan array card platform for host response assay that could be with a two hour turnaround time. Yes.

MS. ELKINS: And what are your thoughts on increasing the specificity overall?

DR. CHIU: Increasing specificity, I think, is important, but my feeling is that to some extent there's going to be a limitation in that. Because again, we're limited with our access to essentially gold standard samples. You know, unfortunately, you know, biology happens to be somewhat, just happens to be somewhat ambiguous in the sense that very rarely do we get clear cut cases.

And in fact, I would argue that it's actually better to have these sort of ambiguous cases. So for instance, to give you some example, the host response is never, is very rarely as clear as you would like it, for instance.

So you have a patient with a fungal infection may have, for instance, the host response that suggests a bacterial fungal infection. We know that these host responses overlap. A viral infection overlapped quite a bit with autoimmune and noninfectious etiologies.

And so my feeling is that we want to incorporate all of that data. And we don't necessarily want to develop a test that has 100 percent specificity, because our feeling is that's actually going to be much less useful. Those are not the cases you encounter in real life.

MS. ELKINS: Can you describe a threshold for the number of reads required to designate a sample as positive?

DR. CHIU: Yeah. So that's been published. And if you look in our literature,

we've already established thresholds, and these thresholds are established in the validation. So for instance, we found, for instance, that with metagenomic sequencing having just 3 reads, non-overlapping reads, to 3 different viral genes, is sufficient to specifically identify a virus.

Over 7 years, and this is a paper that we have published. I'll be happy to provide the reference. The specificity of our metagenomic assay has been 99.6 percent. So it's very, very high.

MS. ELKINS: All right. And we have already discussed age groups. So last but not least, can you discuss how close you're getting to finding a biomarker for long COVID?

DR. CHIU: That's a great question. I think it's not just us, but like several groups or many groups are actually working on this. I would say that we're working on it. We definitely have what would appear to be a cytokine profile that can distinguish long COVID from other conditions.

My feeling is that you don't necessarily

want a biomarker for long COVID specifically, per se, you want a biomarker that the distinguished long COVID from mimic or look alike diseases, including chronic fatigue syndrome and Lyme disease. So we're working on that. Thank you.

SPEAKER: Last question, I guess, can you comment on whether you went back to those samples that were not diagnosed and you use your models to learn whether there are some viral, let's just put an example, you learned that your model shows that it's a potential viral infection, you went back to those samples and look for potential mark or genes that will show that they are viruses that were not yet identified.

So in other words, have you data mined your samples that you were not able to track traditional pathogens?

DR. CHIU: Yeah. We haven't done that yet. Because it's sort of a chicken or egg problem. Right. Because you can identify a signature that may suggest a novel viral infection. And yet, unless you can actually

culture, or no one's going to believe you, unless you actually culture, are you able to identify the pathogen itself.

One thing I can say is that we have been able to identify, for instance, a viral signature associated with emerging pathogens. So I don't talk about it in my talk, but we identified, for instance, two new bunya viruses or emerging bunya viruses. One is Potosi virus and Lone Star virus.

And those bunya viruses are interesting because they have, to date, they have not been associated with human infections. They were identified by doing mosquito and tick sequencing. And so we've actually identified them in a patient, and when we looked at the host response, it was very clearly a bunya viral host response.

So it suggests that in the future this might be a useful technique for identifying new viruses. Thank you very much.

DR. MCGIVERN: Thank you, Dr Chiu. So our next speaker is, it's a pleasure to welcome Dr. Shirit Einay. Dr. Einay is a physician

scientist in the division of infectious diseases in the Department of Medicine and in the Department of Microbiology and Immunology at Stanford University.

Her research program focuses on understanding the roles of virus host interactions in viral infection and disease pathogenesis. Dr. Einav's research program also includes translational efforts to apply this knowledge for the development of broad spectrum, post censored, antiviral approaches to combat emerging viral infections, and methods to predict their progression to severe illnesses.

Please join me in welcoming Dr. Shirit Einav.

DR. EINAV: Thank you, David, for the introduction, and Karen for inviting me to be here today.

MS. ELKINS: And I'm sure our AV folks are good. There you go.

DR. EINAV: Thank you. All right. So it's a pleasure to share some of our work in the

area of targeting cellular kinases to combat multiple emerging viral infections today. So there are over 200 viruses that are known to cause disease in humans, and of course, the list keeps growing.

And when you look at currently approved drugs for treating viral infections, we're quite limited. We're essentially targeting around 10 to 15 viruses, depending on how you count. So clearly, there's a very large unmet need in this area.

The prevailing antiviral approach relies on targeting viral function, specifically viral and schematic functions, usually. And this approach has shown a tremendous utility in treating chronic viral infections, such as hepatitis C virus, and HIV.

But its utility for treating emerging acute infection has been more limited. So the spectrum of coverage that's typically provided by this approach tends to be narrow, and this is, of course, due to the high diversity of viral protein

structures and sequence.

And so it's usually in the order of few viral serotypes all the way to few related viral infections. We know that it takes about 8 to 12 years to develop a novel drug, and a cost of over $2 billion. So combined with the narrow spectrum, this makes the development of such approaches quite a slow and expensive process.

Of course, this is not scalable to meet the large unmet clinical need that we are observing with emerging viral infections. And if you take (inaudible) here, for example, the active moiety of Paxlovid, yes, there was a very rapid rollout to the clinic, but this was not a new chemical entity. In fact, it relied on the revitalization of already existing compound that Pfizer developed for the original SARS COVID.

And since we all know we cannot predict the future threat, this approach does not provide the global health protection we need and the national security readiness we hope for.

Treating viral functions as monotherapy

is also often complicated by emergence of viral resistance, and in the context of HCV or HIV, we solved this issue by using cocktails of individual drugs that target distinct functions. But again, this is not a feasible approach for emerging viral infections.

So when I was a post doctoral fellow at Stanford, I was mentored by Jeffrey Glenn and Stephen Quake, and I was focusing on a nonstructural transmembrane protein and 4B protein and hepatitis C virus. This is in the period where both David and I were working in the field, and there were no antivirals available at the time.

One of my project ended up with the discovery of the novel target, and DAI, direct acting antivirals against these targets. So while there was a lot of excitement with that came a disappointment because this compound essentially targets mostly genotype 4, which is not prevalent in this country.

And treatment was associated with rapid

emergence of resistance. So when I established my own lab around 2011, I decided to try to think a little differently about targeting viral infections and moved to focus more on targeting host functions.

So as you all know, viruses replicate in a host, and they rely on cellular machineries for every step of the viral life cycle. So my goal here was really to try to identify those cellular functions that are required by multiple viruses and use those as potential targets for broad spectrum antivirals.

The hope was that this would also introduce an opportunity to repurpose already approved drugs that target these functions. So combined with the broader spectrum, this could reduce the cost and time for drug development and provide readiness for future outbreaks.

And then since cellular functions are not under direct control of the virus, genetic control, this may also increase the barrier to resistance. Another potential advantage is that

the repertoire of targets that this approach provided larger than when direct acting antivirals.

So viruses typically have anywhere from 10 proteins, if you take virus like dengue, all the way to 200, if you take herpes viruses, majority of these proteins are not enzymes. So not surprisingly, current approaches rely on largely targeting viral proteases, or polymerases.

Whereas if you move to host targeted approach, now you have a repertoire of 20,000 the whole human proteome. So with that comes a lot of diversity and indeed, multiple groups, including ourselves, have characterized very broad cellular functions that could be targeted.

So how do you discover such host targeted broad spectrum antivirals? So I'm going to share two examples. This came from our work and the first goes from the unification of the target, then to the discovery of the drug.

So the approach that we've taken was to focus on cellular proteins that are involved in

intracellular membrane trafficking. The thinking behind this was that studying these virus cells interactions would not only solve some underexplored areas in viral infection but will also potentially identify such common mechanisms shared by viruses, because every virus essentially has the traffic intercellularly during their life cycle between various intracellular compartments.

So in my lab we've been focusing on various cellular machineries that are involved in regulating intracellular trafficking of viruses. And one of our earlier projects focused on interactions between viral proteins and a group of adapter proteins shown here are AP-1, AP-2, and (inaudible).

These are cargo specific adapters that are involved in either cluttering mediated or other intercellular trafficking pathways. And what we've shown is that viruses such as HCV use these adaptors sort of as buses to shuttle intracellularly. And what you're seeing here are live cell imaging of individual viral particles

labeled green via tetracysteine labeling.

And you can see that the green goes together with red, which is (inaudible) labeled adaptor proteins. We then discovered that there is a group of kinases called NUMB associated kinases. There are 4 kinases in this family listed here by (inaudible) and SDK 16. And these kind of stimulate those virus source interactions.

So what we've observed is that they could basically regulate those temporally distinct stages of the viral life cycle, where these virus associated interactions happened. So they could almost be viewed as master regulators of viral infection and perhaps attractive targets for broad spectrum antivirals, because we've seen this across multiple RNA viruses.

So this is sort of the model that we proposed and then validated in a large number of papers. Some are shown here. Interestingly, we discovered that there are already existing kinase inhibitors that are very potent suppressors of these kinases, even though these kinases may not

be their primary target.

So for instance, erlotinib, anti cancer drugs, erlotinib is known as an EGFR inhibitor, but if you look at its kinome profile, it actually targets GAK, one of the NUMB associated kinases with very high potency, with KD of 3 nanomolar.

And indeed, we've demonstrated that this is, some of these compounds essentially elicit in vitro activity, again, 7 viral RNA families. We've also demonstrated in vivo activity in a dengue mouse model, this is data that was generated in my lab, and in John Dye's lab, in an Ebola mouse model. So in both models we saw protection from lethality as well as separation of viral replication.

So this data was generated right during the time of the 2014 Ebola outbreak in West Africa, and it was selected by the Gates Foundation to be included as 2 out of 3 arms of an adaptive clinical study that was planned to be launched.

Sadly, by the time we had access to the

drugs and IRB approval and everything, the outbreak started to resolve, which is good for the outbreak, not good for the approach. But this was sort of then saved as a drawer protocol. This technology was also licensed by a local biopharma, but raising funds by local investors didn't work out, so this was never tested clinically.

Of course, when COVID hit, we tested this against COVID selectivity, and this is for SARS-CoV-2 were higher. So this was not relevant, but some selective, chemically distinct NAC inhibitors that we've been developing have shown some promise.

So I'm going to focus the rest of the talk on the more recent work that we've conducted and here, this was early in the pandemic when like everybody else, we were also very motivated to identify some repurposed approach that may be able to suppress the replication of this virus.

And Marwin and Sirie (inaudible) to the lab at the time, we're spearheading this project. So what we did was to approach the Stanford high

throughput facility and see what's available. We wanted to focus on compounds with known targets, and active bioactives, and FDA approved drugs.

And you could see the libraries that were selected. We complemented this collection with self assembled set of kinase inhibitors. And overall we screened over 3,500 unique compounds. The assay we used was based on immunofluorescence. These are Vero cells that are labeled with GFP.

We then infect them with SARS-CoV-2 in the presence of the compound, and then read the GFP signal after 4 days. And what you could see here is the cumulative data. So uninfected cells on the left, as you can see, there's high signal, because these cells are alive. Upon infection in the absence of treatment, essentially there's complete cell lethality.

And on the right, as you could see, the majority of the compounds showed no effect. However, outliers indicated in those colorful dots show essentially rescue of cells from SARS-CoV-2 lethality.

So if you zoom on this group of compounds, you could see the list here, and 7 of them have shown activity in sub micromolar range when it was tested via plaque assay. These compounds essentially target diverse cellular functions, and for the rest of the talk, I'm going to focus just on one of them called lapatinib.

So what is the therapeutic potential of lapatinib for SARS COVID 2 infection was of course the first question we asked. And what you could see here is that those response curves, now this is the plateau stage in Vero cells on the left, and in (inaudible) cells on the right.

The black curve demonstrates effect on antiviral activity, and the blue curve shows the effect of the drugs on cytotoxicity, and you could see those response effect with EC50 values at.5 and.7 micromolar, and EC50 greater than 20 micromolar.

Lapatinib was able not only to suppress viral replication, but also to rescue cells from SARS- CoV-2 induce laterality in this orthogonal

fluorescence based essay. To study lapatinib in a more biologically relevant model, we collaborated with (inaudible). I'm sorry there's some technical formatting issues here. Sorry.

So we used an organoid model that was developed by the DAS Lab. These are essentially stem cells derived from lung samples that are then differentiated into proximal and distal lung epithelial cells. They're grown in monolayers and can be propagated. And what you could see in this model is again those response effect with EC50s that were essentially comparable to those measured in the cell lines.

And this is how this looks via immunofluorescence. You could see in the center that cells infected with SARS-CoV-2 demonstrated staining for the nucleocapsid labeled in green and lapatinib essentially reduced this staining.

More recent data, in collaboration with Luis Martinez at Texas BioMed, used in vivo model, these are wild type mice, and they're inoculated intranasally with mouse adapted SARS-CoV-2, that's

labeled with (inaudible). And this is done in animals are treated with 200 mg per kg of lapatinib given via oral lavage daily.

This is below the approved dose that's given to patients, to cancer patients. And treatment was given over a course of 4 days post infection, starting a day before. So animals were first subject to bioluminescence measurements. And as you can see here, infected animals demonstrated increased bioluminescence in their lungs. And treatment with lapatinib reduced this signal, particularly after 4 days of treatment.

And these are controlled uninfected mice. Lapatinib also demonstrated an effect on weight loss, as you can see on the left. And it also reduced mortality in this animal model. And this effect correlated with suppression of viral titers measured in the lungs and in nasal turbinates, as you can see 4 days, that was essentially close to the lower limit of detection of the assay.

So basically this shows that lapatinib

suppresses SARS-CoV-2 infection in vitro, in the human lung organoid model as well as in this mouse model. So we of course were curious to determine what is the relevant target. Now, lapatinib cancer targets are ErbB1, which is EGSR, and ErbB2. But when you look at the kinome profile of this compound, it also appears to bind 5 additional kinases.

One of them is ErbB4. So what are these ErbB's, these are receptor tyrosine kinase, they're on the cell surface, and they form homodimers or heterodimers. We did confirm binding of, I'm sorry, catalytic activity of lapatinib against ErbB4. You could see with an IC50 of 28 nanomolar in this in vitro assay, which was essentially comparable to its activity on the cancer target, ErbB2.

So we then asked which of these 7 candidate targets may be relevant for SARS-CoV-2 infection. And to start addressing this question, we conducted an siRNA base screen where we depleted cells for these kinesis individually and

measured viral replication.

This was done initially in Vero cells, and you could see that suppression of the individual ErbB's reduced viral replication. This was then confirmed in Calu-3 cells, and you could see about lock reduction in viral replication via plaque assays.

So it appears that ErbB's are required for viral replication. To further confirm this, we also studied the effect of two chemically distinct pan-ErbB inhibitors, ibrutinib, which is another approved kinase inhibitors, and sunitinib, which is investigational. And once again, you see those response curves with no toxicity in the concentration range used.

So this shows that ErbB's are required for SARS-CoV-2 infections. We validated them as potential antiviral targets, both genetically and pharmacologically. And this is truly a class of compounds versus a single ErbB inhibitor that were discovered.

To determine whether this antiviral

effect correlates with functional suppression of ErbB's, we looked at the phosphorylation level of the ErbB's in infected cells and treated cells, and on the bottom you could see the nucleocapsid staining could see that lapatinib does dependently suppress the nucleocapsid expression as shown in the orange curve.

And then lapatinib also dose dependently suppressed phosphorylation of the ErbB's as measured by looking at both the phosphorylated form and the total protein form and measuring the ratio.

So this shows that we're seeing target engagement and modulation in this concentration range. To determine whether ErbB's are targets that are mediating the antiviral effects of lapatinib, we conducted these rescue assays, and by mistake I included the data with Venezuelan equine encephalitis virus. But we showed similar results with SARS-CoV-2.

So here we're ectopically expressing wild type ErbB4, or catalytically inactive ErbB4,

and what you could see in the green column is that wild type ErbB4 was able to rescue or reverse the antiviral activity of the compound, whereas catalytically inactive ErbB was unable to do so.

So this truly establishes the link between sort of the molecular target and the antiviral effect. So I've shown you that ErbB's are validated antiviral and molecular targets of lapatinib.

So what is the mechanism of antiviral action? This is still a subject for ongoing work, but I'm going to share some of the findings. So first thing we tried to do is to pinpoint a specific stage of the viral lifecycle that lapatinib suppresses.

And to do so, we conducted time of addition experiment where the components added at different time points in the course of a 10 hour experiment, which is equivalent to a single viral replication cycle in this model. What you could see here is that when the compound was present throughout the 10 hours, 0 to 10, there was

significant suppression of viral replication as measured in cell culture supernatant.

But by plaque assays, when the compound was present in the first 2 hours, there was significant suppression, suggesting that the compound may be suppressing early stages of viral entry. But when it was added at later stages of viral replication, such as the last 5 hours, or even last 2 hours, we still observed an effect indicating that the compound suppresses temporally distinct stages of the source copy to life cycle.

And this, of course, has potential implications for drug resistance and so on. We probed the role of ErbB's in viral entry, and again just sharing some of the data we demonstrated, it's difficult to see, but we're tracking here individual SARS-CoV-2 particles that are labeled green by staining the nucleocapsid.

And we're asking how are these internalizing into the cells by looking at their colocalization with Rab7, which is an endothermal marker. SARS-CoV-2 is known to internalize into

this compartment. And what we could see is that lapatinib suppressed the colocalization of viral particles with Rab7, suggesting that it impairs internalization.

We also looked at internalization of the receptors. So if you look on the right, you could see that the red arrow points to the effect of SARS-CoV-2 infection on a level on the cell surface measured by flow cytometry. And you can see that infection causes internalization of the receptor because the virus uses it to enter cells.

Lapatinib reverses this effect, as shown with the green arrow. The same pattern we observed with ErbB2. So it appears that lapatinib suppresses the internalization of both receptors. And then we also demonstrated binding of SARS-CoV-2 spike S1 domain on 2 ErbB's via coimmunoprecipitations.

So at least some of this shows that ErbB's regulate SARS-CoV-2 internalization and viral induced receptor internalization. And that this effect may be mediated, at least in part, via

this interaction of virus source interaction.

So ErbB's are activated by a large number of ligands, and they're known to signal in pathways, including p38 map kinase, ATTM (inaudible). These are pathways that were studied extensively in the context of acute lung injury and lung fibrosis in noninfectious models. Such as neomycin, idiopathic fibrosis, and so on.

So to determine what is the impact of SARS- CoV-2 infection and ErbB suppression on these pathways, we studied the phosphorylation of some of these downstream factors, and what you could see is that at 1.5 hours and 24 hours post infection, we see induction of phosphorylation of these downstream signaling factors by infection.

So if you compare the second to this lane, to the lane on the first, the first lane, you could see induction by the virus. And importantly, suppression of ErbB reduce this phosphorylation.

So it appears that lapatinib suppresses phosphorylation of signaling pathways downstream

of ErbB's. And as I mentioned, these are signaling pathways implicated in injury and fibrosis.

We then looked at the effect of lapatinib on cytokine production and this is again done in the adult lung organoid model. You could see that infection with SARS-CoV-2 increased production of proinflammatory cytokines, including TNF-a, IL-1B, and IL-6 as shown by multiple groups before.

But lapatinib treatment does dependently suppress production of these inflammatory cytokines. Lapatinib treatment also demonstrated effects on tissue injury in these adult lung organoid models, as you can see that in uninfected ALO's, we see very nice architecture of what we call alveolar like structures, and intact membrane staining of clotting-7, which is a tight junction molecule.

Upon infection with SARS-CoV-2, we see complete disruption of this architecture and membrane staining, whereas lapatinib treatment

restored a normal phenotype. So this shows that lapatinib suppresses SARS-CoV-2 induced inflammation and tissue injury.

And what I argue is that this is not only by suppressing viral replication, but also via a direct effect through its activity on those signaling pathways that I mentioned. And indeed, a paper in idiopathic pulmonary fibrosis mouse model has shown an effect on lapatinib. This is a noninfectious model, of course, and you could see effect on fibrosis.

To study the utility of lapatinib in combination drug treatment, we studied its effect on SARS-CoV-2 infection in combination with existing direct acting antivirals. And whereas in combination with mulnupiravir and remdesivir, it showed additivity, its combination with nirmatrelvir, the data I'm showing here, showed significant synergy.

So what you're looking here, we treated cells with either compound alone or with multiple combinations to create this checkerboard matrix,

as you can see on this Z and X axis. And what you're seeing is this mountain or peak above the plane of additivity. This is using a max synergy algorithm, and this is consistent with synergistic effect.

And when we quantify the volume under the curve, it's in the range that is predicted to be significant in vivo. And this synergistic antiviral effect was not a result of synergy in toxicities. You could see on the right showing plane of additivity.

So this is quite exciting because lapatinib and nirmatrelvir combinations have the potential to perhaps reduce Paxlovid rebound, which currently complicates about 40 percent of treatment courses. It may help address resistance concerns that have emerged, yet, thus far, mostly in immunosuppressed patients.

It may allow to reduce the dose. It may provide better tissue injury protection due to this direct effect that I've mentioned. And since lapatinib by itself is a CYP3A suppressor, it may

overcome the need to include ritonavir.

So in terms of the broad spectrum potential, we have tested lapatinib against thus far 5 viral families, and it has shown activity. The majority of these are antiviruses with the exception of pox virus, which is, of course, a DNA virus.

Just to share some of the data. Alpha viral infection model diseases, Venezuelan equine encephalitis virus transmitted by a mosquito and causes severe encephalitis. It's also a biothreat agent since it could be aerosolized.

Here, we collaborated with Artin Narayanan and George Mason University, and using a neurovascular unit model that attempts to capture the 3D blood brain barrier architecture, we observed a log reduction or so of viral replication. This is treating a donkey, which is the wild type VV virus, both on the vascular chamber and on the brain chamber across this BBB like membrane.

Similarly to what I've shown you in the

SARS-CoV-2 model, here again we're seeing effect on reduction of proinflammatory cytokines, as well as protective effect on BBB integrity, as measured here, the FITC-dextran permeability assay.

In a mouse model, the Narayanan lab has demonstrated protection with lapatinib against the V challenge. In this case, it was the vaccine strain of VVTC83. Wild type mice that were treated again with lapatinib had protection from lethality and this effect correlated with several log reduction in viral titers, both in the serum and the brain.

