Validating Complex Biology: How Arrays Can Complement Your Next-Gen Data Webinar
5 December 2012

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

Today's webinar deals with a topic that can strike fear in the heart of almost any researcher – data validation. With the growth of data in terms of studies comes the potential for more errors or misinterpretations in data analysis. Researchers therefore need to be aware of possible pitfalls and dangers of overestimation of their findings and it behooves them to use so called orthogonal methods to validate their data. With the emergence of sequencing-based studies that require scientists to identify rare genetic phenomena using relatively new methods, it is increasingly important to complement these studies with orthogonal technologies such as microarrays. During this webinar, we will discuss the importance and relevance of validating data and provide advice for validation next-gen sequencing results using microarray technology.

We have a fantastic panel of experts on the line with me today. They are Dr. Rafael Irizarry from Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland and Dr. Don Baldwin from Pathonomics LLC in Philadelphia, PA. It’s a great pleasure to have you both with us today.

Before we get started, I have some information that our audience might find helpful. Note that you can hide or resize 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 or additional information about technologies related to today's discussion or to download a PDF of the slides.
Each of our speakers will talk briefly about their work. After which we will have a Q&A session during which our guests 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 your viewing console and clicking the submit button. If you can’t see this box, just click the red Q&A widget at the bottom of the screen. Please remember to keep your questions short and concise. That will give them the best chance of being put to our 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 Affymetrix for their sponsorship of today's webinar.

**Slide 3**

Now, I'd like to introduce our first speaker for this Dr. Rafael Irizarry. Dr. Irizarry received his bachelor’s in mathematics from the University of Puerto Rico and went on to receive a Ph.D. in statistics from the University of California, Berkeley. He joined the faculty of the Department of Biostatistics in the Bloomberg School of Public Health in 1998 and was promoted to full professor in 2007. For the past 10 years, Dr. Irizarry’s work has focused on genomics and computational biology problems in particular and the analysis and preprocessing of microarray, next-gen sequencing, and genomic data in general. He is one of the leaders and founders of the Bioconductor Project, an open source and open development software project for the analysis of genomic data. A warm welcome to you, Dr. Irizarry.

**Dr. Rafael Irizarry:** Thank you. Thank you, Sean, for inviting me and asking me to do this. So up on the screen, you can see my webpage rafalab.org. You can get more information about what I’m talking about today basically by finding the papers, the relevant papers. For those on Twitter, I also have my Twitter account up there and finally there’s a link to a statistics blog that I participate in where we sometimes talk about genomics issues. Okay, so let me get started.

**Slide 4**

So I just want to start by reminding everybody of the wide number of applications of microarrays and I’m assuming that people know how they work, the basics of how they work, I’m not going to go over
that. But I do want to remind everybody that microarrays can be used for a multiple number of things. I’ve listed some up there. Gene expression is perhaps the one that people know the best. Genotyping also receives a lot of attention from all the GWAS studies. There are several platforms. They all are competitive, at least the ones I’ve listed up there, and pretty much all of them have a product for any of the applications.

All right. Now I’m talking today from the perspective of a data analyst and one of the things I want to point out is that for microarrays because they have been around for 10 to 15 years as a standard tool that data analysis tools are quite mature. There’s a project called Bioconductor that runs on our platform. They’re both open source and free where you can get pretty much complete pipeline for analysis of microarray data for several of the applications. If you don’t know about it, I encourage you to look up the Bioconductor Project and see the available tools that are there.

Slide 5
[0:05:26]

Okay. So I’m talking about two things today with respect to the topic of the webinar. One is data reuse, I will talk briefly about that maybe five minutes and then the more important topic of independent validation or verification, what does that mean and why is that microarrays serve as good independent validation of next-gen sequencing.

Slide 6

So I’ll just start, I’ll just give you a very quick example of how I’ve used this. We recently published a paper on DNA methylation comparing cancer, whole genome DNA methylation measurements comparing cancers and normals and one of the referees asked us to show the relationship to gene expression. Instead of running our own experiments, we went to the public repositories and to make this bigger, this is what the referees were asking for.

Slide 7

This is a very simple figure showing gene expression plotted against methylation.

Slide 8

If you look at the caption, you will see that the microarray data was from GEO dataset GSE8671. That’s not our own data. We actually got this from the public repository so it didn’t cost us very much other than the work of us downloading it.
Slide 9

So these are the two big ones. GEO is the American database and Array Express is the European one. They do pretty much the same thing. They basically store all the data from microarray studies and other studies too. Now there is some sequencing data out there too.

Now if you look at the – it’s not very clear in the slide, but if you look at the number of samples that have been uploaded to GEO, it’s at 830 some thousand. Almost a million different samples are up there. So there’s a lot of data and Bioconductor has tools for downloading them automatically, which is something I use almost every day. I’m always checking for new data and downloading them if I find it relevant to what I do.

Slide 10

One of the things that I’ve done with that data is create this web tool with some collaborators that we call the Gene Expression Barcode. Basically, what we did is we download it all the Affymetrix HG-U133 data and annotated it, re-annotated it because the annotation on these databases isn’t perfect, and basically reanalyzed it in a way that we could provide users gene expression calls for every gene on every one of these samples.

