

Well good morning, everybody. My name is Christophe Lambert. I'm the President and CEO of Golden Helix, and I'd like to welcome you to the second in a series of broadcasts that we've been putting on and will continue to put on in achieving genome-wide success. The topic for today is Study Design for SNP and CNV studies. This is a topic that's particularly near and dear to my heart, because over the last couple years we've been doing many genome-wide studies in collaboration with our customers, and we've run into a lot of challenges that really could have been addressed by study design. Today's Webinar has almost 550 people who've enrolled in it, and I think the topic is one that's of high interest to all of us, whether we've suffered from the challenges of batch effects and other problems with experimental variability or whether we're contemplating creating a new study.

So today's presentation, first I'll show a little bit about some of the impacts of a poor study design, just what kinds of problems we've been facing, and if you've been facing them as well you're not alone. It's the rule rather than the exception.

I'll talk about what are these sources of experimental variability and how we might minimize them, and then we'll wind back a bit to talk right about the beginning of study, as you're contemplating whether you do a population or family-based study, how you might calculate power in a study, choice of platform consideration. Then we're going to spend some time talking about plating strategies and how to really go about creating a good experimental design. Then we'll show some examples of when things have gone well and things have not gone well with regard to design of experiments for genome-wide SNP and CNV studies.

We've looked at 30-plus studies now and 28 out of the 30 that we've looked at have had some sort of experimental design problem. For the most part they've usually been sizable problems. So only in about two studies that come to mind that we've looked at over the last two years where design of experiments principles were not a problem. So we've seen this in very famous labs and in government, academic and commercial. This is a ubiquitous problem that we all suffer from is challenges with our experimental design.

There are often some good reasons why we've had to make certain decisions about perhaps borrowing controls from another experiment and so on, but it's led to drastic consequences in the analysis downstream. Some of the things that are the biggest recurring problems are when the cases/controls are not balanced and randomized across plates. I'll often use the term randomization, but you'll see when we talk about experimental design that sometimes it's more important not to – random is not necessarily good enough. You do have to have a balanced design across your plates of equal numbers of proportions of your phenotypes as well as other experimental variables that could cause biases within your study.

Borrowing controls is a good way to cut costs, but whether it's a SNP study and especially a CNV study, if you've run your experiments at different sites and done all the

cases in your site and the controls from a colleague's site, there's going to be inevitable negative consequences.

In family-based studies, whether its trios or nuclear families or even extended pedigrees, we've seen tremendous problems when you don't keep your families together on plates. Most of the tests that go on are comparing family within a family unit, and the unit of a plate is the most consistent thing from one experiment to the next, so it's key to keep those together.

Then a recurrent problem is we look to some large multi-center studies where data is collected, samples are collected at multiple sites, and then the multiple core labs are involved in genotyping. While it seems nice to spread the experiments around and be even-handed in terms of core labs getting their share of the genotyping, when you do that there's real problems that can be created.

What are they? Endless struggles with batch effects, high Type I error. In the genotype studies, when we've had multiple sites and not paying attention to plate and batch effects sometimes we've had to throw away 40, 50 or more percent of the genotypes, and particularly for copy number studies, where the log ratios are highly subject to these batch effects. They can be severely compromised to the point where the best we can do is perhaps just find really large copy variance, and if you look at the literature that's mostly the types of studies that have been published, where we're looking at massive deletions or gains. And we can say this from firsthand experience that the analysis of these studies takes many times as long when you're constantly having to work and rework, drop samples, redo your analysis, etc. because there's been problems in experimental design.

So what are some of the impacts we might see? This is the Wellcome Trust Study in the top and a GAIN bipolar study. Both datasets that have been publicly analyzed. If you do a principle components analysis of the log ratios on the right and the allele frequencies on the left, the allele frequencies, looking at the decomposition with Eigenstrat are the first two principle components in plotting the cases and controls overlaid. They look like they're drawn from the same distribution, and in fact they're this British population. This, however, in some of the smaller components you'll end up seeing differences between cases and controls that are sizable.

Now in the log ratios, when you do a principle components decomposition, you see very strongly that there's large differences between the case and controls in the two different colors, as well as there's big subgroups within the case, within controls, presumably corresponding to sites and batches. We've seen this in the GAIN bipolar study. We've seen this in 28 out of 30 studies we've looked at. This is the Illumina platform, Affymetrix platform, and while there haven't been too many large-scale array-CGH experiments, it's expected that these same kind of batch effects are a real issue.

Now as we look at CNV association, here's a real example from a real study that's kind of representative of the problems that you'll see when you're looking at copy number

variation. Now, the segment mean histogram I'm showing here is the mean of a copy number variant region that spans nine markers in an Affy 6.0 array, and there were four sites. In blue is a case site. In yellow is a large set of controls, and in purple and green are two additional control sites.

Now if we didn't have these two additional control sites and we looked at the cases in blue versus these controls in yellow we would conclude, ah, there's a big difference in copy number. There appear to be perhaps losses and gains in one versus the cases are sitting at a single unimodal distribution.

However, you see that the case distribution is the same as these two other control distributions from these different sites. And interestingly, this was a known copy variable region, but the batch effects themselves, the differences between these sites where the genotyping was run at different times under different experimental conditions created a very large and significant but spurious association, and it wasn't the only one.

We've seen this crop up again and again, and we've developed techniques, and perhaps you've heard some of our previous Webinars on correcting for batch effects. The challenge is often the copy variable regions, because they're more variable and they don't behave like the rest of the distribution, they're more problematic to correct. So really what we want to bring to you today is our learnings on what it takes to minimize the source of experimental variability.

So hopefully this problem is one that's very minor when you get to doing your genome-wide CNV studies, as well as it can't be understated how much batch effects really do impact SNP studies, and we'll show some pictures later of that as well.

So what are these sources of variability? To some degree we can try to minimize them, but we also have to understand what they are so that we can in our experimental design appropriately randomize and balance our plate designs, so that we can mitigate their impact.

Plate to plate variability by far has got the largest impact. So typically we're doing 96-well plates. We may allocate a few wells for experimental controls as opposed to case/controls. So we might have perhaps 90, 92, 93, 94 samples that we're having on a plate.

Now what's causing this variability from plate to plate? There's DNA amplification/hybridization procedures that in some cases can take many hours or even overnight runs, and perhaps the time that those steps are run, perhaps the reagents that are used or the quantities thereof end up that you have very large differences from one plate to the next in terms of the binding intensities that you get for the A and B alleles for genotyping arrays, and similarly you see in the array-CGH differences as well.

Environmental variability is a large factor, and I was looking back at the literature. Ozone has actually been known to have a big impact on hybridization. Early papers back

as far as 2003 have shown this, particularly in the context of array-CGH, but it's something to be concerned about for genome-wide SNP arrays as well.

One approach is to have environmental control, either putting an environmental hood, you can buy an ozone hood and put it around your experiment, as well as your HVAC systems can be rigged up to remove ozone. I've seen some day and night pictures from the same site, where they controlled for ozone and things like wave effects and all sorts of other aberrations seemed to clean up dramatically once you get rid of ozone.