So this is yet a second unrelated viral disease model where we were able to show activity both ex vivo and in vivo. And to determine whether viruses could escape treatment with lapatinib we passaged VV in the presence of lapatinib, or a direct acting antiviral, ML336. And what you could see in the pink curve is that virus that was passaged under the DAA essentially was suppressed initially.

But then this was followed by a rebound,

which coincided with the emergence of a resistant mutation in the nonstructural protein that's known to be suppressed by these compounds. And in contrast, virus that was passaged under lapatinib continued to be suppressed.

When we obtained the supernatants that was derived after the tenth passage and used that to inoculate naive astrocytes, we saw that virus that was passaged under lapatinib continued to show susceptibility to lapatinib, whereas the virus that was passaged under ML336 essentially lost its susceptibility to the direct acting antiviral.

So this sort of shows that the genetic barrier to resistance of lapatinib is higher than direct acting antivirals. And this is something we have seen with some of our other host targeted approaches.

So the model that we sort of proposed and have been validating is that, as I've shown you, ErbB's are involved in regulating existing stages of viral life cycles in the context of

SARS-CoV-2. I showed some of the data in viral entry. We're looking at other stages right now.

Moreover, they regulate signaling in those downstream pathways that are implicated in tissue injury, acutely and chronically. So by suppressing those ErbB's, we could sort of achieve a triple role, not only suppressing viral replication, but also the inflammatory signals, as well as potentially protect from tissue injury.

I don't know how we're doing with time, but I can -- so yeah, we're good. Okay. So what is the repurposing potential of lapatinib. So it is approved globally for advanced ErbB2 positive breast cancer in combination drug treatment with antimetabolites or antiestrogen treatment.

It is dosed orally once daily. And it appears to be accumulated in the lung. And this is something we've seen in our mouse model. In terms of its safety profile, the most common side effects in the first days of treatment is diarrhea. And yes, there have been severe side effects that were reported, including liver

toxicity and QT prolongation.

However, these were always in combination with drug treatments, with antimetabolites in sick metastatic breast cancer patients, and not within the first month of treatment. And there is some concern about drug interactions. As I mentioned, it suppresses the cytochrome P30A4.

And its synthesis can be scaled up easily. It is a small molecule. So what I've shared with you was a couple of proof of concept for the potential utility of targeting cellular kinases to combat acute emerging viral infections and propose a few approved drug candidates for repurposing either individually or in combination drug treatments.

But of course, there are numerous challenges. And maybe this is something I can leave for the discussion. So we're far from achieving the goal of having an established (inaudible) portfolio for pandemic preparedness.

But when we think about the future,

perhaps a best approach would be to both develop and stock host targeted broad spectrum antivirals as the first line of defense. And then couple that with direct acting antivirals, and ideally if there was a portfolio, NIH is currently trying to create where we have small molecules for each of the major viral families that could accelerate rollout, as we've seen with the case of nirmatrelvir.

And I'd like to end by thanking the people in my group, particularly Sirle Saul and Marwah Karim, who've done the large bulk of the work. Our funding agencies and collaborators. And thank you again for the invitation and for listening.

DR. MCGIVERN: It's time for questions now. While people are thinking of the questions, I'd like to ask a question. So how late can you administer these drugs in the animal models? Can they be quite symptomatic?

DR. EINAV: Thanks for the question, David. Yeah, thus far, we've essentially studied

them at the prophylactic model. I mean, we administered them 12 hours prior to inoculation. We are currently actively doing the postinfection, you know, studies. So that's definitely going to be more biologically relevant for treating acute infections. Yeah. Thank you.

DR. MCGIVERN: Thank you. Do we have a question in the room?

SPEAKER: Thank you for that. Thank you for that. It was great talk. You know, we are Center of Biologics, and so we are interested in, you know, protein drugs. We see combination of small molecules. Has anybody tried mixing monoclonals with small molecule, and whether that would be beneficial?

And then another corollary to that, have you looked into using this compound against, for example, hepatitis C, where cirrhosis and fibrosis is a big issue? And so this would be a well suited compound for that.

DR. EINAV: Thank you for the question. Yeah. We did not try to combine these agents with

antibodies. And I don't know if there have been publications where that has been done. I can certainly look it up. It's interesting.

In terms of chronic hepatitis C, I think there are challenges. And again, I want to emphasize that the host targeted approach is, and I'm going to go back to my slide I skipped about challenges. I mean, for an acute infection where we are, we are looking at treatment duration of days, I think that these would be safe.

For a chronic viral infection where you need to treat over many months, I am not sure. So I would take that with a grain of salt. But for mechanistically, I think that there is definitely potential, and as yeah, as you saw, I mean, for idiopathic pulmonary fibrosis, it has been used.

So those pathways are really common pathways for multiple injuries within infectious and noninfectious etiology.

MS. ELKINS: Does this one work? Ah, good. I have the online selection, and a couple of these arrived before your challenges slide, and

before another discussion. But let's go over them again. So with regard to repurposing drugs, what do you expect in terms of dosages and change of dosage, and what are the implications for safety when the dosage is changed?

DR. EINAV: Yeah, thank you. So thus far, we've been able to show an effect in vivo with dosing regimens that were lower than the approved doses in humans. I want to say in general the safety is a concern. But a few comments about that. We do treat all noninfectious diseases with host targeted approaches. Right.

So why would treating viral infections be different with that respect. I mean we're seeing toxicity in direct acting antivirals as well. So it's not necessarily specific to host targeted. It's all about finding a concentration range that would be sufficiently inhibitory for a viral replication, yet not toxic in a cell.

Again, I emphasize acute viruses as opposed to chronic viral infection. So we are

talking about treatment durations that are in the order of days. We're also seeing redundancy in cellular machineries. For instance, some of the NUMB-associated kinases, we see that they have overlapping substrates to some extent, so you inhibit one, it may be sufficient to kill a virus, but not yet cause toxicity.

So and then the last thing I'd say about that is that, ultimately, I think it's really the severity of the outcome that should dictate the threshold. I mean what you can tolerate for a lethal infection such as Ebola would be very different from what you would do for a rhinovirus. Right. So it all has to be dependent on the context.

MS. ELKINS: Thank you. Next is thanks for a great presentation. You mentioned the lung to plasma ratio of 2, was that total, and how different was the time course in kinetics?

DR. EINAV: So the lung to plasma ratio in terms of, I assume, lapatinib concentration. So we actually, it's even greater than that. Let

me see if I have the slide here. Yeah, so here it is.

So this is what we measure in our animal models. After 8 hours of treatment, you could see concentration of 100 micromolar in the lungs, whereas in the plasma, it's essentially tenfold lower at that time point, so quite significant.

MS. ELKINS: And returning to the safety question, could you discuss further from your point of view how the safety profile of broad spectrum antivirals compares to that of narrow spectrum antiviral therapies?

DR. EINAV: It's a tough question and eventually, it's very dependent. It's dependent on the moiety. We've seen liver toxicity from decimvir. We're seeing, I mean so the profiles change. It's really the mechanism of the drug rather than the spectrum of coverage that it provides.

MS. ELKINS: Others in the room?

DR. MCGIVERN: Thank you, Dr. Einav, for an informative, thought provoking presentation.

DR. EINAV: Thank you.

DR. DEY: Good morning. And moving forward, this is Ranadhir Dey. I'm from Office of Blood Research and Review, and it's my pleasure to introduce today's next two speakers. First one is Dr. Viswanath Ragupathy, Dr. Ragupathy also works in OBRR, which is Office of Blood Research and Review. He's a staff scientist and he's virologist with almost 20 years of experience in HIV and AIDS research.

He has published several articles in high impact scientific journals and has been actively involved in international collaborative research studies, including those focused on the use of NG's and the IML technologies.

Our next speaker will be Dr. Gabriel Parra. And he's a principal investigator at the Office of Viral Products in the OVRR at the Food and Drug Administration. And his research focus on the epidemiology, genomics, evolution, and immunity of viruses associated with gastroenteritis.

So with this, I would like to please welcome Dr. Ragupathy.

DR. RAGUPATHY: Again, come back to that first one. Sorry with that technical glitch.

Thank you. Next presentation is unbiased metagenomic exploration of transfusion transmitted infections with nanopore sequencing. Early morning, Charlie Chiu set the stage about this matter, metagenomic sequencing, which makes my life easier to make you better understand about this technology.

So with that, the outline of my presentation is I will go over the background about metagenomics and the issues with the blood safety and how we mitigated it, and also some issues with this technology and challenges with this technology. And also I'll touch upon some information on regulatory insights.

Yesterday, Dr. Marks have highlighted how scientific research supports the diagnostic, the role of biologic products, and I will give you some examples of our research studies supporting

the approval of in vitro diagnostic devices. And also I'll go over a little bit on the methods used and conclude with some results and discussion.

So blood safety has significantly improved in recent years. The emergence of new infectious agents that can be transmitted by blood products continues to pose a significant challenge. (Inaudible) currently used in blood testing centers for pathogen detection are highly specific to known pathogens, making them less effective for identifying the uncharacterized microbes.

Alternative approaches such as pathogen reduction or inactivation, which you heard in yesterday's presentation, where they have shown significant improvements with these technologies. These are broad and nonpathogenic specific actions shows promise.

However, certain methods are not equally effective across all the types of microbial pathogens. For instance, (inaudible) viruses, such as hepatitis E virus and hepatitis A virus

are particularly resistant to these methods. So this is the world we are living in.

And here, I'm showing you with different viruses that have been either emerging or reemerging in different parts of the world. I've highlighted a few parts, which are currently affecting an epidemic which caused the, both the economical impact, as well as morbidity and mortality.

While I was finalizing these slides two weeks ago, CDC Infectious Disease Journal published a report on a patient with unknown rise of liver enzymes with a negative for all the hepatic agents and also some parasitic infections.

So when they did their metagenomic sequencing they found out this human circovirus which is causing, associated with hepatic failure, with the association with the pathogenesis found that individual, immune compromised individual.

So the challenges associated in blood safety, I have kind of grouped them as an emerging viruses. Those are hepatitis B, HCV plasmodia,

and CMV, and emerging viruses are HIV, COVID-19, Jakob's disease, SARS, flu, transmissible spongiform encephalitis, Zika, and Ebola.

And reemerging ones are West Nile bacteria, (inaudible), malaria, and babesia. Sub emerging are HGV, which I'll touch upon, and TTV, I'll touch upon latter part of my presentation. And SEN-V virus, and other infections.

NAC assays are highly accurate in identifying all these infections at the acute stage where you have a (inaudible). However, the residual risk because of this window period, or a new emerging variance is unavailable.

So can we ask questions, can NGOs enhance the blood safety by improving the detection of emerging uncharacterized pathogens. The use of novel technologies, such as NGO's metagenomics, allows comprehensive analysis of all nucleic acids present in the blood sample, providing a valuable tool for diagnosing infectious disease of unknown origin.

This method is known as agnostic NGS and

does not require a prior knowledge of the genome sequence to be analyzed. However, diagnostic approaches faces the challenges while bacterial diversity in clinical samples can be identified through 16S, kind of an universal signature. No such universal gene exists for viruses, which are highly complex due to our error prone replication process.

And I kind of put in here the band metrics associated in characterizing some of these microbes for the known pathogens. It's kind of informatics is quite simple. But with unknown viruses, it's highly complicated, which involves more additional robust tools required to characterize the genome.

So with that, we have some studies and how that supported our regulatory approval of NGS devices. So our office in OBRR is responsible for regulation of HIV in vitro diagnostics, including donors staining assays, which involves HIV and other viruses, and some parasites. And also HIV diagnostics, viral load, and drug resistance

monitoring.

So as you all know that Next Generation Sequencing has been there since 2005, with the 454 raw sequences, and around 2012 commercialization of short sequences have been rolled out. We have approved two genomic based assays that uses Asana technology.

So we envision that at some point these asset sponsors will intend to use these NGS technology. So around '20 time points we initiated evaluating this NGS technology with external collaborations on understanding the complexities and applications of NGS in genomic research.

We published three study results with the multiple applications of this technology. All three are development of either a proficiency panel or reference materials. The first one is specifically for HIV-1 drug resistance, where both the industry and academia collaborated in this effort.

At least 11 centers have participated,

including our laboratory, and this panel is currently available in (inaudible) care for commercial use.

The next publication is there is a need for highly divergent HIV variants are required for artists to perform better and also for the multi (inaudible) vaccine manufacturing. So again, in collaboration with NIAID and Duke University, we developed a highly diverse HIV reference panels. And these panels are also currently available in an (inaudible) with Duke University.

And lastly in collaboration with NIBSC, we developed reference panels for advantageous agents characterization where this panel is spiked with the 25 viruses. And we are part of this collaborative effort where both industry and academia are involved in development of this panel material.

So with the past five years, we gained sufficient experience both in terms of methodologies and also in bioinformatics. In 2017, two NGS device sponsors submitted the

pre-submissions seeking agency guidance on appropriate regulatory pathway and the proposed performance studies.

Just to note here, so there is no predicate available at the time when the sponsor submitted this application. So this is a kind of novel application for us. During the interactive review we provided our guidance on regulatory pathway, and also we mentioned to the sponsor that we asked them to submit their NGS data file in an organized way, that we provided them a template for an independent internal assessment.

And also, we provided that in our guidance on how to submit these NGS, large NGS data files in our secured format. And also we provided the guidance and agreements for utilizing external reference databases used in their analysis.

So in 2019, Sentosa SQ HIV genotyping assay for monitoring HIV drug resistance was authorized via direct de nova pathway, because there is no predicate exist. Special controls

were established to provide a safe and effective use of this IVD HIV genotyping assay using NGS technology for drug resistance monitoring.

In 2019, another device sponsored filed their metagenomic sequencing application at CDRH, and we were involved as a consult review for that application.

So to better understand the complexities with this evolving new NGS technologies, Multi Center International Collaborative studies on, and use platform was performed in our lab to evaluate the assist assay performance.

Specifically this one is use of nanopore sequencing platform for characterizing the advantageous agent. This effort was actually put forth by Dr. Eficon in another office at CBER, in collaboration with both industry and academia. And this study is almost now complete, and we'll be expecting this publication very soon.

So with that, the goal of today's presentation is unbiased characterization of transfusion transmitted diseases, where we

specifically focus on cataloging both the known and unknown viruses with this use of unbiased sequencing approach. And also using the nanopore platform.

And also to develop bioinformatic tools to identify the microbial diversity of bacteria, viruses, and fungus in biological specimens. So these tools we try to implement them in our CBER hive, so that in future applications, we could be able to use those tools for our independent assessments.

So it's kind of, that slide is kind of mixed. Okay. So our methodologies or two parts. One is an experimental and a bioinformatics. And for our experimental part, we have analyzed some clinical samples which are sourced commercially.

And also virus panels are prepared at different viral thresholds, lowest is 1,000 copies, and highest is 100,000 copies. And total nucleic acids are extracted from 200 microliter of the samples. Something is missing there. The panels are also prepared, evaluated individually

as HIV, HCV, and HBV.

Kind of representative viruses for our transfusion transmitted diseases. And we also mixed all these viruses as a virus cocktail and evaluated them. So nucleic acids are extracted using Zymo kit, which extracts total nucleic acids, and cDNA is synthesized.

And we didn't use any gene specific primers here. And so samples are purified using (inaudible) system, and laboratory preparation, of course, we followed the manufacturers instructions provided in the nanopore sequencing platform.

This nanopore is kind of a quite rapidly evolving method. So when we started this work, we used R-9 full fill, and by the time we completed, that kind of evolved with an R-10 full fill. So that's what I'm going to show. I have results from both R-9 and R-10 platforms.

For bioinformatics, we use the half model, and we profiled all the reads above quality seven. And for adapter trimming, we used a tool called Porechop. And for the quality control, we

filtered out all the reads that are less than 200 in this pack. But also filtered out all the reads that have a quality score below 7.

And we also removed the host because we used the clinical material from a host human. So more than 95 percentage of the genome is host genome. So that needs to be removed for better evaluations. And for the alignment, we used Minimap. Also we included multiple databases. I'll touch upon why we have included multiple databases. And for samples, Mega 11 are used for statistics and phylogenetic analysis.

So here is our clinical samples metagenomic profiling. This heat map shows you again these are all known viruses we wanted to catalog and whether to see whether these viruses can be metagenomically or without using any gene specific primer, we can able to detect these viruses.

What you see in the first one is HIV, and you can see we do see a high hit for HIV. And the second one is hep-C. You see it's high for

hep-C. And third one is hep-B, and you'll see it's high it's for hepatitis B. And last one is GBC virus which is also infecting with HCV individuals. I'll show you the additional data on that.

So once we've got this, it's the next step is to characterize that genome in depth. So what we did is we kind of derived our consensus sequence from these viruses and then build a phylogenetic tree using that. So with HIV clinical sample, we were able to find out it belongs to the (inaudible) B, and for HCV it is genotype 40A. And for hep-B it is genotype C.

And for all these three genomes, we were able to obtain near full length of the genome, which is indicated in a kind of linear color coding at the bottom of this phylogenetic tree.

For our West Nile virus, we were not able to get a full length of the genome. Per our investigation, we found out that it's due to a sample integrity issue. However, with a small fragment, what we received, what we observed when

we did an phylogenetic analysis, we were able to see that closely kind of clustering with lineage 1A.

So a little bit on pegivirus. So the pegivirus is the one which we identified with an HCV infected individuals, and once we got an each for this pegivirus, we derived our consensus sequence and generated a phylogenetic tree.

And you can see from this phylogenetic tree greater than 90 percent nucleated (inaudible) was observed with pegivirus genotype 2. So what are these pegiviruses? They belong to the family, Flaviviridae family and it's a single negative strand RNA virus with 9 times 4 kilobase genome in size. The pegivirus genus includes 11 species, pegi A2K, with the pegi C infecting the humans. Historically, the pegiviruses are called as a GBV virus or hepatitis G virus, and currently it's called as an hepatitis pegivirus 1.

So today there are only two known pegiviruses are identified. That is pegivirus 1 and pegivirus 2. Although not linked to the liver

disease, this pegivirus infection with HIV or HCV can slow the disease progression by immune modulating. Specifically, in HIV, these pegiviruses kind of down model at CCR4 and CCR5 receptors.

So in studies of general population and healthy donors, the prevalence of the pegivirus is found to be 0.8 to 44.6 percentage. Through agnostic approach, we were able to identify the full length of this genotype to pegivirus.

On the cartoon below is to show you how closely these pegiviruses are associated with HCV. So HCV genomically has a single polyprotein, and both the pegiviruses also you see a single polyprotein, both structural and nonstructural elements are quite similar, except the untranslated regions on both five prime and three prime, you'll see some differences with this pegiviruses.

So next virus, which we identified as the anellovirus. In our first standardized protocol, we were not able to identify these

viruses. So we slightly modified our protocol introducing rolling circle amplification at the stage of cDNA synthesis.

And by doing the so we were able to kind of profile the hepatitis B and this anelloviruses. So you can see from this profile and also phylogenetic analysis indicate that the Toscano virus is closely associated with our clinical samples, which are sequenced, which is BB9.

So anelloviruses are our single negative stand circular DNA belonging to the family of (inaudible), with a genome size of 2 to 3.9 KB and encodes 4 overlapping fragments, which is indicated in the cartoon below. The most studied genus of anelloviruses, in fact, humans, is the Toscano viruses.

And there are 29 species of this virus identified today. High prevalence in human population at greater than 90 percent of the individuals are infected with this anelloviruses, both HCV and non HCV infected individuals.

They are considered commensal and have

not been associated with any specific diseases just by their ubiquity in the human population. Recent studies have revealed the structure of analog virus particles, which include unique spike domains, that may help the virus to evade the immune system by exposing diverse epitopes as immunological decoys.

Through an agnostic approach, we were able to also identify the full length of this virus. We have deposited both this pegivirus and anellovirus in an SRA NCBI database, which will be released to the public at the end of this year because the publication is still due.

So this is general statistics for combined evaluation of HIV, HCV, and hep-B in different virus thresholds. So on the X axis you see different virus thresholds, and kind of the first row indicates the total number of reads. The ones with the red and orange are HIV reads. The ones with the green and blue or hepatitis B virus. And one with the lavender are HCV reads.

So the first row indicates the total

number of the reads. So these number of what you'll see there is the number of reads mapped to the virus decreases as far as the virus dilution is increased. So this is expected.

And also the number of days is also, (inaudible) also decreases with the virus dilutions. So these two are quite important for, you know, getting the establishing the limit of detection or improving the accuracy of these assays.

And for the read quality, when we profiled this HIV, HPV together, we were able to see a greater than 10 for the quality of the reads. And of course, the read length is always greater than 1,000, enable for the comprehensive genome coverage.

So next is we wanted to measure how low this virus can be measured. So we kind of diluted the virus with the human background as more than around 100,000 viruses or 10,000 genome copying equivalence, or 1,000 base pairs.

So what we have found out is at about

10,000 genome copy equivalents, we were able to consistently detect these viruses. And also we could able to characterize the full length of this genome. At 1,000 copies, we do see the hits, but some of the replicates missed detection of these representative HIV or HPV, RA, CD.

So we kind of stick to a 10,000 genome copy as a limit of detection. And when we ran the replicates of, three replicates, three independent replicates of this at that viral load, and we were able to consistently profile these viruses.

So this slide also shows you the repeatability of HIV, HPV sequencing at 10,000 genome copy in sequence. And what you'll see is all these replicates, there is not much variations. And for our reliable virus detection, at least 20 reads are required. I think that that is one of the questions early morning with Charlie Chiu's presentation.

So for our reliability detection, we found that 10,000 genome copies at 20 reads is sufficient. And also the base count of 10,000

bases with the read lengths greater than 1,000. And also we noted the quality scores of these reads are greater than 15.

So in our previous experiments, we kind of analyzed these three representative viruses individually. In our next experiment, we mixed all these viruses together and we kind of wanted to profile them to find out whether we could able to get that same inference at the 10,000 genome copy equivalents.

So this heat map indicates the three different replicates of our cocktail, and you can see all three replicates, we were able to detect the expected HIV, HPV, HCV. And also here we have included multiple databases. So the use of these databases for taxonomy profiling enhances the accuracy and comprehensiveness.