Slide 11

So just to give you another example of how we had done this, this is from another paper where we basically are showing gene expression calls for something like a hundred genes for 47 different tissues types, each tissue type containing several samples. This is all data that I did not create in a lab—I don’t even have a lab—but I was able to do this because of these databases. The data is there, you can download it, and you can create this. You could also go to our Gene Expression Barcode and download the sort of already analyzed and cleaned up data.

Slide 12

Okay. So now moving on to the main topic of webinar, which is validation and verification.

Slide 13

I want to start with some motivation using some data from last century. What you see on the slide here are estimates of the speed of light obtained from different labs. So every point is the estimate that was published and it’s surrounded by – it has a little line that
represents the standard error that was reported in the paper. So if you look at first two points, you can see that there’s a problem because according to -- if you interpret that correctly, you will say that the speed of light changed in one year and the significance of that is enormous. The P value is going to be super tiny here because those little bars are standard errors.

So how does it happen? How does something that is clearly incorrect get published like this? So if you look at where the points are, it’s actually quite remarkable that these labs in the 19th century, end of the 19th century could even do this. They were measuring the speed of light and they were getting it pretty close. If you look at where the points are, it’s 92.8 and 93.3. Those are actually very close.

But the mistake they made is that they didn’t realize that the measurement they were taking in their own lab were not as orthogonal as they thought. They weren’t as independent as they thought. There’s something about them that was similar that made the standard error, the variability seem too small. So if you were to “validate” the first lab result in another lab where the measurements are now independent, then you don’t – it doesn’t validate. Now if they had joined forced and looked and noticed this variability then they would have drawn a much, much wider standard error because that is the correct standard error to be drawn here.

If you just look at the across experiment variability, it’s much bigger than the reported variability. Now why is that happening? Because this internal validation that fools you into thinking that your variance is small than it really is.

Slide 14

So with microarrays, you have a similar problem and also with sequencing, with all these technologies that you have this problem that you can get fooled by what appear to be very precise measurements when they really aren’t and to highlight this, I’m showing this plot here where you can really appreciate the probe effect. What I’m showing you here are measurements from two genes so there’s a top column and a bottom column, those represent two genes. The left part of the plot have histograms and they’re showing you all of the values from that gene from the public repository. This is from an Affymetrix array and you can see that there is a big bump towards the left that appears to be the samples where this gene is turned off and it’s roughly about at 4. This is log base 2 intensity from the microarray so it’s 2^4 around 16.
Now the other gene, the same hump that appears to be related to lack of expression is up at 7 so it’s 8x bigger. So you have in one case the gene when it’s not expressed shows numbers that are around $2^4$ and for the other gene it’s around $2^7$. So basically what’s happening here is that these two are being measured in different units. We don’t know exactly what those units are but they’re not directly comparable. So what’s the consequence of that?

**Slide 15**

If you take two samples, that’s what is on the left graph here, graph on the left, you take two samples. Say you take two technical replicates and you compare them and you validate. So let’s say one gene showed a gene expression of 14 in the log scale and you do it again and guess what you get 14 again and the gene that showed a 4 in the replication, you got a 4 again. So you think is great and you get an R, a correlation of 0.996 so you think this is the best technology that has ever existed because you basically got the same answers. But the problem is that what’s driving this is not biology but the probe effect.

So now if we do a different kind of validation where instead of comparing the absolute numbers, we compare relative numbers so now I’m going to make a comparison of relative expression. So I’m comparing case versus control and then I have a replicate, a technical replicate of case versus control and now instead of absolute values measurements, I am showing log fold changes, the correlation drops to 0.5.

Now that’s actually what you should get because most genes are not differentially expressed so you should see a big circle in the middle that’s not correlated and then the points that are outside of zero those should correlate and that’s actually what we see so that’s actually pretty good. If you see to the upper right corner, you see 3 point and if you look to the lower left corner you see like 7 points that are correlating. They’re being validated. So this is slightly better validation because once you – remember with the probe effect, you remove this artificial correlation, this lack of orthogonality and you get something that’s a better value.

**Slide 16**

So now let’s move over, move on and do something even more orthogonal which is to compare two different platforms. So now, we’re comparing Affy and 2cDNA arrays. This is even more orthogonal so even a better validation. Again, we see this big circle in
the middle of genes that aren’t differentially expressed, they should not correlate. That’s a good thing because there’s nothing to correlate. There is just noise. Then you see points, the blue points are the validated points. These are the ones that should be correlating and they are.

Now you see if you look at x=0, you will see a series of black points that are outside the red circles. Those points would validate if you use Affy to validate Affy but now they’re not validating because - I’m sorry if you used two-color arrays to validate two-color arrays they’re high and in Affy they’re not, they’re at zero. So those are genes that for some reason the one platform is biased in some way and it gives, it throws this error and by comparing to the different platform, you catch that. So that’s how microarrays are a very good way to validate sequencing because they’re independent. They are even more independent than Affy and two-color arrays because now you’re moving not just across array platforms but across completely different sequencing technologies.