The other biggie is DNA concentration. There's a nice paper by Diskin et al. in nucleic acid research cited below that shows that departing from the supplier recommended concentrations seems to be the one of the largest culprits in the so-called wave effect, where if you do a median smooth it's often easier to see of your log ratios. You see these waves. This occurs; we've seen it in Affy, Illumina, Agilent, NimbleGen platforms. If you don't get the concentration as close as possible to the supplier recommendation these waves are an inevitable result.

Now this paper shows a method for correcting the waves. We've seen in some cases the data is just irredeemable, despite computational approaches. So again, anything we can do to characterize our concentrations and keep them consistent and within the desired ranges of the supplier, of the DNA vendor, and they'll each have their own recommendations, you'll get the best possible results.

Now another thing, we've seen studies where three or four different DNA extraction kits were used and different sources were used. Use the same extraction kit. Use the same method. We've not seen as much data from studies where the capture was saliva or buccal samples, but it appears that they tend to be more problematic in general, and maybe that's just getting the concentrations right, getting enough quantity of it. Maybe it's also any kind of other material that comes along with extracting DNA from saliva.

We've seen one study where granulocytes were extracted from the whole blood. So you didn't have any T-cells. It turns out T-cells undergo rearrangements and you get certain artifacts in copy number studies. It tends to be experimentally difficult and expensive to extract the granulocytes, but if there's a way to do it and you can do it, the most consistent would be if you could just take DNA from granulocytes.

Cell lines are subject to artifacts, and often we're using cell line data whether it's from HapMap samples or elsewhere to use that as control samples for looking at plate to plate variability and looking at reproducibility. Unfortunately, even some of the HapMap samples, NA18540 here that's been sent out and run on Affy 500K and Affy 6.0 and Illumina, you see that there's five chromosomes here that have an extra copy.

So if we're going to be using a cell line as some sort of a reference, just characterize it very well before you use it or, ideally, characterize for yourselves a sample that's been acquired in the same methodology as the rest of your population, because we in fact saw a study recently where there were some 40-plus plates, some 1,500-plus samples, and a

cell line sample was used as a common reference for every plate, and it had wave effects on every single plate. It didn't have these type of artifacts and it made it very difficult to use it as some sort of an experimental baseline to compare, to try to characterize what's going on from plate to plate.

So to summarize our recommendations...it's don't do these bad things. Solve the problem at its source. There are things you can do if you've got a study where these problems have unavoidably crept in, and we'll talk about how to design the new ones so that they don't have these problems. But in summary, DNA extraction, try to extract at the same site, same source. Use the same extraction kit. There may be some evidence that how long you leave the blood on the counter before you freeze it could even cause some differences. Cases and controls, collect them at the same time. And cell lines, we've seen enough problems with them to not recommend them unless you have to.

With plating, we're going to talk in some depth on this later in the presentation, use of design of experiments principles to randomize your phenotypes, as well as experimental factors such as your site and so forth across plates. Place the same well characterized male and female sample on each plate. So we advocate using a number of your wells in your plate for a common pair of control samples, so that we can characterize the plate to plate variability and potentially be able to fix it.

Also, typically in most core labs they'll run duplicates within and across plates to assess variability. Many times those are run and we get back data from plates that only contain 89, 90, 92 samples and we wonder where the other four are. Make sure that you get back those control samples from your core lab, so that you can use it in your quality control procedures later.

It's essential to keep families together on plates and all this plating, don't leave it up to the core lab. If you just send all your samples they may get randomized, but they may not be randomized in a balanced fashion and you really have to take this under control. Some core labs are really being alerted to this and are doing a great job, but it's best for you to really pay great attention to the plating yourself and send things pre-plated, as well as make sure that you have a very consistent procedure for drawing the samples and putting them on the DNA chips.

It's interesting. Next time we'll be talking about some quality control procedures. We've been observing an interesting phenomenon in studies like the Affy 500K or perhaps where somebody's running multiple chips on the same samples. With the Affy 500K you've got NSP and STY chips, and we figured out a way to confirm; if the NSP and STY come from the same person there's some correlation approaches you can do. And we found that one to two percent of the samples in some of the studies we've looked at have actually been mismatched. So errors can occur in the actual putting the DNA onto the chip apparently.

So this is something that you should put the most perfectionist lab tech person on putting your samples on the plates because of those types of errors. Humans make errors about

one in a hundred steps of any procedure whatsoever, so anything you can do to try to have two sets of eyes on that procedure. Some of these studies, you think those errors correspond to \$15,000.00 in genotyping easily. So keep that in mind as well.

Now on the genotyping, we haven't had a chance to really verify that there are any differences in manufacturing lots, but it stands to reason that you would want to buy them all in a consistent lot. Similarly, with your reagents and so forth, run all your samples at the same site over a short period of time. One of the biggest sources of variability besides plate to plate is of course plate to plate across different sites, where you have different machines, people, operators, temperatures, etc. Control those environmental effects that we talked about.

Then you're going to need to think about redoing the bad samples. One thing that we've come across as we've been doing the CNV studies is the SNP quality control measures are not sufficient to assess CNV performance. They're necessary. Really low call rates for SNPs will lead to bad CNV results generally, but there's a number of measures that you want to keep in mind and we'll talk about it in a forthcoming Webinar, the derivative log ratio spread, which is measuring the pairwise delta between each set of adjacent points. That's a very good measure that can really find a lot of problems, but it's not sufficient.

We've found often looking at the outlier distribution is key. Wave effects is key. And we've been doing our best to figure out a way to use all these single sample quality control measures to find problems downstream in segmentation. We've found some of our customers have reported that all their SNP QC measures look great, and then they go to do the copy number and 20 or 30 percent of their samples have a large excess of copy number variance not consistent with biological reality, but consistent with problems of wave effects, outliers, etc. So it's just having clean QC for your SNP metrics is going to leave perhaps 10 or 20 percent of your samples being problematic when you're looking at your copy variation, and we've seen this for both Affy and Illumina platforms.

Now we're going to wind back a bit and talk about, okay, you're thinking about doing a study from the get-go. You're thinking about acquiring samples. What are the things you need to think about? Are you going to do a population or are you going to do a family-based study?

Now family-based studies have, on a large scale, most of the public datasets are not family-based studies. There's good reasons for that. It's expensive and difficult to capture parental and child genotypic information, consent, etc., especially for diseases that occur late in life like Alzheimer's. It's difficult to get parents of Alzheimer's people who are themselves 80 years old without digging up grandma.

The pros is we're seeing more and more of the rare variant hypothesis, particularly for CNV seems to be perhaps where some of the missing heritability is. So when you can look at the family structure you can understand a lot more about the rare variance, particularly de novo ones that are not in the parents but appear in the offspring.

With a family study you can look for both linkage and association and better detecting of genotype errors, and control matching is much more robust to population admixture, and of course environmental factors tend to be more consistent within a family than with a random population. It should be said also that extended pedigrees are the most powerful, particularly for finding rare variance as well as doing linkage studies, but they tend to be the hardest to obtain to get these multigenerational extended pedigrees.

The downside also is when you're capturing families and you're doing, say, a trio study, for the sample size you sort of have to have three individuals captured versus in a case/control study two individuals captured. So there's sort of a loss of power in that sense.