And this is also another question came today morning about the bias of this. So this use of these multiple databases kind substantially mimic those biases. For blood metagenomics, most of the viral database produce consistent results.

This again, it's some general statistics for the virus cocktail. And what you'll see shown above, there is no significant difference with the replicates measured. And N50 read length remains constant for all the replicates tested, which is important for differentiating from viral targets and from non viral targets.

And finally, we have also evaluated, so this nanopore sequencing as leveraged as to kind of start and restart approach. So where we can run this sequencing for up to a certain time point and then you can stop it, watch it, rerun it.

So we wanted to profile it and see how early we could able to detect the virus and how early we could able to detect entire genome of the virus where the coverage is equal to just one read.

So what we noted is in less than 5 minutes, we were able to see detecting all HIV, HPV, HCV when the viral threshold is around that 10,000 copies. At about close to half an hour, we were able to get the full length of this genome.

So in summary, agnostic approach of detecting bloodborne pathogen requires the minimum viral load of 10,000 copies. Full length hepatitis G or a pegivirus, or (inaudible) virus in HCV positive samples, despite its transmission through or transmissions.

Routine screening for pegivirus in blood donors is not currently recommended due to the fact that virus is nonpathogenic in nature. DNA and RNA sample preparation modification enhances the sensitivity of metagenomic approach, as in the case of identification of anellovirus and hepatitis B viruses using rolling circle amplification.

Anellovirus was identified as a coevolving virus in HPV positive samples, despite their presence in the blood products. Anelloviruses are not routinely screened in transfusion safety due to the absence of specific L traits. Though their potential impact on the recipients virome is still under study.

So finally, metagenomic characterization

that is ability to detect a wide range of pathogens, significantly enhances the blood safety by providing a comprehensive insights into the microbial communities present.

With that, I would like to acknowledge my supervisor and their ability to provide a constant support of my both research and regulatory activities, and Karen is our fellow in our lab who performed all the replicates, and the validation experiments, and other members of our HIV team.

I also wanted to thank CBER-hive, who are continuously provided that support, Louise, Ilya, and Sean Smith. Whenever I approached them to implement these tools or in any future applications are filed, we have all these tools implemented for our informatic analysis of our independent assessment. I'll stop here and thank you for your attention.

DR. DEY: Thank you. Thank you. Thank you, Dr. Ragupathy. Thank you. Please hold your questions, because we'll move to next speaker.

Our next speaker is Dr. Gabriel Parra, please.

DR. PARRA: Well, thank you to the organizers to give us, to give me the opportunity to present some of the work that my group has been doing for the past few years. And thank you all of you for being present here. And those that are virtual also. Also, thank you for your interest in hearing our research.

Okay. Today I'm going to be talking about some work that we've been doing to understand norovirus antigenic evolution and vaccines, at this time. And before I get moving, let me give it a try. Okay, here. Antigenic evolution and vaccine design.

Probably some of you already experienced, but just let me put everybody in the same context here. But norovirus is a major cause of gastroenteritis worldwide. As I say, some of you might experience it, and you know exactly what I'm talking about, the norovirus disease is characterized by explosive, uncontrolled diarrhea and vomiting.

The symptoms usually self resolve in healthy individuals within 1 or 2 days. But the imprinting is there. You will never forget that you had norovirus, right. But the in vulnerable populations, they can develop severe disease, like malnourished children in the third world countries, the elderly, and immunocompromised individuals.

And there is no vaccine or specific treatment for norovirus. Norovirus is an RNA virus. So it changed a lot. The capsid is quite simple. It's a nonenveloped virus. The capsid is, hang on one second, the capsid is icosahedral capsid. This whole capsid is formed by one single protein. There are in general, not always, but in general there are 180 molecules of VP1 arranged in this icosahedral particle.

VP1's usually form dimer and trimers to make this particle. This is shown here, a dimer of VP1. You can see this is the axis. So the one monomer is here. The other one is here. And what is interesting about this protein is they have a

very unique structural confirmation, which is only shown in caliciviruses.

The VP1 has two structural domains, the shell domain which is shown here in blue. And this domain forms the scaffold for the capsid to make this array. And there is an extension of these from this shell domain, which is the protruding domain, and this structure has been only seen in caliciviruses.

And it's quite interesting because this is a part of the virus that interacts with the host. And one interesting fact about this is full loops. So there is a lot of potential for variation in that particular part of the capsule. And I'll discuss that later.

A few things. We don't have a cell culture system or robust cell culture system, so obtaining viruses is quite difficult for us. We rely sometimes in stool samples and epidemiological data. That's the reason now my lab is actually quite interested in genomics and the technology, so we can capture information.

And when we express those part VP1 into a (inaudible) system, we actually can make virus like particles. There are no viruses, but they resemble very much the vision of the virus. We don't have an animal model to study norovirus.

But what we know from norovirus, and this is our infection working model, is that the virus is transmitted through the fecal oral route, enters into the system, bypasses the acid from the stomach, and reach the gut. And when they reach the gut, infect enterocytes and enteroendocrine cells.

And it replicates. It replicates in very high tires. Develops disease, and through the disease is going to be as create a bunch of viruses every time that you either vomit or have diarrhea. And that goes to the next person.

The infection in this enterocytes or enteroendocrine cells is facilitated by the expression of certain carbohydrates that many people in this room actually express. 90 percent of the population seems to express these

carbohydrates, different carbohydrates.

So the virus needs to use different strategies to attach. These are carbohydrates from the histo-blood antigen, or we call them for short ABGA. These are not the receptor, but what we know that they facilitate infection.

Once an individual is infected, this is a classical immunology from book, right. Immunity or immune response from book, I meant to say. The individual will develop, two weeks later, very strong in homotypical responses for antibodies. And these antibodies will wane after the infection occurs.

This antibody seems to protect amount between 2 to 5 years, depending on what kind of data you look at. So reinfection can occur. But despite that many people have high antibody titers in the population, reinfections occur very often.

And those that had norovirus before, multiple times, you know exactly what I'm talking about. So how can high titers not protect you from infection. So one of the things that we know

is that not all antibodies are actually protective. And I'm setting the stage. We have no clue what the other antibodies are. The ones that do not protect.

But basically, what has happened here is using control experimental challenge volunteer studies, what we have learned from these studies is that individuals that were challenged with the virus and they developed disease. Sorry and they did not develop disease after challenge.

What was found is that this individual have very high titers of HBJ blocking antibody. So if you have high titers of this, that blocks interaction of the virus with these carbohydrate, you more likely will be protected. While those that do not have those type of antibodies, you will be not protected.

So what I told you so far is a complex interaction of waning immunity potential, or motific immunity. I haven't talked about this, but the virus actually changed, and I told you that there are multiple different HBGA's that that

people can express. So virus can change and attach stronger to certain carbohydrates. So and also an enhance of viral replication based on this difference on attachment.

But one of the things that happened with the virus actually changes generated viral antigenic diversity. And that's exactly what my lab is exploring and trying to understand. Human noroviruses, there are more than 40 different genotypes. What are genotypes. Genotypes are viruses that cluster together in a phylogenetic tree, and they are defined by changes in the VP1.

Just to give you an example, the different viruses from genome group 1 and genome group 2 can be classified based on 50 percent of difference in the capsid protein. So we're talking about more than 250 mutations just in between these two viruses.

Just to give you a perspective that these are very, very, very different viruses. Then each of these genome groups can actually have multiple clusters of viruses that I'm just showing

you here. And those are the genotypes. And we have more than 40 of those that can infect.

So you can imagine that in the lifetime of a person, if these are very distinct viruses, you can be reinfected with multiple of these viruses. And as I showed you, as I told you before, the immunity is somehow homotypic.

So there are multiple chances that the individual can be reinfected during the lifetime. But one interesting thing happened in norovirus and I'm not going to go into much details about why that, but one interesting thing happened is that there are 40 different genotypes, or more than 40 different genotypes that can infect humans.

One single one, G24, is the most predominant worldwide. And it can account for more than 50 percent of the infections worldwide. And in some cases, can be up to 80 or 90 percent. And that's true for all the years, or most of the years that we've been looking in the literature.

So why is that. The reason of that is

this particular genotype actually presents the chronological emergent and replacement of variants, very similar to influenza H3 and 2, or SARS-CoV-2, as we have seen in recent years.

But basically what happens is that one cluster of viruses, I'm just giving you an example and the first one I've been able to see here, Farmington Hills virus, emerged into the population, have multiple changes. It spread through the population very rapidly, it spread through worldwide very rapidly. Caused large outbreaks, and a few years later will be replaced by the next one just because we'll exhaust most of the population.

What is interesting, and as you can see here, there are multiple viruses that emerged in the last 25 years. Two things to consider here. And I'm going to go into detail, although my lab is very interested on those patterns, is the chronological emergence is not linear in terms of the phylogenetic tree.

So you can have viruses from 2007

circulating very close to 1995, and so on, right. So there's no linear or linear evolution on these viruses. One thing. The second thing is that not all these viruses actually cause pandemics. So only 7 out of more than 10 that emerge, the other ones are just locally distributed.

So we don't know exactly why. But one of the things that we know or what we saw, our working model was that this virus emerged with changes in the capsid, and that's how they escape the immune population, or herd immunity.

When I, when we started looking at this data, we actually didn't have enough data to support that hypothesis. So one of the things that my lab did is we, using genomics, and I'm going to go in detail, we selected 25 viruses that we may (inaudible) out of more than 3,000 sequences that will look, we selected these, we selected this representative for each of these clusters.

We make an antisera for each of these, each of these columns is a mouse immunized. And

we tested blocking assay for all of these. And what we have found is that, yes, indeed there is a lot of homotypic responses, and there are some instances of cross activity.

But for the most part, early on, the viruses are quite unique. And later on, the virus that emerged around 2006, there are some cross activities. And I'll go back to that later on. But one of the things that we were able to use, to do with this data is we were able to perform cartography, cartography mapping.

We were able to use that categorization mapping to determine the antigenic distance and using the antigenic distance, we can actually correlate with genetic, with, here. With antigenic differences. So as you can see here that we have a correlation a pair of viruses that have X amount of distance with X amount amino acid differences.

So just to give an example here, Sydney virus from 2012 with (inaudible) virus, 2018, have the most antigenic distance, and one of the most

genetic distance. And so on, right. And using this information we were able to quantify the amount of amino acids changes that you might require to change this.

I know that is not perfect, but this is the best correlation we have. I'm going to go back into some of these data later on. But what we have been able to define is that 18 residues are most likely the threshold for these antigenic difference among these viruses.

And one thing that I want to mention is that what we have found is also that some of these amino acid changes occur in synchrony. So they are coevolving. And I'm just not going to show you that right now. But we have all this data.

So using that information, we asked the question what are the antigenic determinants of viral antigens and how many antigenic sites or epitopes are actually involved in these viral antigen evolution. Using genomics structural information with the experimental research, and I'm just summarizing multiple years of work here,

but what we have found is that the 38 residues which are shown here in different colors, are the ones that are changing as the variants emerge.

Most of these changes occur here close to the ABGA mining site of the protruding domain. This is a protruding domain. Each of these changes or clusters which we call antigenic sites, each antigen site most likely contains multiple epitopes. These are variable. The antigenic sites are all surrounded by, surrounded the ABGA binding site, which makes sense, right.

So the antivirus are targeting this area. These episodes are, or these antigenic sites are changing, and that's the way that the virus escaped these ABGA blocking antibodies.

Like I say, I'm summarizing. Almost every slide is a lot of word here, but I just want to bring the point that most of this data, when we did these experiments were done with mice. With mice, monoclonal antibodies just for because it was simpler. It was easy to do the experiments. It was cleaner to trying to understand.

But all these, many of these epitopes lower in fields have been confirmed using very complicated cell culture system which used stem cell derived enteroids. She was able to show that these epitopes are actually involved in viral neutralization or dissenting any size in viral neutralization.

And also with collaboration, in collaboration with Vanderbilt, we and others have shown that this variable and antigenic sites are represented in the human repertoire for antibody responses.

So what we have found in mice is actually reproducible to humans. So then the question is, okay, we have all these viruses are changes. Can we actually elicit a cross-reactive protective responses to different, for norovirus. And I'm not going, I'm not even going into different genotypes.

Let's just focus on (inaudible) for now, right. And we have a dilemma here, right. Because what we have found is a beautiful word

that another person in my lab can tell, is we, using different approaches, we determined that the most immunodominant of all these sites are A, G, and C.

But A, G, and C actually map very close to the ABGA, but they are one of the ones that are more variable sites of all of these variable sites. So what I'm showing you here is just all the patterns, amino acid patterns that are represented in at least three viruses, in more than 25 years of evolutionary virus.

So we're talking about 47 different patterns in one residue, in one antigen inside, and then close to 45 patterns in another, and so on. Right. So this is just a tremendous variability that we're dealing with.

So we asked the question, maybe the antibody responses needs to go to more conserved area. Everybody showed good results. I'm not going to show you bad results, but I can tell you we spent multiple years here trying to elicit cross reactive responses to very highly conserved

regions of the capsid.

And we for the most part, failed. I'm not going to go into much detail on that. So then the answer is well, we need to, we need to try to elicit immune responses to this immunodominant site.

Serendipity or no, when we were trying to characterize these human monoclonal antibodies that we been working with in collaboration with Gene Crow from Vanderbilt, what we have found is that antibodies that target epitope A, antigenicity A, or antigenicity C, they actually have different degrees of reactivity.

Some of them were quite specific. Some of them had some reactivities, some of them have a unique reactivity patterns. But one of them, one norovirus Noro 123, was able to cross block, cross neutralize and cross react with viruses from 25 years. So there is some hope then, right.

So if this actually targets these residues and is refractory to the virus evolution, there might be. So we asked a simple question.

How many of these antibodies are in the human population. Because when we test sera, they seem to be quite (inaudible).

But how you test that without the fact that individuals might have been infected with that particular virus already. So we went back in time, right. So Chelsea Belinski in my in my group took a, trying to address this question. And she took advantage of sera samples that we have from individuals that were infected in 1971.

These individuals were residents from a hospital here in Maryland. The individuals were infected in January 1971, with one, there was one particular norovirus. With modern techniques more than 40 years later, we were able to determine that this is a G22 virus.

In December 1971, these individuals were infected with G26 virus. So two things. We have, and we have sera after the second infection, so two things. People were infected. People were, people never saw virus that were able to test for, right.

Because these are sera from people, that's on in '79, and we start testing viruses from the 80s and in the future. And what we have seen and I'm not going to go into much detail on data, we have a sera sample from 24 of these individuals.

This is a collaboration we have with my former mentor, we brought those samples here and what we have found is 3 out of 24 individuals have very strong cross blocking responses to virus, to future viruses, right?

That these people never saw. So we asked, so we don't have samples from these. But we, is there any way that we can actually determine where that cross reactivity is coming from. And what we have done is design a new ASI that we call Haka, like the New Zealand dance for the rugby.

But this stands for HBGA blocking antigenic competition assay. It's very simple. The name is more complicated on the ASI actually. So this is very simple. So in HBGA binding,

blocking ASI, what we do is we have the VLP's, we have the sera. If the sera can block this interaction that is going to be bound to the carbohydrate, the sera is able to block the VLP binding, right.

But in the Haka, in the Haka ASI, what we do is we include a competitor, a recombinant protein that has two features. One feature is that the recombinant protein should not be able to bind to the carbohydrate, so it doesn't affect the binding of our testing material, right.

And the other feature is that we can modify it, we can actually deplete the epitopes from this particular protein that we want to study. So in other words, this competitor will suck all the antibodies except for those that the epitope that we modify.

So for instance, we can use this protein to modify the antigenic site A, or antigenic site G, so then we should be able to sort all the antibodies that are non A or non G. And we did, of course, a lot of testing. And basically what

happened is if you put recombinant protein that is only the domain, the domain will not affect the antibodies that are blocking antibody, so you will have a perfect or ideal blocking assay.

If you put the whole P without any modification except for the binding, the P will reduce your titer of antibody. If you put virus like particles that have mutated sites, then you will be able to determine how much of that site actually is involved in blocking the site.

And the long story short is what we have found is that 2 out of 3 of these individuals are actually, their crucial activity to future viruses linked to antibodies that target ANG.

So the next question is okay, can we actually learn how to elicit these antibodies. So we went back to the drawing board again. This is the last slide, the drawing board. And we grabbed multiple monoclonal antibodies that we have from mice. And we tested the reactivity patterns.

And what we have found is that these reactivity pattern is very dependent on the

strain. So for instance, mouse that were immunized with viruses from 2004 has almost no cross reactivity. Very specific. Viruses that were immunized with 2012 have some level of cross reactivity.

And now we're trying to understand why is that, why the antibodies are refractory to this. But one thing I can tell you is that it's not black and white. So it's not that one antibody lose binding for one single amino acid or is resistant to multiple. The answer is some antibodies can be very, very affected by one or two mutations, while other ones can have up to five mutations and is still binding.

So we're looking now, we're doing two things. We're doing machine learning, the model we're building, machine models, to try to understand why these patterns occur. And we're also sequencing with the rhabdo B cell repertoire analysis trying to understand this pattern of cross reactivity, and hopefully we'll learn how to better immunize.

This is work from Michael Landiver. The people in my group have posters, so if you are more interested in this research, please visit their posters. And this is just a summary.

And I need to acknowledge people. Lauren, Kentaro, Joe, Kelsey, Michael, amazing people. They are always working. I try to describe many years of work here in only a few slides, so please go to their posters. Of course, a lot of collaboration and not to forget funding that supports our research. Thank you.

DR. DEY: Thank you, Dr. Parra, thank you. Very exciting seminar. Now, Dr. Ragupathy, could you please come to the stage? And now you can, we can have questions here for both the speakers.

DR. PARRA: I don't know if I can answer questions, I guess. I told them everything I know.

DR. MCGIVERN: Do we have questions in the room?

MR. KUMAR: While others thinking, I

have a quick question for Dr. Parra. Is this norovirus is foodborne or waterborne?

DR. PARRA: Both.

MR. KUMAR: Oh, it's both.

DR. PARRA: Yeah.

MR. KUMAR: Okay.

DR. PARRA: It's actually all of the above. So you can find it in water, contaminated water, in food. It's also person to person transmitted. So yes.

MR. KUMAR: Okay. Thank you. Thank you.

MS. ELKINS: And while we do the online questions, anybody who is doing a flash talk, please cluster over on the righthand side, and it will help if you can get yourselves in order of the talks because your slides will come up in order.

Now back to the most immediate. Online questions, I think, this is for Vishi. Can you share what software was used to remove the host genome?

DR. RAGUPATHY: So we used our internal CLC genomics workbench to deplete the host genome.

MS. ELKINS: Next, I think rolling circle amplification forms are very long single stranded DNA strands. Is that an issue, or can you convert it to double stranded for sequencing?

DR. RAGUPATHY: So the approach here is kind of unbiased and also the threshold what we are handling is with a very, very low (inaudible) of our target of interest, with a high background of human sequences. So introducing this rolling circle kind of enhances the detection of these pathogens.

MS. ELKINS: Thank you. Let's see, norovirus. Could delivery of neutralizing antibodies to target norovirus drive selective pressure forcing them to evolve more rapidly or making subsequent vaccinations more difficult?

DR. PARRA: Yes. That's the reason we're trying to understand.

MS. ELKINS: Kind of the point.

DR. PARRA: Yes.

MS. ELKINS: Next, have you studied whether or not the antigenic sites are coordinately expressed on the surface of the variant among variable genotypes, and whether this impacts the differential coverage of specific sets of antibodies?

DR. PARRA: The answer is yes. We're looking into that. I don't have an answer for that yet.

MS. ELKINS: Let's see. Sera prevalence of animal noroviruses in humans is known. Do you think there's a contribution of animal noroviruses adding variants?

DR. PARRA: The answer is I believe it's not. I believe these are human pathogens and that's one of the reasons that it was quite difficult for us to develop an animal model for human noroviruses. There are some people that have found sequences of viruses in animals in people. But we don't have true proof that those are interspecies transmission. In fact, the phylogeny suggests that these are extremely host

specific virus.

MS. ELKINS: And finally, in a big picture question, in your opinion, how far away are we from developing a norovirus vaccine?

DR. PARRA: I will not answer that question just because of my regulatory role. So well, the only thing that I can say is that there are multiple companies very interested in developing these vaccines. So at this point it's just a race. But I'm hopeful that at some point, we will have a vaccine.

MS. ELKINS: And that covers the online list. Others in the room.

DR. MCGIVERN: So I had one more question for Vishi. So what do you see as important obstacles to the routine use of metagenomic MGS in clinical diagnostics?

DR. RAGUPATHY: So I would say there is no obstacle. We can, we can, now the technologies are evolving with rapid turn around times, which will be one of the key question here with the advent of this nanopore sequencing technology.

And also I've shown that as quick as less than two hours or three hours, we could able to get the information.

DR. MCGIVERN: Thank you. I think we have one more question here.

DR. KANAI: This is for Dr. Parra. I just wonder if you found anything else interesting in terms of the synonymous variations.

DR. PARRA: In terms of what?

DR. KANAI: Synonymous, the so-called signing.

DR. PARRA: Yes. But it might require a few minutes to answer that. But we can pull out that. In fact, it's one thing that I can say, though, it's remarkable how the known G24 noroviruses, mostly explore synonymous mutations to evolve, while the non G25 noroviruses use the opposite, right. And that's the reason that the non G24 norovirus actually accumulate mutations over time. Yeah.

SPEAKER: I have a quick question for Vishi. So you have used the 1-2 platform, right,

and you have extracted the total nucleic acids, if I understand correctly. So my question is if you have just used the RNA and did your assay, what would be the average read length for the 1-2 platform?

DR. RAGUPATHY: Could you repeat your question? I didn't get your first part.

SPEAKER: So if you have, if you have used just the RNA for making the libraries, the way you have proceeded your, processed the samples.

DR. RAGUPATHY: Mm-hm.

SPEAKER: And use the 1-2 platform, what would be the average read length if you have just used RNA?

DR. RAGUPATHY: So RNA cannot be directly sequenced, which you know. So we need at some point conversion into the cDNA. So what we observed is mostly around 1,000, 2,300 base pairs with the use of whether it's a DNA virus or an RNA virus, that's what we see with the representative samples evaluated.

SPEAKER: Okay. Okay. Thank you.

