

*Christophe Lambert:* I'd like to thank the organizers for the kind invitation to speak today. Over the last two to three years we've been doing many genome wide copy numbers, studies, combined SNP and copy number studies and we found a number of lessons learned, particularly with regard to data quality issues.

It's really a garbage in, garbage out story and I guess I would like to really share our exploration of the issues that we've come up with. Now, in these 30 some studies that we've looked at, about 28 out of those 30 times there's been some sort of problem with experimental design that's lead to large biases in the data causing differences between the case and controls, whether it's borrowed controls, splitting trios up between different plates, and especially not doing a balanced randomized design of experiments across your plates leads to large type I, analysts struggle with batch effects and really, copy number studies severely compromised. And of course, you spend a lot of time on statistics banging your head on the data.

The general take home message is let's not try to just use statistics to solve the problem, let's try to solve it at its source. Try to keep things very consistent in terms of DNA extraction, in terms of especially using design and experiments, principles for plating and with genotyping, be very careful to control our environment and do appropriate quality control metrics to not only capture problems with our SNP calling, but as we'll see later, you need different metrics to detect problems with copy number variation.

Now, I thought I would give a couple of examples of Q-Q plots in designs that were good and not so good. On the left is the Wellcome Trust Q-Q plot, before any quality control for SNP study. Now there is some HLA associations that inflate type I error, but even if you take them out, you can see this large inflation of the Q-Q plot.

Now, here's a balanced randomized study, both of these were at the Affymetrix 500K, this is a different disease. Am I lazing your eye there? Sorry. And you see that the Q-Q plot without dropping a single SNP, no Hardy Weinberg filters, no for call rates and you get a perfect Q-Q plot because there's a balanced randomized design. Now, if you look at the Manhattan plot, of course, there's that HLA region in the Wellcome Trust, but there's all this other inflation that 10 to the minus 8<sup>th</sup>, 9<sup>th</sup>, 10<sup>th</sup>, that are due to problems with genotype calling where you have a common set of controls all genotyped together and then the case is genotyped in another big batch.

Whereas type I error is under control in the other study. Now when you look at copy number variation, it gets much worse. The inflation is huge if you do log ratio associations, however it's still pretty much under control when you look at the log ratio associations in a randomized balanced design. So, there still seems to be some inflation of type I error, it maybe just be some, we're not totally sure why that is.

If you look at the decomposition against the first two principle components, and color code by plate, you see there is still plate effects where each cluster of samples in this well randomized study is very different, however because the case and controls are balanced in each plate, when you do comparisons between them, the plate effect is basically canceling out. Perhaps here in this right corner you see one of these plates in green and blue, there's a little more, either cases than controls or vis versa. I forget which was colored which.

We've found, as we've done CNV based QC that you must incorporate both SNP and log ratio quality control metrics and so, all of the usual SNP based QC metrics are essential in your copies number based study, but there's certain quality problems that you can't detect just by looking at your SNP data, so often in a genotyping center you know, you'll run all these samples and they pass the quality seal of approval and then you find 15, 20 percent of your samples have problems with their log ratios whether it be wave effects, or skewed distributions of various sorts.

And so, there's one in particular SNP based QC method I would like to mention where often NSP-STY pairs that are supposedly from the same person are not, some sort of mix up happens in the sample handling and we've looked at a number of 500 case studies and I guess there will be fewer of these in the future, but I guess if you are looking at 500 case studies, often 1 to 2 percent of the samples are mismatched and we've developed a new approach. I've yet to publish it out. Where you can basically correlate the nearby SNPs between NSP and STY. And you see a very clear signal separation between mismatched samples and matched samples where the correlation is very high when the sample between NSP and STY is matched.

It's important to look at the mean X and Y intensity. On the left, this is NST versus STY. If you see large differences like out here, here's an example of a male and a female of one is NSP and one was STY and they were called the same person, so that is who we

were able to verify our method shown previously works. And you'll see occasional mosaic cases where there are less than two full copies of X in the female example and it seems to agree between both NSP and STY. You also see cases with XXY and XXX any large study has these irregularities and they would be candidates for exclusion, as well as the occasional mislabeled; here's a female who was labeled as a male in blue here.

Cell line artifacts. Here's one of the hapmap samples that we look at, one of those 270 samples, NA18540 has 5 chromosomes with extra copies, be very careful using hapmap samples as reference or cell line based samples because these types of artifacts can occur. Once you've done all your processing and get log ratio and do copy number segmentation, after the fact, you can discover certain samples have a huge excess of copy number calls. And this is a biological phenomena, it's a data quality phenomena and you can look sometimes at the distribution of calls within a single individual and find out that there's not a good signal separation between losses, neutrals and gains.

And so, what we were concerned about was how do you detect this type of stuff early on after the data has come off the lab and not wait until you are doing your written analysis six months later to discover you've got quality problems? Many times SNP call rates don't, they can have 99 percent SNP call rates and you still have these data problems with copy number.

The derivative log ratio spread is just a standard deviation of adjacent points. It's a good measure of the noise of the platform. And you see there are these outliers here where on this X axis, high derivative log ratio spread. It does tend to correlate with low call rates as well as the segment counts, it's somewhat correlated, but we aren't able to capture all the poorly performing either excess to few segment samples just with looking at the noise derivative log ratio spread.

Wave effects seem to be a contributing factor to increased numbers of segments found, with, it doesn't really matter what segmentation algorithm you use. The PennCNV has the tool called genomic wave. It's got a nice metric to measure this and we found that if you have a really high and low there's kind of a sign for the metric wave quantity, there's a good candidates for exclusion are re-running.