Now in population studies it's easy to capture them and that's what a lot of our genome-wide studies are. The statistical analysis by far is simpler. Family-based analysis, there's a whole host of challenges with analyzing that data, but there are some advantages to it. You do get more power. I was mentioning more power per sample when they are unrelated. So if you've got to spend \$1,000.00 at the clinic and with the genotyping for an individual, you get more bang for your buck with a population study, particularly for common variance.

Now it's more difficult to detect genotyping errors, although I wouldn't dwell too much on that, and there's of course increased risk of false positives from population stratification. Rare variance are more difficult to find because in a family you tend to assume that a particular rare disease is probably coming from the same rare deletion and you'll see it across all the affected members of a family, whereas knowing that there are many, many rare variants that could be independent causes for a disease you may not have the power in a population study to see the preponderance of a rare variant that you might in a large extended pedigree.

We're increasingly seeing some of our customers using hybrid studies, and they seem to bring the best of both worlds. You can basically take, say, a trio design of affected offspring and unaffected parents, and then add additional controls that are unrelated to those trios and do a case/control analysis on the offspring versus those controls, as well as do all the family-based studies. You tend to get power similar to population-based studies, and it seems that most of the pros and none of the cons other than perhaps just the complexity of doing the analysis, and you have to use methods in particular control for relatedness such as variance adjustment and so forth.

Another big question as you're designing a study is: how many samples should I pick? What should be my design? Should I go with trios, nuclear families, extended pedigrees and so forth?

Those are good questions and the answer is often it depends. It depends on disease prevalence. For instance, here's a nice graph that Nan Laird and Christoph Lange published in *Nature Reviews Genetics* showing, in this case, a trio design with a rare

disease can actually have more power versus in a common disease in blue here, a case/control study can have more power particularly for higher allele frequencies.

Interestingly for very rare allele frequencies, in either case the case/control studies do worse and the family-based designs do better. So if you're looking at rare diseases and very low, minor allele frequencies, there's a lot to be said for family-based studies.

Another way to think of power, often we don't always have an intuitive sense of power, is what happens to your test statistic as a function of sample size? So if we look at the log10 of the p-value on the Y axis here versus the number of cases and the number of controls in a case/control study, and we have two by two tables here where we're talking about the frequency under dominant model comparing the AA versus the minor allele, BB versus AB, at different fractions you can get very different results where doubling the sample size can actually double the exponent. You can go ten to minus two to ten to the minus fourth, down to where you have the lower odds two by two table, where the case/control fraction is much closer and the gain in statistical significance as you increase sample size from, say, 100 to 1,000 is very modest compared with these other two by two tables.

So when you're designing your experiment, and sometimes we just don't know what the underlying disease model is, in many cases we don't, so this gives you a bit of a sense that there's a lot of factors that matter in some of the other ones, things to think about, and we'll talk about simulations in a moment, is the mode of inheritance, penetrance, the effect size, allele frequency for the disease and the marker's allele frequency, the prevalence, heritability, sample size, and of course your study design. Is it case/control or a family-based study?

Now one useful piece of software that we use ourselves and many of our customers have is the PBAT package from Harvard, from Christoph Lange and his coworkers there. He's built a simulation module into his PBAT package and we've built a commercial package. There's also a free version that's frankly a little more difficult to use, where you can not only do family-based designs, but also case/control population designs and simulate all these different factors, allele frequencies, genetic attributable fraction and so forth and specify the family structure, how many offspring, what fraction of them are affected and so forth.

You can basically look at the power to find given effect sizes under different sample size assumptions and so forth. This is often required when you're justifying a grant, that you've done some sort of a power study and you can make some sort of statement about your chances of success of finding something in your genome-wide study.

So what they essentially do is Monte Carlo simulations with the various parameters set, and you get output which basically is power under different parameter assumptions. It's sort of beyond the scope of this Webinar to go into training on all the details of that, but if you're interested in checking out this functionality we have it in our software.

Now choice of platform is something that I'm not going to make recommendations about a particular vendor per se, but some of the considerations that you would think about in evaluating vendors and platforms, there's many platforms from different vendors for different purposes. Content and coverage is important. If at first just don't know what you're looking for, you want to have the best possible coverage of rare variants, common variants in CNV as well as SNPs across the genome. The X and Y chromosome are also of interest.

However, you may be doing a targeted study, looking at a candidate gene for instance, in which case you might be using a custom array, and then you'd also want to know what kind of coverage it has in your particular area of interest.

Often cost per sample is a big factor. You might want to get the Cadillac of genotyping arrays, but you can't afford to do as large of a study on that. So there are often tradeoffs in your content and coverage versus what you can afford to spend on a study.

Now as you know, the Illumina and Affymetrix platforms for genotyping have this dual use of looking at both SNPs and CNVs, but there's also platforms from NimbleGen and Agilent among perhaps others that are specifically developed for CNV interrogation. I believe there's also an Affymetrix array, a 2.7 million array coming out shortly that's designed specifically for CNVs.

So if you're mainly looking at CNV and you don't care so much about SNP, you might consider some of the array-CGH platforms. They do tend to have better signal to noise than arrays that are built for dual use. I've looked at many different studies and that seems to be pretty consistent.

There are things to consider when you're trying to understand the assay. How simple is it? If there are lots and lots of steps that go on those are sources of variability, and some of the newer platforms that are coming out are actually simplifying their assays and this is leading to much better quality results.

So you'll want to assess the consistent reproducibility of results. If you're considering doing a very large study you might ask for some sort of pilot project, where you run a couple plates replicating the same samples, looking at plate to plate variability, maybe doing it on a couple different machines if you're going to be running a bunch of machines in parallel, and look at how consistent and reproducible your results are. So array-CGH platforms are worthwhile considering if you're going to do a genome-wide study of CNV only. There's some very nice data we've seen coming out of both Agilent and NimbleGen platforms.

Signal to noise at a per-probe level is interesting. Sometimes the higher density arrays have less probes per marker. So interestingly, some of the older arrays like the 100K array, which none of us really use anymore for most purposes, actually has the best signal to noise because it's got 40 probes per SNP marker. However, the coverage would not be sufficient for looking at most things, particularly the genome-wide copy number.

A thing that often happens is people are interested in wanting to have the latest and greatest chip even though they've run 30 to 50 percent of their experiments on the last generation.

Imputation can only get you so far for SNPs, and for CNVs there really isn't any methods for imputation. So sometimes you have to live with using the same generation. It's best I'd say, frankly, to use the same chip for your entire experiment if at all possible. It's really challenging to try to merge those.

Software support is of interest. We thankfully support all the major platforms that we've mentioned. So I don't think it's a pretty big consideration these days. The software support is pretty good for most platforms.

Now I'd like to spend some time talking about plating and go into some depth on this. It'll still be a little bit of an overview. I'm not sure you'll be able to walk out of this Webinar knowing how to do this exactly, but this is something that we can assist you with as part of the services that our company does.

You've first got to select your phenotypes and experimental parameters that you want to distribute over your plates. So case/control is going to be key to randomize. If you've got quantitative traits you're going to have to think more carefully about that, perhaps thresholding them and so forth. Things like if you've acquired your DNA from multiple sites, thinking about how that might be randomized

Then we've got to just count how many plates we're going to need and define experimental units, and we'll show this pictorially in a minute, but an experiment unit would be like a case drawn from site number one with blood extraction method number two. Then you'll have maybe 200 of those and 300 of another experimental unit. Then we've got to calculate quotients and remainders and divide these things evenly over a plate.