SPEAKER: At some point I guess there used to be a lot of work on the (inaudible) cell, algo saccharides and the connection with the norovirus. What happened to that research? I didn't see anything anymore. So is it not a receptor or anything?

DR. PARRA: No. It definitely is not a receptor. So we know that when we express some genes that will facilitate the expression of those carbohydrates in cells, they are not. They do, they do act as attachment factors that facilitate. The main problem with this is, the saws in the field is this might be proteinase receptor that the HBGA's facilitate binding.

They approach and this is probably a multi step entry process. The main problem we have is that we don't have a traditional cell culture system to the classical receptor, viral receptor studies. In norovirus, the in murine norovirus, the receptor is a protein, is CD300. So yeah. So we feel that it may be, yeah.

SPEAKER: Yeah. Because I remember those times they used to do that.

DR. PARRA: Correct.

SPEAKER: Depends on your blood type, maybe.

DR. PARRA: Correct.

SPEAKER: Maybe your ear infection.

DR. PARRA: Correct.

SPEAKER: Will be, you know, yeah.

DR. PARRA: That type of research is still ongoing and there is certain, there is certain some correlation of people that express certain carbohydrates in their saliva or in their epithelial cells are more prone to get infected. But that's not the sole requirement, right. So then, and we have more than 10 different carbohydrates, more than 40 different strains. Even we know that even one mutation changed that pattern, right. So yeah, it's a, it's a complicated story.

SPEAKER: Thank you.

DR. MCGIVERN: Do we have anymore

questions online or in the room?

SPEAKER: This is a question for Vishi. One of that, there are several issues involving metagenomics, and so the argument could be made that like 10,000 copies is not sufficient really for molecular diagnostics. I mean, we know that PCR has this sensitivity down to less than 100, or even like 1 to 10 copies for certain assays.

And so, and it probably is not sufficient for blood transfusion screening if you're going to use it as a way to rule out infection, because then if you don't have that sensitivity. So that's my first question. How are you going to deal with this complaint about low sensitivity? Are there ways to actually increase sensitivity to be equivalent to PCR?

The second question I had is and there's a big complaint is that the implementing in the clinical lab, because I run a clinical lab that runs clinical metagenomics, you really need kind of the simple to answer system and it doesn't exist. There are a huge amount of manual steps,

as you know, that are involved.

DR. RAGUPATHY: Yes.

SPEAKER: And their points basically a failure, and essentially like our clinical metagenomic assay has 300 steps. It takes our CLS, they need to be, our clinical laboratory scientists need to be trained for three months before they're proficient at running this assay.

So it's not like putting the sample in an instrument and getting a result, which is ideally what you would want. And then the last question I had is, you know, clinical metagenomic assays are plagued by contamination, and it may not be as true for viruses, but certainly for bacteria. Many of the commercial companies that are running clinical metagenomic assays, we detect bacteria that are simply part of contaminants of the reagents flora introduced.

And the question I always get is how do you interpret this with respect to a patient. It's pretty clear that we identify, you know, organisms that may be pathogens but are also

potentially colonizers or commensals.

And you know, having a clinician, you know, I'm an infectious disease physician. Having a clinician actually try to interpret that is a big mess. And so the way we've dealt with it is we've actually done selective reporting where we're only reporting what we think are pathogens.

And so being able to put findings in clinical context is also a big challenge with metagenomics. So maybe if you could address those three questions.

DR. RAGUPATHY: That's a great three questions. I'll try to address one at a time. So for the last two parts of your question, it involves good laboratory practice. So with a well trained laboratory setting, and structurally followed protocols.

And the question of contamination and even with the issues of multiple manual can be substantially minimized. However, that said, we always recommend a fully closed automated system so that can kind of eliminate source of errors.

So wherever possible, automation always helps to minimize the error.

The one device which we kind of approved in 2019, that one is very, very minimal user involvement other than moving the plates from the thermocycler to the automated sample handling systems. So that's all the user involvement, all the rest is all automated.

And regarding your question on sensitivity, yes, metagenomic suffers with our sensitivity unless we introduce a small enrichment step. Or if there isn't some type of small enrichment is incorporated. I agree with you that we can still push this sensitivity. Without doing any enrichment like limited piece cycles, which is in the case of alumina experiments, pushing the sensitivities too hard with this nanopore platform.

DR. MCGIVERN: Thank you.

MS. ELKINS: And we do have one final question online. There are some data, this is for Gabriel on noroviruses. There are data on

commensal bacteria that regulate norovirus infection and immune responses. Do you have thoughts on how that impacts the development of neutralizing antibodies?

DR. PARRA: And the answer is yes. There is some literature out there as far as I remember. Most of that literature has been build with murine norovirus in control experimental settings. I'm not quite sure we have enough information at the real live human infection.

So I prefer to refrain to give a comment on that. But it's probably very likely. It's very unlikely that it might affect neutralizing antivirus, but everything is possible, I guess.

DR. MCGIVERN: We'd like to thank, since there are no more questions. I would like to thank all of the speakers from this morning for their presentations and the attendees for joining session 3. And next up are the flash talks followed by the poster presentations. Dr. Colon-Moran.

DR. COLON-MORAN: Good afternoon. I'm

Winston Colon-Moran. I'm an OTP in Dr. Nirja Bhattarai's lab. The issue we're working with is on immunogenicity of viral vectors, which is a significant problem and, okay, there we go. And so it's a major challenge for in vivo gene therapy. And that's because AAV, which is existing in nature, it infects the general populations.

There is preexisting immunity and there are responses against the viral vectors once they're introduced. So the host immune response can impact both the safety and the efficacy of gene therapy products.

And so there are a number of strategies that have been developed to deal with this issue. One of those strategies is to reduce the vector dose. Unfortunately, this can render the vector susceptible to low levels of neutralizing antibodies and reduce efficacy.

Another way to deal with this is to modify the capsid and the capsid is the most immunogenic part of the AAV vector. You can do

that by eliminating immunogenic T cell epitopes. The leading CPG regions from the vector genome, the standard of care is actually immunosuppressive drugs. But these have their own problems.

There's patient variability. There are side effects, interference with gene therapy delivery, and you can still develop an immune response after removing the drugs. So our strategy is to learn from nature. We look at viruses, and we look at the mechanisms they use to persist. They have evolved, obviously over millions of years, to inhibit their immune response.

So we're looking to identify some regions within viruses that do that, and we have recently identified an 11 amino acid peptide from the human hepatitis B virus, E antigen. And that potently inhibits the T cell response.

So we are working on engineering that into the vector in order to test it for safety and efficacy in the future. So that's everything I have on the poster 19. So if you have anymore

questions, please come visit. Thank you.

DR. MCGIVERN: Thank you. Thank you. Dr. Alan Baer.

DR. BAER: Hi, everyone. My name is Alan Baer, and I'll be presenting our group's work in developing a novel and simple manufacturing platform for adeno associated viral vectors in gene therapy.

So viral vectors are commonly used, or commonly produced using cell lines. Jumping around here. And the vector market is currently estimated at 6.47 billion and is projected to rise to 9.5 billion by 2026. At issue, however, is that viral vector manufacturing process is complex, labor intensive, and expensive.

And this is a bottleneck, and it contributes to delays in product development, and increases therapeutic cost, and accessibility. Viral vectors are commonly produced using cell lines. And manufacturing is generally divided into upstream and downstream processes.

Our group is looking to simplify a

critical segment of the upstream manufacturing process involving the cell expansion and vector production step. Currently, cumbersome flasks and bioreactors are used for growing cells to generate viral vectors.

We're looking to develop a simplified, scalable, and low cost static gas permeable cell cultivation system called G-Rex. So we're looking to condense the manufacturing process. G-Rex is currently used for production of T and NK cell therapy products. And our group is looking to adapt tech 293 cells for AB vector production.

And with this system, we'll be capable of expanding 5 million cells to 20 billion cells using two plasticware sets. And our data suggests the G-Rex system can be used to manufacture AV comparable to conventional systems, as shown on the figure on the right, where we're comparing a traditional Erlenmeyer flask system to the G-Rex, looking at viral titers.

And our ultimate goal is to improve AV vector manufacturing by reducing cost and

handling. Thank you.

DR. DEY: Thank you. Dr. Robert Dorosky.

DR. DOROSKY: Hello. My name is RJ Dorosky and I'm a research biologist in the Stibitz lab. And today I'm going to talk about my work on live biotherapeutic product purity assays. So LBP's are biological products that contain live organisms, such as lactobacillus. And they're intended for use as drugs.

LBP purity is an important safety parameter and typically involves demonstrating that the product is free of contaminating organisms. Traditionally, this is done by plating the product onto a nonselective medium. But one major challenge that sponsors often run into is that the product organisms will, their growth will interfere with detection of contaminants.

So our approach to address this challenge is to use antibacterial approach, specific antibacterial proteins to selectively inhibit the product organisms, to reveal the

presence of the contaminants.

To that end, we evaluated the antibacterial activity of four bacteriocins and we found that plantaricin S exhibited broad activity against lactobacillus and non-lactobacillus strains that we tested.

But importantly, we found that the lactobacillus tested were far more sensitive to these bacterias and plantaricin S in particular, than the non lactobacillus. And so that suggested that we may be able to use these bacteriocins in a purity assay.

And so to test that, we ran several mock purity experiments with a lactobacillus delbrueckii strain. And so for these experiments, we took this preparation of lactobacillus delbrueckii lab preparation and spiked it with known contaminants and plated them on M11, or M11 that's supplemented with the bacterias and plantaricin S.

And as you can see, L. delbrueckii was completely inhibited by the plantaricin S, and

this allowed for the detection and enumeration of both of the contaminants. And importantly, the assay worked well enough to meet the system suitability criteria for USP 61, which is typically employed for aerobic contamination detection of LBP's.

And so I'll be at poster 39 if anyone else wants to talk more about this. Thank you.

DR. MCGIVERN: Dr. Byung Woo Kim.

DR. KIM: Hello. My name is Byung Wu Kim and I'm a postdoctoral fellow working in the laboratory of Dr. Zhaohui Yes in the division of gene therapy. Moving on to the slides, I'm pretty sure many of you who are familiar with cell biology are aware of some known functions of human serum albumin in cell culture and differentiation.

And these functions include stabilizing proteins and growth factors serving as a carrier for nutrients including lipids, fatty acids, hormones, and other hydrophobic molecules. It can also function as an antioxidant. And lastly, promoting cell adhesion and attachment.

However, from a manufacturing and regulatory perspective, there are several challenges associated with the use of albumen. First of all, lot to lot inconsistency. Issues like impurities, contaminants, or unintended posttranslational modifications can affect cellular microenvironment and lead to produce a reproducibility issue.

Qualification requirements may vary between countries and regions, which can raise these safety issues. And lastly, the high cost, manufacturing cost is another hurdle.

In order to address these challenges, we have been searching for a suitable replacement for HSA and we finally found this chemical compound called Soluplus. And using this Soluplus in place of HSA, we differentiated, we successfully differentiated pluripotent stem cells into hematopoietic cells.

And the bottom panel of the slide is just showing some of my slides relevant to this work, which I'll be presenting in the poster

session. So if you're interested in it, please visit poster number 41. Again 41. Thank you.

DR. MCGIVERN: Dr. Stepan Surov.

DR. SUROV: Okay. Yeah. Hi. My name is Stepan, and my presentation is dedicated to reversal of direct factors and inhibitor activity by factor Xa variants. Oral anticoagulants targeting calculation factors Xa are used to prevent and treat thrombotic disorders. However, they carry their risk of uncontrolled bleeding.

So a rapid, effective, and safe reversal agent remains on medical needs. Several genetically modified calculation factors and variants were proposed to achieve this reversal. And properties of these variants often compared to currently approved drug, which requires high doses and is associated with thrombosis.

Using computational design, we engineered novel factor Xa, whereas that retain enzymatic function, and we aim to test the ability to reverse the action of factors Xa inhibitor drugs. So in our published study, we have shown

that our variants demonstrated effective reversal of apixaban and used bleeding in the mouse tail clipping model, at doses lower than those for Andexanet alfa.

And as a continuation of the study, we are showing that in thrombin generation assay in the presence of all factors Xa inhibitors at low and high concentrations, both variants are more efficient than Andexanet alfa in restoring their inhibited calculation to physiological levels.

So in conclusion, our factor Xa constructs can evade the effect of factor Xa inhibitors, apixaban rivaroxaban, edoxaban, at considerably lower concentrations at both low and high doses of factor Xa inhibitors than Andexanet alfa. The only specific reversal therapy currently approved.

Thank you for your attention. Please feel free to stop by my poster or email me in case if you have any questions. Thank you.

DR. DEY: Dr. Tapan Kanai.

DR. KANAI: All right. Okay. Good

afternoon, everyone. And myself, Tapan Kanai, I work on Dr. Gorman lab. The title of my talk is defining the epitope of influenza HA vaccine elicited monoclonal antibody using cryo-electron microscopy.

As you know that current vaccine is more stem specific, and immune response of and directed toward the more (inaudible) region, there is more variables. Therefore our current clinical trial focused at targeting the immune response to the more (inaudible) stem region.

That resulted in a series of antibody binding the similar epitope. The goal of this project to use the structural biology to define the critical epitope feature for distinguishing the broad and potent antibodies which are binding to the stem from the others.

And I'll talk about the two antibodies that are isolated from the stem clinical trial. That is 1G01 and 1V06, and (inaudible). 1G01 is more potent than the 1V06. And here is the, we develop a cryo- intensity map for the complex of

HN1, 1G01, and view the atomic model and which confirms that both antibodies binding to the lower stem regions, and source (inaudible) to epitopes, overlapping epitope features of the distinct orientations, and binding to the timer.

Then we realize that what is the special orientation that these two have. Then we overlap this to our atomic model aligned with (inaudible) and found that both have a different approach angle to the viral membrane.

Upon comparing these two epitopes of these two (inaudible), we found that there is three distinct binding region found for the 1G01, as compared to 1G06. And one is the overlapping region with the 1G01, where both heavy chain and light chain interacting with the HA timer, HA2 of the timer.

And the other two nonoverlapping regions that is found for the 1G01 that is either heavy chain or light chain interact in the C terminal end of the HA1, or the N terminal of the HA2. And I do believe that the C terminal and the HA1

involving the viral obligation while the N terminal of the HA2 involved in the viral fusions.

And the tables so the detailed quantification of the different parameter that have been contributed by the (inaudible). Though these analysis shows that the 1G01 is more potent than the 1G 06, we speculate, or we suspect that the angle of approach of this (inaudible) to the viral membrane may have some role.

Therefore we did the docking of this two atomic model to the membrane bound HA. And surprisingly, we observed that the angle of approach for the 1G01 historically unhindered approach while 1G06 may be occluded by the viral membrane.

Therefore, from the study, we conclude that higher BSA, additional hydrogen bonding, and sterically unhindered approach may altogether contribute to the more potent or the greater breadth for 1G01, it's competitor, 1G06.

And during our analysis we found that there is no specific difference in the sequence

has been observed. Our future study will be confirmed the difference in binding affinity of these two (inaudible) to the soluble versus the membrane bound HA.

And if you're interested about the (inaudible), please stop by the poster number 53, and thank you very much for listening this brief presentations.

DR. DEY: Thank you. Thank you. Thank you, Dr Kanai. Dr. Sabarinath Neerukonda.

DR. NEERUKONDA: I'm Sabarinath. I'm a staff fellow in Brennan Ellsworth Lab in office of blood. So we compared four different cell lines for their ability to support anaplasma phagocytophilum infections. And these cell links include HL60, which is widely used to propagate anaplasma in vitro, as well as a monocytic cell lines U937 and K5622. U937 and PHP1.

And finally we also looked at a multipotent progenitor cell line, K5622. So we infected all these four cell lines with cell free anaplasma and we spun down the cells onto glass

lights. In addition, we also isolated DNA to quantify the number of bacterial genomes.

And finally, we looked at the expression levels of the receptor p-selectin glycoprotein ligand-1 by testing blood. Here you can see that, you can see that the bacteria replicates in HL60 cell line, which is widely used to propagate anaplasma.

And the cells that are marked with asterisks has the bacteria containing vacuoles, whereas both U937 and the THP1 cell lines did not support bacterial replication.

In addition, we also found a novel cell line, K5622, to support bacterial replication. And here on this graph you can see the number of bacterial genomes quantified by qPCR. Both HL60 and K5622 supported bacterial application, whereas monocytic cell line C937 and PHP1 failed to support bacterial replication.

And finally, we also looked at the expression levels of receptor, PSGL1, in all four cell lines. As you can see, all four cell lines

express PSGL1 at variable levels, suggesting that the ability to support replication is not due to the receptor expression levels but at a post entry host factor level.

We also looked at whether the entry into K5622 cell line is sialic acid dependent. For this, we either treated or not with neuraminidase which cleaves off a cell surface sialic acids, and then we infected the cells with anaplasma.

As you can see, when the cells are treated with neuraminidase, the cells fail to support bacterial entry and replication. And here on the bar graph on the lower right, you can see that when the cells are treated with neuraminidase, we found no bacterial application by QPCR. Thank you.

DR. MCGIVERN: Dr. Atul Rawal.

DR. RAWAL: Okay. Good morning. My name is Atul Rawal, and I'm a postdoc with Dr. Dumisana's (phonetic) lab. And today we're presenting our work on using machine learning to identify HLA variants for either symptomatic or

asymptomatic COVID-19 disease.

Okay. So the HLA has been known to play a crucial role in the human body's immune response to different diseases. Literature has shown a possible role that HLA could play for COVID-19 as well. So to investigate this, we took a machine learning approach on applying different machine learning models on HLA data to see whether we can first predict a patient will have symptomatic or asymptomatic disease, and then identify specific HLA alleles that are more probable for symptomatic versus asymptomatic disease.

The chart here shows the, the chart on the left shows the different machine learning approaches we took. And then we chose the top performing model to apply explainable AI to generate the feature. Oh. Sorry.

The feature relevance part to highlight the alleles that play the most important role or the most impact versus the lowest impact. So based on these charts, especially the lower chart right here, we're able to identify specific HLA-B

alleles, which will either be asymptomatic for COVID-19 disease, or symptomatic for COVID-19 disease.

Based on our results, we share that HLA-B4001 has the highest association with asymptomatic disease, and HLA5101 has the highest association with the symptomatic disease. So HLA4001 will be protective against COVID. And then 5101 is addressed for COVID.

These results show that AIBS analysis can help physicians in clinic to provide better personalized treatments for patients based on their HLA and their clinical outcomes. Thank you.

DR. MCGIVERN: Thank you. Dr. Erica Silberstein.

DR. SILBERSTEIN: Hi. My name is Erica Silberstein. I am a staff scientist at the Laboratory of Emerging Pathogens, Office of Blood Research and Review. And the title of my poster, which is number 49 is single cell transcriptomics reveals the immune landscape of the mouse colon during chronic Trypanosoma cruzi infection.

Chagas disease is transmitted by T. cruzi, which could be transmitted by triatomine bags, blood transfusion, organ transplantation, and from mother to baby. So 30 percent of the, 20 to 30 percent of the infected individuals will develop chronic Chagas disease in the chronic phase.

The parasites persist in three main sites, the heart, the GI tract, and in the skeletal muscle. The goal of this research was to study the immune cell landscape of the chronically infected mouse colon. And also to determine the mechanisms associated with parasite persistence and pathogenesis.

And to do that, we infected mice. We waited three months until they became chronic. We then inspected the colon tissues. We prepared single cell suspension. Then performed cell capture and library preparation using the VB Rhapsody system. And after sequencing, we conducted the biostatistics analysis, bioinformatics analysis.

So and we found 18 cell populations. And in the infected colon, we observe the expansion of TNNK cells, and also B cells. Further, sub clustering of the TNNK cells population revealed the different T cell subtypes, including T helper cells, T regulatory cells, T cytotoxic cells, and T follicular helper cells.

Using this chart communication analysis, we predicted that the CCL signaling pathway network is activated in T cells and macrophages with the CCL5, CCR5 ligand receptor, they're representing the highest contribution to the CCL communication network.

Also, we conducted some flow cytometry experiments. And we were able to identify a higher percentage of CDA positive and CCR5 positive cells in infected animals.

In conclusion, we observe recruitment of T, NK, and B cells to the colon of chronically infected mice, and we also found that the CCL signaling pathway could potentially be implicated in cell recruitment and control of parasite

growth. Thank you.

DR. DEY: Dr. Emily Smith.

DR. SMITH: All right. Hi, everyone. My name is Emily Smith. I'm a postdoc fellow in Paul Carlson's lab in OBRR. And today I'm going to talk to you a little bit about the investigation of novel phage resistance mechanisms in vancomycin resistant enterococcus, or VRE.

So antibiotic resistance is a critical public health concern and vancomycin resistant enterococcus, or VRE, is a multidrug resistant bacteria that causes severe disease in humans when it impacts the urinary tract, the bloodstream, and the heart, for example.

This can lead to hospitalization and sometimes death. The species that cause the most severe disease in humans are enterococcus faecalis and faecium. Now, as an alternative to antibiotics, we can use bacteriophage or phage therapy to treat VRE infection.

And phages are viruses that kill bacteria. And phages can use a variety of

different ways to infect and kill bacteria, and in turn, bacteria have developed mechanisms to resist phage killing.

So the two main questions of my project are, one, what resistance mechanisms could interfere with phage therapy to treat VRE infection, and how can we design an effective phage cocktail to combat these very resistant mechanisms.

And so we used an unbiased method called transposon mutagenesis to create a collection of mutants in two faecalis, or two faecium and one faecalis strain using a Nisin inducible transposase and a Mariner transposon contained on the plasma shown here.

We were able to create a collection of mutants to test these questions. And so we grew these transposon libraries in either BHI broth or auger with and without the presence of Nisin in order to induce the transposase and collected bacterial pellets for DNA extraction.

And we plan to use transposon sequencing

with these samples. And so we're in the beginning of this work. Our future plans are to use these transposon libraries in phage infection experiments and compare before and after phage infection to ultimately identify specific genes that are responsible for phase resistance in VRE.

Then we plan to make specific mutants in VRE, and test these in vivo using a newly developed mouse model for urinary tract infections. The ultimate goal is to use this data to inform future phage therapies and combat this resistance that we see in the clinic. Thank you.

DR. MCGIVERN: The last flash talk speaker is unable to join us today. So this concludes the flash talk session. Now it's time for the poster presentations and lunch. The odd numbered posters are going to be manned and so please go and visit the posters, enjoy the poster session, and join us back here at 2:00 p.m. for session 4. Thank you and thanks to all the speakers.

