
Christophe Lambert: Looks like we still have got a few people coming in. I know it was a bit of a trip to get from the conference center to find your way here. Thank you all for coming. Well, good afternoon everybody. I'm Christophe Lambert, the President and CEO of Golden Helix.

And I'd like to give a presentation today and achieving Genome-Wide success, from SNP studies to CNVs to expression, quantity and trait loci. And hopefully, I'll be able to convey today some of what we've learned over the past three years in doing Genome wide studies of all sorts for both SNP, CNV and the just beginning to see the dawning of people running association tests between expression data and genome-wide SNP and Genome-wide CNV studies.

So, I'll just give a few words about our company and then I'd like to dive right into the science. For those of you who were at my IGUS talk a couple days ago at Turtle Bay. By the way, if you don't know about IGUS, it's a great conference of, generally it's field with statistical geneticist who talk about methods. I think next year it will be in Boston. So, I spoke there briefly for about 15 minutes. There will be some overlap for those of you who've been there.

And I gave a webinar a couple weeks ago that some of you attended. There will be some overlap, but there should be something new for everybody, hopefully. And if there is nothing new, as, me some questions about some things that you're pacing and maybe we'll come up with some new discussion.

So, I'll talk a bit about study design just because I find it so necessary to stress it. And then talk about quality control of our samples. Various approaches that we've taken to plating to really minimize the problems that we've seen in some of our studies such as the many problems we see with batch effects. We'll talk about addressing those issues and then show though we can address batch effects to some degree, there is still some association pitfalls in the world of CNV that we have to address.

And we'll talk a bit about rare variance and I'd like to cover at the end just a little discussion of a paper that one of our customers published on expression QTL. So, really, the way that we've been able to learn all that we've learned has been being able to work with a great number of collaboration partners, strategic partners over the years.

We've been in business since 1998, which is about 11 years now and I've gotten a lot of gray hairs over those years as we've learned the ropes of genetic association studies.

The customers we've got are world wide. They're academic, commercial, biotech, government I believe we are in six continents. Haven't got Antarctica yet, but really at the end of the day what matters to us is what matters to you, making sure that you are generating significant findings, publishing and here's a snapshot of some of the journals by and which our customers have published studies using our software.

So, getting right into the science, study design is something that one would think you would take it for granted that you'd think about your study design up front and be worried about issues and experimental error and so forth, and yet, we found in about 28 out of the 30 some studies that we've looked at have had experimental design problems of one sort or another. And these are studies that you know, leading institutions have done, whether it's Harvard, the Broad Institute, or some biotech or pharmaceutical study, there's the problems of cases and controls not being balanced or randomized across plates.

Borrowing controls, while it's a nice cost savings has dire consequences down stream. Some of the worst things that can happen in family based studies is splitting up your families on different plates. We've looked at one study where the fathers are on one set of plates, the mother is on another set, and the child on a third set. We've also seen studies like the Framingham study where there were multiple generations collected in multiple phases over a number of years and those different phases were genotyped at different times and on different plates.

And when you have all those sources of variability, they lead to painful struggles with batch effects. We know because 28 out of 30 times that's what we end up spending our time fighting. And we're hoping communicating some of our thoughts on experimental design that the next wave of GWA studies will hopefully have mitigated a lot of the problems.

So the problems generally are high type of error. When you are doing genome type studies, sometimes with multiple sites, you are throwing as much as 50 percent of the genotypes away to get a good Q-Q plot. Not to say that that's a recommendation, but sometimes that's, it gets as bad as that. Copy number studies are

the words affected by problems of experimental design and inevitably your statistical analysis is spent beating on batch effects.

There is plenty of seats up front, feel free to come forward, or maybe you just want to be close to the door in case I get boring, and so I'll try not to be boring. The biggest source of experimental variability it seems, in our experience is plate to plate variability. And that's where you're running a 96 well plate typically whether it's affy or lumina, both platforms have these problem in spades. It seems to be due to, probably, my guess is the amplification, the hybridization process, perhaps differences in reagents, environmental variability is something we particularly have seen in a [inaudible] experiments. There's actually some papers, well characterizing this.

In fact, there was a signature genomic came out with a paper recently talking about their, what they've done to mitigate ozone problems in [inaudible]. They originally discovered their ozone problem, their lab was next to a train station and data quality varied with train schedule. And I guess these locomotives produce a lot of ozone and you could see severe degradation of signal quality.

Some people have asked me, well, is the problem as large for say affy or lumina chips. I don't know. I don't know if anyone has ever measured it, but it's pretty cheap to try that. Just get an ozone generator and turn it on and run a chip and then turn it off and air out the room and see if there's a difference.

So, there's a paper by Diskin et al, they produced a way to measure both wave effects and they quantified the fact that these baseline waves that you see in lot ratios, this is copy number data. And these are the intensities that also make up your genotype calls, can be dramatically impacted by DNA concentration and keeping as close to the supplier recommended DNA concentration seems to be the best. Although, both in a study we've looked at as well as, an anecdotal report from one of our customers suggests that other issues, perhaps just DNA quality itself that are independent of concentrations seem to also lead to problems with wave effects. And then Cell line artifacts are another problem where you can have implications of entire chromosomes due to the mortalization process.

And we've often seen people like to use say half map cell lines as a reference samples from plate to plate for comparisons of data quality from one perspective it's good. They are well

characterized. People have done even next generation sequencing studies and it will be great to have super well characterized samples of references, however the differences you often see between different blood sources, whether it's the extraction kit or the quality of quantity of DNA that comes out of saliva or whole blood, they just can differ.

