

Achieving Genome-Wide Success in SNP and CNV Studies

Golden Helix Webcast

February 18, 2009

Dr. Christophe Lambert

Alright, well we're right at the top of the hour and I'd like to welcome you warmly to our presentation today. I'm Christophe Lambert, the president and CEO of Golden Helix.

Today, we'll be talking about achieving genome-wide significance in SNP and CNV studies. This is the first in what we intend to be a series of webcasts. This will be a presentation that will go into some depth, but we intend to go into greater depth in a series of more detailed webinars on many of the topics that we're going to be talking about today as well as potentially additional ones.

I'd love to hear your feedback about this webinar and what you'd like to see in addition to learn about with regards to finding associations in SNP and CNV studies.

Just a couple words on our company. We've been around for over ten years now. We really focus on genetic association software, services, diagnostic development. A lot of what you see today is really the fruits of our ability to work with our customers as well as our collaborators. Many of them are listed here, in particular, where we've done a lot of recent services doing genome-wide copy number and SNP studies and really the success of our many customers listed here among – which is just a portion of them – is really in their ability to get results published.

So that's really – the purpose of this lecture series is conveying to you what we learned about genome-wide SNP and copy number variation studies, SNP associations studies, in general. It doesn't have to be genome-wide. It's really with the hopes that you'll be able to do what we've done and go beyond what we've done in your quest for finding significant results.

So today's topics are kind of what – we'll alternate between SNP and CNV portions of the talk, first talking about quality control in the context of SNPs and then CNVs and then association testing in the context of SNPs and then CNVs.

We'll talk a little bit about some events, regression techniques, and then what happens when you do finally a study that falls short of genome-wide significance, can you still find some useful information in it and really capitalize on that investment you made in one of these expensive studies.

So "Genotype Quality Control in the Presence of Batch Effects" is the first topic. And really, the batch effects are the operative word. If you've got a beautiful set of genotype where the experiments were all done at the same place, a very minor variation, these issues are less prominent. But we've had to deal recently with a lot of studies where the case- controls came from different sites or run at different times, different experimental conditions, and do you really get good quality genotypes in the presence of differences from plate-to-plate and batch-to-batch, site-to-site.

The problem, of course, is that because of these sources of variability, you have systematic errors that confound association testing and the genotype calls, although we wish they weren't, for a lot of the markers, they are really subject to these plate and batch effects, which I'll show shortly.

Some of the existing approaches that are out there, and we're certainly going to tap into them, there's different calling algorithms. A lot of these are in the context of, say, Affymetrix arrays. But we'll be talking about sort of a general case.

Achieving Genome-Wide Success in SNP and CNV Studies

Golden Helix Webcast

February 18, 2009

Dr. Christophe Lambert

BLRMM, Birdseed, and CRLMM are some of the calling algorithms that are currently in vogue. Often, even after you use a good calling algorithm like one of these, you deal with SNP filtering approaches, filtering for Hardy Weinberg, minor allele frequency, call rate, association on batch. Cluster plots are certainly an important of quality control at a SNP level. Then at a sample level, you may want to exclude sample that just have problems either due to some population structure, poor call rate, heterozygosity, cryptic relatedness, or just a generally bad sample.

So in thinking towards what would be an ideal solution, we want really high accuracy calls, good per call accuracy estimates, have our calls insensitive to batching. As we'll see today, none of these methods really get us there. So we're really making do with the best that we have at the moment. Be able to identify the problematic samples and exclude them.

Now, in terms of calling algorithms, the one we've been using based on both a paper by Lynn, et al., recently, showing BLRMM and Birdseed. Their performances exceeded by CRLMM for a couple reasons. One, it seems to be less sensitive to plate/batch effects. But as we'll see in a moment, it's really not immune to that.

What was really nice, when I was reading Lynn's paper, was they were showing the correlation in the confidence rate. So when you have a given call, like "AB," for a given sample, they'll give you a numerical value, like .96, that says their confidence in that call. That's highly the actual error when they look at HapMap samples with known calls in the predicted confidence, actually, is highly correlated.

So that gives us a really good way to filter for bad calls. They tend to have lower drop rates and better accuracy, as shown in this plot here.

Also, this use of CLRMM's corroborated by a couple presentations, Nancy Cox, who works in a group of genotype quality control within the GAIN consortium, she presented a comparison to those three approaches, and CLRMM is really the one that seemed to have the best operational characteristics. This is for AFFY 500K, AFFY 6.0, AFFY 5.0.

The most important point, though, I make when you're doing genotype calling is run all of your samples in a single CLRMM run as opposed to, for memory considerations, for instance, running all your cases put together or a subset of your samples together and then another subset. You get real problems. I'll show that in a moment.

We were doing a study and we actually ran the samples both ways. We ran CLRMM of all our samples in one batch and then we ran them – this study had come from multiple sites and we ran each site in a separate CLRMM run.

Now, when we – when we ran it separately, the confidence is actually that CLRMM estimated were higher than if we put them all together. But it turns out it's sort of a false estimate of confidence. Low minor allele frequencies aren't well represented in some of the batches.

As you can see in this lower plot, when we called our calls by batches, in blue here, it turned out these came from the cases. They were called "AA" when clearly they looked heterozygous in the cluster plot here. This is actually the raw cluster plot. When we put them all together in a single batch, the calling cleaned up nicely.

Achieving Genome-Wide Success in SNP and CNV Studies

Golden Helix Webcast

February 18, 2009

Dr. Christophe Lambert

Now, in other cases, you will get funny offsets, say, between one batch to the next, and then that will be reflected in the lower confidence in CLRMM's confidence estimate. Then that SNP would potentially be excluded downstream in your filtering criteria.

Now, what then is the approach we use? So this would not necessarily be a one size fits all approach, but in the cases where we've had multiple batches and real differences between batches from different sites, this has been our approach. We run all our samples at once. And beware, you're going to have probably a lot of hassle installing CLRMM and you want a 64-bit Linux box with lots of memory to be able to run any more than the – I think with about 1,200 samples you can do in 16 gigabytes with AFFY 6.0. But it really goes up from there.

Drop the calls. You have a missing value put in for a given call if the confidence is less than 95 percent. Then once you put in those missing values, assess, per SNP, the call rates and drop those less than 99 percent confidence. Dial that up and down 98 percent, 98.5 might be acceptable. It depends on the quality of your data.

It turns out this is the most important parameter. It kind of, of course, interacts a bit with the previous one. But if you just stop at this point with this rather stringent filter criteria, I found even if you do some of the other standard things of dropping SNPs that are – have large departures from Hardy Weinberg or a low minor allele frequency, or even looking at batch associations, pretty much, once you put this filter in place of the call rate, you really don't get many spurious associations in these extra filters are, in some sense, optional.

Of course, you want to look at Q-Q plots, verify they're well behaved. And after the fact, when you found significant associations, examine your cluster plots and make sure that there's not some weirdness going on like I showed on the previous.

Some other thoughts on this subject. You can go with less stringent criteria. I'm not saying this is the holy grail of SNP filtering. But for instance, if you look at the Wellcome Trust study where they looked at seven common diseases. In each of those studies, they had about 100 highly significant results that when you looked at the cluster plots, it basically turned out to be problems. There's some like 700 of those across all their different studies.

So if you're worried you're filtering out too much, you might have a less stringent approach, but then you'll be having to look at these cluster plots one at a time and make sure that they're looking okay.

What we find too is, with our software, you can keep all your SNPs in the a project and turn on various filters and off various filters to basically examine if what-if scenarios if you looked at different criteria.

