Hello everyone and a very warm welcome to this Science/AAAS webinar. My name is Sean Sanders and I'm the editor for custom publishing at Science.

In our webinar today, we will investigate the current applications of exome sequencing in clinical research and its impact on the future of healthcare. The exome encompasses approximately 1% of the genome yet it contains about 85% of disease causing mutations making exome sequencing easier and cheaper than whole genome sequencing for identifying disease-causing variants. Utilizing the latest next generation sequencing technologies, exome sequencing has been successfully applied in the clinical research setting for identifying common single nucleotide variants, copy number variations and small insertions or deletions as well as rare de novo mutations that may explain Mendelian complex and rare genetic disorders. Recent advances in rapid and low-cost exome sequencing make it an attractive alternative to traditional targeted gene panel sequencing for clinical research, while maintaining the possibility of discovering mutations in genes previously not associated with a disorder.

To delve more deeply into this exciting topic, I have two exceptional panelists with me in the studio today, Dr. Christian Marshall from The Hospital for Sick Children in Toronto, Canada and Dr. Christian Gilissen from Radboud University Medical Centre in Nijmegen in The Netherlands. Thank you both for joining me in the studio today.

Dr. Christian Marshall: Thank you.

Dr. Christian Gilissen: Thank you.
Sean Sanders: Great to have you and before we get started, I have some important information for our audience. Note that you can resize or hide any of the windows in your viewing console. The widgets at the bottom of the console control what you see. Click on these to see the speaker bios, additional information about technologies related to today's discussion, or to download a PDF of the slides.

Each of our guests will give a short presentation followed by a Q&A session during which they will address the questions submitted by our live online viewers. So if you're joining us live, start thinking about some questions now and submit them at any time by typing them into the box on the bottom left of the viewing console and clicking the submit button. If you can't see this box, click the red Q&A widget at the bottom of the screen. Please do remember to keep your questions short and concise, as this will give them the best chance of being put to the panel.

You can also log in to your Facebook, Twitter, or LinkedIn accounts during the webinar to post updates or send tweets about the event, just click the relevant widgets at the bottom of the screen. For tweets, you can add the hash tag, #sciencewebinar.

Finally, thank you to Life Technologies for sponsoring today's webinar.

**Slide 2**

Now I’d like to introduce our first speaker, Dr. Christian Marshall. Dr. Marshall completed his undergraduate and graduate training at Simon Fraser University in Burnaby, Canada. Following completion of his Ph.D. in 2005, he undertook a postdoctoral fellowship at The Hospital for Sick Children in Toronto. Since 2009, Dr. Marshall has been a research associate in genetics and genome biology at The Hospital for Sick Children, where he evaluates new genomic technologies and applies them in disease gene discovery. He is also the program manager at the McLaughlin Center at the University of Toronto, responsible for the strategic planning on scientific initiatives in genomic medicine. Welcome, Dr. Marshall.

**Slide 3**

Dr. Christian Marshall: Thank you. Thank you Sean, and it’s great to be here to tell a little bit about some of the research we’re doing at The Hospital for Sick Children specifically in using whole exome sequencing in clinical research.
I just wanted to start by sort of giving you an overview of some of the genome-wide molecular studies in clinical research at The Hospital for Sick Children. When I started back in 2004, we were running a lot of microarrays at the time and looking for copy number variation and over the years, this has amounted to somewhere around 10,000 arrays a year run in the genome center. This is sort of followed by clinical microarrays that have been run in the diagnostic lab at Sick Kids Hospital and that’s somewhere around 4000 arrays. Now we’re moving into whole genome, whole exome sequencing technology. So in the genome center, we’re probably running over a thousand exomes a year and also going into whole genome sequencing.

So these are data from both the Center for Applied Genomics, which is the Genome Center at the hospital and the Sick Kids clinical diagnostic lab. Sort of within this a lot of joint research projects between the Genome Center, the diagnostic lab, and also clinical genetics and a lot of these right now are sort of going towards exploring whole exome and whole genome sequencing and the utility of this in the future diagnosis of pediatric cases.

So with this in the next slide, I just wanted to show why we’re doing so much sequencing etc. is because the cost is coming down and this is just showing that sequencing is coming down. It’s becoming less expensive than existing genetic testing that’s done. So on the left you’ve got sort of this cost range over the last couple of years of the reagent cost of genome sequencing coming down. On the right-hand side, what I have is some specific disorders and the amount per individual that has been spent on patients or on subjects in Ontario to have gene panel sequencing done for some specific disorders.

At the top of this, you have disorders that are caused by a lot of diseases or by a lot of genes primary ciliary dyskinesia or spinocerebellar ataxia. These would be upwards of $10,000 or $9000 to do panel sequencing. At the bottom of this, I’ve also put in microarray analysis, which is about $900 to do per individual. So the cost of whole exome sequencing using iron proton AmpliSeq in our facility is currently about $650. So you can see that already the cost of doing a lot of these experiments is less than traditional panel sequencing or microarray analysis.
So I just wanted to go over some of the projects that we are doing and kind of give an outline of what I’ll be talking about today. We have several ongoing research projects. These are aimed at providing evidence for the introduction of whole genome sequencing in the future clinical care at Children at Sick Kids Hospital. This is both for panel testing or to replace panel testing and essentially doing a whole genome experiment but also to look at the genes of interest depending upon the indication and also clinical research genomes and I’m going to talk a little bit about some of the work we’re doing in our autism genome project.

The whole genome sequencing is something that’s not yet feasible for to do for us in-house. The cost and turnaround time is high and so right now in the interim, we’re using whole exome sequencing and for this, we’re using ion proton sequencers mostly with AmpliSeq enrichment. It’s good for an alternative to targeted gene panel testing as part of our clinical research and it’s also part of this developing clinical research exomes for the autism genome project.