And so if you make those difference correlate with your case control status like, oh, we're going to take a bunch of cell line controls and then whole blood cases, and you run your experiments, you will see dramatic association effects that are not biological in terms of mechanism, but rather are in terms of the experimental variability.

And so, the recommendations that we make are don't count on statistical methods to fix your data after the fact. Let's try to solve it up front, so with regard to DNA and DNA extraction. You want, if possible, you want to do it at the same site, the same source. Someone was asking me, well, I'm doing children and I have to do saliva. It's not going to be acceptable to do whole blood extraction, in that case I said, well, and it was a family study, well in that case, we'll do your parents also with saliva and make sure to use the same extraction kit.

I was talking to one of the vendors of DNA extraction kits and one of the things to look for if you for some reason have to do extraction from different sources is hopefully use the same vendor and hopefully they have a different kit for the different types of DNA that can get as consistent as possible results.

There is some evidence showing that there is obviously things like DNA degradation that can occur, so if you got samples that were in the freezer 15 years that were your cases versus new controls, that can cause differences. And so, another thing we have seen is cell, we see there is a difference between whole blood and, which has T cells and B cells in it and just extracted granulosa sites where there is no T cells. Turns out there is T cell rearrangements that happen that can give you spurious associations and I'll show that a little bit later when we talk about quality control.

Now, probably the most important part and only two studies again, actually did this, used good design and experiment principles to randomize and in a balanced way, randomized a phenotype across plates. And I'll talk more about that. We advocate placing a well characterized sample of even better two, a male and a female on every plate from the same DNA source and also running duplicates

within and across plates to assess variability. Of course, as I said, keep families together. If for some reason they have such extended pedigrees that they go over multiple plates, then at least keep your nuclear families together and don't leave it up to the core lab to decide, sometimes they do the randomization but often their choice for duplicates across plates is half map samples because they've characterized them so many times that maybe you have to do a bit of that. But just take control, take charge of experimental design yourself otherwise think about it, we can spend a millions dollars plus on a large study and just with a little bit of statistical work up front, we by missing doing that, we just have analysis headings down stream.

So, this is a message I've given many times and we're beginning to see some nice studies coming out that are doing this and it makes all the difference in the world. Things like buying your arrays in a single manufacturing lot might not be obvious, but it can make a difference, including things like reagents so if you are getting reagents from multiple vendors over time. Multiple lots, those are always variation that can happen.

As I'll talk about later, when you're rewriting your samples, the ones that don't pass quality control typically the genotype centers are highly focused on genotyping. Of course, the same arrays can interrogate CNV, but as we'll discuss later, the quality control metrics that you assess genotype quality can pass with flying colors and you can have atrocious CNV quality for certain arrays, so there's certain metrics that find specifically necessary to assess for CNV to rerun a sample or that the sample might have data quality problems.

So just to give you an example of designs I would not do versus would be like the Wellcome Trust study. We had seven case data sets and two batches of common controls. Here I just had one of the common control batches compared against, I think this was type I diabetes, yeah, type I diabetes genome wide association test and you see there is this inflation of associations. And here we did not do any SNP quality control. So you see, paper after paper that says well, here is the bad Q-Q plot, before QC and then we filtered for Hardy Weinberg and we filtered for low call rate and then you see Q-Q plus that looked like this.

Well, this was a Q-Q plot of a study that was properly randomized with a balanced design. We did not drop a single SNP and you get a perfect Q-Q plot. So there still is experimental variability, but it's not correlated with a phenotype so you, it doesn't mess up your

association studies. There is actually two big P values here that are real and so that's nice to see when you actually have genome wide significant findings. When you look at the Manhattan plot, part of the inflation in that Wellcome Trust study is real associations in the HLA region, but also there is this huge excess of association, 9^{th} minus 8^{th} , 9^{th} , 10^{th} . All over the place that when you look at the beautiful plots in the nature paper, they had to by hand, take all the significant findings and look at the cluster plots and literally assessed hundreds of cluster plots manually by eye. Now, they would say, dropped every bad SNP whose P value was say better than 10^{-10} and minus 5^{th} or 6^{th} or whatever cut off the used, but there's still going to be spurious association in those even more nominal significant regions.

And I found, actually ran an interesting multi marker association test based on [inaudible] on the Wellcome Trust data and there's regions that are not significant at all with single SNPs that when you do multiple SNPs 10 of the minutes 10 with genome wide significant finding with you know, three or four markers and then you go look and you find it's a plate effect or a batch effect where rare genotype errors can kind of go together in an apparent [inaudible] and it's just genotyping errors.

So, while we can do some degree get away with, you know, merging different studies, we have to do extremely stringent QC criteria and we end up throwing away a lot more genetic information than if we had done a well randomized study which you see the Manhattan plot for this GWAS, the same one that had the perfect QQ plot, the genome line findings are way up on the top of the screen above the top of the screen, but you see there is occasional 10^{-10} to the minus 4^{th} or 5^{th} P values, but type one errors really under control. And this is, again, with no filtering for call rate, Hardy Weinberg equilibrium.

There was QC on the samples though that they were all high quality samples. So, if you think there's problems with SNPs and you say, but I can address it with quality control with the CNV associations, if you take the intensities, the log ratios and do association tests on your case controlled phenotype, the inflation is ridiculous. It's 10^{-10} to the minus 200^{th} , 300^{th} P values all over the place, whereas when you do it a well randomized design, again, the Q-Q plot is much closer to the line $Y=X$.