So I'm going to give you an example here. This is actually a real study we designed with a customer. It has 4,000 samples, 2,000 cases, 2,000 controls. The things that we were most concerned about was the case/control status, the site, the DNA extraction method, because we've just seen that those have the biggest difference – cause the biggest issues.

Now there are about 12 or 13 other phenotypes that are of interest, but you can only randomize so many things. We picked these. Notice it's a nested design. In theory there could have been 24 different variations, two times four times three, but there's actually only 12. Some sites didn't use all extraction methods and so forth.

Now with 4,000 samples we were going to put them across – these are 96-well plates, but we're targeting using 90 wells per plate, because six other wells are reserved for various kinds of replication and those two male/female common controls that we want to put on every single plate. So you see here, for instance, Experimental Unit 1. There was 407

who were cases from site 1 with DNA extraction method 1. You divide that. It comes out to nine per plate with a remainder of two, and so two plates get one extra.

So what you're designing with these plates is not just random. We've seen some random designs, where you just randomly assign things. Randomness will create irregularities in the counts in the plates and actually create spurious associations. So you need to get as close to balanced as possible. You see here the remainder. The largest remainder was actually 44 here for experimental unit 4. So I think we started with experimental unit 4, distributed 18 per plate. Then we had to spread 19 at random. Then we went to the one with the next smallest remainder, and you'll see some experimental units; there were less experimental units than plates, so we either had one or zero. We shuffled multiple times in the assignment of these things so that we would get as close as possible to 90.

It was just a bunch of Excel wizardry to do this, and in the end you've got a number of plates that have 90, a number of plates that have 89, 88, 87. We'd actually recommend running the plates that have 90 first, and then you can use the plates that have these spares to do some reruns of the samples that are bad. So it's actually not a bad thing to have these remainders.

Now the next thing we want to do is what we do is randomly assign actual samples. So for a given experiment, we see we're going to pick nine experimental units on plate one. Which specific samples we pick is just done at random and we do a specific random assignment. Then what we want to do is not only check that there's no plate association with the three variables that we've randomized, but also things like gender and several other phenotypes that I've blinded here and so forth.

So what we did is we took our 40-some plates and we created a binary, call it a case/control status, if you will, whether or not something came from a particular plate. Then we compared case/control status versus that gender versus that across the 40-some plates. And this is a plot of the p-values. You notice there are a few closer to zero that we need to look into a little bit more, and so we zoom this in.

We saw just a handful that were less than .05 and actually, if you think about it, we did 43 plates multiplied by 15 different phenotypes. By chance alone you can expect some of these to be significantly associated, and these are pretty nominal associations. And if it wasn't a case though, we could randomize and verify that all these other phenotypes of interest are also not showing any association.

This step that I've shown here is the most important thing you can do in your study design process, this randomization. I'm going to show you a study that employed this type of procedure. It's not this one, actually another one that was actually done by a group that had done a lot of clinical trials and were very used to creating experimental designs with very careful design of experiments principles versus a Wellcome Trust study that there were some good reasons why they weren't able to randomize, perhaps, but if you're familiar with the case/control studies the Wellcome Trust did where there were

seven diseases and a common set of controls, Affy 500K, the controls were all run on one set of plates and all the cases were run on other sets of plates.

Then when you run associations on the SNPs and you plot the Q-Q plot, expected versus actual on a log₁₀ scale here for the p-values, and you don't do any quality control on the SNPs you see a huge inflation of Type I errors. Many of you have seen these plots again and again, sort of before SNP QC, after SNP QC, where the after looks a lot like this one on the right, where things follow the line. But this is not an after plot. This was a well randomized study using design of experiments principles, where we did not drop a single SNP, and it's actually a perfect Q-Q plot.

So the randomization really alleviates the spurious associations in these SNP studies that you see bad Q-Q plot after bad Q-Q plot and they drop 10, 15, 20, 30 percent of their SNPs to get a decent Q-Q plot, where the differences between the plates and because the plates weren't randomized are creating spurious and incorrect calls that are systematically biased by subgroups within the study.

So this was kind of amazing to me to see. There actually were two real associations at ridiculously significant levels, and these SNPs had very good quality control properties. Now you will end up dropping SNPs that have bad Hardy-Weinberg and so forth, but they're not going to cause a problem largely for your association study.

If you look at the same two studies, Wellcome Trust versus this customer study at the Manhattan plot, now part of that inflation of Type I error was due to this HLA region in chromosome 6, but also you see tons, literally thousands of genome-wide significant results. What the Wellcome Trust people did, they really tailored the CHIAMO method to try to mitigate this, but then they had I suppose post-docs and graduate students looking at literally hundreds of the most significant associations, and then they had to by eye look at the cluster plots, say, "Yeah, that was a problem with batch effects." So when you see their final beautiful Manhattan plots, by hand people were looking at cluster plots and that's a painful process, whereas here in this genome-wide plot, the genome-wide significant ones are way above the ten to the minus eighth axis here. Then we basically have Type I error totally under control, without filtering any SNPs whatsoever, not to say you won't, but this is what it should look like if you've done a good job of randomization.

Now the same two studies, we've looked at the copy number, log ratios, ran an association test on the case/control status versus the processed log ratios, and you see massive Type I error, a huge departure. It's genome-wide significant at ten to the minus 200, 300 across the whole genome, and whereas for this study it's much closer to the line.

For some reason, even in the well randomized studies we've looked at there still seems to be excess Type I error. We don't fully understand this phenomenon. Maybe it just is some slight non-randomness. I checked to see if it was gender associated like – what do you call it – where the segmental duplications, where you'd have perhaps a gender association in certain markers that have segmental duplications in X and in one of your

autosomes. Factor that out, it doesn't fix it. And when you do associations on the first couple principle components on the log ratios, it actually is associations across the whole genome. So it doesn't seem to be one particular region of the genome. There seems to be some systematic shifts that occur. Maybe the randomization wasn't perfect, but bottom line is you should see much closer to the expected values by doing randomization than if you do your cases and controls separately, which unfortunately is most of the studies we looked at.

Here is the same study looking at log ratio associations. Other than a T-cell artifact in chromosome 14, you just see all the signals totally overwhelmed by the batch effects, whereas it actually looks pretty nice in this randomized study.

So looking a little more closely at this "good" customer study, there are plate effects. So we haven't removed the fact that plate data differs from one plate to the next, and we've color coded by plate here. So all these different colors just represent different plates and you see orange here, pink here. There's these clusters of data points that behave similar to the rest of their neighbors on the plate.

However, when you look at the same first two principle components in a decomposition of the log ratio data, you see that the case and controls are well distributed in all these groups. Notice this orange group. There might be a little excess of – I forget which was which – the green, perhaps, cases in that. Maybe that's accounting for a slight departure from ideality on the Q-Q plots. It's hard to say. But the bottom line is these plate effects are killer unless you do this type of randomization that we've been talking about.