So that’s the key point that I want to make. If you use sequencing to validate sequencing, you might be fooled into thinking that you validated when you really didn’t because there’s some kind of error, some kind of systematic error that is repeating itself every time and making it seem like you’re validating when you’re really not validating.

Slide 17

All right. So I want to give you one last example where I used microarrays to validate something that we are seeing in sequencing and I also want to make a second point here that’s very important and it’s the fact that there is biological variability in things like gene expression and other outcomes that are of interest among people in genomics. This one is particular to gene expression and basically what we are trying to show is that biological variability is a natural reality and sequencing does not eliminate that.

Slide 18

The motivation for writing this paper was that we were seeing a lot of studies being published with just one sample because sequencing was somehow magical and removed all variability, but in fact it doesn’t remove perhaps the most important variability, which is natural uninteresting variability.
So just to make that point, this is from an old paper, we’re making this point on microarrays and you have two genes. You can see that they’re labeled gene 1, gene 2 and for each gene you have two panes the right and left. On the right, you see data from 12 mice from two strains so there’s the blue strain and the green strain and on the left pane you see the result of pooling all those samples and creating technical replicates. So you can see the technical replicates for arrays is incredibly precise. Arrays are actually very, very precise and the variability that we see mostly comes from nature. That’s what this plot is showing. You can see on gene 1 varies a lot, gene 2 varies less and you don’t see that if you just pool and do technical replicates.

Slide 19

So you can see that with sequencing as well. This plot now is showing you what we did for this plot and it’s also an example of using published array data. We took sequencing data from RNA-Seq that was looking at something like nine individuals and for each gene, we computed the variability across individual. So each point is a gene in this plot. We did it for arrays and we did it for sequencing. This is the same HapMap samples that were done in sequencing and array and what you see is the correlation between the variability seen in sequencing and the variability seen in arrays because it’s a natural thing. It’s there, it’s being measured correctly by both technologies. So you see the genes that vary a lot in sequencing, vary a lot in arrays.

Now you’re going to see that two plots that’s because we did it for two different studies. It’s two different HapMap sets and you see the same thing in both. So the points, the purple – sorry I don’t know what color that is – pink and light blue points, I’m going to show you now those genes, I’m going to show you a close-up of those genes and here they are.

Slide 20

So there’s sequencing and array are top and bottom and then there’s two genes and there’s triangles and circles because they come from two separate studies. What you see is that the first gene, the gene on the left doesn’t vary much across individuals. Each point is an individual but the second gene varies a lot and you see that in both technologies. It is not problem with arrays, it’s not a problem with sequencing, it’s a reality of nature. That is an important reality that we can’t ignore going forward.
Slide 21
[0:20:31]
So that is where I’m going to stop and I will take questions later. Thank you.

Sean Sanders: Great. Thank you very much, Dr. Irizarry. We’re going to move right on to our second speaker for today and that’s Dr. Don Baldwin. Before I get to introduce Dr. Baldwin, just a reminder that you can still submit questions, just typing them into the text box on the left your viewing console and then click submit.

Slide 26
So Dr. Baldwin completed his Ph.D. at the University of Florida in Gainesville and his postdoctoral training at Pioneer Hi-Bred International in Johnston, Iowa. He then moved to the University of Pennsylvania where he created and directed a microarray resource facility in the Perelman School of Medicine, leading the core’s development into the Molecular Profiling Facility for array, deep sequencing, and bioinformatics services. Dr. Baldwin is co-founder of Pathonomics LLC, which provides tools and services for profiling complex collections of microorganisms found in a variety of clinical and environmental specimen types. A warm welcome and thanks very much for being with us, Dr. Baldwin.

Dr. Don Baldwin: All right. Thanks very much, Sean. I appreciate the opportunity to share with you and the listeners today some projects that I’m working on as part of our new start-up company and hopefully illustrate using these projects how we believe that arrays, probe arrays and deep sequencing technologies will continue to complement each other.

Slide 27
So as with anybody trying to do molecular biology these days in the modern era of genomics, we are faced with a wide array of possibilities for experimental design. I’ve illustrated that on this slide where we start often by considering how many analytes do we really want to study. So perhaps in one experiment, it’s only important to do 5 or 10 or 12 different things, maybe their RNA expression levels or polymorphisms in the genome or sites of methylation and the experiment is limited to those analytes and a variety of different sample sizes. On the other hand, we might want to turn to an omics technology where we are very interested in screening the entire genome or proteome or methylome and therefore need a completely different platform that’s allowing now highly parallel
analysis of lots of different analytes. Then of course there are those experiments that fall in the middle where you want to target a certain number of analytes but that number is quite large. So several hundred analytes to be screened to determine an expression signature perhaps.

Well once we’ve made that determination as to how many analytes, the decision tree gets even more complicated because now we have to decide how many samples and so often it’s a bit of a gray area. Do you choose to do fewer analytes but more samples or for the same cost and turnaround time perhaps a larger panel of analytes but a limited number of samples. So there are many platforms that provide us with the sort of molecular biology platform that we need to conduct these experiments and I’ve listed just a subset or a few of them here.