(Recess.)

DR. FRIEDBERG: Hello. Welcome back, for people who've been here. Welcome to the room for people who have not. My name is Daron Friedberg. I'm cochair for this session. And this is my cochair.

DR. RODRIGUEZ: Hi. I'm Marisabel Rodriguez.

DR. FRIEDBERG: The first session, I'm sorry, the session is divided into 2 parts. The first 2 speakers are external, and we'll take questions and answers after they speak. And the last 3 speakers are from CBER, and we'll have a panel on that for questions.

So hold your questions for those and we'll introduce them. It's my pleasure to introduce our first speaker, Rommie Amaro from UC San Diego. She's a professor of molecular biology, and today she's going to be, oh, I should say one more thing. She did a calculation of the spike protein on SARS-CoV-2, which now has made the cover of the Essentials of Glycobiology. So that's a claim to fame.

Okay. So Rommie will be speaking on computational microscopy of viruses. Take it away, Rommie.

DR. AMARO: Thank you. Thank you very much. All right. So I'll just go ahead and get started. Thanks for having me today and thanks for that nice introduction. Okay. So what I am going to tell you all about today are the methods that we're using, which are pretty much entirely computational, in order to understand the structure, and dynamics, and mechanisms of viruses and viral infection, as well as how antibodies and other molecules, you know, can interact with the different components on, you know, in viruses.

And so the methods that we use, primarily what I'm going to tell you about is a technique called molecular dynamic simulations. And I like to think of it as a computational microscope because that's exactly how we use it. So nowadays, you know, biological data is very diverse and in order to answer any biological question in a meaningful way, I think that most of

us appreciate that you really need to use multiple types of experiment, right, to interrogate a particular question.

And so biology is multiscale. It requires multimodal experiments, and that doesn't, you know, I think that really holds across most questions. So nowadays when we're trying to investigate viruses, it's common for us to use these computational techniques as a way to integrate many diverse sources of experimental datasets.

And then extend these datasets with physics based simulations. And so what we can do is take different types of, for example, structural biology data, which can be acquired now using different instruments at different resolutions.

So for example, we can combine high resolution x-ray crystallographic information together with single particle cryo-EM data together also with things like cryo-electron, methods like cryo-electron tomography to really

sort of understand structure from the molecular to subcellular scale.

And then beyond just the structural data, we can further augment that those datasets with data coming from mass spectrometry and other types of methods. For example, glycomics, lipidomics, and genomics.

And so we can now what we do is we take all of these different experimental datasets, and we bring them together to build a highly detailed 3 dimensional model of a biological system. And then what we do is we approximate that system down to its many atoms. All right.

So we basically assume that each atom is sort of like a hard sphere and then we define this potential function. And I promise this is the only equation that I'll show you. I work really hard to take the equations out of my talk. But so this equation basically just, this potential function basically just describes the interaction that each atom has with all of the other atoms in the system.

And then we, all we're basically doing is we're integrating Newton's equation of motion over time. So we start with some particular configuration which we build based on experiment. We then integrate one time step. We get a new structure, we integrate again, we get another structure.

And we do this numerical integration millions and billions, and now trillions of times. And what that allows us to do is build up sort of a dynamical movie or what we call a trajectory that describes the systems motion over time.

Okay. And this is sort of, this allows us to get sort of new views into the structure, and dynamics, and systems that is currently inaccessible with direct experimental techniques.

And so the other thing I'll just mention here is that these methods have become ever more powerful due to a real just huge growth in compute architectures that have taken place over the past couple of decades. Most notably, the development of graphical processing units made by companies

like NVIDIA, were really sort of game changers.

And you know, now it's really common to think about, you know, and there's a huge focus, of course, on machine learning and artificial intelligence. Our field has been using these same architectures now for, you know, almost 2 decades or 15 years or so, thanks to the gaming industry.

So the same types of chips that they used to make really cool, you know, graphics for these video games actually are highly amenable to the types of scientific computations that we carry out.

Okay. So in today's talk, I know I don't have too much time, but I wanted to sort of just touch on the methods, generally, and give you some examples of how we've used these types of simulations to inform on structural dynamics of viral targets, on cryptic epitopes, which are important for biologics evaluation, as well as for vaccine design.

So in the first part, I'll talk about the structural dynamics. So I think most of you,

everyone in here by now is probably very familiar with the SARS- CoV-2 virus. It looks like a golf ball with spikes sticking out of it. And in fact, those spikes are called the spike protein.

They're very important. They sit on the outside of the virus. And so they're the first point of contact that the virus has with human cells. So they play a really key role in infection. On top of that, they're also highly immunogenic molecules. So they play a very important role in all of the approved vaccines that we currently have for COVID-19.

And so, you know, experimental structural biology provided very quickly at the start of the pandemic, provided really key information about what this spike protein looked like. There was a wonderful paper by Jason McLellan and colleagues at the National Institutes of Health that was put into the bio archive, actually, on Valentine's Day of 2020, and published in Science about one month later. And now has been cited, I think, over 15,000 times.

It basically, this paper basically provided the first high resolution structure of the SARS-CoV-2 spike protein. They had resolved it in what's called a one up confirmation, where part of the spike protein had one of its domains, the receptor binding domain, in the up confirmation.

And this was really important data. A couple of weeks later, another fantastic structural biology group, the Group of David Geisler at the University of Washington, published a second study that largely recapitulated the findings of Jason's group. But they also used symmetry to create a model of the of the closed spike, where all this, the RBD, the receptor binding domain, and the spike were in the down confirmation.

And this RBD plays a really key role in the infection process because it's actually the first, it's the part of the spike protein that makes contact with the receptor on the host cell called ACE-2, or angiotensin converting enzyme 2.

So you know, we and others are really interested to understand how, you know, how this interaction was taking place. So one of the limitations to the experimental structural biology is that if parts of the structure are highly dynamic or flexible, then they're very difficult, if not impossible, to actually structurally resolve.

And so when you look back at these structures, immediately you'll notice that there are missing loops and other sort of entire domains of the protein that are just missing. So we use computational methods to basically rebuild these or to sort of complete these models and provide these missing loops, et cetera.

And we could do that using a number of different computational structural biology techniques. Beyond the head part of the spike, which is where, you know, a lot of interesting things take place, it also is connected or tethered to the viral membrane by a stalk domain, which also is very flexible and therefore, it was

not present in the experimental structural dataset.

So we used computational methods. At the time we were using Ipasir layer and another program called Modeler, that actually predated Alpha Poll, believe it or not. But if you repeat these experiments with Alpha Poll, you find largely the same structures.

So we basically, we predicted what these other domains that were, that they couldn't resolve experimentally what they would look like. Beyond just the protein itself, another angle or aspect of these viral glycoproteins that's really important to understand are N and O linked glycans.

And N and O linked glycans are a very important post translational modification. They help proteins fold, they help with trafficking, they do a lot of different things. But they are really highly dynamic. And again, what that means is that the structures that, the (inaudible) structures that became available, they actually

didn't have these glycans actually on them, because they move too much.

But so what we could do is link to glycomics data, and again, this is some of that mass spec data that I was talking about. The group of Max Crispen, again early on in the pandemic, I think by April, provided this paper along with Parastoo Azadi, at the University of Georgia, there were several papers that basically use glycomics methods to determine at each of these different N-linked glycan sites.

And there's 30 some odd N-linked glycans on each of the chains of this trimeric spike protein. They could actually determine sort of the molecular recipe of these glycans. And we could rebuild them. So these are some of the different oligosaccharides that they, you know, that they find here in this.

The mapping corresponds, the colors correspond to the images on the right. We could actually rebuild these into the structure. And then so we have this complete model that we've

built that has all of the domains, all the missing loops, the correct protonation states, et cetera.

And then the end link glycans, and we then are, we then simulate the structure to try to enumerate the sort of, you know, the dynamic and what it looks like. And I'll just say, so what you can see, what I hope sort of jumps out at you is these, so the glycans are those colorful bits on the structure that looks sort of like ornaments on a Christmas tree.

You can really see how they're moving around quite a bit. We also have solvent and water that has ions, and that's sort of like a buffer condition. I don't show you that. We sort of stripped that out of this movie so that you can actually see the protein.

But the other thing I just want to say is, you know, they're, you know, when you see this, when people see this, you know, it strikes them, it's a beautiful image, you know, it's a pretty picture. But it's more than just a pretty picture.

What we actually are doing here are showing the time dependent dynamics as predicted by these molecular dynamic simulations. And why that's important is because these molecular dynamic simulations are actually numerical statistical mechanics.

And so they are computed, these different positions are computed in accordance with the laws of statistical mechanics. And so what that means is that we can link microscopic, time averaged microscopic properties that we've computed in the simulation, directly to experimentally testable macroscopic observables.

So we can compute from these simulations, we can compute things like free energies of binding, entropies, heat capacities. There's all these things that actually are, that can be directly, again, sort of correlated to or compared to experiment.

And so these speculations are expensive, but they are also very useful in actually informing on sort of mechanism and dynamics of

these proteins. And so okay.

So what did we learn. So one of the things that was sort of, of course, high interest was that we were able to show the world what the spike protein actually really looks like. So on the left, in light blue, is essentially what the structural biologists see.

They basically are able to see the protein, just the protein. On the right is what, though, is what the (inaudible) would see if they could see the glycans. And so now we're actually showing the glycans here in these blue, are the positions of the end link glycans.

These sort of puffs that you see actually represents a composite image of all the different snapshots that are sampled from one end glycan over about a microsecond of dynamics. So what you see is that these glycans are basically moving around very rapidly on the surface of the structure of the spike protein, and creating a very good shield that basically hides the spike protein from the human immune system.

The other thing that these simulations showed was why this spike actually adopts these different confirmations with the one up confirmation of the RBD and all down. And so on the left, I'm showing the slow spike, and what you see is if you're looking top down so that the top two panels here are showing the top view. And on the bottom is sort of the side view of the spike.

What you can see is this blue domain here, this light blue domain, that's the receptor binding domain. That's that really important part of the spike protein that has to make contact with the ACE-2 receptor on the host cell.

And in the closed confirmation, what you can see is that this receptor binding domain is largely covered by glycan. And in fact, if you look from the side, you can see that it's actually sort of tucked down beneath that glycan's field.

That's in stark contrast to the open spike, which I'm showing on the right. The open spike presents a very different picture to the host cell. You can see now in light blue that the

entire top surface of the receptor binding domain is exposed and ready for binding.

And in another sort of striking image you can see from the side how now all of a sudden the receptor binding domain, which previously was buried, is now sort of popped up and ready for binding to the host cell.

One of the other interesting things we found was that beyond shielding, so everybody kind of knew that these glycans would play important roles in shielding, because we have known that for some time. But what was unique about this study was that as we were working with the protein, we found that there were two N-link glycans.

When we rebuilt their side chains into the, into this one up confirmation, the glycan side chains actually filled up the void space that was created when RBD moved up. And so it occurred to us that these glycans, these two glycans, maybe they were doing more than shielding. Maybe they were actually helping to act like a structural support of the RBD.

You can see here how they sort of look like a kickstand on a bicycle helping to prop up that receptor binding domain in the up confirmation. And so this was actually tested with experiment by Jason McLellan and coworkers. And we actually found that when you mutate these residues, actually the ability of the spike protein to interact with ACE-2 is highly reduced because it's sort of pulled off that structural scaffold.

So this was actually one of the interesting things, and I think one of the reasons why we made the cover of the Glycobiology Textbook, was because there's a really nice example of using these computational simulations to really learn a whole new element of biology, and to show, I mean, because we showed for the first time that glycans would actually do more than just shield. That they actually played a role in the viral infection mechanism itself.

And I know that, you know, the title of this session is biologics evaluation. I just

wanted to say one of the really nice outputs of these simulations, as one can imagine, is that you can look at the solvent exposed surface area of these viral glycoproteins to actually determine which epitopes are exposed at particular times. And to help rationalize why some antibodies can bind to some structures and not others.

There's many examples of this now in the literature, and I just wanted to show this really interesting one. I mean, I think to me these viral glycoproteins in particular spike is just so fascinating scientifically. What it can do.

And also our human, our response of the adaptive immunity. One of the things I found really interesting that was published in late 2020 was this finding of an antibody, S2M11, that actually was able to recognize that even the closed version of the spike protein, and it actually did that by using a ordinary epitope that would, or yeah, the epitope actually it was formed across three different subunits. And the antibody actually just binds to this tiny little bit that's

exposed at the center top of the spike, which I just find so fascinating. Okay.

So beyond just looking at sort of the, sort of the local dynamics around these experimentally characterized space. Right. So the closed state and the open state, we can also use like what I guess I'll call advanced or enhanced sampling simulation.

Now, to start with, to actually look at sort of mechanism of opening. Okay. So before we were just looking at the closed state and then the open state, but here what we're doing is we're starting from the closed confirmation of the spike. And you'll see this movie playing here, again, the RBD is shown in light blue.

We can use something called weighted ensemble molecular dynamics. To actually sample the whole opening pathway. And this is the biological phenomenon that happens on the order of seconds. But here we can actually retain ties to the structure while we look at that mechanism.

And what's interesting about this is

that if you just looked at the, there was one additional glycan that we found, actually, when we looked at the opening mechanism, and it's shown here in hot pink, that in our previous study when we just looked at the dynamics of the closed state of this like, this pink glycan is just sort of hanging out by the side down by the side, and it doesn't look like it's doing anything interesting.

When we simulated the open state of the spike, it's the same thing. That pink glycan is hanging out at the side, doesn't look like it's doing anything interesting. But when we actually do the work to look at the opening process, what we see is that this hot pink glycan is actually doing all sorts of things to actually promote the opening of the receptor binding domain.

It's making all sorts of interactions. And then at the,end actually helps sort of kick it open, almost like a crowbar. And so Jason and his colleagues went in and created more mutations. And what we found was that this glycan at position 343 was actually even more important than the two

structural glycans that we had previously found.

And in fact, if you mutate this, the way this glycan, the virus actually has, can't even infect cells. And that's been something that's been replicated now in many other labs. That was actually also kind of an interesting finding that was derived from simulation.

Now in the second part of the talk, I want to talk about cryptic epitopes. So one of the things we're really interested to do is sort of push these molecular dynamic simulations to longer time and length scales that are of interest to biological, you know, questions. Right.

So we now can look and use cryo-electron tomography experiments and there was a whole bunch of beautiful data that came out in the late summer of 2020, that used these high resolution techniques. So cryo-electron tomography with the milling to actually catch infections of, viral infection, basically, in situ.

And basically the data that we can get out is sort of really what the viruses look like

as they're infecting cells. It tells us through a molecular map that we can use to rebuild the virus. Again, sort of computationally.

It gives us the distribution and patterning of the spike proteins, as well as can inform on which ones are open, or bent, and so forth. So we use that data together with additional computational simulations that I won't go into detail about, to build the delta variant of the virus.

And this had about 30 different spike proteins. We built the membrane. We also, it contains a few copies of the E protein, as well as hundreds of copies of this M dimer, which is sort of like also giving some structural stability to the virus over its life cycle.

And then what we're doing is basically, again, we have this structure and now we can simulate the viral, the spike proteins. And if you squint, I don't know how big the screen is there, but you can actually see some dynamics here.

This is actually now a system, it's about 305 million atoms representing a full virus of like, you know, standard, you know, size, with a membrane about 90 nanometers in diameter. Probably about 120 end to end diameter.

And we can look at its dynamics. And so in addition to studying SARS 2, we've also studied flu. And so here we have large scale simulations of the influenza virus. Influenza, as you can see, is much more crowded.

There's two proteins that are of interest here in flu. One is hemagglutinin, and then the other is neuraminidase. Here the hemagglutinin is shown in blue. And the neuraminidase is shown in red. And we can build different strains and then compare their dynamics.

And what we find is that when you, when we simulate these glycoproteins in the context of their crowded environment, of their real sort of instant (inaudible) environment, instead of in these dilute solutions, essentially, we see a different landscape of dynamics.

And here, I'm trying to draw your attention to hemagglutinin. And this blue molecule, one of the things you'll see is that one of the chains actually sort of begins to open, and to breathe open. And in fact, here's another view of that. And one of the HA's on the surface, there's hundreds of copies of them on the surface of the virus.

You'll see that originally it starts like the cryo-EM structure. All of the domains are very tidy and well packed. But over time, you can see this breathing. Look at that breathing here. You can see that happening in this chain on the lower left and here from the side.

You can sort of see this breathing motion. And here the glycans are colored yellow. So what we found was that when we actually simulated in the context of the full virus that we sampled this breathing motion, as I mentioned. So here, what we're looking at is sort of what we call a time series analysis. This is showing basically how open the structure is.

It's like a distance metric for opening. It's closed at the beginning, close, close, close. And then sometime later, it achieved an open state. And then even achieves what we call a super open state, where you can really see this sort of like this blade open.

And this was really curious because it turns out that around the same time, an experimental group, the group of Ian Wilson and James Crowe, found an epitope called flu A-20, which actually was found to bind to an epitope that was on the inside of one of these HA's.

And this is really interesting. And I'll comment from their paper that I just want to read here. It says because everybody when we, you know, when (inaudible) give you a structure of a protein. Okay. And if you look across all of the hemagglutinin and spike structures, they mostly look the same. They're like very tidy, they're very packed, and you know they look very pretty.

But the truth is that these proteins are so much more than that one snapshot that they're

able to characterize experimentally. And we know this as scientists, but it's still so hard to imagine sometimes, you know, what these different intermediate states could look like and so forth.

But here from their paper, they say, therefore, the HA molecules on the cellular or viral surface generally have been considered to be stable primers. With the trimer interface regarded as inaccessible and thus not targetable by the immune response of therapeutics.

That's generally because people have been shown that cryo-EMAIL structure of everything all pecked up, that's how we imagine it in the actual cells. But here what they go on to say is the ability of flu A-20 to convert in vivo protection strongly suggests that HA molecules are dynamic and more heterogeneous in their confirmations than we have observed previously. And that the trimer interface is partially or transiently accessible.

This phenomenon known as breathing. And this is exactly what we sampled with those

molecular dynamic simulations, the scale of the virus. We actually saw the opening of this hemagglutinin. And here's this flu A-20 antibody in the closed state. So here you can see a tightly packed trimer. It's total totally clasping. You can't make any sort of like finding mode.

In the super open state, it is able to accommodate very nicely this antibody. So we were able to sample this dynamics, which was cool. And then as I was talking to Matharu Kanekiyo at the NIH Vaccine Research Center, I gave a talk there, and we were explaining well something else that we see in the data, because we always talk about HA.

But there's also neuraminidase. And one of the interesting things we saw about neuraminidase was that in simulation, it actually had a very big (inaudible). Like so, and again, to me what is still so remarkable is that when people think about the neuraminidase molecule, it's like they generally, structure folks, they think about it, it looks sort of like a bouquet or

like a flower bouquet where you have this sort of head domain, where all the kind of interesting neuraminidase enzymes are.

And then it's on this long stalk. And we never really thought too much about it actually bending, but we actually sample really large bending motions also for neuraminidase. And it turns out that they had found an antibody from patients that they were calling the dark side antibody because it recognized an epitope that was sort of like underneath the neuraminidase, that was on sort of, that was closer to the inside of the virus.

And these, our simulations actually helped them to rationalize how this epitope would actually get exposed. Because when the head bends away from the stalk, there's a different exposure of the epitope. And also you can really, we could clearly model that these antibodies could actually fit inside and bind even in the crowded surface of the environment, just of the virus surface, just based on sort of the structural dynamics that

changed, and that we sampled with the simulation. Okay.

So I still have, I think, just about 10 more minutes because we started 5 minutes late, but I'm going to zip through this last part so that we can get to the questions. You know, a lot of what we do is really focused on basic research. I love discovering stuff just about general biology. For me, that's sort of so much fun.

But you know, we're also interested to do things that are applied and directly useful. And so we started to, we became involved in a project to look at immunogen design. And so of course, all of us, unfortunately, are very familiar with the fact that SARS-CoV-2 is, you know, able to mutate like crazy. Of course, we've given it an enormous number of shots on goal for doing, this because we're not controlling infection.

But in any case, and there's been tons and tons of mutations in the S1 domain of the spike protein, which is pretty much everything

that I was talking to you about, the RBD's, the NTD's of the spike protein.

And these mutations change the available, you know, epitopes and our adaptive immune response and leave us more vulnerable to being infected again. So but we became involved with the project with Jason McLellan and coworkers to try to design an immunogen that would go after some other epitopes on the spike.

There's one down in the stem region, and then there's a fusion peptide epitope, which is shown here. But then there's also this what we call an S2 apex epitope, which is on the part of the spike protein. I don't want this to restart. Here, let me pause it.

So what we're looking at now is that spike protein that I've talked to you about before, but we've pulled off the S1, because actually S1 does come off. It's basically shed off of the spike protein in the process of infection. And inside here is, the F2 domain of the spike protein is highly conserved across like

many of the beta coronaviruses.

So we were trying to see if we could help them design an immunogen that would be S2 derived. Now the challenge is that this is a highly unstable part of the molecule. And it's unstable for the purposes generally of infecting the host cell, because once you lose that S2 cap, those, you know, the F1 cap, those domains open and then the central helices are like, what did they say, they're like spring loaded.

They're like, literally, spring loaded, and it comes, the top comes off. And then these helices basically like polymerize upwards. And if you look mechanically, it's this amazing system. There's all sorts of like structural strain that's evolved into these proteins.

Anyway, but what that means is that when you take off S1, S2 is highly unstable. And so many folks are already familiar with the 2P mutations that are part of the spike protein. This was super important for the initial all vaccine development because it stabilized the

protein, and the one up confirmation, which was super highly immunogenic. And that's what got into the mRNA vaccines.

But then over time, Jason and his group also made this construct of a spike called PeptoPro, which had an additional set of proline mutations, and it had really good experimental performance. Like most labs, many labs, you know, when they're doing various SARS experiments, will use these stable constructs because they express better and so forth.

And so the issue, though, is that if you, you can't just directly necessarily use it, it's not a very good immunogen because it's just sort of like, it's just not stable. So it could be expressed, but it doesn't maintain a oligomeric state. And unfortunately, because of that also, they're not able to characterize the high resolution structure using cryo-EM, because it's just too flexible.