**Slide 7**

So in kind of getting back to this and looking at the spectrum of analysis that’s being done, there’s obviously single gene analysis moving on into gene panels depending upon the disorder, exome sequencing and genome sequencing. The issue of course is that as you move towards the right-hand side of the screen here, you’re increasing your amount of data and also increasing the complexity and the amount of, and the need for interpretation. So this is obviously an issue.

**Slide 8**

Well I guess the flipside to this is that you’re also increasing the amount of tests. So as new genes are discovered, there might be a single analysis that’s done that might move on to a gene panel as there’s more and more genes that are responsible for disorder and these are being discovered all the time and being added to panels and so sometimes it’s really hard to keep up. Right now, there’s genetic testing available for over 2000 rare and common conditions and the list of this is growing. So like I said it’s hard to keep up to this and so, the important thing is doing an exome or a genome is essentially doing one test.

**Slide 9**

So this is kind of moving into why do you do whole exome sequencing. Well compared to panel gene, it is a single test. You can
kind of streamline your experiment or your lab just do one test for any of the indications. You often have these predesigned kits and workflows which are pretty easy to work with and of course gene panel negative, if the gene panel is negative, you can reflex two related genes or phenotypes or as new genes come out that are responsible for the disorder and being discovered, you can go back and look in these data.

Compared to genome sequencing, why would you want to do exome sequencing? As Sean mentioned in the intro, the majority of disease causing variants are sort of residing in exons right now. This could change over time as we learn more and do more genome sequencing. But right now, this is where the best chance of finding these mutations are. There’s less data to transfer, analyze and interpret so this is an important consideration and it’s more cost effective with a faster turnaround time.

Slide 10

This is just to show an exome sequencing timeline comparison going from library preparation to the cluster prep to sequencing and then to analysis. Back when we started doing a lot of exome sequencing where you’re using a solid platform and with the ion proton, you can see now that sequencing is relatively fast. So you can do something in a couple of hours or you can sequence for a couple of hours. In the past using solid, you would be doing quite – it would be about three weeks for sequencing to be done. So that’s one of these huge advancements in the new technology and the other advancement now is using the AmpliSeq for the target enrichment. This has sort of allowed us to do the preparation or the library prepping capture relatively quickly compared to a target seq, which is a hybridization technique. You can do this within a day. So a lot of advances in technology and library preparation and it’s made whole exome sequencing extremely fast and cost effective.

[0:10:30]
Slide 11

So just to dive into a little bit deeper into this on the ion AmpliSeq Exome, as I said it’s very fast. It’s simple, it’s reliable, it’s an hour of hands-on time, very efficient and uniform. So we’re right now sequencing two exomes with a V1, V2 chip and this typically gives us over 90% of the bases covered over 20x and a relatively automated analysis. So you can get annotated filtered variants relatively quickly using ion reporter.
Just to look at some more specs on the actual library that takes about eight hours to do, there’s almost 300,000 prime repairs across 12 different pools, very low input of DNA so this is another important consideration, you can use 50 nanograms. It covers a lot of the genes, a lot of the coding sequence of the genes. Amplicon size ranges between 225 and 275 base pairs. So in the grand scheme of things, constructing the library, preparing the template and running the sequence and analyzing the data, you can have DNA on a Monday morning do this and you can have sequence by Wednesday. So this is a really quick turnaround time.

**Slide 12**

In terms of exome sequencing analysis, this is just kind of a really general flow of what’s happening once the sequence has been done. The sequence you generate the reads, you can remove the poor reads from there, do the alignment, variant calling. At that point, you might want to do a lot of coverage analysis and see how well your exome has been covered or your targeted genes of interest.

Moving on into annotation, we can get variant effects in the frequencies and the populations or internal databases that you have and then that kind of more analysis with something like ion reporter or custom pipelines where you want to filter down and do variant filtering and then eventually variant confirmation perhaps using Sanger sequencing. So the analysis that we use is we use ion reporter in some instances but also have custom pipeline with interpretation as obviously it needs to be based on the disorder that you’re looking at.

**Slide 13**

This is just to show ion reporter. It does have a lot of really nice features. Basic filtering, you can view the variants and adaptations relatively quickly and you can kind of drill down deeper when needed.

**Slide 14**

So the next part of the talk, I just wanted to show or outline different studies that we have ongoing right now that we’ve started relatively recently. One being this kind of panel research study, this is a retrospective sample set where what we’ve done is take in – we are taking individuals that have been submitted for traditional or standard of care panel sequencing. You can see just the different numbers here of 25 of the first individuals that have been run ranging from different divisions, nephrology down to immunology in
the different number of individuals there and also the different phenotypes that we’re looking at.

So these are actually all been – we’re exploring whole genome sequencing, these have all been sent to complete genomics for whole genome sequencing but we’re also taking advantage of the ion proton in the center and doing whole exome sequencing as well.

**Slide 15**

How would we analyze these? So again this is sort of this situation of doing an in silico panel. You take your whole genome or whole exome sequence. It is a known gene? Is it one of those ones that was on the panel? If yes, then you might do a little bit more genetic analysis and validation in a certified lab and that might lead to the diagnosis. If no, maybe it’s a gene that’s related or something that’s been recently discovered and that happens in some cases and you can do a refinement analysis and angle back and look at the expansion of phenotype and see if it is something that is related to the phenotype in that individual. Then of course, if no, some of these things are still negative and maybe this is a novel disorder, you can go on to more gene discovery mode but you have all that data available for use. Really, what we’re using the exome sequencing for is a rapid and low cost approach to sequence a lot of genes really quickly.

**Slide 16**

Just some examples on some of what we’ve done in these first 25 individuals. This is just a chart just to kind of give you an example of three different situations, the first one being from nephrology with a diagnosis of FSGS. The gene panel result happened to be negative. In the whole exome sequencing, we were able to find a mutation that explained the phenotype and it just didn’t happen to be on the panel and this was in PLC1. So this fit the diagnosis and it was positive.