Now, with this filtration criteria, minor allele frequency, alone, tends to drop 150,000 markers, but those would probably not be significant 'cause you don't have much power with low minor allele frequency and you may end up throwing out a couple hundred thousand others with an AFFY 6.0 array, leaving maybe 500,000 markers. The point is those will be highly confident and you can potentially amplify those out with imputation methods to another couple million more.

So it's kind of – there's trade-offs involved in these filtration approaches. And as with anything, just document what you do in your publication and make sure your cluster plots are good for the significant findings you find.

Sample quality control is kind of another issue. In general, we drop samples that have a low average confidence across all the SNPs for a given sample. You can basically average the CLRMM confidences. You can also look at the call rate. Call rate's not necessarily equal to confidence either. So what we tend to do is examine the distribution. If you have 99 percent confidence on average, but then there's one sample that 97 percent, you may drop it. On the other hand, if you have a little quality calls, your tails might be – what you drop might have a lower confidence.

So it is, of course, sort of situational depending on the quality of your experiments.

Another thing, looking at population stratification, if you merge your data set with HapMap genotype calls and you use Eigenstrat, you have your Asian in green, your Yoruba in blue, and your Caucasian in orange here. Then we projected – so this is along the first two principal components – we then plot the samples for the given study and we see they're supposedly expected to be Caucasian, but there's a few leaning towards African American probably and there's even one dot here up in the Asian.

In a study like this where there's just a few of those, what we tend to do is either drop them or try running associations with and without them. Sometimes, I think our focus on population stratification is a little overblown. Of course, if you've got a mixture of two entire populations, it's probably best to study them separately.

You can also, of course, run Eigenstrat to correct for the potential population stratification.

There's some other QC approaches we do in copy number that might actually apply nicely to SNPs, like finding gender mismatches, mosaicism, and cell line artifacts. But we'll talk a little bit about that in the next section.

Just a word on Q-Q plots. If you're not familiar with Q-Q plots, the idea is that you have an expected association and versus actual, expected being just random by chance alone. If you do a million tests, what's the distribution. So the expected value of the P is one over the rank. So you expect a line following $Y = X$. But if you don't do any QC, you see, of course, a lot of false positives due to these batch effects.

If you do some QC, you may still see – for instance, for diseases like rheumatoid arthritis where there's a large HLA region, you could see an amplification in the upper right here of a lot of significant findings that are just dozens of significant markers, say, in the HLA region. So if you can knock that region out in Chromosome 6 and then see what's left and you see a nice linear relationship, perhaps, with a few interesting genome-wide significant peaks, you know you've done a good job of quality control.

So these are really useful tools that go along with both – in a sense, you go back and forth between association testing and quality control to some degree. It's not like something you can just do all up front and expect to be done.

Now, switching again, switching a bit over to CNV analysis. So we've been doing genome-wide scans for copy number variation on literally thousands of samples across multiple studies. There's a whole lot of learnings we've had on dealing with quality control issues in that context.

Then we're going to switch back once we've kind of finished on the quality control to talk a little bit more about association testing.

So if we thought batch effects were a problem in SNPs, they were probably an order of magnitude worse in copy number variation studies. I plotted here – if you take log ratios and do an association test, say, Affymetrix versus Illumina, both platforms have this issue if you don't address them, which is there'll be massive significant associations across the whole genome, 10 to the minus 100th, 200th, everywhere. Basically, if you don't deal with plate and batch effects, they're going to hopelessly confound your association testing.

Some of the existing approaches – we've really been kind of beating this drum for sometime. But earlier on, most people were not even considering that there were these large differences and they're just going ahead and making copy number calls. Of course, it's been recently noticed and there's a paper in *Nature Genetics* where they're trying to deal with the batch effects by correcting for them downstream in association testing. It's one result, one approach. A lot of the studies that we've seen also in seeing analysis had to focus on really large CNVs or common ones that have high quality prior information. So in thinking of where we wanted to be, we wanted a really good normalization procedure that fixes the batch problem before association testing, be able to detect CNVs wherever they are, regardless if they're common in healthy people. They may be common in your disease population and you need to discover them.

There's special issues with sex chromosomes of how you deal with the copy number difference in, say, the X chromosome, one in males, two in females, and, of course, have a way to exclude and identify – identify and exclude problematic samples.

So there's different normalization approaches for different platforms. The Affymetrix approach of normalizing against the reference sample, taking a log of the A and B intensities, so you basically take the median of A plus median B for a given SNP across your 2,000 samples in a study and then you take your intensity for A plus B, take that ratio, and you get a ratio that you take the log and you get a so-called log ratio.

Now, we've played with do you take controls only as your reference or case- controls. We tend like – especially because of potential batch problems, we tend to put all the samples as their own reference, particularly for large studies.

For Illumina data, there's an option to use an external HapMap reference that has the advantage of it's the same reference every time. But we found when you normalize with HapMap samples, you get, for instance, plotting the log ratios for males and females in the X chromosome. On the upper one, we normalized within sample using all the samples versus normalizing with HapMap. You see the spread of the distribution is much worse. You have less resolution normalizing with HapMap. It's perhaps no surprise 'cause those HapMap samples were run at a different facility than your samples.

So there is an option in BeadStudio in the successor to BeadStudio to normalize within sample. We highly recommend doing that.

I won't talk a whole lot about 2-color experiments. But we've been getting some experience with those, less in the context of GWAS, more in the context of cytogenetics, where you have a single sample reference and a two-color experiment where you do a – the experiments are run,

Achieving Genome-Wide Success in SNP and CNV Studies

Golden Helix Webcast

February 18, 2009

Dr. Christophe Lambert

compared on the spot. You can either use a single sample or pooled samples. There's pluses and minuses of those. As well as when you're doing a paired analysis like that, correction for probe-specific binding affinities can certainly improve results. You've heard of correcting GC content. There's also some higher order corrections. For instance, modeling the binding affinity as a function of the entire probe sequence.

Looking at HapMap data, for instance, we came up with a very nice quality control procedure where you basically take the average intensity across each chromosome, plot a histogram of it, and you should expect zero for the autosomal means. I don't have the X or Y in this plot. We'll show that a little later.

You see there's certain chromosomes that have outliers. Here is green one, one of the samples of Chromosome 3. In this orange one is Chromosome 4. Well, it turns out if you look at NA 18540, it actually had outliers for Chromosome 4, 7, 9, 14 and 21. So this person – if this was a real person, they wouldn't be alive. So it's clearly some sort of cell line artifact. But so when you're dealing with cell lines, it's not uncommon to see this type of odd behavior. These are samples you'd probably want to exclude or at least exclude those chromosomes from analysis.

It's perhaps another reason to be careful using HapMap samples as reference because NA 18540 is not the only suspect sample.

When we've looked at other studies, normally, you don't – that are from whole blood – normally, you don't see this type of behavior. It tends to mainly be a cell line artifact.

An important quality control step is looking at anomalies in what you're reported gender is versus your actual gender. So if you plot the mean intensity of the log ratios for X versus Y for your samples, you'd expect the males to have a higher Y intensity, and there shouldn't be any Y intensity for females, and, of course a lower X intensity 'cause males have Copy No. 1 for their X chromosome.

You end up seeing, often – sometimes you'll see a dot in the middle of one of these clusters where there's a misclassification. Then sometimes you see some phenomena like this where there's a reported female who doesn't quite have Copy 1. They don't have any Y intensity, but they have a low X intensity.

In this sample, is a HapMap sample, NA 18540, actually, is reported to have mosaicism in the X chromosome. Sixty percent of the cells are Copy No. 1 and the others are Copy No. 2. So you kind of get this mixture. So these are potential candidates for exclusion as well.