So we went at this question of can we use simulation, these same types of methods that

we use to look at the S1 domain. Could we do that to understand S2 dynamics. To understand how it opens and then to design mutations that would prevent that opening from happening to create a stable construct.

So we used, again, the same methods I talked about before. These weighted ensemble simulations to understand how best to actually open and let me play that again. What you can see is that, you can see it sort of falling apart. It's sort of opens. And here it's going slomo. You can sort of it, it begins opening by sort of splaying at the top, and then it sort of rapidly unzips.

And so there are, you know, again, this is as many of these fusion proteins, it's a trimer. And so there's three sets of helices at the (inaudible) interfaces that mediate the opening. So you know, we could actually really look at sort of what interactions were happening in opening.

And what we saw was that chain A breaks

away from chain B initially. Then chain A breaks away from chain C. So like one whole chain kind of comes apart. And then those last two chains separate. And then when we were looking at so, okay, so we understand the mechanism of opening. But now like, let's look at the interactions and try to figure out if we can design mutations that would stabilize, you know, stabilize it in the closed state, to stabilize the apex epitope in that too.

And what we immediately saw. So just, and I'm sorry because I don't have too much time. So I'll just say what we're showing here is looking at sort of the interaction profile over the dynamics of each of these, there's like rings of residues that form through the central helix. And they're like residues coming in from these central helixes.

And what we saw were there were two positions in particular at the top. This position 991, and then two down from that at position 998, where it's really yellow. And yellow means it's

not really making good contacts. And these were also some of the first to unzip, because it's unzipping from the top down.

So we went in there and we said this is where we want to mutate first to try to strengthen these mutations to slow down opening. And so we actually put in tryptophanes, and it turns out now you've seen all those, the V9991 line and the T998 line actually go to pink, from yellow to pink, which just means they're really sort of making a lot of contact now.

And in fact what we see is that when we put in two of these tryptophan mutations, so tryptophane at this top position, 991, and also 998, we see that there's a synergistic dynamic coordination of residues that stabilize the flow state.

And there's all sorts of interactions that are happening. I have one last movie to show you. But this is again sort of looking at this. This is the mutations that we made. And you can see how it's, how these sort of rings kind of come

together to create this sort of much more stable network of interactions.

Not only with each other, but also sort of, you know, in the flanking residues above and beneath and around also. So there's all manner of mutations are happening. And the other thing I'll say that sort of I think fun for me and fun for us, is that it's a way of doing immunogen design where it's very information rich.

So there's, you know, there's a way of doing protein design where you just, you know, sort of make a lot of mutations, and then you sort of see what you get. This actually, you know, we go through first learning about the protein, understanding the mechanism, and then designing these.

And so with just these two suggested mutations, so very, very few experiments, we didn't make any other suggestions. We were able to develop this construct that was significantly more stable experimentally, in terms of differential scanning. Sort of like temperature

studies.

As well as it was stable enough that they were able to resolve that high resolution crystal structure or cryo-EM structure. So this is now the most stable S2 construct that we know about. And is sort of being, you know, it's going on, could be used in other types of experiments now.

And again, this was the work, the structural work, again of Jason McLellan and colleagues in his group. Okay. And so now I will be really happy to take questions. I know I presented a ton of information, but hopefully was able to give you a view of how you can use these physics based simulations together with a plethora of biological datasets to actually, you know, better understand the system, do biologic discovery, and actually do, you know, design work also.

I just want to thank my group. I have a wonderful group here in La Hoya, and just all of the COVID work was just, you know, sort of a very

special opportunity for folks to come together and work on some, you know, an amazing, very team oriented science. So it was difficult, but it was also very rewarding.

Lorenzo Casalino (phonetic), Christian Seitz (phonetic), Erica Honic (phonetic), are three folks who really worked really hard on the on flu, and also Lorenzo, and Zandra, and Jason, and Ling were really instrumental for the SARS 2 work.

And of course I want to thank funding, both monetary and also for computing, because these are very compute intensive experiments. And so with that, I'll be happy to take questions and discuss.

DR. FRIEDBERG: Thanks very much, Rommie. Do we have any questions in the room? I think we have one. All right, Mustafa. Is it not on? Just come up here.

SPEAKER: So I find that the movement of the glycans fascinating. It's almost like there's a current or a wind.

DR. AMARO: Mm-hm.

SPEAKER: Obviously this structure probably dictates that, but what about the environment, different, for example, the age groups, very old people versus very young, their sera might contain different ions. Do you think they also affect these movements? Therefore, expose the protein epitopes differently.

DR. AMARO: That is an excellent question. Thank you for that question. Yes. So you are absolutely right that the dynamics of these, oligosaccharides, essentially these glycans can definitely vary with different conditions.

Interesting that you mentioned this in the context of the spike accessibility. That's something I hadn't thought about. We know for very similar molecules like mucin, so I didn't talk about it, but we're really interested also to understand how the virus is actually interacting with like the host cells like (inaudible), and also with things like mucins and heparin sulfate, and all that.

And there, it's also, the environment is super important, and we, you know, we know from our studies, and I think now there's a couple of others that depending on the degree of, for example, dilation, so how much, like how many of these might be decorated with sialic acids.

These can coordinate to different extents, calcium ions. And when this happens, it actually can occur to, if you have enough calcium there and the right groups, it actually can have a like a local phase change, actually. Where it can almost form like a gel like structure in the vicinity of the virus, which of course will affect all, I mean that also affects all manner of different biological properties, including exposure, but also just infection and so forth.

So yeah, there's so much, but there's still really a lot more to be learned there.

DR. FRIEDBERG: Thanks. Online questions.

MS. ELKINS: Nope. This is actually mine.

DR. FRIEDBERG: Oh.

MS. ELKINS: And probably a stupider version of the same question. You mentioned solvent effects. But what are your baseline computational models assume in terms of (inaudible), that is salt, pH, buffer molecules, and how much can you vary that within the computational approaches?

DR. AMARO: Yeah, that's an excellent question. And sorry. Yeah, I sort of breeze through that, but I am, I'm going to, I'll see if I can find a slide. So in our constructs like of this, where it it's really sort of just the dynamics of the spike, this is really sort of what we do here when it's just a single protein is we try to link to in vitro conditions.

So it's usually like 150 millimolar sodium chloride. We're looking at pH 7.4, and then just water. So very simple sort of really like almost like a benchtop experiment. But we can definitely, and one of the things that as I, you know, we can already see hints of and that we

know is that as we make the biological environment more complex, that the dynamics change. Right.

And so we also in work that I didn't tell you about today, because I was focusing more on sort of biologic evaluation, which I thought meant more like on antibodies. We're also really interested to understand what happens, you know, understanding the dynamics of these viruses in the, you know, sort of like in the institute context, inside the human, and also in aerosol.

We're really interested in understanding the structural dynamics of different types of viruses in different aerosol conditions. And so in these experiments, and so I'm showing here an image of the SARS-CoV-2 virus. This is the purple and blue in the aerosol.

Here we have mucins, we have albumin, we have ions, we have lipids, we have different types. And what we've done is basically created a computational mimic of like lung surfactant. And so we can actually make, but to answer your question, sorry, that was sort of like a long way

around to answer your question.

Depending on sort of like what experiment we're trying to mimic, we can tune the ingredients to what we include computationally. And we can actually now include quite a range of substituents that mimic fairly closely the actual, you know, sort of in vivo or in situ conditions.

It did take, though, some work. For example the mucins, this was sort of these red molecules here. Those are again very difficult to structurally characterize because they're so highly dynamic. But we've been able to make models, and this is work that is just now in press, sort of developing the sort of the basic constructs that can be used in simulation of these complex scenes. I hope that answered your question.

DR. FRIEDBERG: Thank you very much for that. And I'd like to move on to our next talk. Thanks, Rommie.

MS. ELKINS: Whoa, whoa, whoa.

DR. FRIEDBERG: Oh, sorry.

MS. ELKINS: We have a couple online questions.

DR. FRIEDBERG: Oh, you have online questions.

MS. ELKINS: Do you want to go ahead and take a look at the online ones and let me know if you have trouble seeing those.

DR. AMARO: Okay. I can see them if we have time. I'll try to answer these quickly because I know everyone's on a schedule.

MS. ELKINS: Sure, go for it.

DR. AMARO: One question was about the furin cleavage site. It says the furin cleavage site of SARS-CoV-2 spike is in a disordered loop. And it sure is. Did you find anything interesting in that region in your simulation study?

So oh yeah. So we haven't actually published too much about this, but we expect to, I think, over the course of the next year. So yeah, I guess I shouldn't say too much about it at this time. There are a lot of very interesting dynamics, you know, it's a highly charged area,

which also means that it attracts particular types of molecules even beyond (inaudible).

That there are likely interactions with the host glycocalyx. And then depending on whether or not it's been cleaved or is still intact, there are definitely sort of differences in the dynamics of the spike. And I think, and others also probably suspect that dynamics at the furin cleavage site are probably really important for epistasis and understanding sort of the mutational range of these host proteins, you know, as they're sort of evolving in populations.

But so that was a non answer question to your question. So yeah, there's interesting things, but we're still learning about it.

And then to the second question from Carol Weiss (phonetic), do the glycans protect the spike from proteases and consequent degradation. That is a good question. And I guess also, or Karen, oh no, she the one I answered.

That region like for example, you know, do they protect it from proteases generally.

Possibly. Again, also that furin cleavage site, you know, is actually quite exposed from the glycans, which is also interesting.

And so it has its own set of dynamics that does leave it exposed even in spite of all of the glycans. So it's, you know, these glycans have evolved certainly with the host protein, and all of the other factors. So you know, in sort of like I'm trying to understand the mechanism of other proteins, and how they interact with it. That's the last question, so thank you.

DR. FRIEDBERG: Thank you very much, Rommie, for such an interesting talk. Our next speaker is Frank Delaglio, and while he gets up here, I can say I've known Frank for 30 years now. And it's been a pleasure knowing him. He's always supported us computationally from the, for our NMR studies.

And today he's going to tell us about chemometrics and machine learning to enable applications of NMR and biomanufacturing.

DR. DELAGLIO: Splendid. Thank you very

much for this privilege to present the work of our colleagues at NIST. Very good. Thank you. So the work that we'll present today, it's just this slide. Oh, there. Okay. Okay, very good.

So the work that we'll present today was conducted at the Institute of Bioscience and Biotechnology Research, which is a small institute in the neighborhood here, that's run jointly between NIST, the University of Maryland, College Park, and the School of Medicine.

And the work that we'll see today is work from a circle of groups at NIST IBBR who develop measurement methods to support biomanufacturing. And the folks on this slide are the folks who are developing NMR based methods to support biomanufacturing.

So we'll cover some different areas where we're applying NMR to biomanufacturing topics, works at different stages of completion, and in particular, problems that have associated computational challenges. And we'll try to emphasize a few kinds of analysis methods that

might be generally applicable to other folks in their work.

At IBBR, we're concentrated on two things, structural biology and biomanufacturing. And both of those topics are quite familiar to everyone here. Monoclonal antibodies have become the largest growing kind of therapeutic. And they account for about $150 billion a year or more in medical sales.

And among the reasons that these are important is because of the fact that on the one hand, we can engineer a monoclonal antibody that will target very specifically just about any target that we can come up with. But also since monoclonal antibodies of a given class are very similar, the methods for expressing them, and purifying them, and testing them, and delivering them are very similar.

And so knowledge, biomedical knowledge that we get about one particular antibody therapeutic can give us an advantage with the others. Recognizing the fact that these

monoclonal antibodies are so commercially and medically important, NIST, which loves to make reference standards and reference data, made the NISTmAb. The monoclonal antibody reference material.

And we'll see applications that take advantage of the NIST map and its existence. Now, in the NMR labs, we developed an NMR method that we call NMR spectral fingerprinting. And it's a straightforward idea. We take an NMR spectrum, and the details of the NMR spectrum reflect the structural details and chemical details of the underlying molecular system.

So for example, in this case what you see is an NMR proton carbon spectra, it's what NMR folks like to call hydrogen atoms. Protons and in a spectrum like this one, each one of these peaks represents one CH group somewhere in the monoclonal antibody. Position of that peak in one of the dimensions is the NMR chemical shift of the hydrogen atom. And the position that peak in the other dimension is the NMR chemical shift of the

carbon that it's attached to.

Now, generally speaking, we don't know specifically which particular atoms and amino acids each of these peaks belong to. But even if we don't have assignments available, we generally know the amino acid type that's associated with these different signals.

And just about anything chemically or structurally that happens to the underlying molecule will be affected and observable in the spectrum. Because these positions of the peaks are determined according to chemical bonding partner, hydrogen bonds, dynamics, the solvent, all of these details.

So if anything changes about the underlying structure of the antibody, peaks will move, they'll grow, they'll split, they'll disappear. So as an example, we have two cases here. Two kinds of spectra, two related antibody samples in overlay. And one of the spectra is drawn in red, and the other spectrum is drawn in blue on top of it.

So we have molecular systems where there are almost no changes. And if a person analyzes this data by eye, it's very difficult to under -- well, to find anything quantitative in the data just from a rough visual comparison.

So what we would like to do is find numeric methods where we can take spectra like that and convert them into some kind of answer. So for example, over here, we have 2 collections of replicate spectra of 2 kinds of samples that are very difficult to analyze by eye.

But I'm going to describe how we can use a technique called principal component analysis to directly take this stack of spectra on two sets of replicate measurements and turn it into an analysis like this, where we can make a determination about whether the two samples are substantially different from each other, or substantially the same.

So we're going to use PCA as a method to quantify this kind of spectral similarity. And here's how it works. In order to explain this

method, I'm going to refer to a test case where we take the NISTmAb monoclonal antibody and oxidize it with peroxide. And the effect of that is to oxidize methionine side chains on the surface of the antibody.

And when these side chains are oxidized, it modulates the therapeutic effect of the monoclonal antibody. It may reduce its effectiveness. It may increase its immunogenicity. So being able to monitor the oxidation state of these methionine's is an important aspect of biomanufacturing.

Now, there are many proteins that will bind to a monoclonal antibody. For example, the protein A. So we're going to prepare two kinds of oxidized samples. One, the NISTmAb oxidized all by itself. And then the other, the NISTmAb bound to protein A before it's oxidized, with the assumption that protein A is going to cover up, protect certain side chains from oxidation. So the pattern of oxidation in these two kinds of samples will be different.

Here's what the actual spectra looked like. So it turned out that the particular protocol that we use to oxidize these sample, oxidize them both almost completely. So the mAb that's oxidized all by itself is 99 percent oxidized. And the mAb that's oxidized in the presence of protein A is 95 percent oxidized.

So the region of the spectrum where the methionine signals show up looks empty on rough inspection. But if we go very low in the contour levels, we can actually see some small residual signals.

So in gray here is what the intact native form of the antibody looks like. In blue is the signal leftover in the NISTmAb that's oxidized all by itself. And in red, you see there are a few more signals. That's the sample that was oxidized in the presence of protein A.

So some of those side chains were protected by protein A. So we can see a few more signals. But these contour levels are at five times lower than the ones here. So these signals

are very small.

We would like to find a way to take the entire spectrum and use it for analysis purposes. And the method that we're going to use is principal component analysis directly on the matrix of spectral intensities.

And the way this method works is that we consider that each spectrum in a collection is represented as one object in a high dimensional space. And the coordinates of that object are simply all of the intensities in the spectrum.

So if the spectrum has 100,000 points, we can represent it exactly as one object in 100,000 dimensional space. And if you think about this representation, spectra that are similar means the shape of the spectra, the intensities of the spectra are similar, so they're going to lie in related regions and close together in this high dimensional space.

But spectra that are mostly the same, but differ in a few particular features, are going to tend to align along lines and curves in this

high dimensional space because they'll have one set of coordinates in common. Those are the parts of the spectra that are similar. And some other collection of coordinates that vary in a continuous way.

In principal component analysis, we take this huge high dimensional representation of the spectral series and more or less simply fit it to a series of multidimensional straight lines. And this gives us the direction of maximum variance in this high dimensional data and lets us orient the projection so that we can take data, for example, like our case where we have replicate measurements on two samples and compute the first principal component, which helps us determine a direction of maximum variance.

And we make a projection down to a lower number of dimensions to get a result. So we can inspect the results this way in the form of a scatter plot, or as I'll explain, we can also express the results of this kind of analysis as pseudo spectra.

But in an analysis like this one, we will project our data down to a small number of dimensions by PCA and look for clusters and use the distance between them as a way to gauge the difference between these different classes in a quantitative way.

So here's an example of the case that we've been talking about. Here's a scatter plot of 3 sets of replicate spectra, native NISTmAb, the oxy, and A oxy samples that we've been talking about. And what you see here are the principal component spectra that corresponds to this principal component scatter plot.

And as you'll see, the first component is very much like an average spectrum. And the remaining components will highlight the regions of the spectra that are different between the different classes. And so what you see is, in the presence of this oxidation, the signals, which are drawn in red and blue, are in the higher components.

So differences associated with oxidation

are not limited to just the methionine region of the spectrum. They're actually dispersed throughout the whole spectrum. So there's a very strong and robust collection of information in here, if we could find a way to get it.

Now, another way to think about principal component analysis that should make sense to a spectroscopist is as a spectrum decomposition. And I think this is a nice way to understand it, as well. So using a kind of simulated version of our example where we have replicate measurements on two samples, and we have a stack of those, and that's the data that we're going to analyze by PCA.

So to compute the first component, what we do is find a linear combination of all the spectra in the series. And we choose that linear combination in a way to match the entire series as a whole, as closely as possible, in the least square sense.

The coefficients that we use to mix the spectra together to generate that linear

combination are the PCA score values. Those are the values that we see in the scatter plot. And the linear combination that we generate is the PCA loading spectrum. So that is the linear combination of all the spectra in the series according to these linear coefficients.

And we extract these by different kinds of linear algebra methods. But in a case like this one where we have two samples, the first component spectrum looks very much like an average spectrum across the spectral series. So we have these score values. We have this loading spectrum. We can take this loading spectrum and now subtract it from all of the spectra in our original series and that will generate a residual.

We can repeat this procedure as many times as we like. So we can take this residual and find a set of coefficients to make a linear combination of all of these planes. And in this case with 2 samples, 2 kinds of samples, that kind of spectrum tends to look like a different spectrum between the 2.

Now, in this case, we have 2 different samples. If the samples don't change during measurement, and if there is no drift during the measurement, then 2 component spectra should be sufficient to describe the entire series, no matter how many spectra there are.

And if the series needs more than 2 components, it's a diagnostic that tells you that the sample is changing during the course of the measurement.

Now NMR spectra are generated in the time domain. We measure them in the time domain, and we convert that data into a spectrum by Fourier transform. In a case of two-dimensional NMR data, we start with a stack of one-dimensional measurements and then we Fourier transform all the rows, and then all the columns, and that gives us a beautiful two-dimensional spectrum.

The Fourier transform is a linear operation. So the information content in all of these cases, the original time domain data, the interferogram, and the spectrum, the information

content in all those cases is the same. So this linear transformation of the data doesn't change the results that you get from principal component analysis.

This is all linear combinations of the data, as well. Now, as a little pointer that will be interesting to folks who do NMR, one of the ways that we can make these measurements more practical is to use a measurement technique called nonuniform sampling, which basically means we randomly skip over some fraction of the data that we would normally measure in a conventional experiment.

And we use special compressed sensing reconstruction methods to fill in the blanks and generate a good looking spectrum. Because if we try to Fourier transform data that has gaps in it like this, we get a spectrum, but the spectrum looks kind of crazy.

It has these things that seem like random artifacts. They're deterministically, they come in a deterministic way from the pattern of

points that we keep and the pattern of points that we skip.

But as I said, since the information content in all of these cases is the same, we can apply principal component analysis to a nonuniformly sampled spectrum that's been reconstructed only by ordinary Fourier transform. And a result like this can't be analyzed by eye. But we can extract the same information out of it that we could in a fully reconstructed spectrum.

So the take home message here is that not only does principal component analysis give us the advantage to analyze a spectral series directly without having to do any peak detection or complicated preprocessing, but it can also be applied to extract information that's not possible to extract by eye.

Now, to show some examples of where this information that we extract by PCA can cross over into issues of biomanufacturing, we'll take a look at some biophysical measurements on these oxidized samples. And so what we see here are melting

curves with respect to oxidation time. And we're looking at a change in the temperature of these two melting transition points.

And with longer and longer oxidation time, these temperatures get lower and lower as the protein gets destabilized. Then it's easier to melt. So not surprising, but we see this with a time resolved biophysical measurement.

We can also gauge the activity, the binding activity of the monoclonal antibody. So we've chosen three kinds of proteins, protein A that we already talked about, which binds the FC domain of a mAb. Protein F, which binds to the fAb. And protein L, which also binds to the fAb.

But these two proteins, A and F, cover up some methionine residues on the surface, which protein L doesn't. It binds in a place that's not near to any methionine residues.

So protein A activity decreases with oxidation time in the activity measured by SPR binding strength. Same thing with protein F. But protein L, the binding doesn't change.

We can also measure details of protein stability and aggregation by the method of relaxometry, also called water NMR. We can measure the water signal and how it behaves. And in particular, the relaxation properties of the water signal, which will reflect the proteins that are dissolved in it.

And so, as an example, we can measure relaxation time with respect to oxidation time and see that it changes systematically. One of the beautiful things about these kinds of measurements, they're conducted with low field benchtop NMR instruments, and you can actually take an intact vial or a loaded injectable syringe and put it intact inside the NMR instrument and make a measurement.

Not only nondestructively but noninvasively. So it's a very cool technique. So we have these three different ways to characterize the stability and activity of the monoclonal antibody protein. And then look at these beautiful results.

The information that we get from relaxometry, from binding activity by SPR, and from thermal unfolding, all of them can be correlated in a quantitative way with the score values that come on a principal component analysis.

And we didn't have to do anything complicated. Just take the sample, which is already dissolved in water, put it in an NMR tube. Make these measurements, take the spectra, stack them up, put them in PCA.

Now to continue with the topic of low field NMR, talk about another set of projects that is underway in the group, and as a follow on reference material to the NISTmAb monoclonal antibody reference, NIST is now also introducing the NIST show a live cell reference material.

It's a show cell that's been engineered to express the NISTmAb. And it's been engineered to express the NISTmAb in quantities that are significant in a biomanufacturing context.