In the second case, cone-rod dystrophy, we were able to find both the gene panel and the exome sequencing agreed. There’s variance in PROM1 that made the most sense. I would note that we also detected some other variants in related genes that might also be contributing to the phenotype and one of those being in the x linked CACNA1F.

The third example is what I’ll go into a little bit more detail but this was a clinical diagnosis of Adam’s-Oliver syndrome, which is an extremely rare disorder. There isn’t really a gene panel for this per se
but there’s a couple of known genes and four known genes in fact that are thought to cause this. So when we looked at the whole exome sequence data, we did not find any mutations in those four genes but we did find one and I’ll expand on that in a second.

So in general, our experience so far is there is very good concordance between calls from the gene panel and proton whole exome sequencing and in some cases the whole exome sequencing is picking up variance outside panels that maybe contributing to the clinical presentation.

Slide 17

So to expand a little bit on this Adams-Oliver case, this is the prior diagnosis. It’s characterized by this absence of skin at the scalp vertex and also some limb defects. Whole exome sequencing did not reveal any causative mutations in the known genes. We had very good coverage of these four different genes and they were unable to explain the phenotype. So in this case, what we did is we want back and sequenced the parents as well to get a better idea or to go into sort of gene discovery mode.

Slide 18

Doing this trio analysis in ion reporter, it quickly got us down to about I think there’s 12 punitive de novo changes. We were assuming a dominant model at this stage and using IR4 interface to filter and find this really high concerted mutation that was de novo in exon 8 of this gene ACVR1. This was a plausible candidate. So when we go and you look a little bit more, the mutations in this gene are associated with the disorder called fibrodysplasia ossificans progressiva or FOP and a lot of these features overlap with Adams-Oliver’s syndrome. After revisiting the phenotyping and going back and looking at the clinical presentation, we found that this was consistent with the variant of this disorder. So in this case, it was the exon or the trio analysis revealed something that we weren’t really looking for and it kind of helped us to go back and interpret and look at the phenotype again.

Slide 19

For the last part of the talk, I just wanted to kind of briefly touch on our autism genome project design and how we’re using or thinking about ways to report this more as a clinical exome in the future. This is our Canadian Autism Genome Project. We used high-resolution stent microarrays as well as whole exome sequencing and some whole genome sequencing as well. We’ve run over 2000 high
resolution. By high resolution, I mean over a million SNP microarrays and we do copy number variation analysis on these individuals. We have 200 genome sequences, whole genome sequence where we’re doing copy number variation comparing to the arrays and also the single nucleotide variance and the indels as well. Then for exome sequencing we have about 600 individual sequenced. A lot of these are on the solid 5500 XL platform but now we’re moving a lot of these on to the ion proton platform for exome sequencing.

At the end, what you get is a combined high-resolution genome analysis by integrating a lot of these things in some cases individuals will be run on multiple platforms and then we go back and look at the genotype/phenotype correlation by looking at the families and looking at segregation. Really, one of the things that we want to do here is development of a sort of a clinical research exome report for the autism genome project.

Slide 20

Just as a subset to this, we have a really nice cohort from Newfoundland from a clinical geneticist Bridget Fernandez. So all of these individuals have been seen by a clinical geneticist and in this sort of pilot, we’ve got 75 trios in the Solid 5500 platform and also 75 trios with the ion proton. This is a decision tree that we’re going through to decide what would be interesting to look at.

So if you take the annotated single nucleotide variants and indels putting some priority on whether it’s a de novo variant or not, if not going into whether it’s potentially an autism gene or it’s involved in cognition and then we have a candidate list of about 125 of these AST genes. If not and we’re at a higher frequency in the population over 1% and this goes on as sort of our further research gene discovery secondary findings, etc. If it does fit this criteria then we’re looking at whether the variant is deleterious, whether it’s a loss of function mutation related to the phenotype or if it’s a variant of unknown significance so usually a missense mutation in some of these genes of interest. So it might be really rare what a missense mutation and we have to do further analysis. So these would be variants of primary interest and then we go back and look at the segregation within the family and the genotype-phenotype correlation.

As we’ve said before, we’re using a list of around 125 autism candidate genes also looking at the de novo variants as well. Typically, we’re finding about 25% of cases have a variant. This is
either a loss of function variant in one of these genes or a de novo mutation that maybe related to the disorder. Some examples I’ve listed here under exon included.

Slide 21

This is an example or I just wanted to show that it’s not only de novo mutations that we’re interested inherited roles or inherited variance will play a role in autism. This is a pedigree here. We would have just sequenced the trio but we’ve gone back and done segregation analysis by sequencing the rest of the individuals with Sanger sequencing. So the whole exon sequencing revealed a stop mutation in NRXN1. NRXN1 has been implicated in autism before but also or at least haplo-insufficiency of NRXN1 in not just autism but also in other neurodevelopmental disorders.

So in our interpretation here, this happened to be inherited from the mother. This is maternally inherited. It’s a stop mutation and it maybe the pathogenic variant in this family. But when combining it with our array analysis as well, so this is copy number variation analysis coming from microarrays, the interpretation can depend on the technology and so the array has also found a 15Q deletion, a 500 kb deletion. This is a known association with autism with variable expressivity and this has been inherited from the father. So our interpretation now might shift slightly maybe a pathogenic variant from the NRXN but the 15Q deletion might also be contributing to the phenotype.

Slide 22

I think as things progress in terms of exome sequencing, some of these copy number variants will be able to be found but right now we’re still using microarray analysis and exome sequencing to get a better picture of what the high resolution genomic makeup is of many of these individuals.