**Slide 28**

As many of the listeners are aware, they range from say Sanger sequencing all the way through to highly parallel microarrays to the deep sequencing technologies. Something, some decisions are quite clear-cut so of course you would not tackle a whole genome sequencing project using Sanger technology these days. That would be something you would run on a deep sequencing instrument like the Iron Torrent or a HiSeq. But similarly, if you just want to know that you cloned a cDNA in frame in your expression vector, you wouldn’t use and Iron Torrent or a HiSeq platform to do that so most decisions like that are not so clear. The possibilities for using say microarrays or probe arrays or deep sequencing could have many different places that they fit into this decision tree and for about the same amount of money you might run either an array or a sequencing experiment.

So we have to then consider not only turnaround time and cost as I’ve diagramed here getting larger as you proceed across the tree for number of analytes and number of samples, but also the length of the spreadsheet. That is the cost in time and money that you incur for conducting the analysis. So all of those factors have to be considered once you’re in between platforms and technologies to start one of these large-scale molecular biology projects. We’re faced with the same thing in this company that hopes to apply these technologies for microbiome and metagenomic screening.
All right. So I’m going to start by looking at my definitions for the classes of genomics platforms that we’re considering and the first is the determinate platform. An example of this is probe arrays. So in a typical microarray experiment, we have two components, the target which is the complex sample that usually carries a fluorescent dye but otherwise represents the, you know, complex pool of abundances for lots of different analytes, perhaps RNA transcripts.

The other independent component is the probe array here diagrammed as a flat piece of glass with an oligonucleotide coming off of it and this microarray has defined positions for each probe and each probe is many copies of the same synthetic DNA at a known location. These arrays of course can be quite sophisticated and I’ve illustrated at the bottom here an Affymetrix gene chip that contains millions of probes all at discrete locations in a determined, predetermined that is, content and layout.

Slide 30

So with a determinate platform like microarrays, you have independent measurements. So the probe content on the arrays of course independent of the sample that you’re hoping to measure. The array is designed first before you ever collect the sample. So the consequence of that is that the measurements of abundance for one analyte doesn’t affect the measurement of another. Sorry the graphics are a little displaced here, but you get the idea that if in your target pool you have transcript A, which is quite abundant and transcript B which is fairly rare, then hybridization of all those A transcripts to the A probe is going to give a lot of fluorescent signal but that signal does not have or very minimally affects the signal at probe B. So you independently measure those two abundances and the result from one doesn’t affect the other.

Slide 31

Compare that to a dynamic platform. So for this sort of technology, I am using RNA-seq that is the application of deep sequencing to profiling RNA abundance and the components here are similar in that we have a complex pool of things typically will get called the library. So in this library, you would have a representation of the abundances of all the different RNAs and the measurement is again a microarray platform, in this case the Illumina HiSeq is a randomly assembled array of features each of which is a representation of one component of that library that’s being sequenced, but it could also apply to Iron Torrent or 454. All of them are either randomly assembles arrays of templates or arrays of reaction centers, wells or
channels say in the packed biosystem, which in a highly parallel way conduct sequencing reactions.

Slide 32

This is a dynamic platform because the array content, the measurement in each of these arrays of reaction centers depends on the composition of the library. That is in the case diagram here, when you randomly sample a library that has a composition of A and B transcripts then the things that will be measured on the array are A and B and the relative number of spots or wells filled with each transcript will represent that library. That ratio may well be different in a different sample or a different replicate of the same sample.

So the consequence of this is that randomly sampling the population that you’re hoping to measure consumes measurement capacity such that repeated measurements of the most common analytes will use up your sequencing capacity at the cost of measuring the most rare things. So that’s not a problem when you have millions of different locations available say on this Illumina HiSeq channel and the relative difference is not that big. So here it’s 6:1. But what if it’s 6,000,000:1 or 6000:1 and what if it’s not just A but tens of genes that are in the high abundance class and thousands of genes in the low abundance class? You can see how the sequencing capacity of any of the deep sequencing platforms quickly becomes consumed by just remeasuring over and over again the most abundant components of the library and not measuring the most rare of the components of the library.

Slide 33

So in comparing determinate and dynamic genomics platforms, these are the key considerations and the considerations that I think allow us to see how the two different approaches complement each other. So in a dynamic platform, you’re often looking de novo content, in other words taking a sample of mostly unknown things and you’re doing a screen. You’re going to discover what’s present in that sample. Whereas on the determinate side, you have an assay that has predetermined content. You have to know ahead of time what the probe sequences will be in order to create the probe array and those then are the things that will be measured in your complex sample.

[0:30:24]

In a dynamic format, the coverage will depend on throughput -- the throughput, sorry, will be dependent on the coverage. In other words, if you want to cover a particular sample with many, many
measurements then you will be unable to put lots of different samples in because as you add more samples to the same measurement capacity, you’ll decrease the amount of coverage per sample. Whereas with a determinate platform the coverage is independent of the throughput. You can do as many samples as you like. Each sample is on a separate assay and that assay is exactly the same across the entire experiment.