We still have seen in a couple well randomized studies we've looked at, there seems to be some inflation that we don't fully understand. It may just be some residual lack of perfect

randomness. I'm not sure that's the case though because when we, it's not any particular locus that seems to be contributing to this inflation. If you do a principle components analysis on the log ratios on this well randomized study and do an association test of the first couple components across the genome, it's significant everywhere. So, it seems to be still some systematic variation can occur even when you randomize well, but certainly when you look at these log ratio association here type I error is pretty much under control and you are not seeing these ridiculous 10^{-200} P values.

Now, when you actually start distorting and making copy number calls, your type one error will reduce, but what you see is just as the case can be with genotyping, it's worse with CNVs where you have to use multiple markers. The distorting can be biased again by the shifts that occur plate to plate that, and I was taking to some people at Broad Institute saying, "Well, what do you do with batch effects?" And they said, well you got to randomize and they said they regressed the heck out of all these batch effects and that will not work if you haven't randomized your phenotypes because if your phenotypes are just perfectly correlated with your major sites or batches, you'll basically subtract out both the batch effect and whatever case control difference there is.

So, here's a plot of the principle components, decomposition of the log ratio covariance matrix plotted against the first two eigenvalues and you see in color coded by plate the different samples, so indeed in this well randomized study, the plates are very different, they are shifted in their intensity. This is my contention that plate effects are the biggest issue. However, we color code by case control status, you see that each of the subgroups that are the plates are, have balanced quantities of case and controls.

Looking at this orange cluster in the lower right, it looks like there might be a bit of an imbalance, perhaps that is explaining some of the inflation. So, and sometimes those imbalances occur because you have to drop a lot of add samples from certain plates, so you do your best to get a good balance design to start, but then still there could be some residual bias.

So, on experimental design, I just thought I would walk through how we could go about doing it. And this is an actual study we helped one of our customer design where this is a nested design. There's 12 experimental units. Term on the unit would be something like all the samples who are say cases from site one DNA extraction method one and this study there was 407 of them

and you see some experimental units, there was only say 28 who use extraction method, three site one and controls. Now, we'd advocate not using multiple DNA extraction methods and randomization on plates is not going to remove effects due to DNA extraction methods, but at least we can do a good job at randomizing our case control status over the plates as well as mitigate any kind of site and extraction method bias' towards them also being on the plate, because potentially, we could regress on DNA extraction method and site if they're not also correlated with plate, but if we have two sources that are confounded together, along with confounding our case control, it's going to be more difficult to regress on differences due to extraction method or study.

And so, what you do and we just do this in Excel. There is probably software they can kind of lay out all this experimental design stuff for you, but the gist of it is, if you have 407 experimental units, divided by, in this case, 45 plates, we had 4000 samples, we'd have 9 per plate with a remainder of 2. And what you basically have to do is allocate the reminders kind of in a random fashion and in a balanced fashion to add up to, in this case, we targeted having 90 samples per plate, holding back another six for various quality control procedures.

This was run actually, it's going to be run at CIDR and they have four that they reserved for their quality metrics and we reserved two for repeating a male and female on every plate. So, notice it's not a pure random design to have imbalances, so you're actually making sure that your imbalances between any one group are at worse by one sample and then to verify that we've done a good job, we typically run association tests. I thought I had that slide here, but I don't.

You run association tests per plate on each of your phenotypes and not just the three your randomized for, but any other phenotypic data that you might have. Is it limited to how many sub groups you can do and still get a good design, and so, it's worthwhile just to make sure there's not severe imbalances in your plate regard to other phenotypes that might be of interest. Some examples of those would be things like maybe you had like a case control status, but you also had a age of onset and other quantitative or sample based criteria.

So, after you've run your studies, quality control is where you spend most of your time. It's not exciting in some sense, you spend 80, 90 percent of your time there messing with the data, but

it's where your time's best spent. And that's where we've striven to create tools that really make that a lot faster and easier than writing a bunch of scripts and also enabling ways to visualize it. So all the visuals you see here are easily creatable in our software. So, as I said earlier, we found you need both SNP-based QC and CNV based QC measures whether you're doing a SNP study or a CNV study. Maybe you are going to study CNV of your G loss six months from now, but if you don't look at these CNV based metrics, you are not going to be running the samples that are problematic and then you'll be pulling your hair and trying to run samples six months later. Just think of all the things that could have changed in terms of __, reagents, experimental site that you might use and so forth.

So in SNP based QC the standard metrics that we're all familiar with like sample call rate and heterozygosity, but there's some others that they may not have thought of that we'll talk about in a moment, as well as various CNV based QC metrics. So, for SNP-based QC, we've used, of course, different calling algorithms sometimes, this was [inaudible] and birdseed and depending on your thresholds for confidence pit with C realm you could have higher or lower sample call rates. So in this case, people often reported C realm often gives higher call rates than birdseed, yet I picked a strict threshold for confidences so on average the call rate is a little lower.

I actually found if you do a consensus call, where mismatches are dropped and then if you've got missing for one and the call for the other, you put them in, you can actually increase your call rate. And so with this consensus set of calls, we got up to 98 percent on this particular study. A lot of the samples actually had some problems.

And X heterozygosity, actually a customer wrote a little script for this that we have available at our website. You just get the fraction of heterozygotes in the X chromosome and you see the males here who are homozygotes and then the females who have more heterozygotes but then there is this tale of samples that have this funky middle of the road heterozygosity rate that typically are associated with some sort of quality program.