So I'd like to now open the floor for questions and answers, and hopefully this has been informative. There's obviously plenty more we could have talked about. One area that I think I've probably given short shrift to is the actual selection of the samples. For instance, how do you match up your cases and controls, draw them from similar distributions? Hopefully a lot of that stuff is common sense, but it's worth thinking about those things as well.

Please type your questions into the question and answer pane and I'll get to them. While you're doing that I'd like to make the offer to you that I'll be traveling in the next three weeks giving talks in Malaysia, China, and then it'll be IGES and the ASHG conference in Honolulu.

We're going to send out an e-mail shortly with just a little Web form or PDF form asking you a bit about your study, and if you'd like to spend 20 or 30 minutes with me just talking for free about your study, either one you've done already and what you're about to face in terms of problems and challenges and I'll give you some advice on that, or you're starting to think of doing a study for the first time and you'd like some pointers to some of the resources we've used and some guidance we'd be happy to do that.

It's also in light of self-interest that we've fought with so many studies that have had these problems, and we just don't want to have to fight with them any more. We'd love it

if every study from now on really employs the best possible design of experiments principles whether we analyze it or not. We just want to see the whole field do better. So please come see us at IGES and ASHG. We'll be at booth 512. That's two to the power of nine.

We anticipate doing a couple more webcasts in the months to come. The next one we hope to do is more in-depth on quality assurance for SNP and CNV studies. You may want to look at some of our past Webinars on our website. We've talked about these topics in some depth, but we intend to go in even greater depth as we've learned and keep learning more and more, doing studies on behalf of our customers and in collaboration with them.

So thank you so very much and we'll now go to questions and answers. Again, you can type it in the pane. If some of you have to leave early we understand. Sometimes these questions tend to go for some time, but we'll just go on until we've answered all of them, and thank you so much for attending.

Here is someone asking, "We specifically tested various DNA concentrations and we did not get any specific difference in genotyping results. Do you have any comments?"

This paper by Diskin et al., they showed some real differences in DNA concentrations. I've also looked at a study that characterized the DNA concentrations, and I didn't see a great correlation either. So I'm not so certain it's entirely DNA concentrations. It may have to do with amplification, poor quality of DNA to start, which leads to problems.

But the thing I didn't mention is what I can confirm very solidly from their paper is the wave effects very much do map to differences in the GC content. GC content is simply the number of Gs and Cs or the fraction of them for a given portion of the genome. And interestingly it's not just the probe itself. It's GC content, but the GC content of the megabase or so surrounding the given marker. So I don't think we've totally nailed down everything that there is to know about wave effects, and I appreciate that you've seen it's not just necessarily DNA concentration.

"What causes the wave effect?"

We've found correlation with GC content. There's been found correlation with DNA concentration, but perhaps not all the time. Is it some super coil DNA phenomenon? Is it the fact that it seems to not just be the probe intensity itself. I've actually done some regression tests, where you can take the GC content of the probe, the 25-mer or whatever from your array, and then take the GC content from the surrounding megabase and about half the variability is the probe itself, but half is kind of the environment around it. So I think this would be a very interesting research question to try to answer experimentally just what's driving that.

"Where can we carry out pre-study power calculations? Is there a free software or not?"

Well, the PBAT package from Christoph Lange, and it's on his website at Harvard, is free and will continue to be so. We've entered into an exclusive arrangement to commercialize the package and work with Christoph Lange to improve on it over time. So you can also license our commercial package for experimental design.

In terms of actually doing the power calculations, another thought about actual plate layout. There's probably some good design of experiment software. We just used Excel, where they can lay out the entire experiment, but I haven't had personal experience using them.

“Is it possible to get a copy of the presentation?”

We will be making a recording of this available shortly afterwards.

“If we can use the same experimental design for bacterial or virus genotyping association of some particular phenotype study.”

I'd say yes. Really, these are universal design of experiments principles. It seems very unfortunate that these principles have not been employed in most of the studies we've looked at to date. It's good to have a statistician who holds your feet to the fire on experimental design. I think perhaps the really high genotype calling rates kind of lulled us into a false sense of security that we didn't have to use design of experiments principles, but it's just a best practice to do whatever type of study you're doing.

A colleague of mine, who we've worked with some studies has said, “We definitely need to publish these findings.”

I'm so guilty of not publishing enough. I apologize. I've got a new statistical geneticist on staff who's taking on more and more of this work, and hopefully we'll be getting some more publications out.

“What do you mean by Type I error?”

A Type I error is a false positive. It's where something is associated when it really shouldn't be, as opposed to a Type 2 error where something is not associated when it really should be. So a Type I error is just all these excess of associations that are not biological reality, but are false positives.

“Was the customer study you were showing using the same Affy 500? If not, may it be due to platform difference.”

Well, as you know there are some early access arrays for the 500K. These were not that. These were both sort of standard Affy 500K, and they were actually both done probably about the same time a number of years ago. So it wasn't a matter of the 500K suddenly got a lot better. It actually was just the differences really were accounted for by the experimental design.

“What is linkage versus association studies?”

Linkage is really looking at the transmission of risk alleles through the family structure. They really go back 40 years. The earliest ways of looking at differences between individuals came from linkage studies, and you basically need family-based data to do it. It's probably a whole major topic. We've actually done a couple linkage studies using some publicly available software.

What you get out of a linkage study is broad peaks, where you kind of get a ballpark range of where you know something is going on genetically that's being inherited. It may be a SNP. It may be CNV. It may be segmental duplication. Who knows? But you don't get a specific locus. Then you would have to, after finding a big, big linkage peak, go and do a targeted genotyping, say, within omega base region or something that was highlighted by linkage.

So the PBAT package that Christoph Lange developed can do association in the presence of linkage. So you can do a SNP by SNP comparison looking at the family structure with a large collection, for instance, of trios or extended pedigrees. So there's where you could also do a linkage study with the same data, and it usually involves selecting a set of a few thousand tagging SNPs across the genome and running it in a package like Merlin.

For association studies, you can run them of course on the case/control and that's basically comparing. For a given set of case and controls you'd look at perhaps a two by two or a three by two table, comparing the fraction of case and controls in each allele group, and see if it's statistically different between the cases and controls.

“What is a wave effect?”

Again, in array-CGH this was first characterized back in maybe 2003 and we've seen it in pretty much every array type we've looked at, but we've seen studies that have no wave effects. If you apply some sort of a median smooth that's the easiest way to see a wave effect, and it's basically usually a GC content correlation that goes on, where your intensities rise and fall and it's not due to copy variation. It's due to some sort of anomaly of binding affinities and GC content.

“How would you manage a situation where you have multiple quantity to phenotypes and how would you randomize them across plates?”

A very good question. Typically what you have to do is discretize your quantitative traits. So you might divide them into high and low or maybe even up to quintiles, fifths. There may be a natural cutoff that you know above a certain threshold typically is that person is much worse off than another. And so much as we had like multiple sites had four levels, well, your quantitative trait you perhaps have four or five levels, and then you would define which experimental units are at which level, and then you would do your

best to randomize those across the plates. I think that would probably be the best approach.