In a dynamic platform, typically the analysis starts with identification. In other words, you have to first figure out just what it is that you measured before even considering quantitation. Whereas on the determinate side, you predetermined the identity of what you’re measuring and barring some sort of cross-hybridization or noisy signal, your analysis typically starts with quantification of the analyte.

Then finally, I’ve listed a few examples of both the dynamic and determinate approach. So in a dynamic system we often think of mass spectrometry for proteomics, which also has this property of measuring the most abundant things, most commonly and then rare things not as commonly. And then on the determinate side in proteomics of course, the old standby’s like ELISA assays and multi-well plates or now increasingly the production of arrays with antibodies. In the RNA and DNA world, the dynamic platform certainly is RNA-Seq which we’ll continue to talk about today, the use of deep sequencing technology to do transcript profiling and the determinate side will be RNA profiling on a probe array like Affymetrix or Agilent or NimbleGen.

For those of you interested in say methylation, there’s again the example continues. The dynamic platform might be the Sequenom EpiTYPER assay, which allows you to identify a region of genomic DNA that you want to assay and then this technology will tell you whether any methylation is present without knowing ahead of time, which basis might be affected. Compare than then to the Illumina microarray for methylation, which uses probes to a variety of known sites susceptible to methylation and then determines whether or not a methyl group is present in a particular sample.

So I think I’d like to illustrate using this small diagram the complementarity between a dynamic and a determinate genomics platform. So often, you would use the dynamic platform like deep sequencing or regular sequencing to do analyte discovery and it’s the
discovery of the things that are present in a complex sample that leads to production of the determinate platforms. So the probe array depends on having that prior sequence knowledge in order to create the probes.

The determinate side then often allows focused profiling with low cost, high repeatability across a large cohort of samples, something that sometimes is more difficult when you’re using an open ended or dynamic sort of platform. So often, the determinate side is a lot of probe arrays run in a highly repetitive format on many different samples in order to screen samples to find phenotypes, clinical cohorts, other conditions, time points that might be most interesting and most productive to feed back into your deep sequencing reaction. So a sort of large scale, high throughput sample screening to identify the conditions that are most interesting and then deserve more attention back on your de novo dynamic platform.

Then as Rafael nicely illustrated, both technologies can be used for validation. It’s important to cross over the technology barrier to try to use a completely different method to validate results that you got in your screening or discovery stage.

Slide 35

So as an example of what I’ve just discussed, I want to describe for you the PathoChip which is a project that I’m working on to build a probe set and deploy it first on microarrays and then on sequencing that measures the presence of all sequenced viruses in the public domain plus several hundred pathogenic microorganisms. So our PathoChip is an Agilent custom microarray that has DNA probers 60mers printed on the surface of the array with an average of about 10 specific probes for every accession and an accession being a virus or a particular species of bacteria or fungi and so on.

[0:35:08]

Our current PathoChip probe collection covers over 3200 viruses and several hundred other organisms on standard lists of human pathogens including a variety of bacteria, fungi, protozoans and helminthes.

Besides these probes for specific DNA regions that should be unique to each of our accessions, we’ve also designed probes that target regions that are conserved between two or more viruses as well as probes that saturate the genome of particular organisms that we’re interested in.
So as an example of the conserved probe design, here is a diagram of say four genomes, four different somewhat related viruses and even though the overall genomes of these viruses is not related, there are regions that are conserved. So for example, target C1 is present in both virus 1 and virus 2 and so we have probes that target that region C1 and should be able to pick up the presence of virus 1 or 2 in a sample.

Where we think this will be useful is in the case of say virus 4. So pretend that virus 4 is something not in the current sequence database and therefore represents an unknown virus not previously characterized. Well having probes to target C3 that is known from other organisms, other viruses allows us the chance to hybridize to that genome from virus 4. So we hope that using these comprehensive collections of conserved probes we’ll be able to pick up regions in unknown virus genomes.

Also, we’ve saturated a variety of viruses so we’re interested in viruses that cause cancer and so for a number of those viruses, we’ve chosen probes that will saturate the entire genomic sequence. In this case, I’ve diagramed cytomegalovirus where we have several hundred probes to the genomic region of CMV in the reference sequence and then hybridized that to a sample that contains a particular strain of CMV AD169.

So probe 1 shows the hybridization level for a sequence that’s present in AD169 but probes 2, 3 and 4 are regions of that particular strain that are quite variant so they’re different than the reference sequence and so you see the probe signal level on this probe set array is lower. The probes in region 5 to 6 are to a region of AD169 that’s completely deleted so of course no signal. And what we think we can use these saturation probes sets for then is to sort of map what strains and what versions of a particular virus might be present in a sample compared to the overall reference sequence.

We’ve used our probe set deployed on arrays to do typical sorts of spike-in experiments. This is just a diagram showing that it’s a quantitative response using whole genome amplification on our samples with a couple of different approaches showing that we can detect a spike in PhyX 174 genome against the human genomic DNA and also illustrating on the right that we don’t even have to do
amplification. In this case, an Epstein Barr Virus infected cell line was used to just extract total DNA and hybridize it to the array and we can detect the EBV that’s present even without amplification reaction.