We've come up with a new method that we get to publish if anyone's interested in using it, let me know. If you're still doing Affy 500K data and there's plenty of it out there, both historical and some people are doing new studies. If you do a correlation between SNPs that are genetically near by on the marker map, you

know, within, you take the closest pair of SNPs that are no further than 2500 base pairs apart between NSP and STY and you do a correlation of the allele count, 012, just a numerical correlation. If you for some reason, for sample handling you'll have two sample, NSP and STY chip that you'll think are from the same person and it turns out through mishandling, one's from Mary and ones from Mark. Well, if it's from Mary and Mark, X heterozygosity will show that they are a different gender, but if it's you know, Mark and Peter, then they'll have the same heterozygosity rate, but this correlation metric will show that they are not from the same person.

And so in a number of 500 case studies we've looked at, anywhere from a little over 1 percent of the samples are mismatched. And then if you start joining that data together doing SNP and copy number associations, you know, you are really putting all sorts of noise into the equation. And if you think about 29 samples out 800, it doesn't seem like a lot, but it could be over \$10,000 of money down the drain of samples. So, another thing to think about is your trying to improve the quality of your experiments would be some sort of process of redundancy, whereby two sets of eyes would be following the process. You know, human beings make about an error 1 in 100 times going repetitive tests, but if you have two independent sets of eyes, it could be parallel the liability of 1 in 10,000.

Cryptic relatedness is important to look at. I use Plink for that and you know, we integrate with, it becomes an export to Plink formats. And you can see certain samples that are, there's a PI_HAT statistic and some people criticized the F identity by descent method that they've got. But for quality control it works quite well.

You can detect identical samples that have been run. And we had no twins in this study that we knew of. In fact, the age of these people were 10 or 15 years apart, so clearly they weren't twins, and yet they had the same genotype, so again, sample mishandling can occur when you have PI_HATs around the 0.5, those are like first degree relatives, siblings, or parents and offspring and then you get into cousins and so forth.

Sometimes certainly while you want to be removing these duplicates, it's good to do a relatedness graph to see if there is any kind of extended family relatedness that could really confound results. And you know, you perhaps could drop first degree

relatives and those might be judgment calls depending on your study.

I don't talk too much about family based studies, but we do a fair number of them these days. Both for SNP and CNV and looking at homozygosity. As you know, family based studies are perhaps one of the more powerful ways to look for rare variance. When you, again, look at relatedness, we've drawn in these black bordered squares, individuals that are supposed to be part of the same family and we see occasional individuals who clearly belong to a different family.

Sometimes you also see streaks where there seems to be a correlation between certain samples and a bunch of other samples and that often indicates some sort of contamination. Now, with population stratification, one thing we like to do is download the Hapmap3 Affy data when we look at Affy studies from the Hapmap3 website. And then just append those to your genotypes and even if you are doing a 500K or you can, there's enough markers that are in common that you can just join the matching ones and do a principle components analysis, ala Eigenstrat and plot the first two or three components versus one another.

In blue you see this was one of our studies. And down here is where the Seth population is, as well as, an Italian population. The nice thing about Hapmap3 is it's got a bunch of other populations sub groups of in Africa and Mexico and so forth, so there's a group of samples here that actually had been previously unknown ethnicity and it's quite close to a Mexican group, but not quite, so you can not only, potentially exclude outliers, but have a sense of just what ethnic group they're from.

Now, when we get to CNV based QC one basic measure that is useful is the derivative log ratio spread, which is simply the standard deviation of each pair of adjacent points divided by some square root of two normalization factor. And when we looked at the call rate versus that log ratio spread, you do see in general that low call rate SNPs also have high standard deviation for the pairwise differences. And but another important quality problem is an excess number of copy number segments found with whatever calling algorithm you use.

And that is also a lot of the high, there is a lot of derivative log ratio spreads are connected with excess, but derivative log ratio spread is not sufficient to capture them all. And so, we explored a number of other metrics to try to figure out just what kind of

measurements can we do on samples to figure out what we're going to have down stream copy number calls.

Wave effects are often contributors to excess copy number calls. These waves look like gains and losses and various calling algorithms really can't necessarily tease that out. It looks like a real signal difference. There's a paper I mentioned earlier by Diskin and Al, they've incorporated a tool into PennCNV called genomic wave. That's a PERL script. Beware, PennCNV doesn't work on Linux you need to use red hat or certain other distributions. So you can run this tool and get total garbage if you run on the wrong Linux distribution. But we found their measure of waviness is predictive of excesses or lack of copy number calls. They're algorithm to fix the data though, while visually you see the waves get knocked down, the signal appears to get knocked down as well, so you get lower wave effects, but then lower signal separation and when you look at before correction and then after correction for waves on the same study, we've kind of lost or multimodal distribution of different copy number segment averages.

Here's neutral, loss of one, gain of one, loss of two, gain of two pulse calls. So, again counting on that you can just use a statistically tool and correct for wave effects, don't count on it. Better to try to get the right high quality DNA at the right concentration and do everything you can to prevent having those problems in the first place.

Now, when we did some regression models on some different quality scores versus the number of segments found, we could explain some of the variability by variable odd ratio spreads, some by the wave effect, but the extra measure that we added that boosted the R square from like, I don't know, 40 percent to something like 80 percent was looking at the extreme value distribution of log ratios. And so, you look at say the top and bottom one percent, maybe even the fraction of those or the delta of those to see if there's a skewedness and we see that we really capture, say below a threshold of, in this, depending on your data and single to noise, but this was a minus one, maybe minus 0.75, really seemed to capture all of the samples that had high segment counts.

And similarly with the upper threshold and lower threshold we were able to capture these high segment counts. Sometimes it's just kind of capturing the fact that the data's got a high variance which would have been captured by that derivative log ratio spread

measure. Here's, notice this is the same scale of plus minus of about two units, a bad sample and a good sample. Notice the high variation as well as clear wave effects going on that would just create a nightmare for making copy number calls. But you can also see skewed extremes where in here, if you look at the top one percent, it's here up at the 1.5, versus it's less than minus one on the other side and here it's below minus two skewed downwards versus something around plus 0.75. So when you see that big skew, these are not copy number gains and losses. This is just probably a really bad sample that there's either a quality problem with the DNA or with the chip or somewhere in the experimental process.