Let me go a bit more in to batch effects now and then we'll talk about how to fix them. In an ideal world, we would have randomized our cases and controls on plates. We wouldn't have borrowed controls from other experiments. If we're doing a family-based study, we would have kept the families together on plates so we're not introducing sources of variability splitting up the families.

But why do we see these differences? We tend to – in our experience, we don't see, within a single plate, a 96-well plate, we don't see much variability among the samples on a plate. It's from one plate to the next. It's from one date to the next, one site to the next, machine to the next. Environmental conditions can be important.

We also see that it's not just some systematic shift, up or down, in intensity, although that can happen. That would just be corrected by quantile normalization. But it looks like there's sort of nonlinear sequence dependent shifts in intensity. So that leads to one of our recommendations, which is to consider placing a well characterized male and female sample on each plate so that you can have the same samples on every plate and be able to potentially correct for these funny variations that occur from one plate to the next. It might be temperature dependent or some environmental condition dependent.

How do you correct for these? We've taken a principal components approach kind of modeled after the Eigenstrat idea of calculating population – correcting for population stratification in SNPs. How Eigenstrat works is it basically – you take a numerical matrix of zero, one, or two copies of the minor allele is what you create as the numerical matrix, do a principal components analysis on that. When there's a correlated shifts in allele frequencies by different ethnic groups, those are seen as the largest components and they can be corrected for.

The same idea if you have large – you have batches of samples that have these correlated shifts in log ratios. They basically show up as the largest principal components. We can basically factor out the first Q principal components and basically extract the shifts that occur as a function of batch and then you can do analysis downstream.

Just to give a kind of pictorial, on the upper left, we've got the Welcome Trust, Eigenstrat on SNPs. You see basically the SNPs are drawn from the same population, British subjects, and generally Caucasians, although these outliers are interesting to plot them against the HapMap samples to see what the mixtures are.

But if you do a principal component analysis and plot the first two components on the log ratios, you see there's a big shift between case and controls. The case and controls were run at different times. Those are going to cause all sorts of grief in terms of systematic variation. I've also plotted a similar plot for the GAIN bipolar study where batch effects are a problem as well.

This is a plot on of our customers graciously let us show. They're looking at 8,000 samples from a very large study with multiple phases. On the left, we plotted the average intensity of Chromosome 21, a histogram, for the 8,000 samples. So you see in blue here, one of the phases actually has this funny bimodal distribution, so there's probably some difference in how half of those samples were processed. You see this pink phase is kind of one nice distribution centered at about zero, which is what you'd like to see. Then there's a couple other smaller phases that are shifted to the left. So this was the biggest effect we'd ever seen.

There's also the genetic information was captured over multiple generations. Some of it was 20 years older than other information. So there was a lot going on in this data.

Well, we ran this principal components analysis, and then other than outliers that you see here and there's actually some others off to the scale to the right of the histogram, which we would potentially exclude, everything got nicely centered closer to zero and we can now do analysis downstream without having these messy effects.

So just a couple bullets on what we suggest. We'll show a few more plots describing this. We found SNP quality control procedures tend to not really be terribly correlated with – we've seen samples that pass all SNP quality control procedures, but then have real problems with copy number. So we don't use the SNP quality control procedures. We do things like , of course,

Achieving Genome-Wide Success in SNP and CNV Studies

Golden Helix Webcast

February 18, 2009

Dr. Christophe Lambert

verifying the gender concordance that we described earlier. Perform the principal component analysis. Selecting the right number of components, we'll probably talk in depth at a later date on this whole subject.

In general, you want to look at your scree plots, your Q-Q plots, of association on log ratios. Be aware of the correlation structure, extended pedigrees. You don't want to over correct because there may be some family structure that start messing up by using too many principal components.

In some of my earlier talks, I've been suggesting a lot of components. I'm tending to dial back on that to fewer principal components. Another key thing we'll talk about in a moment is you got to exclude the sex chromosomes in your PCA analysis 'cause that'll introduce artifacts based on the systematic shift of males and females in the X, chromosome or the Y chromosome, in the case of arrays that have those Y markers.

Assess your log ratios for the presence of outliers in terms of looking at the mean log ratio per chromosome. Then after your calling is done, we actually find there's some steps after you've done all your analysis where you want to revisit quality control and potentially exclude some additional samples.

I wanted to show you a really dramatic example of the rattiest single we'd seen. It turned out we had a multi-center study and there was one center that just had all these funky wave effects and very variable log ratios. If you were to try to make calls on this data, you would go crazy.

It turns out because there are _____ samples, there was a correlation structure to the nastiness that occurred in these samples. We were able to correct for them.

Now, in retrospect, we've been doing this study with and without these samples because they really were over the top bad. But this was just an example of how dramatic you can have a problem with quality where principal components can correct for it to a great deal. Here's just another plot of a different chromosome, Chromosome 1, sample from that same study. Green is before and blue is after. I've also overlaid them.

By the way, almost all the diagrams and plots you've seen are coming from our new SNP & Variation Suite, which today is the grand launch for it. We'll talk a bit about that towards the end.

This is a new things that we've just been experimenting with, which is, again, if you put in the sex chromosomes and principal components, what actually happens is, say, for the male and female X, they'll be shifted together into some common midpoint. So you'll actually remove the difference between males and females. In some context, that's not necessarily a bad thing. However, if you want to actually see this difference in males and females, what we found is you can actually calculate the principal components on Chromosome 1 through 22 and then subtract out the first few principal components for all of those, but then apply that same process to the X chromosome and potentially the Y as well.

What you see is a before and after. Before is above. You see the separation is not as good. There's an overlap between the average intensity for X for males in blue and females in pink. Then the separation actually gets cleaner as the distributions tighten up as we remove some of the variability.

This was actually HapMap samples. There was three plates. Even that plate effect of they did the Yorubas on one plate, the Ceph's on another, and the Chinese and Japanese on another plate. There is plate effects even in the HapMap samples.

This type of cleanup that we see here is even more dramatic on studies that have more dramatic batch effects.

A little word on Illumina quality control. A lot of the normalization procedures and calling of log ratios Illumina does for you, and they do a pretty good job, but in a couple of studies we've seen, this problem crops up largely because they don't do a quantile normalization step. It's possible to have certain signal intensities that are very sort of off the charts. For some reason, in the studies we've seen, the outliers are not symmetric. They tend to be negative outliers. So if you were to do a study and do segmentation, you basically are going to get this huge enrichment of what apparently looks like copy number losses, but it's spurious. It's just some problem with calling.

So we're still working on this problem, but be aware of it. Don't indiscriminately send data through without looking at it. One approach would be potentially to Winsorize the outliers, which you basically do something like take two standard deviations in every point, less than two standard deviations, you make it equal to two, the minus-two standard deviation to every point greater than plus-two standard deviations, and make it equal to plus-two standard deviations.

Another potential approach would be like a median smooth, which I've shown here. You see a legitimate copy number loss. It goes down to not quite minus one, which is nothing like the minus-six units we were seeing for some of those outliers. So just a – you may have to do some post processing of Illumina data if you have this outlier problem.

So I've started alluding to the idea of making copy number calls. We use segmentation algorithms that we'll talk about in a moment. But after segmentation, if you do a distribution of the number of segments found, and here's our histogram, so remember, in our calls, we have both neutrals that are Copy No., 2 losses and gains. We see a number of samples. This was, again, in that study that had those really ratty samples. They were finding this real excess of copy number calls. So these would – just by looking at the number of copy number calls, you can know there's some problem with some of these samples and you can potentially exclude them from the study.