The challenge is also if you have multiple ones. As we've shown you can – there's a certain point where if you have too many phenotypes that you try to randomize, you just start getting irregular counts for the proportions across the plates. So pick the ones that are most important to you, and you can probably do three, four. We've done maybe five phenotypic in site or other variables. If you're looking at several thousand people you can do a nice job of randomization.

“Does randomizing on plates help if you have upstream confounds? Example: if you have cases from one site, controls from another, possible differences in collection and DNA storage. To what degree does randomizing on plates help with this?”

Good question. If you've already confounded it upstream, what we're mainly doing with the plate randomization is we know that's one of the biggest sources of variability. If you happen to already acquire all your controls with one DNA extraction kit and the controls with the other, that wasn't a good idea and is something that I would advocate not doing. But at least we can do our best to deal with the plate to plate problem. And as I said, we do have methods to try to mitigate with data processing these batch effects. So all is not lost if you've already got all this DNA and these confounds have already happened.

I'm glad you brought it up. It's true that if you've already confounded things – so if you for some reason have to have multiple sites, what I would advocate is have equal number of cases and controls at each site or equal fractions of cases and controls at each site. So do your best to do that, and at the end of the day if you haven't run your genotyping I would consolidate and do it all at one place.

“Will wave effect be observed working with a small number of SNPs, i.e., less than 100?”

You may not see the wave, because your SNPs are perhaps dispersed over the genome. If they were very close physically, I guess what you would do is basically run a linear regression on GC content versus log ratio intensity, or even the raw intensities you can see this as well, right within the raw data before calculating log ratios. If there's a reasonable correlation, typically we see anywhere from an R squared of 15 to 30, .15 to .30. So you could perhaps observe that there's those effects.

“Do we get the power analysis software with Golden Helix? Do we pay for it separately?”

Actually the family-based module, the PBAT module is where this power analysis resides. So that's a separate and unfortunately most of the PBATs are family-based and there's this useful population-based simulation procedure in there, things we pay royalties to Harvard whenever we sell something that uses their functionality. So we pretty much have to put that functionality in the PBAT package.

“Are there any special considerations for CNV studies when looking at tumor DNA extracted from FFPEs?” So that’s formalin-fixed paraffin-embedded tissues.”

We’ve seen that the data quality from that can often be much more challenging. I haven’t really talked about paired analysis, paired designs. I assume you might be wanting to use array-CGH for that, in which case typically you’re taking a healthy tissue from the same person, ideally, and you put those together with these two color experiments, so you get a delta between the healthy tissue versus the tumor.

One thing that we’ve seen as well with tumor tissue, it’s not always pure tumor. So when you end up processing the data downstream you get a mixture of tumor and healthy. So models of copy number variation that assume discreet states may have problems versus segmentation approaches, which we use, that can basically handle mosaicism where you’ve got mixtures of different cells in there.

In terms of FFPE tissues, we had one person who was working with embryonic research where they’re taking very, very small quantities of DNA. If you don’t deal with it properly, getting good quality amplified DNA you do end up having real problems with data quality downstream.

I think it’d be best probably to talk with your particular vendor when you have special considerations like these. They may have special protocols for how to amplify the DNA in certain cases with small quantities and so forth.

“You alluded to the fact that randomizing based on experimental units can often overcome the lion’s share of Type I error. Did I understand this correctly?”

Yes, it really does, but with the caveat of the earlier question that if you’ve got confounders – or basically why this works is the plate to plate variability is the biggest source of variability, for all the reasons I mentioned earlier. So if you can deal with that, then that seems to be the lion’s share. How large some of the other factors are could probably be better assessed once we’ve done this randomization.

“Besides getting a hood, how can we minimize ozone and temperature effect?”

Surprisingly, a hood is not that expensive. When you think of the tens and hundreds of thousands spent, it’s just a plastic hood with some sort of a fan, vacuum kind of a unit on the thing. Well not a vacuum, but something that removes the ozone.

Agilent on their website has a nice PDF discussing HVAC, your heat and ventilation, air conditioning unit. So there are attachments you can put to basically remove it from the entire lab, and in some installations we’ve seen they double it up. They’ll put both the HVAC removing ozone as well as ozone hood. When you think of how much we spend on genotyping I think it’s a good expenditure to consider.

“Your points regarding randomization raise serious questions regarding the concept of using public dbGaP data as controls and also as cases, to what extent can the issues due to lack of randomization be overcome by census SNP and subject to QC procedures including GEM?”

I'm not sure I understand what GEM stands for, but let's put it this way. The copy number data is particularly challenging to merge these public datasets, and it almost becomes – I should back up – except in the case where you're looking for large rare variance, in which case you're fairly resilient to noise.

In the case of SNP studies, you perhaps read about all the SNP QC procedures people do from looking Hardy-Weinberg and looking at homozygosity and lots of sample procedures. Hardy-Weinberg, call rates, and there's a lot of formulas people use for QC. The one most important parameter in QC is call rate. And a lot of people use like 95 percent call rate. If you increase your call rate to about 99 percent and you use a package like CRLMM, where if you specify a threshold of confidence below which you drop SNPs, you end up throwing away in the bad studies, in the good studies not too many, but in the bad studies maybe 50 percent of your data, but then that 50 percent is pretty good.

Now I've seen, irregardless of almost any other factor, although we still filter on things like Hardy-Weinberg and controls and so forth. So if you really ratchet up the call rate requirement per SNP, then you can probably do a decent job at comparing between studies with SNPs.

With CNVs we've described in previous Webinars our principle components of procedure for removing these batch effects, which, by the way, if you have large extended families it is probably not an appropriate procedure, because family structure, you'll be messing with that. Not to say the situation is hopeless, but there are a lot of challenges.

“Can you explain more about derivative log ratio spread to control bad samples.”

I wish I'd prepared a slide on it, but if you look a derivative log ratio spread, if you just look at the variance of the standard deviation of log ratios, things like large copy gains and so forth will kind of distort the measure, whereas if you look at the pairwise comparison of each pair of points, excepting the transition between a gain and a loss, which there's not too many of those, the point to point difference is something that's a very good measure of the noise of the experiment.

So essentially you can look at a histogram, much like you could look at a histogram of call rates, and you'll see certain samples have very high derivative log ratio spread, and all it is is a standard deviation – I think this is divided by the square root of two for some reason – of those point to point differences, and the really high ones we've seen are also correlated with low call rates for SNPs, but not always, but they are very correlated with problems in copy number detection, where you find an excess on one hand or too few copy numbers. So if you have super-noisy data, a good copy number calling algorithm

will look at the noise and look at the signal, and basically not be able to see any signal within the noise. So a very high driven log ratio spread is a good reason to drop points.

“What is the best DNA extraction method?”

We tend to work on the analysis side and I can't say a particular vendor is the best. The key is just that it can differ from vendor to vendor, so pick a good reliable vendor and use the same one. I can't necessarily endorse any particular vendor, even if I wanted to. I just don't know enough and haven't run enough experiments on different extraction methods to really say what's best.

In part that's also due to the fact that all the other confounders that we look at in these studies it's hard to just isolate the contribution due to DNA extraction method.

“We have samples of both whole genome application and native DNA samples of one study. How many samples should we plate from the same samples to understand if these WG are comparable to native DNA samples?”