Slide 39

So we did a pilot screen and using our PathoChip probe array looked at 20 GI tumors some from small bowel, some from stomach and we extracted DNA from those tumors and hybridized it to our PathoChip. Most of the samples showed no presence of the viruses and the pathogens and so most of the signal is yellow as diagramed in the main hierarchical clustering here. But three of the samples you’ll see that cluster to the left have some hits for particular organisms.

Slide 40

So those organisms are shown here on this slide where I’ve indicated the full change above background level for a particular pathogens or microorganisms in these three samples. Some of those organisms are known to be associated with the GI tract so that sort of makes sense that they be present in a GI tumor. But some of them are not normally seen in the GI tract and one of those is a hit that we got with 6-fold more abundance than the reference signal for Fusobacterium nucleatum. The reason I point this out is that a couple of months after we got this from our pilot experiment, we learned of two independent validation papers published in Genome Research that show fusobacterium association with colon cancer. So we believe this technology is working for screening tumors to uncover the sort of microbiome and maybe even the causative biome of microorganisms that are present in cancer.

Slide 41 [0:40:00]

So our goal has been for this technology are to be able to simultaneously measure DNA and RNA from a complex biological sample. To also use clinical samples that often are in poor shape because they’ve been in formalin fixed and embedded in paraffin, and also be able to expand our assays to a variety of other input samples types like blood or serum or saliva or even in nonclinical samples like from soil or water or plants. Then also to be able to boost our sensitivity by employing some novel whole genome and whole transcriptome amplification methods or having the choice of doing this without amplification.
So given that these are our goals, we also have analysis goals that go along with those. So we went to be able to update our probe collection regularly and change the way the platform is composed as new organisms are identified. We also like to customize the content so that we can create focused panels that measured a subset of organisms for a particular application. We’d also like to, you know, sort of standardize our analysis pipeline so that experiments that have similar goals can use a similar analysis method yet also have the flexibility to customize that as we need to provide custom reports for a particular client’s goals.

So given these goals for both technology and analysis, we as most people have, are facing a choice of arrays or deep sequencing or both.

As I diagramed before, we think we’ll probably use both. We think that we can use the dynamic side of a sequencing project like the Human Microbiome Sequencing Project to do analyte discovery, which then will inform the content of our PathoChip probe set. So we’re current using Human Microbiome Project sequencing data to generate new content for the PathoChip.

In the Human Microbiome, deep sequencing is being used two ways. So on the left is the workflow for discovery of anything in their sample. This is a genomic DNA extraction followed by shotgun library prep and then very sequencing intensive sort of workflow in which only 3.5 samples are analyzed per channel of an Illumina HiSeq generating 10 billion base pairs of sequence in order to discover what microorganisms might be present. The de novo comprehensiveness of this approach is balanced by the inefficiency in that first step of analysis, which includes alignment, assembly and discarding human sequences often means that you’re throwing away half or more of the reads that you just paid for because they came from the human host and they’re not part of the microbiome.

Compare that then to the right format where you are just focusing on ribosomal RNA sequence. So here, PCR is used to amplify 16s ribosomal sequences and then a Roche 454 approach is used to do much less sequencing to identify what bacteria might be present. The limitation here of course is that it’s just bacterial. You’re only
looking at 16s RNA and so there won’t be any representation of viruses or fungi or other microorganisms beside bacteria.

Slide 45

So we’re going to then consider besides sequencing how will we use our array to inform our sequencing projects and we think that sample screening as I mentioned before it’s one of those ways. We’ll use the probe array on the Agilent microarray to screen lots and lots of samples and determine which ones deserve the most attention for higher cost, lower throughput sequencing.

Slide 46

Also, we are developing using these probe sets an enhanced protocol for sequencing. We call this PathoChip sequencing in which we hope to combine the best characteristics of a determinate platform with a dynamic sequencing approach retaining the opportunity to discover new microbiome content but doing it in a way that boosts cost and time efficiency.

Slide 47

So we think our PathoChip sequencing format will fit into the sort of middle here where we’re still allowing the opportunity to discover new things whether it’s a virus or a bacterium but in a way that allows us to put many more samples into a HiSeq sequencing channel boosting efficiency.

Slide 48

Okay. So finally as was mentioned before, we think both platforms will validate, you know, will be used to validate the other. So we think that screening experiments using arrays may implicate a particular casual organism. Say we might discover other organisms in colon cancer by screening hundreds of samples and then of course we’ll have to define exactly what strain or species that organism is and we’ll use deep sequencing to do that.

On the other way around, sequencing say in the Human Microbiome Project might suggest signatures for different phenotypes, different clinical conditions and those signatures then provides the content that is appropriate for a focused determinate array like the PathoChip format to validate independently that signature across a large clinical cohort.
So using determinate approaches and dynamic approaches, we think we’ll be able to do the full cycle of analyte discovery, sample screening, and validation to characterize microbiomes and metagenomic samples.

**Slide 50**

Our services will range from array and sequencing service on samples that are provided to us to development of focus subsets for particular applications to in the future good manufacturing compliance, protocols, as well as clinical diagnostic protocols.