We also saw just a few funny outliers that had very few calls. I suspect also was just a super high noise and so really nothing could be found. So you might notch out the two tails of this distribution. Then just plot it here.

There's yet another quality control procedure you can do is look at the histogram of the mean of the segments found. You should see a nice trimodal distribution both overall as well at a per sample level, which is what this is. When you have a ratty sample, like here, you just don't see these tails resolved. Sometimes you'll see a bimodal distribution and you know there's a problem there. So automated detection of this is actually possible by fitting a trimodal distribution or if your data has even clear differentiation to higher copy number states more modes.

Achieving Genome-Wide Success in SNP and CNV Studies

Golden Helix Webcast

February 18, 2009

Dr. Christophe Lambert

But in general, for the AFFY and Illumina data we've looked at, for the most part, at looking at very small copy number changes. You tend to see pretty much a trimodal distribution. If you look at really large changes, then you can actually resolve other copy number states. But then have to be like these mega-based type copy number variants.

So quality control takes most of your time and effort. If you do it right, a lot of the association testing gets easy. So the problem in association testing is pretty much quality control. It creates spurious associations. Hopefully, we fix that. Then just a few thoughts on how to perform association testing with phenotype complications.

I gave a snapshot of our software here to kind of give you a thought, some thoughts on there's many different genetic models you might want to do, additive, dominant, recessive. Recognize, though, if you're doing each of these different genetic models, there's a certain amount of multiple testing involved. So if you had to pick one, you might do a genotypic test, which is the most general, where that pick up like a heterozygous advantage effect or perhaps an additive model.

Use exact test when available.

Permutation testing is particularly useful when your data departs from the model assumptions of some of the tests.

We can do quantitative trait tests. We see sometimes if you have some outliers that are very large values for your quantitative traits, you'll have an inflation of Type 1 error. The permutation testing can help mitigate that.

Population structure correction, do an Eigenstrat is something that we've integrated. You want to be concerned also about confounding factors. That's something we'll talk about in a moment with regards to regression.

There's many topics I'd love to talk about and I feel like we've already are doing a fire hose today. But you might be interested in haplotype association, runs of homozygosity association, as well as there's a whole other discussion about family-base association.

We do have some webinars on these topics in our archive. But we also plan in the future to go in more depth into some of these topics.

In the world of CNV association, again, we've talked about quality problems. If you don't deal with that, everything else downstream is sort of garbage in, garbage out. But there is an extra challenge in copy number data of accurate determination of CNVs. If you haven't done the quality control, it's going to be a nightmare. There's still some challenges to deal with.

Then the question, how do you do association testing. A lot of the current approaches, again, I said before, focus on finding either larger CNVs to exclusion of small ones and potential to the exclusion of previously uncharacterized ones.

We've done some playing around with the Birdsuite and maybe operator error. I'm sure they're improving it over time. But we found that the batch effect problem is still – is not accounted for before they make their calls. So this is a plot of some of the association testing on their discreet

calls that we did on a bipolar study. It was just too noisy to really – to find much meaningful results.

So what we want to be able to do is accurately detect small CNVs, find genome-wide significant associations, much in the way we've done SNP scans, but doing copy number scans.

A lot of the papers that have come out lately on copy number association have been, "Well, we found a really big deletion or gain in these three samples. Here's our paper." What we're more interested in finding is smaller but more common copy number variants that are actually genome-wide significant and are not just sort of we observe them in three cases and no controls and hope it's real. We want to see a genome-wide significant difference.

So what we've been doing in CNV association testing, we'll show some more slides on this. Again, quality control can't be stressed enough. Use a good calling algorithm to find copy number segments. We then discretize based on a trimodal distribution unless you can actually see more states than that.

Run association tests both on log ratios and discreet calls. You'll generally see, even after quality control, your Q-Q plots will really still have an inflation to Type 1 error. But once you segment, you should get very nice Q-Q plots.

Chris Amos gave a talk, and it's recorded on your website, showing that phenomena a number of months back, looking at lung cancer.

One other thing we do is look for significant peaks that span multiple marker 'cause you're always concerned about quality control. If you have an association that spans a very small region, it's probably spurious. A median smooth is actually a nice way to highlight these significant peaks.

Then after you found a significant region, go back, visually examine the log ratio data in significant regions, look for clear evidence of contiguous spans of gain or loss markers and recognize that the start and end regions may well vary by sample, especially when you look at a fine green.

There's some interesting talks given at ASHG recently, or this past conference, where they did down to ten base pair resolution with genome-wide sequencing. They're finding something like 200,000 common copy number polymorphisms per person.

So it's reasonable to expect that copy number polymorphisms are going to be a lot more like SNPs. What we're trying to do is do our best, given the noise of the data, to be able to make good copy number calls and do these associations.

So when you have log ratios like you see in the upper left, the idea is there's a contiguous region of, say, here's a loss followed by a neutral, and that there's a shift in the mean between those two states. There's different approaches from Hidden Markov models and segmentation approaches to try to basically go from this noisy data to a discreet call.

The problem is that you have little guys down here that might be a gain, might be a loss.

There's some survey papers that have pretty much shown the segmentation approaches appear to perform the best. There are some newer Hidden Markov models. One thing to be cautious

of with Hidden Markov models is that our experience of the data to fine green is they're not always just clean Copy No. 1 center about minus a half and neutral to zero. There's some mosaicism sometimes. A segmentation approach like circular binary segmenting is actually a better way to go than Hidden Markov models.

We actually, for ten years now, have had an optimal segmentation approach that actually exactly solves the problem of finding optimal segments. We've demonstrated, on the same data that Willenbrock used in their benchmarking, that our approach we've called CNAM, actually exceeds the sensitivity and false discovery rate of circular binary segmenting.

So we've been using that in our software. The point is use a good calling algorithm. Some of the – the best ones are a little slower, but it's worth it in terms of the results you get.

You tend to see – when we've been going down to like ten marker resolution in AFFY and Illumina data and so you do tend to get more of a smooshing together of your losses and neutrals, gains. We'll discretize it into a three-state model and actually create copy number covariates. Each time there's a change in copy number, we will introduce a new covariate. You can also use the mean value of the segments as an alternative, particularly if you think mosaicism or something might be going on.

You basically set up a big matrix a lot like you do with a SNP study where you've got your Y, which is your case control status or quantitative trait, and then your X's. Your predictors are the copy number states across the whole genome.

Here was a whole genome scan of bipolar disorder. We found this interesting peak in Chromosome 2 with one of the models. It turned out it actually went up to 10^{-14} when we did a recessive model, but I just had a plot of – I think this was recessive in the sense of the rarer effect of a copy number gain, I believe, was what was going on in this study.

Actually, the more common thing in that particular region was a loss. So here's how you – you look at your data. We plotted about 50 samples overlaid here. You see this looks like a loss. It looks kind of noisy from a distance. But if you zoom in, it turns out there was actually 23 markers where all of them were on average lower. It was actually fairly consistent that it was this break point of these 23 markers in this Chromosome 2. I think it was p22.1 region.

We can also then do a histogram of those 23 markers in cases versus controls. What was funny about this was what was actually driving association was this tail here of a potential gain. So until we actually go do the biology and see if there's really some gains here in addition to the obvious clear copy number loss and copy number neutral, there's some real question marks about whether that association is real or just some difference in the distribution between cases, controls, maybe a quality problem.

So you do have to look at these histograms. You do have to look at the log ratio data. Here, I was plotting losses. I plotted the gains. It's actually a little less convincing. So this would be one potential association that we might put less stock in.