When you're looking at the, we often like to look at the sex chromosomes, X and Y as a quality control metric with Affy 500K here's the NSP versus STY, here's some of those mismatches of where I told you of where the X intensity and the, for NSP is low, of course, bonding to males, but STY corresponding to females, and so this was a sample handling error where two different gendered people were labeled as being the same person for NSP and STY. It was not the case. Now, here's is a case where the agree between the NSP and STY and it's probably a good experiment in what's going on in some X [inaudible] where for a certain fraction of the cells in the sample this happens. This is probably a female who just had less than two full copies of X in a certain fraction of her cells.

And also when you look at X versus Y in this study, you see XXX over here and here's XXY so there can be these outliers with regard to excess copies of presumably inactivated X or Y chromosomes and they can be candidates for exclusion and almost in every study we look at, when you do the reported gender versus the imputed gender, you see these occasional mismatches where this has got to be female and yet, either you know, there was a recording error or perhaps the person was transgendered or something and decided to report which gender they'd like to be.

So for cell line artifacts, that often shows up where you take the mean of every chromosome and you see certain chromosomes have a real high mean, sometimes real low means will be deletions. Obviously, having five chromosomes with an extra copy would not be consistent with a living human being and yet this was one of the hapmap samples that we used per all the time, you know, the 270 hapmap samples in a 18540. And so, you wouldn't want to be using that as part of your normalization procedure with those excess number of chromosomes.

And so again, on segment over abundance, or under abundance, you do sometimes see sometimes very few segments found that usually corresponds with an extremely noisy sample which is usually captured by a derivative log ratio spread metric. If you look at an individual sample you would hope to see something like a triangle distribution or maybe better if the signal is really clean you could see other humps where as if you don't see that, there's usually been a problem with somewhere with the data quality. And our goal, hopefully, is with these various metrics we could exclude these samples before the full downstream analysis of doing the copy number segmentation.

So then how do you put all these SNP and copy number metrics together? Often as you know people pick cut offs and someone says, well, let's pick .99 and then everybody who reads the first publication says, "Well, I do .99 ever after." Well, in some studies, your genotype call rates are lower and you know you kind of have to look at the distribution and some of these thresholds, frankly are a little subjective.

You know, you can calculate some tails of the distribution and do two sigma or something like that, but the approach I would like to do is pick reasonable cut offs, often visual for your different metrics and look for metrics that, count the fraction of the metrics that a sample fails for. Usually you'll have multiple metrics failing for these bad samples and then there will be boarder line samples that might fail one or two metrics and you'll have to look more closely at them. But in general, by looking in a multi dimensional sense of five or six or ten different measures of quality, you get a much better separation between the really good samples and the samples that clearly have under performed.

Per SNP QC is something that you know, the sort of the standards people like to use in of call rates from anywhere from 0.95 to 0.99. You know, genome wide significant departments from [inaudible] controls. We're all quite familiar with that. I would just add to that that when you have a multi site study with real problems with batch effects, if you use a call rate of 0.95, you get still a lot of spurious associations. I found in almost every study that I've looked at, when the per SNP call rate, if I crank it up to 0.99 pretty much remove most of the type one error. The same time, I might drop 40 percent of my SNPs if it's a multi site study.

The single site study usually, I might drop 15 percent, you know, give or take, depending on the quality of the data. So, this

parameter is the most important and in a sense, if you really ratchet up insisting high per SNP call rate, almost all the other metrics see to, they're not superfluous, but there will still be some departures from Hardyweineberg and so forth, but you pretty much get rid of type one error. And you verify this, of course with your Q-Q plots. If you want, if you are worried, oh, I'm throwing away to many SNPs with this type criteria you might also consider, well, okay, so what if I'm throwing away 50 percent of the SNPs, I'm going to do amputation in another million and a half SNPs, but then I know I'll be doing amputation on really high quality SNPs, so you know, I often will try doing no QC and then plot the quality control metrics as a separate track, and then any interesting findings you can then follow them up, or at least like QC, so it doesn't have to be with visual software you can just look at the data, and then pick appropriate thresholds.

So as I alluded to with the Wellcome Trust data, you can have HLA related regions that inflate type one error. So, here's before QC in the first plot and then here's after QC, but then if you remove the HLA region chromosome 6, you get a clean Q-Q plot with just one genome wide significant finding. So, if you see an inflation it might just be due to a highly large number of markers associated in a given region.

Now, with CNV, as I said, the batch effects are the worst, 'cause you're looking at intensity data and if you look at the Wellcome Trust case controls versus one another in a genotype sense, using PCA you see their drawn from the same population and there's a nice overlap. If you were to look into some of the smaller components, you would start seeing some of those batch effects, but in the CNV, the visit two components are totally related to batch effects where there's large shifts between the case and controls and of course, between case and controls. So, when you see that big of a difference, that leads to those huge Q-Q plot problems.

And so the approach, given most of our studies had these problems, we had to come up with a solution and we've taken the PCA based approach, kind of analogous to Eingstrat where the idea is the CNV shifts in intensity will be correlated over thousands of markers and so, you can take the first few components to extract the differences between these subgroups and then having factored out that difference, you get a corrected CNV data that you can then do associations on directly, so you can then go on to do copy number calls with.