So in other words, we optimized this

expression system in the same way that a pharma company would optimize an expression system. So now this is another beautiful test bed for different kinds of measurement techniques to support biomanufacturing.

Now, as folks know, biomanufacturing of these monoclonal antibodies is done in large scale bioreactors in growth media that have 50 or 100 different materials in them. And the properties and therapeutic activity of the end product can vary greatly with small changes in the conditions and protocols.

So there are lots of measurement needs. And a lot of them we could imagine could be nicely addressed by NMR. But a superconducting magnet is not very practical to put in a laboratory where either large scale or small scale bioreactions are going on.

So we have a project in our lab to build a closed loop system or an at line system, where we use benchtop NMR as a kind of universal detector for monitoring what's going on in a

bioreactor.

And what we do is monitor the reaction media, not the cells. And the compounds in that reaction media tell us what the state of the bioreaction is.

There are many things that we could do with information of this type. For example, very straightforwardly, when we measure a collection of NMR spectra over time, that collection of spectra becomes a signature for the bioreaction.

And without any further analysis, you can use it to decide whether one bioreaction is proceeding the same way as a previous bioreaction. So there are all sorts of things that you could do with this, including early detection of failure.

And you could also use it to optimize production of the particular materials. But most interestingly and attractively, if we want to develop methods of adaptive manufacturing, methods where the controls and protocols and materials of the biomanufacturing are adjusted in real time in response to measurements, we need a good

measurement technique that's time resolved.

So as an example, for our early work, here are NMR spectra, an excerpt of NMR spectra, measured on a 100 megahertz benchtop instrument from media collected from CHO cell growth over the span of 23 days. Okay.

And we can work with the spectra directly along the lines of what we've seen already with principal component analysis. And we can also do conventional analysis where we identify particular analytes and determine their concentration from the size of signals in the NMR spectra.

Now, something that we didn't expect or plan on, but actually turned out to be quite interesting, is that at benchtop resolution, we can also see signals from the monoclonal antibody that's been expressed and secreted into the medium. And we haven't done much with this kind of analysis yet. But the interesting thing is here's what that looks like with 500 seconds of measurement at 100 megahertz.

And here's what a similar measurement looks like on a 400 megahertz superconducting spectrometer. And what we've shown previously in other publications is that information that we extract with a 600 megahertz spectrometer and 1D data, is very similar to the kind of information that we can extract from a 400 megahertz spectrum.

These are PCA results of those two kinds of cases. Now when I look at this, I'm delighted to see that this antibody signal, even though it's a blob, it has features. It certainly has information in it. So I'm very hopeful that we'll actually be able to do things to analyze the monoclonal antibody signal as well as the media component.

Now, in our interest to work with and quantify metabolomics data, we've recently introduced an open source piece of software to automatically identify and quantify the systems in complex 1D spectra. And that software is SAND, it's developed by a graduate student, UAU, in Art Edison's group.

So he worked with me to develop this software, and our goal was to get a piece of software that was not encumbered by a graphical interface or anything like that, so that it could be used conveniently in batch mode. And that you didn't have to do anything other than put in a spectrum and the analysis region. And then you would get out a result like this.

So what you see here is the measured spectrum in blue. The model created by SAND in yellow, and the individual peaks in the model underneath. The way SAND works is in the time domain, okay, it models signals in the time domain.

And I won't talk too much about the details of how this is done, but a critical thing that we do, let's kind of take a page from machine learning, that's not generally done in NMR signal analysis, is use to paradigm that folks in machine learning use by dividing data up into training set and validation set.

And so we take our time domain data and

subdivide it in these two kinds of classes. And the training data is used to optimize the model, determine where the peaks are and how big they are and so on. And the validation set is used to determine whether or not we need to add more peaks to the model.

So that's the decision making part. And we use hybrid optimization of mixture of Monte Carlo and gradient optimization methods to work on this problem. And as an example of the results that we can get, these kinds of complicated mixture spectra like urine spectra that we show here, very difficult to evaluate an automated method. Because even when a person looks at these spectra, it's difficult to make actual rigorous determinations about individual signals.

So what we did to make the problem easier, is we took urine and made a four compound mixture that we could spike in in different amounts. And since we put the mixture components in, we know exactly where their signals are. And then we can quantify how those signals grow with

addition of more of the spiked material.

We get beautiful linear behavior. So that's the result from SAND. And then finally, this is a very early result, but it uses some aspects of machine learning that might be generally interesting to the audience. So let's talk about them too.

And as an example of why this is useful in a biomanufacturing context, we'll take a quick look at work from Anu Kaur (phonetic) in our group who's studying the effects of different excipients on mAb structure and dynamics.

And so as an example, your three common excipients, salt, polysorbate, and sucrose, salt tends to have a destabilizing effects on most cases. So in these melting curves, adding more salt reduces the melting temperature. Whereas in the case of sucrose, which increases stability, that tends to increase this melting temperature. And polysorbate in this case has no effect.

Now when NMR chemical shift assignments are available, that is to say, when we know which

peak in the NMR spectrum matches which amino acid in the protein, we can do detailed analysis of the dynamics of the molecule.

So we can do things about the overall dynamics of the molecule, whether or not it tumbles isotropically, and whether an excipient or other change will change the way in which the protein tumbles, because the structure gets elongated.

And we can also measure the internal dynamics of the protein. So in this case what you see is a map of how rigid each particular residue is under a particular circumstance.

So I won't talk about the conclusions of this particular work, but this just gives you an idea of what kind of information that we have available when we have NMR chemical shift assignments available. But to get them, to get these chemical shift assignments.

So what you see here is a proton nitrogen NMR spectrum. So in this case every amino acid gives one signal in this spectrum, and

the position of that signal is the chemical shift of the hydrogen atom and the amide nitrogen that it's bound to. Okay.

So from mAb, this spectrum is so complicated that we can't even label all of the assignments together. So in this illustration, we have all of the assignments associated with the light chain. And here's all the assignments associated with the heavy chain.

So these kinds of spectra, these kinds of spectra, have hundreds of peaks. And for every peak that you see here, probably 1 or 200 other peaks, related peaks from other two-dimensional and three- dimensional spectra, had to be analyzed and curated by an expert. Okay.

So this kind of peak analysis is the labor sink. The time-consuming aspect of trying to apply NMR for structural biology purposes. And so being able to address the peak detection problem in a more automated way is valuable across the board for many kinds of applications, because most every NMR workflow in higher dimensional data

requires identifying the peaks.

And we have lots of software to help us identify peaks, but generally it doesn't do very well in the case of the presence of small artifact peaks, or cases where peaks are only partially resolved, and they're not well resolved maxima.

Now, there are also machine learning approaches to analyzing NMR data in different ways. And universally, they use a representation of the spectrum that's equivalent to a grayscale image. But a human analyst works with this contour graphic.

Now, if we want to, what we're going to do is repurpose existing image analysis tools. And as you can expect, the way that you preprocess the data has a big impact on what you can do with it in a machine learning context.

So I want to say that identifying a peak in a picture like this, it doesn't depend on how big the peak is. It depends on the shape of these features. Right.

So this is not a good representation of

the information, because the dynamic range, the brightness is what represents the information in the spectrum, but we don't really care if a peak is 10 times bigger or whatever. We just want to find that it's there.

Whereas a picture like this, has predictable brightness, its brightness doesn't change. We have a. Black background and contour peaks. So this is a more suitable data representation for reusing existing image analysis tools.

And NMR has a particular advantage in machine learning approaches in that it's possible to generate simulated data that mimics experimental data very well. So as you can see in this case, and we can also use semisynthetic methods.

So the results that I'm going to show you work with actual measured spectra, like the one that you see here, where a computer program has identified the existence of particular isolated peaks. And those are confirmed by an

expert. And once we identify, say, a dozen of these peaks in a particular spectrum, we extract them from the spectrum and make a hundred different variations in them by adding noise, a little bit of line sharpening, line broadening, phase distortion.

So we generate a library of individual peaks that are based on experimentally measured data, and then we can add them together in different ways to generate synthetic spectra.

And the method that we're going to use to approach the peak analysis is the commonplace method of image segmentation for which there are many, many existing solutions. And as you know, image segmentation is the task of taking a picture and deciding what each pixel belongs to. Is the pixel in a dog, is it in a cat, is it in a stop sign.

And because this is such a general and important problem, there are already fantastic neural networks that have been designed and trained to work on this problem. And the fact

that they've been designed is handy, but the fact that they have been trained is critical.

And here's why. Here's a diagram of a neural network technique called the U-net. And what we have here is a convolutional neural network. The picture of the spectrum goes in here, and at the end is an output that tells us where the peaks are.

And what you see here in these different strata are parts of the network that analyze the image at different scales. So each strata further down analyzes the image at one-half of the resolution of the level above it.

So all of these different strata analyze the image for features at small and large scale. And this part of the network learns how to identify lines, and curves, and boxes, and circles, and so on. And that information at all these different scales is combined at the end, and then used to make a classification.

So if we want to reuse a network like this one, all we have to do is retrain the end of

it that takes the feature information and makes some kind of analysis. So we can take these preexisting networks that are already been trained beautifully on millions of pictures, and dogs, and cats, and retrain them with a relatively limited amount of data to do NMR peak picking.

And here's some examples. So in training, we present the neural network with a little image patch like this of several peaks. Since these are generated by adding together individual peaks, we know exactly where all the true signals are. So our ground truth that we want the network to reproduce is just a list of spots where we know the true peaks are.

So here's the input. Here's the target, the training target, the ground truth. And here's what the neural network produces in these cases. So it actually does, I would say, about as well as a person can do. And here's a result on a complete spectrum.

So that could be of interest to you because it's relatively easy to reuse these

existing networks. And even MATLAB has relatively straightforward ways that you can load an existing network and retrain it for an application like this.

So I hope I've shown you that, first of all, that there's lots of interesting ways that NMR can be used to support biomanufacturing. And we have lots of exciting analysis methods to support those measurements. So thank you for your patience. I hope it was interesting.

DR. FRIEDBERG: Thank you, Frank. Any questions in the room? I have one. In the case of the original PCA that you showed us, you didn't have assignments in the spectrum. Do you need to know assignments?

DR. DELAGLIO: No.

DR. FRIEDBERG: So --

DR. DELAGLIO: Because just the spectra go in as they are in their entirety, every point in the spectrum.

DR. FRIEDBERG: Okay. And the blind spot might be a peak that isn't in one dataset for

some reason but appears in another, but it's really there, you just didn't detect it in the first slide.

DR. DELAGLIO: That's right. Yeah. So what we wind up detecting in these cases are subtle changes in the overall shapes of the spectra. So the peaks are in almost the same positions. So you wouldn't easily be able to see the differences by eye, but across the whole spectra, the systematic differences are all there.

DR. FRIEDBERG: Okay. And the last sort of thing on that is as you know, NMR spectra are very sensitive to pH.

DR. DELAGLIO: Yes.

DR. FRIEDBERG: So how do you, does that, does your algorithm or does PCA have a way to filter that out or to raise a flag and say this might be a pH effect?

DR. DELAGLIO: Right. So that's a great question. And when we apply PCA to AMI detected spectra, that's an issue. Most of the work that we do is with proton carbon spectra. And since

those hydrogen atoms are not labeled, they don't really change with respect to pH.

DR. FRIEDBERG: Okay.

DR. DELAGLIO: Thanks again.

DR. FRIEDBERG: All right. Thank you. Oh, you do have online questions. You were shaking your head no.

MS. ELKINS: Late breaking.

DR. FRIEDBERG: Oh, they're late breaking, good.

MS. ELKINS: How do you introduce in 15C13 in mammalian cell culture?

DR. DELAGLIO: With great difficulty. So there are ways to express the mAb in E-coli and in yeast. As we saw in the previous talk, the glycosylation of these proteins is critical to their therapeutic effect. And we only get that glycosylation when the proteins are expressed in mammalian cells. So we worked very hard to make cell lines that can live in minimal media.

The difficult part is that for proteins of this size, we have to label them with

deuterium-2. And that's the sticking point for mammalian expression. So just we solved that problem with lots of painful trial and error.

DR. FRIEDBERG: Anything else? All right. Let's thank Frank again. Thank you, Frank.

DR. RODRIGUEZ: So in our next part of our session, we're going to feature three distinguished speakers from the Center for Biologics Evaluation and Research, CEBR. First we will hear from Dr. Nobuko Katagiri, researcher and regulator at the Office of Therapeutic Products.

Following Dr. Katagiri, we have Dr. Lisa Parsons, a staff scientist and chemistry manufacturing and controls reviewer in the Office of Vaccines Research and Review. Lastly, Dr. Hussain Ezzeldin, senior digital health expert in the Office of Biostatistics and Pharmacovigilance.

After all three presentations, we will open the floor for questions for Drs. Katagiri, Parsons, and Ezzeldin.

DR. KATAGIRI: Let me just, right.

Okay. Thank you very much for this opportunity. And today I'd like to introduce some of the computational part of the research from Kimchi-Sarfaty's group at CEBA.

After giving you some background about how synonymous, why synonymous single nucleotide variations matter, and some concepts related to the codon usages, I'm sorry. I'd like to show some of the highlight of our studies related to the sequence variations.

First, what is synonymous single nucleotide variants, or SNB's. Unlike other two types of SNB's, synonymous SNB alter the codon sequence without changing the code encoding amino acid sequence, because the genetic codes are regenerative, regenerate and the same amino acid can be coded by multiple synonymous codons.

Now every year, CBER receives a number of biologic therapeutics, and many of them contains a sort of, some sort of sequence variations. And that's why it is so important for us to know which variations are potentially

harmful. Among those sequence variations, traditionally, the synonymous variations and codon optimization, which is a combination of numerous synonymous variations, are considered to be least harmful.

However, the studies in the last decades identified synonymous variants, which are shown to be associated with the numerous diseases, including cancers, and those are just a few of the longest. But how can synonymous variants can change the function of protein without changing the amino acid.

The scientists have been learning over the decade, again, that actually synonymous variation can impact the protein expression and function through multiple mechanisms, not only the splicing, but also in many other steps in both transcription and translational steps.

Especially the codon usages being affecting the translation speed due to the different availability of the aminoacyl-tRNA, which leads to the core translational protein

folding. And all these steps are analyzed using several nucleotide sequence related parameters such as the (inaudible) or dinucleotides at codon junctions.

And as well as codons, but also the codon pairs, which is adjacent codons, are recently also implicated in translational reason. Now please note that those multiple synonymous codons or codon pairs which cause the same amino acid or amino acid pairs, appears in a different frequencies in codon.

They are not appears at the same frequencies, and that we call a codon pair usage bias. And this is the important context throughout my talk. In the next several slides, I'd like to show you some snapshot of various codon and codon pair usage table, or CoCoPuts websites that was made in collaboration with Hive, and now widely used by both industries and academia.

Basically, this website gives you various types of synonyms, the sequence related

parameters in black holes, bar graphs, tables, and heat maps for all species in gene bank. Or in 52 human tissues in tissue CoCoPuts. Or in human primary cancer CoCoPuts that provides 32 cancer types and associated normal tissue types.

And this is the example of the heat maps of codon pair usages. Most recently, we added two more new CoCoPuts. One is for mouse embryo which contains various mouse strains, tissues, and embryonic stages, as well as that for the SARS-CoV-2, which has thousands of lineage, lineages starting from the very first report of this genetic sequence to today.

And this website has not been released yet but will be soon. And all about the CoCoPuts can be seen in poster number 2. Now using the CoCoPuts information, we calculate the codon and codon usage biases. Usually, those parameters are counted and calculated as the parameters per gene, which we call sequentially.

Now, we examined if these parameters or codon position has a statistical significance. In

this panel, which is an alignment of this homologous sequence, in sequential method, all the TCA codons highlighted here are counted in the same way. While in the positional method, those 4 in codon 2, and 1 in position codon 3, are treated differently by codon position.

And codon and codon pair usage biases were compared between permutation tests with these two methods. We observed distinct differences of codon and codon pair usage biases, not only between the sequential and positional method, but also at different positions and in different 4 genes.

For example, this graph is a potential codon pair usage biases for the codon pair that cause the amino acid LE throughout decoding region of atom TR13 gene. And as you can see in average, on average, those open triangle TCGGHE appeared at much higher preferences compared to the other, all other common pairs.

But also it is obvious that the range of the frequencies or preferences has a quite wide

range and in some positions those less preferred codon pairs are actually preferred. This positional significance can be effectively incorporated into genetic engineering strategies, including codon optimization.

Now let's change the gear a little bit and then take a couple of prediction tools for evaluating the impact of sequence variants. Many of the publicly available tools are capable to predict the effect of nonsynonymous or missense variance in high accuracy. And a few of them are also capable of predicting the effect of synonymous variants.

However, their outcome is not that great as shown here. Why is it so difficult? We think there are major, 2 major challenges. The first one is very few entries of disease associated synonymous mutations are found in available databases, because most of the synonymous variants are filtered out at a very early stage of the clinical sequencing pipeline.

And therefore, we are manually mining

and curating the disease associated synonymous variance. And this is still our current status of our database. Second, most of the amino acid based molecular features cannot be used for the synonymous variants, although they are very critical for the nonsynonymous mutations analysis.

And therefore, we explore new venues to evaluate the nucleotide changes, as shown here. Mostly at the nucleotide level, but some of them are protein level. In addition to those which have been already used.

And one of the studies on miRNA and sequence variants are seen in the poster number 4. This, the pipeline is already made for this project, and we have created a reliable dataset for disease associated variants and control. And computed nearly hundreds of primary molecular features for each gene, both Y type and variants.

And clean the data. And currently we are examining the machine learning tools, such as Random Forest or Support Vector machines. This project is ongoing, and we are trying to get the

optimum combination of the molecular features and algorithm to reach this goal.

Now most of our product has multiple cooccurring variations. So once we can predict the outcome of the single mutation, how can we make use of them for the multiple variations. To address this question, we took this similar strategies for the first 2 steps. We created a curated, reliable dataset of the double triple mutations, and more mutations. And then scored the selected attributes for each individual variant first. Then we tested 2 pairs of models in how to combine the scores of multiple variants.

One is to take additive or maximum scores. The other one is to treat them as coupled or independently. We asked this because, traditionally, when you have multiple mutations and you see one of them are already known for the CDA effect, automatically, that one has been determined as the main sole course for the phenotype.

But there is not much proven evidence.

And our results showed that additive model is as predictive of the pathogenicity as the maximum model, or slightly more predictive in some cases. And there are no differences between independent and coupled models.

That means that the common assumption may not be always the case. Our approach is highly extensible to a wide range of monogenic diseases, does not require machine learning, and can also be instrumental in assessing the safety of biological therapeutics.

Lastly, I'd like to briefly mention our interesting finding in codon usage and cancer while we are assessing the relevance of our cancer CoCoPuts in diagnosis, drug development, and regulation. We asked how codon and codon pair usage are related with the various cancer patient metadata and asked the association between snips and each genes expression data from the patient, as well as each patient's cancer type, age of diagnosis, and survival information.

Rather unexpectedly, codon usage and

gene expression patterns were associated with 5-year survival rates in cancer patients. In those panels on the left, the green are low risk group, and red, high risk group, has been distinctly separated using codon usage context.

While on the right panel, these two groups were not well separated by using the total nucleotide changes. In summary, we established CoCoPuts, a publicly available resource for codon usage and other nucleotide variations that is periodically updated and used widely by scientific community.

Quality and size of SMB datasets for training, and the choice of the attributes of the variations, are critical components towards generating improved prediction models for identifying functional synonymous SNB's.

Our approach to successfully predict the effect of cooccurring variants shows that additive model can be useful in some cases. And all these approaches, as well as the use of the positional codon and codon pair usage bias information, can

be applied to both therapeutic design and genetic diagnosis.

And this research was carried out in the collaboration of multiple institutions, both inside and outside of CBER. Thank you very much.

DR. RODRIGUEZ: Next, we will hear from Dr. Lisa Parsons.

DR. PARSONS: Hello. Hi. I'm Lisa Parsons, and I work for John Cipollo for the vaccine structure lab in CBER. And today I'm speaking to you on Pythonic applications for the study of glycans. So I'll give you an introduction of glycans, and a little bit about Python, and then I'm going to speak of an example program that I've written.

For those of you who aren't familiar with glycans, it's a general term for sugars that are linked together via glycosidic bonds. So glucose is not a glycan, but sucrose is because it has a one to the linkage.

And sucrose can be linked together with various carbons, and the numbers refer to the

carbons. So you can get lots of different linkages, which is one of the reasons glycans are kind of complicated to study.

The picture on the bottom left, that's an N- linked branch glycan. Those are the types of glycans I study. N-linked means it's attached to a nitrogen on a protein, usually via an asparagine. But in publications, you usually see it as a simple nomenclature, as shown on the right.

So each shape and color represents a different monosaccharide. So for example, the purple diamonds represent sialic acid. And the reason we study glycans is because they're everywhere. So the picture on the left are some common classes of animal glycans. And in the center there you see microbial cell surface glycans.

So you have fungal glycans, and oomycetes, and you also have some gram negative bacteria shown here. And for those of you who've taken biology, of course REM negative bacteria,

they make cell walls which contains peptidoglycan. But some bacteria also excrete capsular polysaccharides, as you see on the far right.

And capsular polysaccharides are actually important components of some of the vaccines that CEBR regulates. Viruses, as you've seen in some of the previous talks, are also, like can also be glycosylated. So here is the SARS-CoV-2 spike protein in green and yellow, binding to the H2 receptor in blue.

And all of those pink blobs are the N-linked glycans. What's interesting about N-link glycans is that they're not all the same. They are structurally and functionally diverse. So there's 9 common monosaccharide in vertebrates, just in vertebrates. Bacteria are a lot more complicated.

And they're arranged in different compositions with different linkages and modifications. So in the glycogen database, I found 780 unique N-link, glycans. And when I say, whoops. What I mean by composition is you have

some glycans that are high- mannose, and so these green circles are the mannose.

And this is what they look like right after they have the sugars put on the protein, and they've come out of the endoplasmic reticulum. And then some of the glycans will get further process as they go through the Golgi. So you'll get complex glycans like you see here.

And the complexity that you see in, say, recombinant proteins grown in different cell lines, depends on the cell lines. So MBCK cells, yes, they will form these complex glycans, but insect cells will form much simpler glycans.