We've done a lot of looking at Wellcome Trust data and haven't as much chance to do the segmentation on that just because we we've been doing – focusing on our customer collaborations. There's some very interesting findings. We've been finding that we expect to be

published in months to come. An important, though, is when you do SNP association and CNV association, we are finding different regions that are associated.

That gives us the hope that maybe some of the dark matter, the unexplained heritability might well be in these copy number variants. Of course, we'll still looking at rather low resolution compared to the ten basic _____ resolution of next generation sequencing. So it'd be interesting to – as those technologies start going – becoming less expensive and we can go genome-wide, we can do more with those.

Here's kind of a plot of log ratio association tests on the Wellcome Trust studies. There's some interesting regions that are showing up. We found 34 to 40 percent of the strong associations seem to be corroborated by previous SNP studies. Notice also, though, in seven and fourteen, there's some region that's highly significant across nearly every disease. You have to beware. There's some T-cell rearrangements that occur. In general, we found across many different diseases the proportion of T-cells and B-cells between case and controls, it's easy for that to differ. Then that shows up as a highly significant association. It may be biologically relevant that your T-cell/B-cell ratio is different or that's there's rearrangement that might be going on in greater number than case and controls. But we've seen it's not very disease-specific. These same regions show up again and again, so beware.

I'm going to skip really quickly through regression techniques. We'll probably be doing a whole other webinar on it. We've really scaled up our capabilities of doing regression with interaction effects, moving windows, both of genetic variables as well as quantitative variables. You can potentially even do a moving window of multiple log ratios or copy number call markers, combine them with SNP markers, combine them with confounding factors that you're worried might be a problem. So there's just a whole lot you can do with regression.

In particular, the whole issue of confounding variables where you can do a – factor out – for instance, you've got a test of – you're doing association tests on diabetes, you know obesity is a confounding factor. You don't want to find SNPs associated with obesity, so you can sort of factor out the body mass index, for instance, and then look for SNPs that are associated with the diabetes phenotype after factoring out whatever body mass index association might be going on.

Finally, I wanted to talk a bit about what about what about when you don't find a whole lot in your study and how do you make some gold out of that straw without cheating, which one concern – we do want to aspire the genome-wide significance, but there's a real danger of selection bias in reporting nominally significant findings. You start looking at the top twenty nominally significant markers and then saying, "Oh, well, that one is involved in a gene that's been reported before from my disease." You pick that up. Eventually, we could start sort of moving all the literature in the direction to sort of preexisting assumptions.

So it's two complementary approaches we've been using is select all the nominally significant findings at a given threshold, pick a fixed threshold, and then look for a significantly overrepresented pathways or ontologies. You can, of course, do this with your significant findings as well. One of the best tools we've used for that is GeneGo's Metacore product. They've just got some great curation of known information about SNPs, genes, pathways, and ontologies. They have an army of something like 40 scientists reading papers and typing in which ones are validated, which ones are questionable. It's a great way to look at your data.

Achieving Genome-Wide Success in SNP and CNV Studies

Golden Helix Webcast

February 18, 2009

Dr. Christophe Lambert

Another approach that's complementary is do a median smooth of your P values, or you logP values, and select regions that tower above the rest by virtue of multiple near adjacent markers being significantly associated and then feed those into some sort of pathway or ontology software.

What we find is, particularly because of concerns of quality, if you pick some low cut, you know a lot of those are going to still be markers that have a problem with quality. But if multiple adjacent markers, which adjacent on the genome, but actually they're not adjacent on the chip, are showing significance. They're probably _____, but at least you know they're not driven by a quality problem. You'd expect them to be more believable if you see 15 peaks at 10^{-4} within the same region than if you just see one, or maybe not even 15, but 5.

So there's, of course, seek replication of any of these nominal findings in additional studies. Be careful of the selection bias.

Of course, if you've only done a SNP study, consider a CNV study and we can help 'cause we've been getting a lot experience with these.

Another things we've seen particularly of value, again, and, of course, all these approaches we've been applying in both SNP and CNV studies, augment your phenotype if data is available. Some case-only quantitative trait locus can be very powerful in finding markers and genes which might mitigate the, say, the onset or the progression of your disease if you have some sort of biological measurements of those things.

Here is just a – I wanted to show you some work we did on bipolar disorder where we took some of these – these actually were either significant or nominally significant and we found that certain pathways in neurotransmitter GABA receptor activity actually were very suggestive, even chloride channel activity of being involved in pathways related to bipolar disorder. This is just a first pass I did actually with a free package called "BiNGO." I look forward to doing some more of this type of analysis with GeneGo where it's just a much richer source of data.

So in summary, we've done a fire hose of looking at many different topics. This is the first in a lecture series. We hope to go into much greater depth on a lot of these topics. Quality control, I think, has been the real in depth message today for both SNPs and CNV. If you get that right, everything else gets easier downstream in your statistical test for SNP and CNV association.

We really hope this has helped you in thinking about who you do your studies.

I did want to make a little plug. We got to earn a living. We have been doing a lot of these services, helping customers do their SNP or CNV associations studies. We're also really exploring possibilities, diagnostic development, been working some projects along those lines.

So if you're struggling with a study or if you even just need help making _____ calls because you don't have a big enough machine, we're here to help.

Of course, today is actually the launch of our SNP & Variation Suite 7 to the public. Our people are working late into the evening to make sure all our demo requests, links, are working in our website. You can try our software. Almost all the plots you saw today and analysis were done with SVS 7. Operators are standing by, so to speak.

Achieving Genome-Wide Success in SNP and CNV Studies

So I'd like to really acknowledge all of our customers, all our collaborators. Some of them we can't name. Some we could name. I felt just best to give a global thank you. There's been a tremendous amount of feedback we've had from beta testers, from our collaborators that we've been – as we've been sort of eating our own dog food and having to use our software to do our studies. We've really done our best to make it as usable and as powerful as possible to do these genome-wide studies.

So I've taken a full hour to present. I kind of expected that might happen. But we'll stay as long as you'd like to ask questions. Feel free to type a question into the "Go To Webinar Questions" pane. I'll begin answering them. I'd like to thank you for attending. If you have to go, I understand. But we'll stay here to answer questions and wish you all in accelerating your quest for statistical significance as well as really doing something meaningful in solving the diseases that are plaguing mankind.

So thank you very much. Let's go to the questions.

Oh, my, there's plenty. Feel free to keep adding questions.

So _____ for Illumina _____ genome chips?

We've largely been using Illumina's calling algorithm. We try to do, in the CNV world, some basic – run through the same sort of workflows as Affymetrix. We were doing it just on a small HapMap sample for comparison and found that Illumina's calls were actually had lower variance. So we've been going with those.

I'd like to try it again in a much larger study, see if we can get any improvement. With that being said, even after Illumina's log ratios are calculated, you absolutely need to do a principal components analysis to deal with batch effects.

I know there are some published alternative methods for SNP calling that people have done for Illumina. I can't say that I have enough expertise to comment about their relative merits.

Someone asked about reviewing the reason for dropping SNPs with minor allele frequencies 0.01. There's a couple reasons. One is the reason you might want to do it is they're not going to have a lot of power to find significant association anyhow. So you're kind of hitting yourself with an extra multiple testing penalty by including those.

Another reason is often just the calling algorithms don't do so well on these markers that do really lower minor allele frequency. So yes, if you have a large enough study, even a one percent minor allele frequency could certainly give you a significant result.

So I'm not a big fan necessarily of having to drop those if you believe you've got a lot of rare alleles that are involved. But just be aware that those low minor allele frequency markers could be more problematic.