Slide 23

So just in summary and observations, we’ve got several research aimed at testing the utility of whole exome and whole genome sequencing in the future diagnostics of kids at the hospital. The cost of whole exome sequencing or whole genome sequencing is becoming less expensive than the current genetic test. We’re using ion proton sequencer as an alternative to targeted gene panel sequencing but also moving into clinical research exomes. So far, our results have shown really good concordance with gene panel testing and it offers the ability to find other variants that maybe
contributing to the phenotype outside of those and the question is when do you use this early in diagnostics as a tool.

Currently testing the yield of whole exome sequencing, as I say, in these more complex neurological disorders like ASD. Finally, the integration of copy number variation I think is extremely important and as I alluded, we’re moving towards whole genome sequencing where we think we can find this but also where exome sequencing, advances in exome sequencing will also help with CNV analysis.

Slide 24

So these all are just people I’d like to acknowledge. Steve Scherer is the leader of our group at The Hospital for Sick Children. Thanks.

Sean Sanders: Great. Thank you so much, Dr. Marshall.

Dr. Christian Marshall: Yeah.

Slide 25

Sean Sanders: We’re going to move right on to our second speaker today and that is Dr. Christian Gilissen. Dr. Gilissen studied computer science at the Radboud University Nijmegen before obtaining has Ph.D. in 2012 in the Department of Human Genetics at Radboud University Medical Hospital. After his Ph.D., Dr. Gilissen continued his work in Nijmegen as a postdoctoral researcher and contributed to various studies on the implementation of next generation sequencing in the clinic. His current work focuses on the interpretation and clinical utility of whole genome sequencing data for patients with intellectual disability. Welcome, Dr. Gilissen.

Dr. Christian Gilissen: Thank you very much. I’m very happy to be here and to show some of the things that we have done.

Slide 26 to slide 27

So I’d first like to start off with giving you a little bit of an idea of the department where I come from. So I come from the human genetics department in Nijmegen and I think it’s a bit of a special department and the sense that we have a really large research section but we also have clinical genetics and we have genome diagnostics all in one and it’s a pretty large department therefore with about 300 people. The advantage is that we have the opportunity to actually develop new technologies in our research department and then actually move them gradually into diagnostics.
But I’m going to start off with showing you the ideas that we had when we started with implementing next gen sequencing for clinical research that’s about three years ago and then we were considering various options as probably most people have. So first, we looked at our current Sanger sequencing at that time. So the Sanger sequencing was very accurate and it was also quite cheap and we had a good turnaround time for single genes very well. But well the results for genetically heterogeneous diseases were a little bit of a problem and we didn’t find a lot so when next gen sequencing was coming around, we were looking at what the options were, what could we do.

Sort of the more obvious thing was gene panels. So targeted resequencing, we looked at that and you see you have a lot of advantages. You can actually optimize your enrichment kit to get the best performance out of it so you can make very dedicated kits. You have a little chance of having incidental findings so findings that you’re not interested in for actually figuring out what’s wrong with the patient and the advantages at the amount of data that you’re looking at is not so much so the analysis is relatively easy but also the interpretation is relatively easy.

On the downside of course, you have to design your enrichment but also you have to redesign it so when you are looking at diseases where not all genes are yet known for then when a new gene is being published, you have to go back to your design and adapt it. Another problem maybe that you have a lot of different designs done for all the diseases. You have the single design and you also have to wait then until you have a sufficient amount of patients such as you can pool them altogether and then do the sequencing.

So we also considered exome sequencing. So one of the big advantages is you don’t have a bias for a particular gene panel and that’s very nice. So if a new gene is being published, you simply go back to your existing data and you have another look and also mainly also for me but also in the laboratory, it’s very nice that everything is the same so you get a very standardized workflow and you can really optimize your workflow to get the best possible quality. Another advantage is that you can reuse the data again so if you have for example done an exome on a blindness patient, you can use that same data to actually interpret the variation that you find in deafness patients that’s’ very nice. Obviously, it’s very simple to add
new genes to your design because you’re sequencing everything. So it’s just in silico enrichment.

On the downside, yeah, enrichment kits aren’t perfect for exomes so you will miss things and it will be very difficult to optimize existing exomes. So if you’re sequencing, if you’re interested in a particular gene and you’re missing a specific exon there due to dropout in the enrichment, it’s very difficult to change anything about that.

Also, a disadvantage is if you known pathogenic mutations that do not lie in the coding sequence you will obviously miss them because that’s not what you’re sequencing. Another problem for us was also that the chance of incidental findings is a lot bigger than a really targeted design.

So lastly then genome sequencing so that’s perfect. You have no bias in what you sequence because you sequence simply everything. You also have little technical biases in the sense that you don’t expect that there’s going to be a really big coverage dropped in some regions due to the enrichment step that you have with exome or targeted resequencing. Another big advantage is that it allows the detection of structural variance such as inversions for example that are really difficult to detect now within the other technology.

On the downside, the data analysis is going to be a really big bottleneck because it’s a lot of data. You have to interpret also, you have to come up at least with some ideas on how to interpret the noncoding variance that you find which is still an ongoing challenge and what was at that time for us at least the killer is that it’s really expensive at that time and very time consuming to do it. So in the end of course we decided for exome sequencing.

So I’ going to show you the two approaches that we decided on using. So the first approach is what we call a gene package approach. So this we do for genetically heterogeneous diseases where we feel that most of the genes that are involved in the disease are already known. So then we sequence the entire exome but then we put top bioinformatics layer of a gene package such that we only look at those genes that are in that package and we’ve done a pilot study on that for five different genetically heterogeneous disorders and for each of these disorders we sequenced 50 exome and then we evaluated that. So I will go into detail about this study later but the results have been published also very recently.
And then the second approach that we used is a little bit of an extension of the gene package approach so actually we also use a gene package but then we do a little bit more. We do these for diseases where we feel that most the genes involved are not yet known and where the disease is pretty severe so we’re suspecting really de novo mutations on. Usually then the disease is sporadic in the family so only the child is affected, there are no other people in the family. So then, we sequenced not one exome but we sequenced three exomes, we sequenced both parents and the child and we tried to detect de novo mutations. Also for that, we actually did a pilot study. We did 100 trios with intellectual disability and also those results have been published last year.