**Slide 51**

Our applications range from screening of human samples not only for tumors but also for other disease phenotypes perhaps to uncover causative pathogens or prognostic biomarkers or even microbiomes that normally occur and are predisposing for a particular clinical condition. But then not only human screening but also veterinary pathologies, monitoring of the microbiome that’s present in biopharmaceutical manufacturing processes that’s present in environmental samples that might have clinical and public health relevance as well as perhaps foods safety.

**Slide 52**

So with that, again I appreciate the opportunity to talk about how I think arrays and deep sequencing complement each other as well as validate. If anyone is interested in talking about this further or in microbiome screening, feel free to contact me.

**Sean Sanders:**

Fantastic. Thank you very much, Dr. Baldwin. Many thanks to both of our speakers for their superb presentations and we’re going to move quickly on to the questions submitted by our online viewers.

**Slide 53**

Just a reminder, you can still submit them by typing them into the text box and clicking submit. If you don’t see that box, just click the red Q&A icon.

So we’re at about five minutes before 1:00 but we did start late so I’m going to run over the hour a little bit and we’ll see how many questions we can get to. So the first question I’m going to put out to both of you. The original question that I planned to ask you talked about major sources of variation but we have covered quite a lot of that so I’m going to change the focus slightly. Dr. Irizarry, I’m going to come to you first and ask about the mention that you made of
various platforms and tissues, one of our viewers asks how you merge and also account for variability beyond biological variability in these samples?

Dr. Rafael Irizarry: Yes. That’s a good question. So the batch effect, something we didn’t talk about today, but I highly recommend everybody learn about it if you don’t know about it already. It’s a difficult problem in genomics in general. We see it in microarrays, we see it in next-gen sequencing as well and it makes things like integration difficult. So what we have done specifically for this Barcode Project if you read the original paper, which you have probably seen in Nature Method in 2007, one of the things we discussed is the problem with the batch effect that it makes it very hard to integrate data from different studies.

What we ended up doing at that point was using a very crass approach, which was to turn – try to make very binary calls or either call something off and on and when you do that, you do get things that are slightly more comparable if you are able to look at the data across several samples. So the general approach I’ll just say a couple of sentence on this, the plot for every gene you plot all the data. You make a histogram basically of all the data for that gene across all the tissues and you will see the batch effect in that especially at the background level where things are off. You will see that some samples are off but they’re like at 4, others are off but at 3 and by turning them on to zeros, we in a way get rid of that batch effect. When you do those a lot of sensitivity by turning things into zeros and ones and it’s part of our current work to try to do something less crass than that. But it is a difficult problem and you can see what we’ve done, the more details of what we did by reading the papers.

I would say that in general when you have any kind of dataset that you’re planning to analyze and it’s big, you always want to be looking at the principal components and S plots and look out for batch effects. We have a Nature Review, genetics review paper on this that can you can also look at for more details.

Sean Sanders: Great. Dr. Baldwin, anything to add to that?

Dr. Don Baldwin: Yeah, I was just going to mention the need to make sure that you examine your data ahead of time to see whether a batch effect is present and some sort of complexity reduction analysis like PCA, which is really useful for telling you that something is going on. So at least then you can make an approach to address it.
I also briefly wanted to mention that these sources of variance are of course, you know, present in sequencing or in arrays and some of you will remember the MAQC project sponsored by the FDA to get a handle on where the sources of variance come in to a microarray experiment using standardized samples. Well that project has been resurrected for next-generation sequencing and there is a large consortia of labs organized by both the FDA and in parallel by the Association of Biomolecular Resource Facilities, ABRF, to use standardized samples to look at the variance across many different kinds of deep sequencing platforms focusing on RNA-Seq. So stay tuned for papers probably coming out early next year repeating this sort of MAQC approach for deep sequencing helping us to see where different points technically and maybe even biologically variation can enter into a genomics project.

Sean Sanders: Great. So let me stay with you, Dr. Baldwin. This question says when using microarrays to validate next-gen sequencing, the matching of transcripts to probes is challenging due to annotation and this carries over to normalization as well. How would you suggest one deals with that?

Dr. Don Baldwin: So often, I just study the signal level without worrying about the annotation. In other words allow the statistical analysis to tell me which probes on the array are doing something interesting with a good P value, good replication within the platform, and then turn to trying to figure out what it was that was being measured. So a little bit reminiscent of the deep sequencing, now you need to identify what it was that you measured. So yes, the first thing to keep in mind is that the identity of the probe given to you by the manufacturer doesn’t mean that it’s exactly interrogating the annotated sequence that goes along with that probe. There may well be other sources of signal.

So just as with the validation of next-gen sequencing using a focused array, we need to validate array using something else. If you’re going to really put a lot of effort into the result that came from that probe set or that signature then it’s always best to go in with TaqMan or Fluidyme or northern blot or Western blot or something else besides microarray or now of course sequencing RNA-seq to really confirm for you that independent of the probe design, the accession, the annotation that you think is responsible really gave you that interesting differential expression result.
Sean Sanders: Great. Dr. Irizarry, a question for you. When you put different datasets from GEO together, how did you normalize the dataset to make them comparable?

