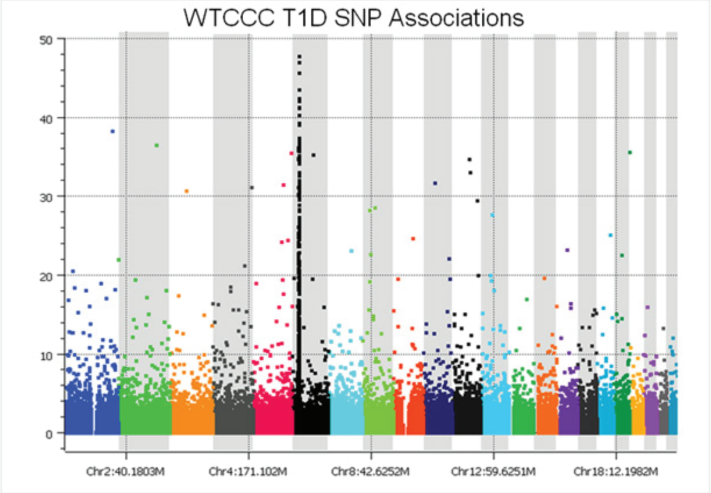
Stop Ignoring Experimental Design (or my head will explode)

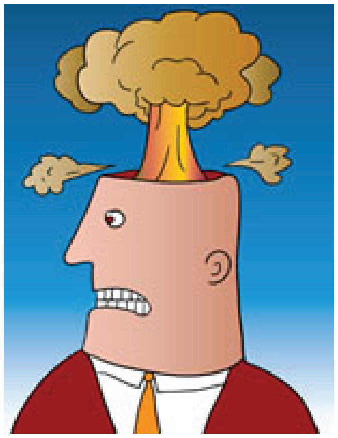
by Christophe Lambert, CEO & President of Golden Helix

Over the past 3 years, Golden Helix has analyzed dozens of public and customer whole-genome and candidate gene datasets for a host of studies. Though genetic research certainly has a number of complexities and challenges, the number one problem we encounter, which also has the greatest repercussions, is born of problematic experimental design. In fact, about 95% of the studies that we analyzed had major problems with experimental design. Namely, some aspect of data collection or experimental order (i.e. plating) is not randomized with respect to the phenotypes of interest. The unfortunate result is endless struggles with spurious associations due to confounding, to the point, in fact, where real associations cannot be distinguished from experimental artifacts. This confounding only gets worse when two or more poorly randomized experiments are combined with the goal of increasing power through mega-analyses. A timely example can be found in the flaws revealed in the recent Sebastiani et al. Science paper, "Genetic signature of exceptional longevity in humans", which appear to result almost entirely from the analysis not taking into account batch-effect driven spurious associations.

How is it that this easily avoidable problem is almost universal? We've seen poorly designed experiments come out of some of the top GWAS research centers in the world. We have been sounding the alarm with increasing intensity over the past couple of years, and I recently spoke on this topic at the CSCDA conference in Belgium in the context of CNV analysis.



Example from the Wellcome Trust Case Control Consortium



Consider the famous Wellcome Trust Study that some say started the golden age of GWAS, ("Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls", Nature 2007). The consortium performed the genotyping for the two control populations in two distinct sets of plates, and each of the various disease studies in their own separate set of plates.

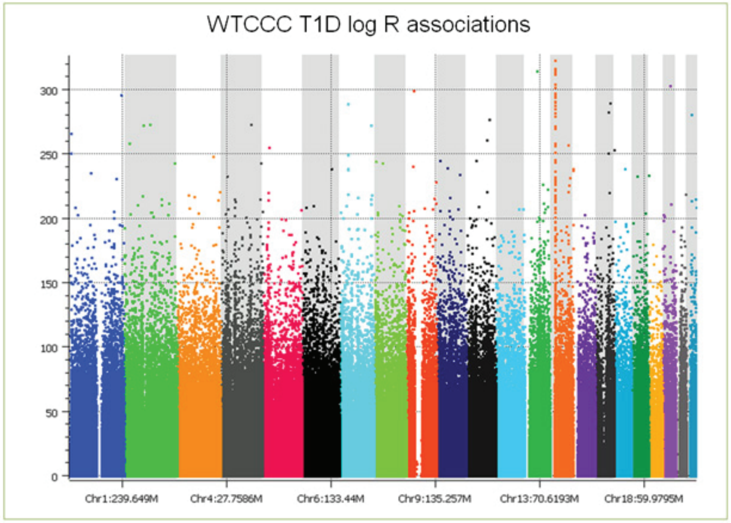
The question that comes to mind most often is why, if this is a problem, are the study's Manhattan plots so clean? This is because the Manhattan plots included in the paper are the result of careful manual curation and are the result of discarding of hundreds of spurious associations. As an example, before SNP quality control, the Manhattan plot of Type I diabetes cases vs. the National Blood Service controls looks the plot at left.

At a glance, the plot might not seem too bad – there is a nice peak in the HLA region of chromosome 6, plus several associations in the upper portion of the graph. However, considering that a p-value of 5e-8 is regarded as the threshold for genome-wide significance, it is clear that there are easily over 100 spurious associations. These associations are directly due to the measurement error varying between the cases and the controls because they were run in different batches, versus block randomizing the seven sets of cases and common controls across all of the plates.