Okay. Just a little bit practical now about the workflow that we use. So we do enrichment with Agilent version 4 kit. Currently, we do sequencing in collaboration with BGI so I should mention that all the results that I’m showing you are a little bit older so this BGI collaboration is something of the last month but before that we did all of our sequencing with solid and 5500 instruments from Life Technologies. But since the few months we started this collaboration and we sequenced sometimes 100 bases with a median coverage of 75. Then BGI is kind enough to also do the basic analysis for us so they do the read mapping with BWA and they do the variant calling with GATK. So this currently all outsourced but just since a few months before that we did everything in-house.

Then the second part of the workflow is really what we do ourselves and that we also try to focus our work on one of the most important things is quality control. I’ll go into detail for that a little bit to give you some footholds obviously, to give you some idea of what we actually do and what you can do yourself. Then I’ll show you how we do the interpretation and what kind of methods we’ve devised for that. Well usually after the interpretation which is in our lab done by two molecular geneticists, we always validate our results that is currently still being done by Sanger sequencing but we’re very quickly moving to doing that with the IonTorrent instrument and then based on those results we write a report.
Okay. So a little bit on detail on the quality control that we do just to give you a really good idea, so basically some recommendations on my part. The first quality control that we do is really on the raw sequence. As soon as we receive the data and we really look at the raw sequence using a published tool FastQC, which I can really recommend and we run this and by default on all of our samples just to check that the quality of the data that we get is good. Also, we run coverage statistics so we try to figure out what the average cost is for our samples and how the coverage is distributed and how the genes that we are interested in are covered and very much in detail. For that, we use Bedtool software and that’s something that especially for clinical research is very important not just to know what you have sequenced but also to know where you have not sequenced and what you may have potentially missed.

So a second thing that I wanted to show you at least is the quality control that we do more or less at the end of the analysis is that we look at the results that we actually get. We look at the variants that are being called and we look at things that overlap with DB SNP and we look at the number of truncating mutations that we find in a single exome and we look at things like the transition and transversion ratio. These are really good statistics to do on all your samples and to detect outliers very easily and see that this is something maybe subtle and maybe wrong with your sample and that it needs to be resequenced again.

**Slide 32**

So this was really on the quality side of the data but I also want to show you something that has been very important for us and I’ve seen for many labs is the risk for sample mix-up and especially if you want to do the sequencing yourself, you will know that this can happen every easily. I’ll give you one example where for example you have to go to slides on an instrument and then simply two slides are switched. If you don’t have any checks in place, you will never figure out that there was actually sample swap and you may actually be looking at the wrong patient.

**[0:35:00]**

So one thing that we do which is a very straightforward test is actually looking at the gender of the patient so looking at the coverage of the chromosome X targets and the chromosome Y targets and then calculating the coverage ratio. You get something like this when you do that for all of your – for a batch of samples. Then we determined some thresholds and you can see very clearly that’s easy to distinguish, your male and your female patients and
this can be checked back with patient database to figure out whether there has been a sample swap or not.

Slide 32

Of course, this only takes care of part of the sample swap but it’s a very easy test to do. So additionally we also have a 12 SNP test where we separately sequenced 12 SNPs on the original DNA and then we compared those results to the data that we got from the exome and this is really the test that should be able to figure out all of the potential samples where offset were made.

Then lastly, one test that is also quite straightforward is when we do actually a trio analysis. We also have a test for that in advance just to figure out whether the trio is indeed a correct trio and whether we have labelled correctly the father, the mother and the child. You can just do that by looking at common SNPS and then calculating a concordance ratio as soon as one of them swaps, you can easily see that the concordance drops significantly so you can pick that up very easily. So these were just some handholds for you.

Slide 33

I’ll show you a little bit about what we’ve done to actually scale this to doing 400 exomes per month, which is what we are doing currently. So for that, we have to come up with a solution for a molecular geneticist and we devised a small software package that helps them to interpret all of the variation that we find in an exome. So this user interface first of all limits all the variants to only those variants in the disease gene package. So if they are looking a patient with a disease indication for blindness, they will only find variance here within the known blindness genes and these gene packages we have made ourselves with a specialist in that field of disease.

So what you can see here is on the left side you can see a quality control criteria that I just explained to you. If anything is wrong with this test, flagged automatically. Then on the top, you can actually see the variance with all of the functional annotation that we do for these variants and on the right side, you have the panel that helps the molecular geneticists quickly filter down to the variant sets that are relevant for this patient.

So and then something that we did very recently really to limit our turnaround time much more is that we’ve implemented a link with a patient database or when a molecular geneticist actually find something that is interesting you can immediately submit this
variant for follow-up. So then when it gets Sanger sequenced and it gets recorded also for us so that we know that these variants were found and the way were interpreted by the molecular geneticist.

A second database that we use is what we call a variant database. When the molecular geneticist makes the interpretation for a variant for example saying okay I looked at this variant, I looked up in literature and I believe that this variant is probably benign, he can actually also submit that information to a separate database. When new samples get loaded and then the molecular geneticist sees these previous results from previous exomes back in the interface so then he doesn’t have to do the interpretation all over again.

**Slide 34**

Okay. Now really going to the pilot study that we did. So this is a study where we actually did 250 exomes in total so for five genetically heterogeneous diseases. So I mentioned already briefly a little bit about the gene package design so we only used known genes so no candidate genes, the genes are updated every three months and they are created by experts for this disease from across our department.