Well I think you could run a plate or maybe less, and then run a principle components procedure on the log ratios and see if within the same plate there's a difference. So the key would just be looking within the same plate, comparing your whole genome amplified versus native DNA. We could talk more if you'd like to when I get back from all my travels.

“You were mentioning about the PC analysis for population stratification. We've done PC for Singapore Chinese samples comparing with HapMap samples, remove the samples which fail the stratification, or even after this when we did a PC and the QC passed samples they still had issues of clustering.”

Once you remove some components there's always going to be some – well not always. What we find is there's kind of this distribution of these experimental differences, that it's not just like, “Well, I fixed three components because I got three populations and it all goes away.” So you may need to use additional components. Again, it's probably beyond the scope of this presentation to really talk about the correction of batch effects. It's something we hope to cover in our next Webinar, but yes, look at doing more components.

Also, you can selectively choose components. Sometimes if you're working with a family-based study you'll find certain components related to the family structure and you don't want to remove those, whereas other components are related to plates and batches.

“Do you have any thoughts on whether the standard normalization methods, quantile normalization for instance, used for SNP genotyping, which were also used in CNV methods, may be inadequate in contributing to the observed batch effects?”

A very good question. I did some analysis at one point a long time ago and I don't even have it anymore, where I didn't do any quantile normalization, and then I used the principle components batch effect correction method to remove the contributions. In a sense the quantile normalization could be distorting the tails of the distribution is one concern, but when I compared the results they looked very comparable to when we did quantile normalization. So we've been just using quantile normalization typically, but there is concern that those methods might be impacting it.

One thing I also tried was using the GC correction approach on waves at the raw data before doing quantile normalization. The Diskin paper does it on log ratios. You can actually try it on the intensities and then go through the quantile normalization and reference calculation procedures to get log ratios, and I got comparable results both ways.

So it could be a factor, but it's not clear that's the biggest challenge. I'm quite certain though that the quantile normalization is not creating batch effects. It's really the experimental factors.

“What in your assumption was the major cause of inflation in Wellcome Trust study because of not randomization of case control? Is that due to differential clustering for genotype or real difference in quality of SNP call?”

It appears to be the clustering difference. The clusters just shift on you between these different plates and batches. So these genotype calling algorithms that are assuming three nice clusters and then the real data has got big shifts between them, which is, by the way, one reason you want to use methods such like CRLMM and call it on all of the samples at once, so that the underlying cluster problems are reflected in a lower confidence in the calls, and as a result of that you'll drop the SNPs that actually have the cluster problems between these batches.

I know some people recommend calling a plate at a time. I think the evidence is getting to be pretty overwhelming that that's not a good way to do it.

“Is there any rule of thumb for selecting a number of principle components correct for batch effects and population stratification in genotyping analysis and whether it's the same for CNVs?”

The approach I've been using is use as many as it takes to get a Q-Q plot that is pretty close to the line $Y = X$. Also, if you're doing a family-based study, be more conservative, especially if the families were done on the same plates. You're going to be comparing within families anyhow, which will be within a same plate, so you're less worried about those plate effects and more worried about if you have extended family structure that a principle components analysis will distort the data.

“How do you calculate the sample size for perspective cross sectional study?”

I guess I'm going to have to pass on this one. I have to pass it on to one of my experimental design experts here, and maybe we can talk at another point to try to get that answered for you.

There's a question about the slide of examples of study designs, SNP, Q-Q plots, "When the cases and controls are run on separate plates, will SNP QC fix the problem?"

SNP QC, it will largely fix the problem, but then you will toss a lot of SNPs. But interestingly, I've used very stringent QC procedures on the Wellcome Trust study and another study where we had these real problems and were looking at the SNPs, and there were no spurious Type I errors to speak of. But then we did a haplotype analysis and it turned out there was some rare, on the rare side haplotypes that actually were due to plate effects that didn't show up until we'd actually run a haplotype association. So there was this ten to the minus tenth haplotype association that we found in the Wellcome Trust data that was not even visible in the SNP study. Then we looked at plotting, the cluster plots and so forth of the genotyping. We could see clearly that this haplotype came from a particular plate.

So if you're stringent enough on QC you can mostly fix problems, but it also does highlight. If we're interested in more complicated association studies like haplotypes, again, we're going to get confounding at another level when we use these genotypes.

There are so many questions I'm going to go to the questions where the people have still stayed on and apologize for those of you who couldn't hang around long enough.

"Could you please comment on effect of number of SNP content on the power calculation of your adjusted Type I error? Is it fixed? Example FDR.05."

The effect on the number of SNP content on the power calculation. I haven't looked too closely at that power calculation. There were a lot of parameters that went with it. So I apologize. I can't quite answer that one.

In general, more SNPs also create more multiple testing adjustments. So if you find something with a p-value of .01 with a single SNP, but then you've done that with 100,000 or a million tests, then you're needing a ten to the minus sixth or seventh or eighth p-value to have statistical significance.

"In terms of using principle components to look at batch effects in CNV analysis in the Golden Helix software, how many eigenvalues do we typically want to look at to identify outliers' structure within a sample set? I sometimes see up to eigenvalue ten or even into the 20s showing small clusters of structure. Do you have any thoughts on how to address these or their importance?"

One thing we've looked at is you look at these different eigenvalues and run association on the plate, whether or not it comes from a particular plate. We have gone into the 20s

and 30s in correcting batch effects for some of these studies that had problems with design of experiments procedures.

So what is driving it? I think in part some of it, too, is GC content that kind of varies as a function of parameters like perhaps concentration or perhaps the amplification or hybridization stages. So the problem is the variability is sort of a distribution of large effects, smaller, smaller, smaller, smaller effects. So it's challenging to pick when have we corrected enough, because if you do too many components you start distorting the true signal.

“You talked about allele frequencies for sample size and power calculations. What about the genotype frequencies? What is the minimum number of individuals for genotype in both groups to do the comparisons?”

In the model I was showing, in the two by two tables for instance, it was a dominant model. So you were looking a major allele versus a minor and the heterozygous. All these power calculations can be done on genotypes as well as alleles. So what's the minimum to do the comparisons? I hope I'm not distorting your question, but I might not be understanding it clearly enough. But the number of individuals is a parameter that you would test at different stops to see what the power is.

“If clinicians would be so careful with design then there'd be no need for us statisticians.”

Yeah, a nice joke. It really is a make-work project for the analysts dealing with all of these challenges of batch effects and so forth.

“Do design of experiment designs ever make it harder to do the layout? Will it increase human error?”

Well perhaps, I guess. One way to deal with human error that we found in the software development field is something called pair programming, where a person makes an error like one in a hundred, whatever, keystrokes, logical thinking steps, etc. So if you have somebody riding shotgun you get kind of a parallel reliability of two people with one in a hundred, and both paying attention you can get a one in ten thousand error. So it might be worthwhile thinking about having somebody riding shotgun, watching the lab technician to try to minimize error in all sorts of fields of human endeavor, where you have that kind of parallel reliability. It's shown that you can just get up to even six sigma levels of error with proper attentiveness on the part of all operators.

“Thanks for the Webinar, very informative. Your point about experiment of design is an excellent one. However, your Wellcome Trust example is a bit misleading as many significant SNPs in the Q-Q plot are indeed truly positive due to the strong 6p effect for Type I diabetes. This can be seen in the Manhattan plot subsequently shown. Can you recommend a book on experimental design useful for GWAS?”