One thing I'd like to do actually is I'll plot my genome-wide significance plot and then I'll have some subplots lined up along the genome browser that will plot various quality control metrics like _____ from Hardy-Weinberg in minor allele frequency and then I can sort of make my own judgment rather than necessarily dropping those markers.

Achieving Genome-Wide Success in SNP and CNV Studies

Golden Helix Webcast

February 18, 2009

Dr. Christophe Lambert

Someone asked, “Do you drop SNP _____ Hardy-Weinberg equilibrium at $P 10^{-7}$. Yeah. I’ve used that as a threshold. Of course, if – but I’ve also – using the criteria I described earlier where you really dial up the call rate, I found that you don’t even have to drop markers that are not in Hardy-Wein equilibrium and you don’t tend to see spurious associations from them.

There’s an interesting publication out recently, a guidance on publication of genome-wide studies, that talks about some of the controversy around dropping things for low Hardy-Wein equilibrium where they say actually a large fraction, like 15 percent or something, of the replicated – I think the replicated genome-wide association studies – actually, the markers were significantly at a Hardy-Weinberg equilibrium. So the assumption Hardy-Weinberg equilibrium is just an assumption and so, again, I keep all my SNPs and then I turn on and off various filters and see what they do.

In general, you really use a Q-Q plot as well as a plot of genome-wide, sort of Manhattan plot, to make decisions about quality control.

What’s been your experience when using HapMap 3 for population stratification correction?

I’m not sure – what I’ve been using is basically the Affymetrix 500K 6.0 270 HapMap samples as well as the ones Illumina has made available. I find it’s very useful to basically merge those with your study as I showed and pick out those outliers.

In fact, the one study we’re doing recently, the outliers that we had picked, they actually knew they had a mixture of some different races. We had directly picked out the ones we said were African American, in fact, where the Asian sample was correctly flagged. Then some of the ones in the middle tended to be some interesting eastern, Middle Eastern, descent populations that were mixed in there. So we basically just used the autosomes, Chromosome 1 through 22, and do a PCA analysis. It does a very nice job.

Why do you not recommend splitting families across plates? Good question.

So in general, there’s the least variability on a single plate. So if you have a – of course, if you have a huge family that’s more than 96 individuals, you’re going to have to split them. But at least keep the nuclear families together. Basically, it’s the variability from plate to plate is large.

When you compare either genotype or copy number variants within a family, all the family-based association tests are pretty much doing counts and comparisons from father/mother to child or looking at the extended pedigrees. So let’s remove that plate effect. So we’ve actually seen the family studies are some of the cleanest in terms of Q-Q plots of copy number variant association because they’re kept together on plates.

We’ve looked at some studies where the families were split across plates and they are very – they’re not – the difficulties are not insurmountable if you’ve already spent the money and done it that way. But if you’re doing them in the future, keep the families together absolutely.

How do you say the number of principle components for adjusting why do you exclude the first few top PC?

So we are including for the top few PCs and you’re basically factoring out their effect and they tend to be – if you plot, as I plotted the Welcome Trust data there, those first components really

Achieving Genome-Wide Success in SNP and CNV Studies

Golden Helix Webcast

February 18, 2009

Dr. Christophe Lambert

correspond to the largest differences across the most markers and then the next component is a smaller difference across fewer markers and it falls down. You basically look at a scree plot, try to pick an elbow, and we've got some tutorials on how to pick the number of principle components. We often take the log of the Eigenvalues and pick an elbow.

What tends to happen is there's not always a very clear delineation point. We've experimented with different values. Often, it's sort of just until you see a scree plot that's not totally crazy. Especially with families, make sure not to do too many because – with large families, the large extended pedigrees. With trios, it's not as much of an issue.

The choice of the number of components is still a bit of an art. So a lot of it is a bit of trial and error. In our newer version of our software, we're making it easier to just pre-compute as many components as you like and then try applying a difference in the number of components and you can experiment with looking at the Q-Q plots of association tests on log ratios. It works pretty well.

There's a question. How many clinical procedural covariates are supported in Helix _____ regression algorithms?

Plink analysis usually involve up front covariate adjustment _____ genetic association tests on the residuals. Only _____ crashes of more than two to three covariate are used.

We can – I don't know that there's really a fixed limit with clinical covariates, probably thousands, other than at a certain point your number of observations is less than your number of covariates. You have ranked efficient regression.

We do have, of course, stepwise regressions, so you can even add a thousand. It's just going to be kind of slow with a stepwise approach with a thousand covariates.

So we should be able to support a lot of them. If it doesn't, let us know and we'll fix it.

Do you throw out any markers before you run your CNV calling algorithm. When I use logistic regression to predict gender using LLR, we get a thousand or so that are beyond genome-wide significance. Is that a good cutoff? Does PCA control for this or hide real problems?

This is a good question.

As I mentioned earlier, the Q-Q plots on log ratios will create lots of spurious associations, even after correction.

We can't necessarily rule out all those markers and yet the Q-Q plot looks way inflated. We found after you segment that then that inflation goes away. So basically, a few noisy markers are not a big deal. We tended to be leaving them in and use the fact that we're segmenting over reasonable spans of markers. The chance you got a few bad ones is there, but they tend to average out in the noise when you do segmentation.

You were saying if you do predict gender using the log ratios, you get a thousand or so that are beyond genome-wide significance. Make sure you've done your principle components analysis not including gender.

Achieving Genome-Wide Success in SNP and CNV Studies

Golden Helix Webcast

February 18, 2009

Dr. Christophe Lambert

And yes, there are some markers that are associated with gender that you can potentially exclude. I haven't been doing it lately, but I can't say necessarily that we've got the final word on this.

Someone was asking me, "What do I do if I want to try your software on my data?" Well, you can download a free demo. It's often good to look at some of our past webinars that describe the use of the software for various workflows. You can download a free fully functional trial and feel free to contact us to get some help to discuss how you might approach the analysis of your data.

As everything, there's always a learning curve and we can help shorten that for you by walking you through some of the steps of analysis.

Someone asking, "How can I run CNV PC correction by Eigenstrat?" So I guess it's – our software has this principle components correction built into it. So if you're not a customer, you might want to become one. But basically, you are – you're calculating principle components on the log ratios and it's very – it's a fairly expensive procedure, but we can do it in a few hours for a genome-wide study and a couple thousand samples.

You pretty much need the memory management to be able to deal with the 40-gigabyte matrix like you have in copy number variation that we've built into our software. So you're not going to be able to readily do this in our or any other package that we're aware of at this time. So pretty much this is something that we've built and you can try to write it yourself or you can come to us.

"Is there a specific kit for DNA extraction that works best?" is another question. Unfortunately, I just don't deal enough with the wet work to really be able to make that kind of recommendation, so I apologize.

But I would say that whatever you use for extraction, use the same kit for your whole study. That's probably the most important thing. I imagine there's a number of credible vendors who will do a very decent job just to make sure you – all your DNA handling is done consistently.

We think it may even – there may even be a difference how long you leave the DNA out before you freeze it could have an impact. So any source of variability in how the sample is extracted, prepared, etcetera, try to keep those all the same.

Okay. So someone asked, "What are the variables you use in the principle component analysis to remove the batch effect?"

So all we're using is the log ratios in the principle component. As I said earlier, you just want to use Chromosome 1 through 22 'cause they're expected to be copy number neutral and then you can actually apply those components to the X and Y chromosomes as well to remove the variability.

The reason that works is because it's the same kind of process that's filing up in all the other chromosomes. It's filing up X and Y and we can basically see those patterns in the first 22 chromosomes and then fix it without actually shifting the log ratios.