So how do these packages look like. Here you see the diseases that we started with. It’s five diseases between 98 and 200 genes approximately per package.

**Slide 35**

I’m just going to show you the results because it will take too much time to show you the entire study but I’m going to show you the main conclusions, which I think are also the most interesting. Is if we actually look now at what we found for these 50 exomes for these 5 genetically heterogeneous disorders, you can see that we have a very high yield in terms of patients that we were able to find the cause of the disease. So for example, for movement disorders we saw 20% and we could relate that back to the known yield that we have by Sanger sequencing for these diseases and then you can see them in all cases at least we perform better. The colorectal cancer group is a little bit different and I won’t go into detail about that here but for the other diseases we can see clear improvement in the yield that we get.

To further substantiate this, we also actually tried to look at suppose we would have sequenced with Sanger all of the genes that are currently being offered in our department for Sanger sequencing,
what would have been the yield then. So if you look at that you can see now in the blue boxes that the yield would have been higher but it is still not the yield that we achieve while doing exome sequencing for these diseases. So that's really confirming that for us exome sequencing has a really added value compared to the Sanger sequencing workflow that we're using.

So a little bit on the gene packages, so here again you can see the five diseases that we looked at and you can see the size of the disease gene packages and this is really showing you the advantage of doing exome sequencing because currently these packages look quite different, they are a lot larger, because new genes have been published and these have been added in the last two years to these packages.

Moreover, from the study we could conclude that if you sequence more than three genes if we do Sanger sequencing for more than three genes or three genes or more then exome sequencing already becomes cost efficient. So with that in mind we actually extended these five packages and as you can see we now have many more packages with different sized gene panels that we're currently using for clinical research.

Okay. This brings me to my second pilot study, the de novo approach. Again, I'm just going to show you the major results. If you're interested in the details you can read the paper in The New England Journal. We looked at 100 patients and 200 parents. All patients had severe intellectual disability and sporadic and there was no syndromic diagnosis and there was a negative family history. Also, you have to realize that for these patients they have reached the end stage of conventional strategies so we already did targeted sequencing of specific genes and we also did genomic microarrays for these patients.

So what we find there and these are results from 2012. In at least 19 patients, we could come up with plausible candidate diseased genes for explaining the intellectual disability. Now since that time, we have been doing a lot of follow-up trying to get even better results on the same cohort and now from June 2013, we have solved 29 and we still have 11 candidates so a lot of genes that were previously
candidates have now been confirmed. I’m just showing you that you can get a yield of about 30% for severe ID by using this de novo approach.

Okay. So I want to finish off with this last example of what I think is a nice illustration of the benefits of using exome sequencing. So this is a female patient four years old that came into our clinic and she had delayed development. She had very little speech and behavioral problems and the MRI was normal. So we immediately decided to do exome sequencing but unfortunately the first results were negative. We couldn’t really find anything in the exome at first.

So then we moved on from our first analysis to what we call our second analysis. So when we do not find any cause in the known disease gene package, we say okay there is no result and we can move on to actually what we call opening up the exomes so simply looking at everything in the exome and trying to figure out whether anything stands out.

So this is the user interface that I’ve shown you before and these were the results basically that we got when we looked at de novo mutations outside of the gene package. Quickly we figured out that one of these genes in this list was just very, very recently published in a paper in Nature genetics and was linked to Coffin-Siris syndrome. Then where we actually looked at the phenotype of these female patients and we compared that to the patients that were described in the paper, it matched almost perfectly so in the end we could establish what was wrong with this patient genetically.

So just showing you that by exome sequencing even if you do not find anything immediately, just by the publications on new genes or gained insight, you can go back to your data and then still get a definite result on your patient.

Okay so with that I want to draw some conclusions, at least our conclusions is that exome sequencing has a higher yield for genetically heterogeneous diseases than using Sanger sequencing approaches. Also for us we concluded that the de novo mutations are a common cause of severe intellectual disability. So still for all of our patients that come in with an intellectual disability, we do a de
novo approach and then the results that I show you are also very reproducible we find and then some future direction.

[0:45:13]

As you’ve seen, we’re increasing the number of packages that we’re offering so we’re still doing exome sequencing but simply we are offering more kinds of analysis that we do on them. Also, we’re more and more opening up the exomes so looking outside of the known genes to see if we can still find out what’s wrong with the patient. We’re moving to a process where genetic testing comes much earlier in the clinical research process. So rather than first doing all kinds of different tests, we much more quicker move to doing exome sequencing immediately.

So the last thing that we’re doing right now is actually we’re doing a proof of concept on using the whole genome sequencing and the results on that I can tell you look very promising. Thank you.

Slide 45

Well I forgot the acknowledgements, my apologies. So I would definitely like to acknowledge these people who really worked on all of these projects especially supervisor Joris Veltman and all the other people that you can see here on these slides. Thanks.

Sean Sanders: Great. Thanks so much and it’s good to see Dr. Hoischen again who was in one of our previous webinars.

Dr. Christian Gilissen: Yes, exactly.

Sean Sanders: Well thank you very much to both of our speakers for the wonderful presentations. We’re going to move right on to questions submitted by our online viewers. Just a quick reminder to those watching live that you can still submit your questions by typing them into the text box and clicking the submit button. If you don’t see the box on your screen, click the red Q&A icon and it should appear.

Slide 46

So we have received a lot of questions so far so I’m going to dive right in. We have about 15 minutes to take them. I’ll start with you Dr. Marshall. What are the advantages of exome genotyping over SNP genotyping and this viewer asks if some of the disease related SNPs on or possibly outside of the exome might be missed?
Dr. Christian Marshall: Yeah. So this is a good question and it does bring up one of the limitations of exome sequencing in that some of these – because you’re targeting things that are in the exons you’re not going to be able to find things that are intronic. But again, SNP genotyping is something that’s better left to do by probably a microarray and in general this is something probably more relevant to common disorders where you’re looking at a SNP association. What we’re looking at here is more exome sequencing sort of hypothesis where you’re looking for a rare variation.