I totally agree and I think I mentioned at that time that, indeed, some of that inflation is indeed due to chromosome 6. If you take it out it's still a huge amount of inflation, as you can see in the Manhattan plot with all those ten to the seventh, eighth and ninth p-values spread throughout the genome. But agreed, the chromosome 6 is a factor.

Can I recommend a book? Probably not so much a book specifically for GWAS, but probably the classic in experimental design is George Box's book. I think it's called *Design of Experiments*. But just a good general book on design of experiments would probably serve everybody in good stead.

“When we're doing data analysis, these SNPs will be analyzed together by looking at their SNP direction or will they be analyzed separately as individual SNPs?”

Well in our studies, mostly what we're looking at with Q-Q plot is just individual SNP, but surely we may be doing SNP/SNP interactions, as I had mentioned, with this haplotype analysis where we saw plate effects creeping up, where we were comparing multiple SNPs. I think it was three or four SNPs and a haplotype. Indeed those issues can come up.

“Does the PBAT software allow power calculations for family-based studies or can it address population-based studies as well? If it's only for family-based studies, can you recommend any other software?”

Indeed, the PBAT does actually do population simulation. It was something I actually didn't know until recently. So it's nice to know it's there.

“How can you determine something is a wave effect and not a true call?”

Well one phenomenon about a wave effect is there's a sort of a curvature to it versus a CNV call will tend to be much more of a discreet jump. So just visually looking at those is one way, which is why it's kind of painful to contemplate visually assessing every single CNV call.

“Using the same SNP genotype data in family design linkage association will provide differences in statistical power or the same power to detect association signal. Bi-allele or multiple allele may provide different statistical power.”

Yes. Indeed, when you're doing these combined studies, both the SNP – sometimes people look at micro satellites to the power with the family-based and the case/control could be different. I'm not sure exactly how you'd assess sort of a combined power, if we have any way to do that, but you could independently talk about if we were doing the case/control side of it and the family-based side of it, what would be the relative powers of the two studies.

“If a well designed trial fails to show an association with a phenotype, does this imply environmental effect, failure of the LD model, or what conclusion can be reached?”

Well as you know, we do come up empty many times in our genome-wide studies. It may be lack of power. It may be that the disease is environmental or it may be that it's yet undetermined genetic factors like epigenetic type things like methylation and so forth, or it could be interaction at the stasis that we just haven't located.

So it's a real challenge. Why haven't we been able to explain for very heritable conditions much of that heritability? There's a great paper that came out a little while back, where people have claimed that, "Well we just don't have a large enough sample size," or, "Phenotypes are too murky." Well someone looked at a big meta and mega-analysis of height, which is highly heritable, although there's probably some big environmental components, as well as it's a pretty clear cut phenotype. You can measure how tall somebody is and maybe adjust for age, and it's thought to be something like 80% heritable, and with all these studies of tens of thousands of individuals, only about three percent of the variability in height was explained with all these studies.

So I think probably the missing link has to do with the network interactions that go on. We're kind of these fault tolerant beings that a given SNP or a copy number variant, we may be fault tolerant to one or two of those or three of those, and it's a combination of many of them, and it's more going to be the biological network that we're going to have to understand somehow as to how we can throw out equilibrium versus thinking about individual SNPs, but we're going to need those individual SNPs to understand the biological network.

"What is the cutoff to select SNPs with lower minor allele frequencies in 2,000 controls and 2,000 cases?"

Well sometimes you want to have a different cutoff of call rate as a function of minor allele. Something like .01 is often what people use for minor allele cutoffs. I don't even care to usually cut things off due to minor allele frequency. One of the reasons people have removed things due to low minor allele frequency is they have, say, Fisher's exact test and the Chi squared test behave poorly, but if you have a Fisher's exact test you can do tests on low minor allele frequencies and they'll do just fine, but you do have to be aware that sometimes those rare alleles are due to genotype calling errors.

"Why do we need to do dominant tests for B allele?"

There's no reason we have to. We could have done additive or recessive, but that's just what was convenient to do some demonstration of the Chi squared p-values and how they vary with sample size.

"In our experience with Illumina 370 CNV arrays, reduction of LLRs in a region we know has copy number equals one. It's very small, maybe undetectable. Are the differences in ____ one-tenth greater on other platforms?"

This is looking at like a hemizygous deletion. We can detect them pretty well, and we've done it on Illumina 370, but we've also seen certain – a lot of it is at a given site. Maybe with a given DNA quality there's particular challenges.

One thing to bear in mind with Illumina data, if you use their HapMap as reference, which is the default, you actually get pretty bad results compared to if you have a large study and you recluster all SNPs. It's BeadStudio and their successor packages feature that you can recalculate the clusters and you get much better results than if you use the HapMap samples as reference, because those were run at a different site at a different time and so forth.

Someone asks, "How do we correlate between a SNP, CNV, methylation? Is this possible in your software?"

Our software can take general quantitative and SNP data. So you could import methylation data and CNV data, and even pour some of the batch effect correction procedures on it. We've actually had one customer who was doing an association on CNVs versus gene expression data and found an interesting finding that's hopefully going to be published soon. So yes, you can analyze that type of data. We haven't focused too much on specific things like clustering and so forth for, say, gene expression, but some customers have found very useful the association testing machinery in that context.

I feel that our questions are growing almost as fast as I can answer them. Maybe I'll just pick and choose a couple more and then call it a day. Thank you all so much for attending this Webinar. A lot of them are just thank yous and people are leaving. Let's see, maybe one last question and then we'll call it a day.

"GC content may influence intensity data, but how should it affect CNV analysis because you're looking at a ratio?"

A very good question. I've never been a fan of GC content correction approaches because you could argue, when you're doing a log ratio in population studies as opposed to, say, paired analysis or something, you'd figure that the same effect would affect both your control samples that uses reference for that particular probe. In other words, a probe is being impacted by a GC content, but it's in both numerator and denominator log ratio and it should cancel, and it largely does by the way, but it's the phenomenon of the GC content in the environment around it or something, where even though you take that ratio, the GC content based wave effect, and it's clearly a wave that correlates well where GC content can still be there.

Anyhow, I'd like to thank you all so much for your questions and attention. This was really a record Webinar for us with over 500 people attending, and we've still got 80 people listening after almost two hours or an hour and three quarters. We look forward to seeing you. Maybe I'll see some of you at the IGES or ASHG conference. I'll be giving

a talk at IGES and we'll be giving kind of a private presentation to anyone who'd like to attend, talking about our software and so forth somewhere during the ASHG conference.

I look forward to anyone taking us up on our offer to just have a 20 or 30 minute consultation with us on a current study or analyzing or a future study you're contemplating. I'll freely give of my advice. If you just send some particulars of what you're doing to our attention, we'll probably have somebody call you to schedule something with me and just inquire a little bit more about your study, and we could get to know each other better, and hopefully we could learn from you as well as hopefully this presentation has been educational delivering information from out side.

So again, thank you very much and look forward to an announcement of our next Webinar and I hope you can all attend. Take care now. Bye-bye.

[End of Audio]