We've tried the approach of fixing the wave \_\_\_\_ employed and for some reason, after correction we find that not only the signal,

not only the waves are dampened, but the signal separation seems dampened and maybe it's operator error, but ideally we just not have wave effects in the first place. And there's a paper by Disken et al that presents this method that suggest DNA quantities as the culprit, but there maybe other evidence that there's other factors.

The one parameter that we found that was capture a whole lot of problems that none of the other metrics did was looking at the extreme value distribution of the log ratios, looking at the bounds for an upper one percent and a lower one percent and that kept these really high averages for the top and bottom one percent, it seemed to capture most of these samples that had an excess number of segments

And when you look at one that had, here's one that had a high score at both ends. Notice it's the same axis, really an axis of evil. This data's nasty. Whereas this is a very clean sample that the variants is much smaller and you don't see these actually as wave effects as well, but you also see this phenomena where there is upward skew and a downward skew of the data, so when you look at those tails not being balanced, often that's an indicator that that's a quality problem.

And surprisingly, again, some of these samples can have good call rates when you look at the genotyping some of them not so good. Now, how do you put all these metrics together, whether it's, we tend to put all the SNP and copy number metrics in one big spreadsheet. We may have ten different ones. We'll pick some thresholds. Often, as you know, thresholds are arbitrary, but then you'll count how many different metrics does a sample fail on? And if it's just failing on one you may let it through, but normally looking at it in a multi varied fashion like that, you see a pretty clear separation between bad samples that will fail many metrics versus good samples that pass them all and then there is a couple on the boundaries that you have to make a judgment call.

Now, how do we correct batch effects? I spoke on this last year basically, we've employed a PCA based approach, sort of analogous to Eigenstrat. Eigenstrat on the Wellcome Trust data from the first two components shows these populations between case and controls are the same. Where as when you look at log ratio it's a huge shift, huge differences between the populations. And so you can correct for those shifts, using a principle components approach, you can basically remove a lot of the batch effects.

So here's kind of a before after. Green is before, blue is after. And here's the Wellcome Trust log ratio associations huge significance everywhere and then you clean it up and then you see a few interesting signals, actually most of them T cell artifacts, but at least there's some real biology there.

Now, I think in my abstract I sort of presented PCA as a magic bullet and over the last few months since I've submitted the abstract, I've found more problems and issues. There's a lot of assumptions behind it. We do need large sample sizes. We assume the first principle components contain the largest, contain the undesirable differences. But this is not true for large pedigrees for large family based studies, it's not a great method. And of course, you have to exclude sex chromosomes in your principle component decomposition.

The biggest problem we found is not all of the batch effects are corrected, in particular the copy number regions had a different variance structure sometimes than the neutral regions which represent the bulk of the data and so, the very thing we want to measure most sometimes has the most lingering batch effects. And so, we're, we have ongoing efforts to address them. The bottom line is better design and experiments is probably the best approach randomizing your case as it falls on plates in particular.

And we've also had some improvements that I don't have a great amount of time to go into in segmentation. Particularly with the Illumina data there is large negative outliers that need to be accounted for and they call it the skew in the distribution up here is and we've been able to tailor our approach to removing those outliers.

Some of them are real as we've found, comparing data and next generation sequencing versus Affy and Illumina platforms. Large negative outliers often actually are correlated with this is a Bentley Illumina study, a lot of the deletions found in this Bentley next generation sequencing study were covered only by single marker in the 500 K or 6.0 Illumina data and when you looked at the distribution of those, they were mostly negative.

So, a lot of those things that look like noise really are single marker deletions. Just briefly to kind of wrap up, when you get to associations beware T-cell regions, there's rearrangements that occur in T-cells that differ between cases and controls due to differences in extracting the DNA and so forth. Biggest problem we've run into is copy number intensities can vary between sites,

where here is cases in blue, controls in yellow, purple, and green. These cases look very different than these control, but these cases look a lot like these controls and so this was really a batch effect drive difference. This was a copy number segment that was standby about nine markers and it looked very exciting until you did the comparison.

Also beware, segmental duplications, you can find these really nice associations and it turns out there is a duplication on the X or Y chromosome and if you have some bias where there is more cases that are male or female you'll see significance in the agreement. So, I would like to thank you all and open the floor for any questions.

So, I'm not filtering SNPs by call rate or markers, I'm filtering samples by call rate, so good point. I would not want to do that.

All this data I've been showing you is quantile normalized and so, quantile normalization will address gross shifts in the intensities, but it just doesn't, these plate effects appear to be where each SNP changes in a different way and so you get large clusters of the m that will behave the same way and then you'll see an adjust in the first principle component and it gets some more in the next and so on, but the differences are not, you know, if it was just a constant shift, we wouldn't see associations. There's just differences in all of the markers, probably probe specific, sequence specific and you know whatever the environmental effects that are impacting the finding of any of these.

So, I guess I should stress that the biggest problem that we see in studies is the plate to plate variability and so that's different than the types of effects you're talking about. Those effects you're talking about are real as well, so, if the biggest take home would be, if we could just randomize properly with our plates as well as, if you know you've got different types of blood, different sources of DNA rather, you can't mitigate the differences in the DNA that are there, but at least you can mitigate how that might be distributed across the plates and you can randomize that.

I would say, if you have to do saliva, well, then also do saliva from the parents. Don't do blood from the parents and saliva from the children. Just whatever you can do to keep it as consistent as possible. And it's not to say that saliva or buccal cells have to be bad, I've seen good data from them. It's just I've seen more variability in that, so good quantification of your DNA and making

sure you have good amounts before you go through the procedure.  
Anything to lower your odds that you are going to have a problem.

*[End of Audio]*