Sean Sanders: Uh-hum. Dr. Gilissen?

Dr. Christian Gilissen: Yeah, I fully agree. I think if you’re looking at common SNPs, you’re more looking at less Mendelian diseases so then genotyping areas maybe the better approach because you can get a higher accuracy and you can also genotype outside of the exome. Of course, if you’re not really sure about your hypothesis that this is really common effects, common SNPs then exome sequencing is maybe a middle ground.

Sean Sanders: Uh-hum. I’m going to stay with you, Dr. Gilissen. We have a number of questions about coverage asking essentially what is the minimum coverage needed for exome sequencing when you’re calling variants?

Dr. Christian Gilissen: Yeah. So that’s a very good question. So I think if you want to call an individual variants, you know, 20x coverage is usually sufficient to make sure that you’ve covered both alleles. But the problem is with exome sequencing that if you get an average of 20x then you have a lot of variability in your target coverage so that’s why we sequence at a lot higher coverage. We go for at least a median of 75x and I think if you want the quality exome where you’re pretty sure that 90% of your targets is covered sufficiently for doing a reliable variant calling, I think that’s really necessary.

Sean Sanders: Uh-hum. Another question for you, Dr. Marshall, a number of people have asked is about validation. So what is the process that you go through to validate small indels in structural variance from your exome studies?

Dr. Christian Marshall: Yeah. So again a very good question. So as everybody knows especially the indels so Sanger or single the nucleotide variants are fairly accurate and obviously if it’s going to a clinical diagnosis and we have to do Sanger sequencing to do the confirmation. For the
indels, as everybody knows in exome sequencing there is a lot more error associated with them. So in general, we’ve been using Sanger sequencing but as Dr. Gilissen said we’re sort of moving towards either, you know, maybe doing a high throughput type of validation for those types of variants.

Sean Sanders: Uh-hum. Excellent. Let me give you another question quickly. Reproducibility for whole exome sequencing, how reproducible is it for the sequencing and for subsequent variant calling?

Dr. Christian Marshall: Another very good question. So we’ve done a lot of work with especially comparing to genome sequencing where we’ve had individuals that have been sequenced multiple times. Again, if you’re using exactly the same programs and exactly the same methodology, the reproducibility is very high especially for the single nucleotide variants. Again for the indels in some cases, they’re sort of mapped a little bit differently in different cases but in general if you sort of have that 75 to 100x coverage, reproducibility is very high.

[0:50:24]
Sean Sanders: Uh-hum. Anything to add, Dr. Christian Gilissen?

Dr. Christian Gilissen: Yeah. Well I fully agree. I mean and everybody is doing these kinds of validations so trying to see how reproducible the results are, I think what we see is that in general the results are very good. We get a high reproducibility and in cases where we have differences that are usually related to things like target coverage to repeat regions where something has gone wrong with the mapping or something as simple as the way the variants are actually being reported so you know, just a one base off or something like that. But I agree fully with Dr. Marshall that what we see at least is that if you have sufficient coverage and you’re not looking at some kind of a bizarre region then the reproducibility is very good.

Sean Sanders: Uh-hum. Great. Let me come back to you Dr. Marshall. In your experience, what type of variation seems to matter the most in explaining diseases looking at SNPs, indels, and copy number variation? Is there any kind of logic there?

Dr. Christian Marshall: Yeah. So again, you know, this is again goes back to what your indication is. If I take it back to something like the autism spectrum disorders then you know, the more complex neurological phenotypes and we saw it in Dr. Gilissen’s presentation as well, we’ve been most successful in the past anyway with looking at copy
number variants as finding deletions etc. What we can really kind of hone in on are probably these loss of function or these variants that sort of take out one copy of the gene for these complex things. You know, right now it’s really hard. I think it’s really difficult to interpret a missense mutation and what that means. But, you know, as we get more and more sequence we’ll be able to interpret these easier.

Sean Sanders:  Uh-hum. Dr. Gilissen?

Dr. Christian Gilissen: Yeah. Well it seems very difficult to say what contributes more. Obviously, we’ve seen in the past that copy number variants area major cause and now we’re really zooming in on single nucleotides variants that cause disease and we’re even moving into de novo mutations. I think we will have to see which variants contribute most to the disease in the future.

Sean Sanders:  Uh-hum. Let me shoot that question at you, this is actually a question that tends to come up quite a lot when we’re doing these kinds of discussions on sequencing is different ratio groups affected by the same disease can exhibit very different phenotypes.

Dr. Christian Gilissen: Yes.

Sean Sanders:  What would be the best approach to elucidate the causative genetic variance to explain this phenomenon?

Dr. Christian Gilissen: Yeah, that’s a very good question. So I should say that we usually work on our own patient cohorts so these are from within the Netherlands, from within our region so there’s not many ethnicity differences. I should say that of course the phenotypic spectrum that you see with different ethnicities is something that I really can’t say too much about since I’m not a clinician.

Sean Sanders:  Uh-hum.

Dr. Christian Gilissen: So I think this would really be a question that you would ask a clinician how to solve this problem for different ethnicities the same phenotype looks very different.

Sean Sanders:  Uh-hum. Dr. Marshall, are you looking at maybe a broader spectrum of –

Dr. Christian Marshall: Yeah. So obviously, the population in Toronto is where we do have a lot. My only comment would be that it is difficult in some instances
when you do have an individual of a non-European ancestry. A lot of
the control data or a lot of the population data out there is from
individuals of European ancestry and so it is important to take this
into account and when you’re looking at a lot of rare variants in
these other ethnicities than to then go do the right experiment to
see if this is just a population specific variance. So, you know, we’ve
seen this a lot where it looks like a really interesting variant. You go
back and you look and it happens to be just a population variant.