You've heard me speak, perhaps in the past about different methods, we tend to, rather than using screen plots or the Eigen values typically look at Q-Q plots so you do it with 5, 10, 15, 20, 25, components until you see that the bulk of the distribution is very close to the line, $Y=X$ and we exclude sex chromosomes from PCA analysis.

Now, before correction you see this ratty green plot, this was really sample. It should probably be excluded, but you can correct for a lot of this. Turns out, there was a couple hundred samples from one batch that had similar distortions, because they were correlated, we were able to pick it up in the first, you know, we had to use a larger number of components for this study and you do see still there are some regions that have some clear gains and losses, whereas the rest of this was just bad data. Here's the Wellcome Trust data log ratio associations before and after PCA.

And so then you'll see that there's still some associations here and there, but you really knock down this noise where there is associations of 10 to the minus 200 everywhere. Now, I'd like to point out some limitations of PCA. And what we've been trying to do to address them. I don't think PCA is the be all and end all to batch effect correction. It does assume modestly large sample sizes, you aren't really going to do so well with 50 samples of PCA.

And it does assume your first few principle components contain undesirable differences. If you are doing large extended pedigrees, this won't be true. You could have family structure showing up in your first few components. And of course, the gender differences between X and Y will pop up in your first few principle components.

So what we do is do components in the first 22 chromosomes and then apply them to all chromosomes. The problem that disturbed me the most is that small batch effects can still remain uncorrected in the copy number regions. So, if you think about it, the correlation of co-variance structure, you're looking at of all these markers is going to be more heavily weighted towards neutral markers because that's the bulk of the distribution and we do correct the intensities in the copy regions as well with this process.

But there is just certain regions that for whatever reason will not be corrected as much and they can show up as associated regions. So, after the fact that you have to look closely at any association and make sure it's not driven by some sort of a batch effect. So, we've

been looking to address this by different methods and it's still ongoing work. Correcting after centering by marker and/or sample means.

With the software, you don't have to pick the first few largest ones, you can take your components and do associations with the family structure or associations and see which components are associated with what and exclude components you don't want to remove. And something I would like to experiment some more with would be calculating components on subsets of markers perhaps say just a copy variable markers and at the end of the day, I keep coming back to, if we could just do better design of experiments at the front end, most of this is not such a problem and I wish, you know, I wish all of the studies we have seen thus far didn't have so many experimental design problems.

One thing to beware of, particularly in the Illumina data, but it's true for other array's as well, the, you can have Illumina data that's especially large negative outliers that maybe real, homozygous deletions have very large negative values. And when you do copy number calling algorithms, because you can have most of your data is at plus minus, you know in a band of say plus minus .5 and then you add to it a minus six outlier, which could, again be a deletion. If you tend to skew your distribution to the left here, so we've adjusted our segmentation algorithm and to any customers who've done segmentation in the past we used to recommend doing ten marker minimum.

Well, often a side effect of that is nine other markers that weren't part of the deletion would come along with that huge negative outlier and you'd get this big negative skewed distribution. And so, what we do is, when we do our segment break point detection, we allow up to single marker break points and we have an outlier removal method where you basically delete the cut points, but not the data point itself and so your permutation testing for comparing adjacent points, is just done on regions that are not single marker outliers.

And we also tend to pick a point 001 as our permutation significance. And now, this is something we haven't implemented directly yet in our software, but I've done it in a cytogenetic setting is the outliers if you just Winsorize them and you are only trying to detect minus one or loss neutral gain Winsorizing is basically you find the upper and lower say 1 percent or .001 percent and then any data point above that you set it to the value of that 1 percent, and 99 percent percentile. And we found working with

some cytogenetic experts who have interpreted the spectrum that that seems to be have the best, deliver the best most believable copy number segments in the array CGH setting when we've looked at array CGH data extensively.

So, we've played with median smoothing your data is a very nice way to visualize it, but as a way to remove outliers, it creates all sorts of false structure that messes up calling algorithms so it's not a good way to go. I thought I would mention a bit about next generation sequencing. As you've heard some of the presenters mention, they sort of fall right now with the short [inaudible] of next generation sequencing is large deletions can be missed, so I think there is still plenty of time and place for the genome wide arrays to be used for finding cytogenetic size and several 10s of [inaudible] sized deletions and gains.

What I've plotted here on the first track, the Bentley paper that came out a little while ago, it was an Illumina next generation sequencing of NA18507s. So it was a hapmap sample. I think it was [inaudible]. And so you can, because there is public data for affy 500K, affy 6.0, and Illumina 1M. We compared all three of those platforms, found the deletion, but the next generation sequencing missed it. And it misses a lot of the large deletions. And interestingly, one of the small deletions found in the Bentley paper, when there actually was some coverage corresponded to a single marker in the affy or Illumina high density arrays.

So a lot of times these single marker outlier are real. And here's a histogram of all of the markers that are found in Bentley called dilations and you see clear skew of the distribution towards being negative, although, there's, as we see, single market probes can still totally be off which is why you don't want to believe a copy number call, necessarily of a single prob. It was a downer for me. I wanted finally a gold standard. Of copy number variations that I could compare all the methods and arrays with.

And I thought, okay, great, finally next generation sequencing, they've done it on the hapmap data and all the big things were being missed, so there's some room to go. And that was with 40 full coverage, I think, too.

So finally I would like to talk about association and wrap up with any questions and answers you might have as well as we'll have our drawing for our Ipod. I've talked about this at length in various presentations, beware the ubiquitous significant T-cell regions.