Achieving Genome-Wide Success in SNP and CNV Studies

Golden Helix Webcast

February 18, 2009

Dr. Christophe Lambert

The next question, "I've been using the internal controls other than HapMap for generating CNV calls. I'd like to know what is the average number of CNV calls that could be found. I get an average of 90 to 100 per sample. These are disease germ line. But similar subject with a HapMap _____ average of 750 per sample."

If you're using internal controls, i.e. – a key sort of parameter that determines how many CNV calls are found is, of course, the size, the minimum size, of region that you specify.

Now, in our studies, 'cause we're going for the real – the small copy number variants, we're using a – we have a constraint in our common _____ optimization that can enforce the segments that are least ten markers long.

If you put that up to 50, you'll find much fewer copy number variants and you'll have a lower false discovery rate, but then you'll also be missing some of the smaller ones.

So I've heard some talks that 750 to 1,500 per sample with a genome-wide study of actual changes. When I showed a histogram that was centered around 3,000, remember, every second one was a neutral one, so there was actually about 1,500 that we were seeing in that AFFY 6 study per sample.

So but again, we're going with a very small yardstick. We know we're going to have a higher false positive rate. But in a sense, by doing – looking over thousands of samples, on average, we're doing better than – we're not just generating random segments. There actually has to be some signal there and they have to be signaled across lots of samples to see significant association. So we can deal with these much smaller copy number regions, find significance, and then go back, look at the data, confirm it's real, to whatever degree you can believe that.

Then with this noisy data at a small copy number region, you're going to want to do experimental verification on at least some of your samples to verify that what you found is true using something like quantitative PCR.

All right. So the next question is, "Can our software for association analysis we use plant association kinetics also _____ provide service and consultation in the plant science sector?" Also, feel free to contact us. We'll discuss the challenges of your study.

A lot of plants are a lot like human DNA in the sense there's two copies of the chromosomes, but there's a few plants like wheat that are probably a whole different ball game where there's multiple copies of each chromosome.

There's also, of course, the whole question of are you dealing with inbred samples and what are the appropriate analysis processes for those. That's a lot different than kind of these studies we're looking at.

So feel free to contact us. We can look at what you got, see if we can help.

But in general, yes, for these _____ markers, you can analyze the data for plant genetics. I think there still may be some limitations in our new visualization of the chromosome browser that is hardwired to be human. That's going to be changed shortly to also support like cattle and rat and mice and basically even import whatever genome map you want. But that'll be coming shortly. The absence of that doesn't prevent you from doing analysis.

Let's see.

Someone asked about, "We use Illumina to genotype in Bead Studio to do our QC. Can we skip the Bead Studio step and do all the QC and Golden Helix?"

Well, we still use Bead Studio to generate log ratios and still advocate using that. But yeah, a lot of the QC I personally do in our software.

"Would you mind briefly retouching on the CLRMM batching versus no CLRMM batching?"

Okay. So it takes a lot of memory, but in terms of batching, what we're talking about is when you run all your cell files, say, from a 500K or a 6.0 study, you select all the cell files at once. The reason is if you have these multiple batches, a) you're minor alleles, if they're really rare, are not – your homozygous rare alleles will not be well represented in some of the clusters, particularly if you're doing a plate at a time as your batch.

So it's going to – it's going to have some problems actually miscalling things as well as the more you can represent – say you have 2,000 samples and you can basically represent all the three clusters of AABB and BB, you're going to make better calls. And if there's some funny systematic shift by batch in those clusters, that will result in a funny, in less confidence, in the call as reported by CLRMM versus when you don't have those shifts in the individual batches, it'll be highly confident of those calls.

So you actually want it to report a lower confidence on markers that actually do differ from batch to batch.

All right. Another question. "Comment on subjectivity of redefining clusters when using DNA of limited quantity and quality. Would you have more than one analyst to examine inter-rate or reliability in other strategies?"

Yeah. This was one of my kind of concerns about wanting to have an objective criteria for assessing cluster plots. I was hearing a very nice presentation from a fellow from the Wellcome Trust who was discussing how they did their clustering. Now, in general, he says a lot of it is fairly straightforward. You can see clear problems, but if you're looking at hundreds of them, you'd probably want to have more than one person doing it and make sure you get some – start having some consistent rules about what it is a good cluster plot, what's a problematic cluster plot.

I hope in the future to be able to automate some of these procedures. But right now, that's kind of where we're at. If you do a really bang up job on quality control, you do tend to have much fewer significant peaks to look at cluster plots. So it's less of a problem than if you're looking at hundreds of them.

Someone asked if imputation will be added to Helix Tree. There actually is an imputation for missing values that we've had for a number of years in Helix Tree, which then SVS 7 doesn't yet have that feature. But if you're talking more about the imputation approaches like Mach and Impute, we haven't integrated those. Some of the licenses of those are going to make it not possible for us to integrate them, so we may have to – we'll have to write our own. We expect to have that at some point.

With that being said, feel free to use these external packages. They've got a nice body of literature behind them. Then you can import those imputed SNPs into our software and do all the downstream analysis.

So we'll just keep pressing on 'til we get to the end of the questions. There's a number more. Thank you all for continuing to stay. I think this is very interesting, a lot of these questions.

"In CNV analysis, if all the data, control plus samples, is used as a reference for normalization, how different is analysis done to detect CNVs?"

When we've done log ratio association tests, the results are highly correlated, especially for the most significant markers, so if you use controls only versus cases of controls. Of course, because the median – most of the copy number changes are rare. The median will tend to be picking up a value that's close to Copy No. 2. So it's pretty much the same no matter how you do it.

Of course, commonly in these, the median is going to be some funny midpoint, say, between Copy No. 1 and 2. So then you won't see as necessarily a nice centering about zero of your Copy No. 2. So that will impact the segment means. Then, of course, that would impact your calling. So that would be one reason why you might do an association rather on a discreet call instead on the segment means.

Hopefully, that helps a bit.

"How much is the reliability of the models from humans to plants? I have not seen any examples from plants."

So I'm not sure I fully understand the question. But of course, there's plant genetics. A lot of the inbreeding, and so on, makes the approach of analysis of that data different. I'm not sure if we're talking about modeling a gene that's in a plant that might impact a human and they share some common evolutionary advantage like yeast or something in yeast models that some genes actually hold in yeast in humans. I guess it's hit and miss would be my thought.

"With regards to haplotype-based association testing, what's the best way to deal with multiple testing issues?"

I'm a real fan of the haplotype trend regression developed by Demetri Zakin, Bruce Weir, and others, where you can do a moving window haplotype across a genome of a number of markers, say, like five markers. Then your multiple testing penalty would be your usual genome-wide Bonferroni adjustment.

We do actually have in our software a regression module permutation testing if you have – you can do genome-wide permutation as well as a single regression permutation. We don't yet have haplotypes in SVS 7, but our predecessor product of SVS 6.4 does have haplotype testing.

But what you're talking about is if you go and make – and then through the multiple testing and the regression, there's the degrees of freedom that accounts for how many different haplotypes you've included. If you're talking about enumerating all the possible haplotypes, I guess then

you do have that multiple testing involved. But I'm not so much a fan of that from the perspective of there's all this phase uncertainty that makes an unambiguous haplotype calls a little shaky. So that's just my thoughts on that.

Someone asked, "You mentioned about ten base pair resolutions was also relevant to go to a one min marker probe generating CNV calls."