Sean Sanders: Uh-hum.

Dr. Christian Gilissen: I would agree. Sometimes when our quality controls are a little bit
off, we sometimes see that this is a patient from a different ethnicity
than what we’re used to so that’s why it stands out. So it’s definitely
ethnicity something to take into account when you are doing your
analysis.

Sean Sanders: Uh-hum. Great. Actually talking about analysis, we have a question
on that. Dr. Marshall, you mentioned your analysis time is about 12
hours I think in one of your slides. Is this normally how long it takes
and does the interpretation of the trio take longer?

[0:55:01]
Dr. Christian Marshall: Yeah. So what I am seeing a sort of analysis there. This doesn’t
necessarily include completely your all the way to interpretation.

Sean Sanders: Uh-hum.

Dr. Christian Marshall: And, you know, essentially what happens is the sequencing is done
this is with the ion proton and we just kind of run the analysis, the
variant call, the alignment in the variant calling etc., overnight. You
know, that end interpretation is something that could obviously take
a long time. It’s, you know, once you do have those variant calls that
are annotated, something might pop out relatively quickly and you
might see something within minutes and in other instances you
really it could take a long time. You could look at it for hours and not
be able to find something. So this kind of go into that next category.

Sean Sanders: Uh-hum. Great. Staying on analysis how important are numerical
qualifiers like P value and SNP score in filtering candidate variants to
causative mutations? Maybe, Dr. Gilissen, you can address that.

Dr. Christian Gilissen: Yes. So I think for us they are very important. We use this course to
actually remove those variants where we think we will not be able to
validate them in the first place and it takes a lot of interpretation burden at first instance from the people, the molecular geneticists who’s looking at the data. So many times, they use these values to filter out variants. They look at what’s left and they can make very quickly. They can figure out which mutation we’re looking for. Obviously, when you don’t find anything, you want to be able to go back and you want to look a little bit less strange in your data. You also want to look at these SNPs and variants whether are the qualities course are usually not sufficient but of course if you find variant then in a gene that is very much linked to the disease, you may still decide to actually try and validate that then we’ve had quite some success with that.

Sean Sanders: Uh-hum. Excellent. Let me stay with you for another question. Using exome sequencing, which is the optimal strategy to identify new susceptibility genes for heterogeneous genetic diseases like perhaps familial breast cancer and ovarian cancer? Would large case control studies be feasible for that?

Dr. Christian Gilissen: Yeah, that’s a very good question. I think actually that’s something that a lot of people are actually working on now. There’s a lot of consortia doing these studies because we all want to know whether this is going to give us new results. So there are some examples known where actually indeed by exome sequencing GWAS cohort, they did find rare variants explaining a small fraction of that cohort immediately. I think maybe these rare variants are not going to explain the majority of these patients for sure but at least it will explain a small portion and it will clean up the cohorts. When you are going to do your typical GWAS analysis, you will have a much cleaner cohort and you will get better results.

Sean Sanders: Uh-hum. Uh-hum. Dr. Marshall, anything to add—

Dr. Christian Marshall: Yeah. No, I completely agree. You know as we get more and more you get more and more sequence, it’s easier. You start to see patterns I guess and so you know you need to— a large case control cohort would be optimal in this case. I mean obviously, there’s budgetary constraints and things like that. But in some cases variants that are in unknown significance category will eventually trend towards probably being causative or at least contributing but you just need the numbers.

Sean Sanders: Uh-hum. Excellent. Well we’re coming to the top of our hour but I’m going to just get a few more questions in about the diagnostic side.
The first one asks what do you think about the need to differentiate between clinical or diagnostic exome sequencing or whole exome sequencing and research whole exome sequencing were the most important differences between the two. Dr. Marshall?

Dr. Christian Marshall: Yeah. So I mean this is something that Dr. Gilissen and I have been talking about quite a bit actually is that it obviously is a completely different level and in terms of the diagnostic exome sequencing, you really need a standard procedure. You need to validate the tests that you’re doing and it’s really a lot of work to go through that. You need to be a lot more cautious in what you can look at and pay attention to turnaround times etc. Whereas you know, in the research, it’s a little bit more, it is research and so you know, patients or subjects have consented for that and you can have a little bit more leeway.

Sean Sanders: Right. Dr. Gilissen?

Dr. Christian Gilissen: Yeah. I fully agree. If I look at what is different between the research exomes that we do and the ones that we want to do in the diagnostics you can see we put a lot more emphasis on quality control, standard operating procedures. But also we’re really thinking about the numbers because of course in research we have a research who is really interesting in his exomes and he can spend like months looking at a single exome. Whereas in diagnostics we really have to do large numbers so we have to really think about how to make this interpretation process going as smoothly as possible such that we can do, get the numbers that we need.

Slide 47
[1:00:13]
Sean Sanders: Fantastic. Great. Well a perfect place to end so we are unfortunately out of time for this webinar today and on behalf of myself and our viewing audience, I wanted to thank our speakers very much for being us, Dr. Christian Marshall from the Hospital for Sick Children in Toronto in Canada and Dr. Christian Gilissen from Radboud University Medical Center in Nijmegen.

Please go to the URL now at the bottom of your slide viewer to learn more about resources related to today’s discussion, and look out for more webinars from Science available at webinar.sciencemag.org. This webinar will be made available to view again as an on-demand presentation within about 48 hours from now.
We're interested to know what you thought of the webinar, send us an email at the address now up in your slide viewer; webinar@aaas.org.

Again, thank you to our panel and to Life Technologies for their kind sponsorship of today's educational seminar. Goodbye.

[1:01:08] End of Audio