This is a plot of the Wellcome Trust. If you, associations in chromosome 14. There's a region that's associated in six out of the seven diseases. And what it appears to be the problem is, there is a difference in the quantity of T cells and B cells when you, when the DNA was collected and of course, the cases and controls were collected separately and there are chromosomal rearrangements that occurred in various, usually when you look it up in the genome browser, say the UCSC browser, you'll find an association chromosome 7 or 14, there's a region, chromosome 22. It will be AB parts, antibody parts and if you find association there, it's usually not specific to your disease. It's more some sort of a data collection thing.

On the other hand, some studies that are not well randomized don't show this up either. But we've seen a lot of our studies, especially chromosome 7 and 14. Chromosome 22 is a region that's significant in some of the Wellcome Trust studies, but not others. The strongest thing will be in type one diabetes and it makes you wonder, maybe there is some T cell thing that could be relevant. Now, of course, what's special about the type one diabetes study is that it was young people because it's early age of onset and so, you know maybe there's some property there, we saw some associations on the age of the individuals with T cell regions that it was hard to tell if it was just differences in population collections or real age effect.

So, here's one of the things that bothers me the most about batch effects was this was a region we found in a customer study where it was nine marker, copy number variant. It was in a known copy variable region. The data was precleaned. You could see loss neutral gain, when you averaged the markers in the region. And when you looked at the cases in blue, versus these controls in yellow, there's this clear difference in it looks like losses and a few neutrals and a bunch of gains. Pretty strong signal. However, there was two other sites that contributed controls. That these controls, varied distribution looks like these cases. So, if we didn't have the advantage of these two extra sites we would have been rushing to do quantitative PCR and fine replication and so forth.

But looking at this and then we saw this in three or four other areas of the genome, the same pattern and it was the same batch that had the difference. Now we were wondering, is it some artifact of all of our data processing, our principle components correction, we went back to the original raw data off the array and looked at the just raw intensity distribution, and there was dramatic differences

between the different sites and so it's just the problems you have from site to site, and I know we like to be politically correct and spread the genotyping love by genotyping in multiple centers and collected all together and it's great for science that we're pooling all this data, but it's not great for science that we're having all these experimental differences that are really clouding our ability to find something.

So, another, you know, we saw this great finding in one of our studies where there was an association at the end of chromosome 1, genome wide significant between cases and controls got all excited, looked at it and turns out there's site mental duplications that occur in this case it was in the Y chromosome where this disease was something like 3:1 more common in males. And so, really, all this was differentiating between genders.

Because the male had Y and both the male and female have chromosome one, we were basically differentiating between males and females looking at the intensity of this region that has the same sequence both on chromosome 1 and chromosome Y. And so you have to really drill into your data to identify that those things are not a problem.

So, finally, well, almost finally, rare variance, there's a new feature coming out in our software soon, Rudy's showing some nice pictures in our booth, actually, better ones than this of that T cell region that I mentioned in chromosome 22. He's got some slides on the Wellcome Trust data where you can plot all the copy number calls across a whole genome study, and of course, that's too much data to look at, but you do get a sense, look there's streaks of blue and red everywhere of common variants, as well as when you zoom in you see the occasional, it doesn't look very large, but it's over a megabase of a gain, you know, the rare variance, so you can sort by case controls status, look for large gains and losses that span megabase, and you see as well as you notice here at the beginning of chromosome 15 a very common and large copy variable region.

Actually, this overlaps the region in that schizophrenia study, that 15Q11.2. I think that the schizophrenia region extends out there and this was not a schizophrenia study. And so there are indeed in the general population, large gains and losses at the front of chromosome 15. so I guess the caution I'd have with rushing to publish rare variant is get some additional public data sets as well as your data set, if you've got a small sample size and look at the regions where you find rare variants to make sure it's not just a

chance occurrence because saying I found it in three cases, zero controls, you know, it's certainly nothing like genome wide significant, on the other hand, have we figured out yet, how do you adjust for the fact that well, the size of a copy number gain or loss, what is the distribution of those sizes and the distribution of them in cases versus controls.

I'm not a big fan of genome wide burden of copy number variants, saying, well, there's more gains or losses in cases than controls because if you go down to a single base pair and ten base pair [inaudible] there's a couple hundred thousand per person. And it's got to be pretty much the same, so where's that magic threshold of size where we somehow are claiming there is a larger burden, not to say there is nothing to it, but I would also be concerned about how batch effects could skew our counts of copy number calls between case controls.

I'm supposed to be giving a talk shortly at a psychiatric genomics conference. We have some preliminary data where we found that you know, a rare variant in three female Alzheimer's cases, zero controls in a large GWAS study. It was interesting, they were kind of large ones that overlapped right in the [inaudible] gene so they extended, some cases a couple megabases, but the region in common was a Thyroid hormone and there was some previous publication, just kind of clinical data showing thyroid hormone levels associated with Alzheimer's in females.

I got a colleague to try to replicate in one of their studies, didn't find anything, anyone doing Alzheimer's here, and would like to look at the region, maybe we could talk afterwards. So these rare variants are there, how much of the missing variabilities are going to be explained by them, I don't know.

Finally, there's a study that came out, Yan Sun published this recently in cardiovascular genetics. A common copy number variation chromosome 6 associated with gene expression level, endothelium 1 and transformed B lymphocytes from three racial groups. And turned out he just took public hapmap data, gene expression data and the raw cell file data and processed it using our software and you could see, it was 115 marker copy number variable region that there was homozygous deletion, hemizygous deletion and copy neutral.