Yes, it's possible to do that. You'll have a lot of spurious outliers that'll tend to be driving the findings. If you do set it to one marker, what you might do is exclude all the ones that are one or two and that'll help suck out outliers.

Also, we've got a multi-variate segmenting approach. We can look at all samples simultaneously. We actually do use a one minimum marker probe for generating calls. What you can actually find is you can find markers that have a significant shift between – across all the markers at a single base pair resolution.

Now, I just want to clarify. In your question, you said about ten base pairs. When I was talking about ten base pairs, I was talking about some genome-wide next generation sequencing where they sequence down to ten base pairs. I'm talking ten markers resolution. I think that's what you meant as well.

So the next question. "Why did you say that the log R ratio of Q-Q plot is always bad?"

Well, if you – what we found is there's a funny correlation structure, I think, that you can actually correct for lots of principle components and have your entire line of your Q-Q plot be below the XY axis, in other words, but we just – we just tend to see some funny curvature in those lines that probably has to do with the correlation structure of these associations.

I don't entirely like that situation. Maybe we can improve it. But I wish I had shown some Q-Q plots of Copy Number Data. Perhaps a more detailed webinar will help – will go into this issue. People have been asking, "How do you determine the number of principle components?" We'll answer this question to much greater satisfaction in a future webinar.

A question was asked, "Could you please repeat your comment on CNV association on chromosome 7 and 14? So we've – certain regions in 7 and 14 are – they're known to have chromosomal rearrangements. If you look up the associated gene, you'll see some like antibody parts. That's not to say every association on Chromosome 7 to 14 is one of these. But you should know that these regions exist. We're not the only ones who've seen these in their studies. There have been some talks I've heard at the GAIN workshop on this as well.

Basically, we're seeing in many different diseases, public data sets as well as some of our collaborators, not all studies though, the one that Chris Amos showed on lung cancer as well, that these regions in 7 and 14 are related to the difference in the quantities of T-cells and B-cells. We actually did a study where they extracted the granulocytes from whole blood and so you have no T-cells. We compared granulocyte versus whole blood and we found those regions were lighting up. So really there's basically a lighting up of certain T-cell regions.

So one might advocate if you're doing a new study and you're worried about T-cell artifacts, you might just extract your granulocytes and do your study on the DNA from them.

Achieving Genome-Wide Success in SNP and CNV Studies

Golden Helix Webcast

February 18, 2009

Dr. Christophe Lambert

So then there was another question along the same lines. “What regions are commonly affected by the T-cell rearrangement in germ line DNA samples in _____ analysis?”

You know, I don’t have them off the top of my head. There was a previous webinar. Actually, I think my talk that I gave at IGES and then we put it up on our website, actually, list those regions and maybe that’s something we can put in our FAQ section. I apologize. I don’t sort of have the exact regions off the top of my head.

So another question. We’re getting close to the very end here. Thanks for staying. We still have, gosh, 77 people who are listening to this. That’s wonderful.

“Could you comment on using the Illumina _____ file as a reference, which is a cell line for calculating the log R ratio versus using mean, median, your own test samples, for reference to calculate your log R or copy number?”

So a good question. In fact, I think I’ll show that slide again right here.

So this was a study we did – this was actually a – I think it was a three ten, three thirty, a three thirty Illumina study, we did actually some time ago. The bottom one was when we used the HapMap samples as a reference and the top was normalizing within sample.

I looked at about four or five studies now where I’ve actually bothered to do the comparison. In every single case, I got a much cleaner separation between males and females, which was sort of a baseline comparison of Copy No. 1 versus 2, by using the in-sample normalization.

So there’s basically – in Bead Studio, there’s a, I think, the analysis menu, you just go “Tools” or “Analysis,” “Re-cluster All Samples.” Then you’ll get basically be able to calculate clusters based on your current study.

Now that being said, you might want to do some QC and drop some of the bad samples, so maybe an iterative approach. You might just re-cluster with all your samples, calculate the log ratios, go do some QC control, find some bad samples, drop them, go back and recalculate the clusters. That would potentially help exclude some of the contribution of some really – some samples to have particular quality problems.

Let’s see.

“I have Illumina data and see the same excess hump of negative log ratios that you showed in your slide. Can you recommend a paper that describes how to perform this quantile normalization?”

So in our manual, when we discuss our whole process, actually, you point to an Affymetrix technical report and a couple papers, I think by Boltstad, et al., that describes how Affymetrix does this quantile normalization.

In the not so distant future, right now, our quantile normalization and workflow is all set up to directly read Affymetrix cell files. What I intend to do is get it so you can also just take the original raw data from Illumina or other platforms and do that same type of normalization.

Achieving Genome-Wide Success in SNP and CNV Studies

Golden Helix Webcast

February 18, 2009

Dr. Christophe Lambert

There's actually a way to do it, which is you can create a text file. I think we support import of an Agilent format as well as the Affymetrix. They're all text-based CMT format. You could put in those log ratios, import them, but we don't yet have the quantile normalization sort of predefined except by reading cell files. That's something I intend to remedy. I'm concerned about this negative outlier problem in being able to do quantile normalization on a genome-wide scale.

Someone asked, "Can the software handle SNP 6.0 3,000 samples at a time?"

Yes. I've done 4,500 6.0 samples on one study you are looking at. So should be able to do that.

"On your slides, Bonfronny's not really conservative. Can you explain?"

Yeah. There was a talk at IGES. It was actually one of the student awarded talks – I think it was IGES – as well as the year before at IGES, a poster. I think it was by Joan Bailey-Wilson, if I'm not mistaken, that they basically did a lot of simulations on genome-wide studies looking at how over conservative is Bonfronny 'cause it's sort of been repeated as _____, "Well, we've got to use a false discovery rate." And yes, there's a correlation structure of the SNPs that would make us think Bonfronny just multiplying your P value by the number of tests is too conservative."

And yes, it is a little conservative, but not much. So that's kind of what that point was. I think I did skip it in the slide just for time. But so it's not just me saying this. A Bonfronny genome-wide significant result actually have pretty good threshold of significance.

Well, I'd like to thank all of you for attending this. Many of you have actually expressed your thanks for the talk. We will be, hopefully, if our recordings work, putting this up on our website shortly and so you can see it again as well as send your – send your colleagues to view this if they weren't able to see the talk and would like to view it again.

By all means, feel free to contact us if you have an interest in software services, any ways we can help, and as well as any feedback you might have on certain areas that you, as we do upcoming lectures, that you would really like us to drill into in more depth.

I'm thinking the next in the series is we're going to perhaps start on nuts and bolts starting from perhaps even how to install bioconductor in CLRMM and doing CLRMM calls, doing the quality control. It may be – it may be too much detail. But on the other hand, when you actually get down to doing it – I have a couple colleagues at the FDA and SAS who are trying to install CLRMM and they had just as much hassles as I did. So here I am recommending it and I know that you're actually going to have a bunch of pain and suffering trying to install it and running of the memory on large studies.

So we plan to do some detailed lectures in the months to come. So your feedback on what you liked and what you like to see would be very helpful.

So you could just send an email to probably support@goldenhelix.com would be a good place to send the email.

Achieving Genome-Wide Success in SNP and CNV Studies

Golden Helix Webcast

February 18, 2009

Dr. Christophe Lambert

So again, thank you so much for attending. I look forward to seeing you at future webinars. Tell your friends about it. We'll close the meeting and thanks again for attending and thanks again to all our customers and collaborators who have made this software possible. Check out the new SVS 7. We're just launching it today to the general public. Look forward to getting your feedback and we'll continue improving things and really help you in your quest to achieve significance.

Thank you so much. We'll close the meeting.

[End of Audio]