And CBER's interested in this because some of the sugars that are produced could actually cause hypersensitivity, like some of the insect cell lines. If they put an alpha-1 3 linked glucose on the sugar.

So glycans, they perform a lot of functions. We've already heard about the folding and transport et cetera. But I thought this example here was interesting because this is a

host pathogen example. So here is our flu virus, and the hemagglutinin binds to sialic acid on the host cell. So you see that happening here.

But likewise, innate cells will bind to the sugars on the flu and recognize them. In our lab, we study these glycans by mass spec, and we typically want to look at the overall glycan composition of samples. Protein site specific composition, and then the linkages. And we do that by starting with a glycoprotein like you see here. And then we cut it up with different proteases.

And we can then look at the glycopeptides by LCMS, or we can separate the glycans from the protein entirely using enzymatic or chemical means and analyze the free glycans. And the program I'm going to talk to you about is involved in analyzing these free glycans.

Moving on to Python, the reason I use Python is because I have a lot of complex data, and I find Python to be a very useful tool. It's a high level general purpose programming language

that was easy to learn and easy to read.

And it has lots of books and online tutorials, a plethora of science based libraries, including for machine learning, which we've heard a lot about. And there's several GUI options. I usually use Tkinter, which is the default.

And this language is 30 years old, but continually improving. And this is an application, just an example that I wrote to take a protein sequence, I wanted to optimize my glycopeptides for LCMS. And when I do that I want to make sure I get like a single glycan site per protein.

And so I can put in my restriction enzymes and decide which one to use for a protein. So that's how short you can make an application using Python. And here's some other ways I've used Python in the lab. I've used it to do sequence analysis of viral protein databases with special focus on potential glycosylation sites.

So this image is an analysis of the H3N2 flu virus over time, and you can see the number of

glycosylation sites increase as it evolves in humans. I've also pulled information from protein structure files here. I was just looking at the average beta factors in different sections of the protein.

I used it to assign glycopeptide fragments and mass spectrometer data, and to coordinate lectin and mass spec data. And also to assign whole labeled glycans in multi experiments. And to prepare a glycan database, which is the example I'm going to talk to you about.

But first a little bit more background. So how we do free glycan analysis with mass spectrometry. So we start with our released free glycans, and then we label them. And this helps to improve ionization to protect labile groups and to enable semi quantitative analysis.

And these labels can come in three different forms. You can do reducing ends, where this is, the label is added to the end that comes off the protein. You can label specific sugars. In this case, just the sialic acids are labeled.

Or you can label every sugar, and this is a classic method where the OH groups on all the monosaccharides are replaced with O CH3 groups. And then once you have your sample ready, you mix it with a matrix. I typically use 25 dihydroxy benzoic acid. And it's dried on a plate, and that plate is sent into the MALDI instrument.

And MALDI stands for matrix assisted laser desorption ionization, which is exactly what happens. The laser hits the sample, and the matrix assists the laser into desorbing and ionizes the sample at the time of flight tube, so that you get these mass, sorry, you get these spectra of mass overcharge with the smallest molecules on the left of the spectra.

And these 3 spectra, because I typically do spots for samples so I can calculate the standard and the average and the standard deviation. So I get an idea of how well my spots were spread.

And after I get the spectra, then I calibrate them, I pick the peaks and the isotope,

and by the isotope, I mean, if you, these nice sharp peaks down here in the spectra actually look like this here. This is a isotopic window. And this happens because of the natural abundance of C13. And every C13 a glycan gets, it adds (inaudible).

But in the glycan library, what you're interested in is just the C12 peak. So I just want to pick the C12 peak and then find the area under the rest of the peak so that I can do semiquantitative analysis. And after we get the peaks picked, then we just compare them to the library, and assign the spectra.

A little more on libraries. They are different species of different glycoforms, so the library should contain glycans relevant to this system you are studying. The glycan masses change when the different labels are used, so you need a different library for each label.

And libraries could contain hundreds of glycans. It's very repetitive and easy to make errors, if you have to type them in by hand, which

makes this an excellent application for Python.

And so I had already started this application to make glycan libraries automatically, but a couple of years ago there was a person in the lab who gave me another extra challenge. She was doing linkage specific sialic acid labeling on SARS-CoV-2 spike protein produced in different cell lines, because she wanted to see what kind of sialic acid linkages were present.

And there's the alpha 23 links or the alpha linked, and the label she's using, if it's alpha 23 linked, the sugar will have a loss of about a dalton, and if it's alpha 26 linked, it has a gain of 27 daltons.

So if you have a molecule like this in your sample, there's actually five different linkage options you could have. So you can imagine how spread your library would be and how painful that would be to turn all those different options into a library.

So I decided to modify the program I already had and add the ability to automatically

calculate all these labeling options. And when I wrote this program, my object, my objectives were to make the input as simple as possible, so the user doesn't have to type or copy and paste on labels.

And so I use a shorthand for each sugar. And if you push that button the calculate, it will give you the label that we want to use in your publications. The calculated math, and that math is based on how you prepared your sample, and the modification you choose, and the ion you choose.

And it also calculates the chemical formula. But you don't have to type in every sugar by hand one at a time. You can make a whole list of them and then put it in as a single file. And it writes the whole database file for you.

My second objective was to allow the user to define new labels with easy math calculations. So I came up with this. So our label that we were speaking of earlier, it's imidation and demethylation. Sorry. I cannot see that from here.

Anyway imidation, you just tell the program where you want it to happen. In this case it's sialic acid, but S is the shorthand for that. And it happens sometimes, and then you just type in the atoms that it gains or loses. And it will automatically calculate the math based on what you put in there. So you don't have to calculate the math.

And my third objective was to allow the user to add sugars and other options as needed. So I have another default menu where it's basically the same thing where you put in the name of the sugar you want, the shorthand you want, and then the atoms it has. And then how many methyl groups it would get if you permethylate.

So now if we go to our double label problem, here is the challenge that she gave me. And I made it so that you could choose more than one label at once. And if you push calculate, so this is the shorthand form that represents this molecule. And it will automatically print out the labels into your library.

And then after you have your library made, you go to the other part of this program, which is to assign MALDI. So assigned MALDI is used to assign those spectra. And the button that's circled, that's how you would get to the library maker.

And once you have that made, you put the library here and you point this to the folder where all your spectra are. And you assign it, and you'll come up with this. Like I said, I always collect free spectra.

So they're all lined up here. And I know that's overwhelming, but if you then go and click the zoom button, you can see the peaks are picked. The labels are, they're assigned. The isotopic windows are there to show you how well they're fitted.

And this is an example of the default GUI, but it allows you to add buttons, which I added buttons up here. There's even an option appeared so that you can pick the peaks you want, and it will make suggestions for you if the

assignment is not in your library.

And then for publication it will calculate the average and standard deviation, and it will also make a table for publication. And if you're interested in MALDI, this program is available online. We published this last year. And it's available at GitHub.

And I'm going to say thanks to the fellow lab members, past and present, for their feedback. And to the organizers for inviting me to talk. And to you for your attention.

DR. RODRIGUEZ: So lastly, we're going to be hearing from Dr. Hussain Ezzeldin.

DR. EZZELDIN: Thank you so much. And you're here for staying that late. Thank you for that. All right. Let's see. Okay. So today I'll be speaking to you about another side of what CBER does. Most of the work you've seen is looking at the work before the products are developed, maybe scientific, maybe some of the basic science, some of the product development.

CBER also does a lot of work on the post

market. So once the products hit the market, then we have to follow for safety, and make sure that these products are safe and effective as have been demonstrated in the clinical trials. So I would like to speak to you today about a couple of pilot studies that we've conducted within CBER on ways on improving and expedited biologics adverse event surveillance.

So this is the outline. I'll give you a quick background, speak about the journey, and that into our pilot platform that we've built. And then, of course, sum up with the conclusions.

This, just to give you an idea about the wide range of products that CBER deal with, from vaccines to gene therapy, genome transplantation. So the number of, or the wide range of participants in our products and the users is vast, from healthy to sick people.

And this is challenging. Add to that the healthcare settings where these products are administered. This also complicates our job to do post market surveillance.

So what are we looking at here? What are the challenges and what are the opportunities that we have? So in the current method of adverse event reporting, we have manual detection of the adverse events. Meaning that clinicians, practitioners have to go out and be on the lookout for these adverse events, detect them, and then go on to validating these adverse events, making sure that they're actually adverse events.

And this requires time, and time intensive processes. They maybe require access to diverse and desperate sources of data. And then of course, added on that the manual reporting step, which is of course labor-some that requires data entry and also requires granularity of the data, which sometimes is not available.

So this is the current way of adverse event reporting. What we are hoping to achieve is this is the future that we're looking at. We're looking at an automated or semiautomated detection of adverse events. And this allows us to create batch detection leverage algorithms to do this.

And score these adverse events with the ones that are more likely to be adverse events, that can be reviewed. And then this takes us to the validation component, where we are able to extract the evidence that supports maybe an adverse event or not. And present this to the reviewers of these adverse events to make it easy for them to make a determination.

And then of course, semiautomated reporting. So once you have this semiautomated detection validation, you can very easily extract all the relevant information, populate it in the individual case safety report, and send this to FDA for review.

So this is the future that we are looking at, and we are hoping to move from an existing manual process that creates burden, and sometimes under/over reporting, depending on the media, to a more innovative approach that reduces the burden and increases the quality of the adverse event reports.

So I wanted to show you here these

couple of reports that we get at the FDA. The one on the left is what we call the vaccine adverse event report, and the one on the right is the med watch, which is for the other type of biologics.

So the one on the left is for vaccines only. The one on the left (sic) is for the rest of the biologic regulated products. And the reason I'm showing you this is that this is what ultimately comes out from our platform. Something that the public is familiar with. We as CBER and our viewers are familiar with it as well.

And what's highlighted here in yellow, oops, sorry. What's highlighted here in yellow, these are what's called required fields. So any report that gets to the agency has to have at least these fields filled out.

But there are additional information that could also be very helpful. Let's look at our journey here. So we started in fiscal year 2019, which is October 2018. We started with working with the foundational data partner, started building phenotypes. Phenotypes is a

fancy word for algorithm that you can use to detect an exposure and outcome as well.

So this is what we call the phenotype. So we worked with our foundational data partner to develop some of these phenotypes and validate them. And working on mapping, how we can use this data into filling an individual case report form.

From that, we worked on operationalization of this platform. So we started on enhancing these phenotypes that we built, and we started thinking about how can we scale this this platform so it can be deployed on more than one data partner.

And to do that, we built a prototype with an organization that falls within something called exchange networks. So if you're familiar with exchange networks, these are health information exchange networks that allow hospitals, providers, to exchange clinical information amongst them.

Primarily for the use case of continuity of care. So if a patient moves from one hospital

to another, a hospital can request their clinical information from another hospital so they can know what this person had received in terms of diagnosis in the past, what type of medication they use.

And this is happening on everyday basis. It happens to you all the time, probably don't know it. But this happens, and this is done for the continuity of care use case. And this happens probably on the range of billions of records exchanged on a monthly basis between hospitals within the U.S.

And we are hoping to leverage these health information exchanges for public health use case. So we are trying to use these networks on, can we go into healthcare providers and ask them to send us information about an adverse event, so we can have a more rich and high quality adverse event report that we can review and maybe form a better decision about whether this is related to a product or not.

To do that, we worked with one of these

health information exchanges called eHealth Exchange. It's a network that was developed by the Office of National Coordinator. I think now they changed their name.

But we worked on a couple of cases, something we called pull use case and a push use case. So from the name, if we are looking at the FDA as the point of view, in the pull use case, we are pulling the data. So we are looking to get additional information about the specific adverse event. And the push use case, the provider is sending us data about the specific adverse event that they have detected.

And I will speak about these two. So we're able to conduct this, these two pilots in fiscal year 2022. And I'll share some results with you today about these pilots. And now we are in the phase of pushing this platform to become part of our standard pipeline within CBER.

So it's not a very easy step, but we're working towards that. So what is the pipeline that we're speaking about here? What's this best

platform that we are going to be using to conduct these pilot studies, the pull and the push.

Our pilot platform has three major components. The first one is what we call data quality assessment. So once we pull the data into the platform, we check for the data quality. And I'll speak about this a little bit.

And then after that, we have what we call the phenotyping component, or the detection component of the platform. And what we do is that we run these fancy algorithms that we have, to extract the relevant information from the clinical data and present this to the reviewers to validate this.

So we have the third component of the platform that's meant for validation and reporting of these adverse event. So this is the platform that we've built to conduct these pilot studies. All right.

So what do we mean by data quality assessment. For regulatory grid data, when you are speaking about, is this data fit for purpose,

you are looking at 3 components. You're looking at whether this data is conformant, meaning that the data is in the standard that you are expecting.

You're looking at completeness of the data. Meaning that do you receive the data that you are looking for or not. And then the plausibility of the data, and I apologize, the format has been a little bit jumbled when they switched the slides around.

But anyway, so we are assessing 3 components for the data, its completeness, its conformance, and its plausibility. And here are some definitions. If you're looking at performance, are you getting the data, for example, in ISBT 128 format. If you're looking at block products. If you're looking at, for example, at completeness. Are you looking, are you finding your vaccine brand or the lot numbers and your data.

And for example, plausibility, if you're looking at, let's say an adverse event rating to a

vaccine administration, is this adverse event happening after administration or before. So this is, is this data plausible or not.

So these are the three factors that we look at for data assessment. Now let's shift to our detection or phenotyping component of the platform. So for that, we built three types of phenotypes. What we call simple or scalable phenotypes.

And the reason for this is that if you look here on the right, when you go down this spectrum, the complexity of these phenotypes is going to increase, but at the same time, their positive predictive value is going to increase as well.

On the other hand, the interoperability of these phenotypes is going to increase. What do I mean by interoperability? If we are thinking about scaling this approach, meaning that we want to use our phenotypes and have a lot if not all the healthcare providers use them, if you are going from a simple scalable, which only relies

on, for example, diagnostic codes, to a more complex one, then on the interoperability of these phenotypes is going to be very hard.

Meaning that less and less people are going to be able to use your phenotypes and implement them on your on their platform. So it's a trade off. You can use simple phenotypes with maybe low predictive value, or you can use more complex phenotype with a high predictive value, but maybe less people are going to be able to implement them.

All right, so these are the phenotypes that we have. We have three of them. So you can see here this is the claims comparable, which are the simple phenotypes. They only use diagnostic codes, and we have like a GitHub library with these phenotypes.

So this, the top one is looking at acute myocardial infarction, anaphylaxis, appendicitis, Bell's palsy, and all the outcomes that we are interested in. More specifically for vaccines.

And then if you are looking at the

second tier, you are adding to these diagnosis codes. You're adding some procedure codes, maybe some observations, sometimes some medications.

And then of course, if you look at the more complex phenotypes that we have, we also include on top of that, terms. And what do I mean by terms? I mean that for the complex phenotypes you're actually relying on notes, with the different types of notes, whether it's like nursing notes, clinical notes, discharge notes.

But at that point you have to implement some sort of natural processing, natural language processing pipeline to extract this information. And if you are able to do this, then you are able to leverage different terms to improve the performance of your phenotypes for detection of these outcomes that you're interested in.

And I know I'm going very fast. I'm trying to get you home, so apologies for that. All right. So well, what I'm showing you here, this is what we call the chart review tool. This is part of our pilot platform that I showed you.

And this is a platform that allows our reviewers to look at different components. It allows them to look at exposure. So they can look at the data that we received from this product. And all this are clinical data that we received from this eHealth exchange, this health information exchange network.

It's clinical data, so they can look at the data, ascertain exposure. They can ascertain outcome. And then they can also based on the data available if they feel comfortable, also speaking about imputability.

Meaning that if they can ascertain causality, if they have enough evidence to say yes, we can say that this outcome happened due to this product. They can also speak about that. And they can also speak about the severity of the outcome itself.

So they can do all this within our BEST platform. And in our pilot, what we've added here in addition to our pilot platform that has the data quality assessment, the detection, the

validation components, now we can communicate with eHealth Exchange, which has members in its network which are different hospitals.

Hospitals that can pull clinical information from, pass it on to our platform so we can conduct these validation and reporting steps. So I spoke about a couple of use cases, the pull use case. And it starts with FDA knowing a certain adverse event.

So if we know a certain adverse event, let's say from an adverse report, we can go into our BEST platform, send the request to eHealth Exchange, which is in orange here. Telling them we have knowledge of this adverse event. We want to get additional information.

EHealth Exchange takes this request. Sends it out to its members in the network, which are the different healthcare providers, and does what's called patient discovery. If it finds this patient that we're looking for, it pulls this additional clinical information, pulls it back, sends it to our BEST platform, and then we can

start our validation step and make a report about this adverse event.

So this is what's called the pull use case. We conducted this pilot study with 11 healthcare providers across the country. And then of course, when we did the study, we were able to look at if you are looking at an adverse event related to vaccine, can you assess the quality of this data that you received from the study, and look at the different types of information.

Do you have the required data fields, do you have optional data fields or not. So if you recall, we are looking at completeness, conformance, and plausibility. So we conducted different number of tests and here on the right we have like 39 tests. Sorry, 21 tests looking at required fields, 44 tests looking at optional, and then helpful fields.

And these are about 330 tests in total looking at the different aspects of data quality, completeness, conformance, and plausibility. And what I'm showing you here is that your green is

completeness. You're kind of pinkish is the plausibility. And the orange-ish is your conformance.

So you can look here at the green. More to the right means that this is complete data. In the middle means that partial completeness. And of course on the left means that this data is really not up to what we expected.

So the more to the right is better. So you can see here the green is probably more to the right. So the data is more complete. We have maybe less luck with conformance, which is expected. So standards is a major issue. And data, and real world data of course.

And then so as we mentioned here, we received about 270 post vaccination adverse events from different, 11 different providers' electronic health records. And then we assessed this as you can see here on the right.

One of the very interesting things that we found out is that if you are receiving an adverse event and you are requesting clinical

charts to review this adverse event, this probably takes between 4 to 6, sometimes 8 weeks.

In our pilot study, we were able to receive this information from the moment we request the data to the moment that we receive it within our pilot platform.

The range here is probably about for the limited case window, meaning that we are limiting the information that we're getting with defining some dates around the adverse event itself, took about on average 52 minutes to receive from the moment that we request the data to the moment that we receive it.

For the entire case period, so if you're looking maybe years in the past, this maybe goes all the way to an hour and a half. And here on the right are the medians. And then you, of course, you have the minimum, the maximum.

So looking at something like this, for example, for public health emergency use case, if you can get the data for a known adverse event within an hour, this is a huge, of course,

improvement on the current methods that are being deployed and implemented.

So we did the root cause analysis to this variation that I showed you for these differences in the completeness, conformance, and plausibility. And we were able to put our hands on some of these challenges. And of course it comes down to the United States core data for interoperability.

We found that there are some challenges with immunization data. We found some missing information with medication compliance. Now, of course, there were some technical issues with the implementation of this infrastructure itself.

Second thing that we did is the push use case. And this is probably the more exciting one is that healthcare providers taking the phenotypes that we developed, running them on their back end electronic health records, flagging cases, which is this automated detection.

So here it starts, the healthcare provider itself, there's an exposure to one

biologic, there's an automated outcome detection. They send the data to eHealth Exchange, which is this health information exchange network.

EHealth Exchange pushes this information to the BEST platform. We received it, and then we can start our review case from at that point. So we did this with Cedar Sinai, which the one that has this pull and push use case, as well as the VHA was another participant in our push use case.

One of the interesting things that we, this is just a view of how we build our phenotype. We look at exposure, we look at the diagnosis, and then we try to add some additional information, observation, procedure, or medication to increase the performance of our phenotype.

But one of the very interesting things that we found is that if you have sufficient information in your data that you received, you actually have much better performance for these phenotypes, which is not really very surprising.

But this actually speaks about the importance of looking at the data that you are

receiving, and really speaks about the importance of us working with the Office of the National Coordinator, that regulates electronic health records, and their vendors, to improve the data quality so we can have high performance operating phenotypes that would allow us to automatically detect cases, send them to our phenotype.

So in summary, we conducted these two pilot studies. We were able to find that overall the data quality meets the general requirement for CBER. So this is very exciting for us to move forward with this approach leveraging health information exchanges.

Of course, there are some variabilities even among the same electronic health record vendors. But there are gaps, of course, dependent on the data standards. But we are continuing to work on enhancing our platform, as well as working with our federal sister agencies to improve the data quality, and of course improve our querying capabilities, and enhance the efficiencies of these adverse event detection logic.

And of course, conduct more evaluation and validation studies. Of course, this work has been supported by many people within CBER. With our BEST contractors. IBM doing these pilot studies.

And then of course, this work has not been possible without the help from eHealth Exchange, and our pilot participants.

There are some references here and that's everything. If anyone is still awake, you have questions, please let me know.

DR. RODRIGUEZ: Thank you very much. So now we're going to open the floor for questions for our last three speakers.

DR. FRIEDBERG: So if anybody has any questions, step up to the mic. And if not, I had a question for you, Lisa.

DR. PARSONS: No.

DR. FRIEDBERG: No? Actually I have two questions for you. How do you handle or know if you're dealing with an impurity in your system? Do you have some code or something to pick up

impurities, or do you do intensities?

DR. PARSONS: Yes. I do have a, it's sort of an average, okay. You have to take the noise into account. I have a noise cutoff basically.

DR. FRIEDBERG: Okay. Just curious.

DR. PARSONS: Yeah.

DR. FRIEDBERG: The second question was how do you, it's kind of related, how do you know if you miss a peak? So you've got the noise threshold, how much above the noise threshold is considered a peak in your case? So I guess below it doesn't count, right?

DR. PARSONS: The user sets it.

DR. FRIEDBERG: Oh. Okay.

DR. PARSONS: I tried to make it, the user can change everything.

DR. FRIEDBERG: The user can do everything.

DR. PARSONS: Yes.

DR. FRIEDBERG: Got it. All right. It looks like we have nothing online. We're kind of

over on time. So I'm very sorry. But thank you very much for your talks this afternoon. It was great. Thanks everybody for coming.

(Whereupon, at 4:35 p.m., the PROCEEDINGS were adjourned.)

* * * * *

CERTIFICATE OF NOTARY PUBLIC

DISTRICT OF COLUMBIA

I, Michelle Begley, notary public in and for the District of Columbia, 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 District of Columbia