And you saw it in all three racial groups and there was this association with this region in hypertension gene expression. And what was interesting was it was true, the trend held and was

significant in the Asian, Caucasian, and African groups. And then when you combined all three he got nine times 10^{-9} the minus nine P value of the logistic regression like log ratio tests. So, if we're looking for new avenues to explore copy number association, you know, even in this small study of expression versus copy number, it appears that at least some nice findings to be found looking at expression quantitative trait loci. Now, he actually picked, I think just seven or so candidate genes, so he wasn't doing an all pairs you know, 10,000 expression versus GWAS and that's something that we hope to be supporting in the near future where you can do that kind of all pairs analysis.

Problem there is how much multiple testing. In this case, there was not that much multiple testing at all. So I would like to thank you for your rapt attention and your attending and I would like to open it up to some questions and answers and then we'll have a drawing for the free iPod to one of the lucky winners in this room, we hope. Thank you very much.

Yes. I often don't use it, if I have a really stringent call rate, but then I'll clock Hardywineburg as a track. And if I do use it, I do it in controls only. We have a function sort of just in controls, but looking at the literature, replicated genome wide findings have rather, not these massive 10^{-200} departures from equilibrium, but 10^{-5} happens. And you know, that's just in part because if you look at both case controls combined, you've biased your sample towards cases and if there's disease causing [inaudible] cases, it's likely to be potentially out of equilibrium.

Other questions. Yes. So are those meta-analysis driven by batch effects? Usually the prudent authors are you know, looking very carefully at their data, verifying that their cluster plots are good and in some cases genotyping with some low through put, high quality you know, PCR based method or something, so I'm not going to cast dispersions on any of them, but I think it could be dangerous if you're not careful, one of the other dangers that really you run into with meta analysis is things like strand flips and so on between different platforms. You know, I was trying to, you know the plot I showed of the principle components analysis, against different Hapmap populations.

The Illumina data has been made public, but only after it's old and processed with the genotype calls from the Brod Institute and they have like swapped around the strands and so on, so that they don't correspond with the calls you get right out of the Illumina

software, so you know, it's a pain to try and get that all to be adjusted if you are just going to straight join up your Illumina study with that study, so no, I think the large signals are part of function of sample size and you know invariably a meta analysis doesn't have the benefit of trying to adjust for batch effects, but if each separate analysis had done a very careful randomization between cases and controls you know, it makes meta analysis easier too because you can really control for, you know, if you have five sites you can have a covariant that can potentially adjust for that, but if each site has 40 plates that are all contributing to a confounding, you know, you don't want to put 40 dummy variables in some sort of digression to adjust for 40 plate effects.

So, thank you. Other questions? Yes. Yeah, could effect, there could be some sort of boundaries where you cut offs could go this way or that way. I usually just start with sample QC first and then, but you could argue as well, do I want to include potentially bad SNPs in say a principle component analysis of population structure, maybe not. You know so, in practice, it doesn't seem to make a large difference because I use the whole auto zone, like I'll use half a million or a million SNPs in the principle component analysis and it ends up always being population structure driving the first couple components and not batch effect or errors in genotyping, so that would be the one you might worry about. And as I mentioned early, in general, I like to do all these, I think I showed about a dozen different sample QC metrics and once I'm sure that I've excluded all those samples that failed that then I go on to do the QC metrics, but if you did it in a different order, you could have some differences in there.

Right, well, usually the problems are see are pretty gross ones that, also, I've done a number of studies, like for instance when I showed you that plot of the NSP-STY mismatches, there's another study where I had to drop 20 or 30 samples. You got more significant after dropping the 30 samples than having them in there, so I'm not a fan of throwing away tons and tons of samples, some people would disagree with me, but I often don't like, I think population stratifications a little overblown, unless you have two very different populations and on these occasional outliers the various papers I've seen on it suggest that it doesn't make such a huge difference, and even if it does make a difference you can always run Eigenstrat, correcting for it. So I tend to not like to drop samples because their ethnicity departs, but you probably want to be particularly careful with anything like rare variance, obviously where you can have rare variant for a particular ethnic group.

And those who spend a lot of time and know there are pitfalls with population structure. I saw a new pitfall today, or the other day at IGUS, Ralph McGinnis from this, I think he said the Sanders Center or the Wellcome Trust, he spoke on how you can find certain samples where just a few of their chromosomes will have say, a Caucasian sample will have African ancestry haplotypes that they just had some distant ancestor and that could show up, obviously, potentially as some sort of a rare variant or rare haplotype or rare CNV so even when you think about using [inaudible] when you're assuming that you are correcting for the large scale ethnic differences, when he showed a sample that for all intents and purposes looked purely Caucasian fell right in the middle of the cluster, but had some distant African ancestry on some pretty long haplotypes on the chromosomes, so I hope that helps.

So, does the person have to be here to win Andy? You think so. We had some people fill it in elsewhere, didn't we? All right, well, any other questions? Thank you. Well, no we do not, but we have some interfaces to export to eagle format and mach. We worked with some mach interface and Josh probably knows better than me, he was doing the project management work on that.

And it was just so slow that it wasn't very feasible. I saw a nice poster out there, I forget who showed it that eagle seems to have the best non performance, but in terms of calling it also is fast, so, we don't have a direct, here's, enter the parameters, send it out, fetch it back, but will probably be able to build something down the road, but for now, there is a way, I believe, to export to that format. And then import your inputted SNPs and then do the rest of your analysis.

Anyone else? Well, we've got you holding on for the big prize drawing. Why don't we bring up the basket, collection basket. How many iPods? Just one lucky winner. All right. Congratulations. Well, again, thanks everyone for coming. And if you're interested in more information, we've got a number of recorded webinars on our website. Try out software and just call us up and we can chat about your study if you are doing a design or if you've got problems with batch effects and want to know how to fix them after the fact, we might help with that too. Take care now.

[End of Audio]