Dr. Rafael Irizarry: Well it depends what kind of datasets we’re putting together. If it’s for example something that is colon cancer, that colon example I gave where there was at least two studies with Affymetrix that were on Geo, we could have put them all together and basically run RNA on all of them and performed a batch correction using either ComBat or SVA, these two methods for batch correction and then continue. But usually for that particular study what we ended up doing was analyzing them separately but that’s what I would do if it is something like that.

Now if you are joining together two datasets, let’s say where the outcome of interest is confounded with the study they come from. Say I want to compare liver to kidney for some reason and the liver data comes from one study, the kidney data comes from another, that is much harder because then the batch effect is confounded with the outcome of interest. Again, what we’ve done as described in this Barcode paper where we basically turned things into zeros and ones on and off. That’s like the only thing we have been able to come up with up to now that we feel completely confident but more confident in. It’s this crass approach of turning things into zeros and ones so that then things being comparable across different studies.

You could try to do that across platforms as well. If you have a way to turn data into zero and one for another platform that isn’t Affymetrix then it becomes more comparable if you do it that way. The absolute numbers are not comparable. Those are measured with different probes so they’re in a different unit and you should be very careful about that kind of integration. It’s a hard problem and again just to summarize what we have done up to now is turn things into off and on calls and also some of the genes get called we don’t know, we leave those out.

[0:55:35]

Sean Sanders: Dr. Baldwin, a question for you. Can you elaborate on the levels of expression that would be required so that you can be sure about the biological, pathological manifestation in the samples you’re studying?

Dr. Don Baldwin: It’s a common question with a very difficult answer in that each assay performs a little bit differently. So this will probably be the
case in sequencing as well. Each probe on the array and each gene captured in a sequencing library will have a little bit different inherent noise, assayability of if you will, ability to detect it and to discriminate change. So whereas gene 1 may be easily discriminated at a 1.5-fold change because it’s expressed in a sweet spot of the dynamic range for your technology say a thousand units versus 1500 units of fluorescence. Gene 2 may be very difficult to measure that same 1.5 fold expression difference because it’s down near the background of your experiment where maybe it goes from 10 to 15 units of fluorescence. So relying on a pure fold change cutoff or a pure threshold for amount of signal often unnecessarily throws away data. The better approach I think is to apply multiple testing corrected statistical analysis to tell you which genes whether it’s from sequencing or arrays consistently change regardless of their place in the dynamic range of expression. Then based on P values, start applying filters for a level of fold change and level of expression, but start with a statistical analysis and a confidence measurement.

Sean Sanders: Great. Well unfortunately, I think that we’re going to need to wrap up the webinar shortly but I’m going to give you both one more question and I hope it’s not too divisive. But the question that came in is would the two of you be willing to head a movement demanding validation of next-gen sequencing data before publication? And I’m guessing this viewer is thinking something similar to the MIQE guidelines that I know that there’s a number of scientists that are already pushing. So Dr. Irizarry, how about you start us off?

Dr. Rafael Irizarry: No, I’m not willing to head a movement but I will definitely continue to give talks and publish papers that explain all these issues and, you know, I hope that others do the same. I think in general, science sometimes get a little overexcited about things that come up that are new like they did about sequencing and they did about microarrays 15 years ago so you see like this flurry of papers that are perhaps a little too optimistic, but that kind of gets self-corrected. So I’m actually a positivist about all this that I think at the end even though we see there’s a lot of noise, the signal somehow at the end comes out. I know there’s examples and I know we can do better but I think that the way to do this is just from the bottom up instead of the other way around.

Sean Sanders: Great. Dr. Baldwin, last word to you.
Dr. Don Baldwin: I don’t think I’d start a revolution to mandate anything in the genomics field because it changes too quickly and I agree that I’m very much reminded of the process we went through with microarrays repeating itself now with deep sequencing. So I think the best way is again from the ground up as Dr. Irizarry said. When I review papers, I insist that they be done with proper replication and if critical enough, validation on a different platform regardless of how much was spent. So just because it cost a lot to do sequencing doesn’t mean that you’re able to ignore biological variance. So I think for the person who asked the question, you can all lead this effort by asking questions at seminars, asking hard questions when you review papers, by getting postdocs and graduate students to consider the sources of variance wherever they come from, enforce good science in genomics regardless of which technology you’re using.

[1:00:17]
Dr. Rafael Irizarry: That’s a good way of saying it.

Sean Sanders: Fantastic. Well, yeah, I think that’s a great way to end the webinar. So it just remains for me to thank our speakers for providing such interesting talks and discussion, Dr. Rafael Irizarry from Johns Hopkins Bloomberg School of Public Health and Dr. Don Baldwin from Pathonomics.

Many thanks to our online audience for the questions you submitted to the panel. I’m sorry that we didn’t have more time to get through all of them.

Slide 54

However, you will be able to carry on the conversation in about three weeks from now by going to blog.affymetrix.com where some of your unanswered questions will be addressed.

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 particular webinar will be made available to view again as an on-demand presentation within about 48 hours from now.

We’d love to hear 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 wonderful panel and to Affymetrix for their kind sponsorship of today's educational seminar. Goodbye.