To resolve this, the analysts devised a set of clever filters for dropping the SNPs that were most likely to have spurious associations. Filtering, for instance, on call rate, Hardy-Weinberg equilibrium, minor allele frequency, and so forth. This was an appropriate strategy given the money had been spent on the experiment and the samples had already been genotyped. However, if the experiment had been properly randomized, no filters would have been necessary. Unfortunately, this band-aid practice of applying filters to SNPs has been followed in GWAS publications ever since, instead of addressing the problem at its source – the experimental design.

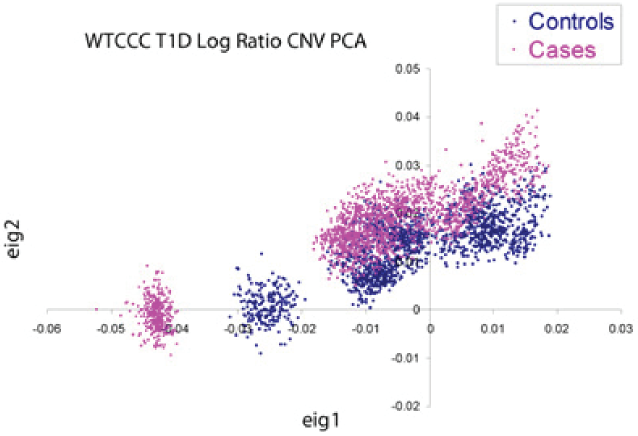
Some might say that filters solve the problem. However, this is not so. Not only must some SNP studies drop 30% or more of the SNPs, but haplotype and gene-gene interaction analyses often surface additional spurious associations even when stringent filters are applied. This gets worse with imputation and mega-analysis projects.

But that's not the end of the story; things get worse still when you examine the copy number intensity data. When association tests are performed on the log ratios, you see the entire genome covered with false positive associations (see plot below), making copy number analysis problematic at best. With CNV analysis, no amount of filtering can remove batch effects. We have developed a number of approaches including principal components correction to treat the problem, however, there always remains confounding that can just not be fixed. Further, these batch-based differences, while alleviated somewhat, do not go away by making discrete calls through averaging across multiple consecutive markers in a copy number region. (As you examine this plot, keep in mind that 5e-8 is considered significant, a point just below the first tick mark on the Y axis.)



If we dig into this further by performing a principal components decomposition of the covariance matrix of the copy number data, we see there is significant non-random structure in the data both with and between the case/control data set. Pictured below are the cases and controls plotted against the first two principal components of the matrix. Devoid of plate-to-plate variation, these clusters should overlap perfectly and be virtually indistinguishable from each other.

A few years later, the Wellcome Trust reran their samples, this time with specialized copy number arrays. Unfortunately the same type of mistake was made again, but with a twist: they put two diseases on each plate instead of block randomizing all eight of them along with their controls (see supplementary information 1 for "Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls", Nature 2010). The authors' rationale for their "phased randomized cohort screening" was not about controlling for sources of variability, but rather as a way to get some studies done earlier (including all of the controls) to "facilitate piloting of data analysis pipelines". When this design was described to a Fellow of the American Statistical Association, he was shocked that such a reputable organization could possibly make such a fundamental experimental design error. Is it any wonder many copy number studies are coming up empty other than some success stories of finding very large rare variants spanned by dozens to hundreds of markers? (Yes, there are other reasons too, but that is the topic of my last blog post.)



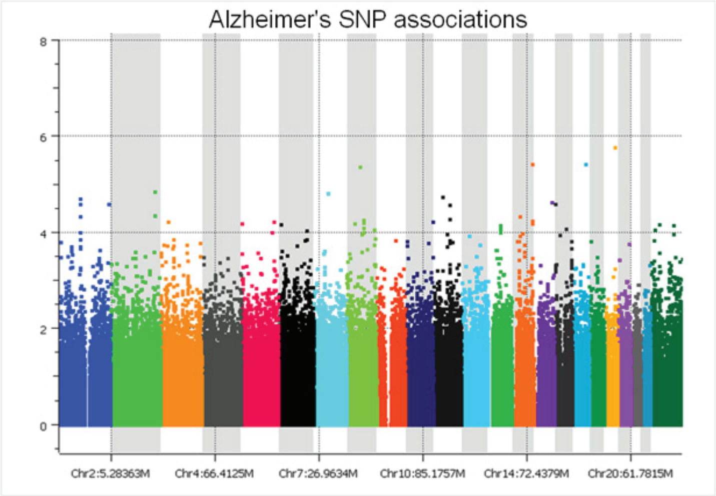
Why does this travesty continue?

One would assume that if researchers just knew better they would do better, and I believe this to be so. However, why is knowledge of design of experiments (DOE) not universal? DOE is not a new field. Heck, Fisher founded the field of experimental design in 1935 in his book, Design of Experiments. Scientists run experiments for a living; one would presume they have been schooled in the need to control for sources of variability when running an experiment. GWAS studies are just another experiment. Have scientists made an exception for GWAS and thrown design of experiments principles out the window? Or do they just not know what the biggest sources of variability are and thus don't know how to randomize for them? Or did industry's promise of 99% genotyping call rates just lead researchers to conclude that accuracy is assured and experimental variability is, therefore, a non-issue?

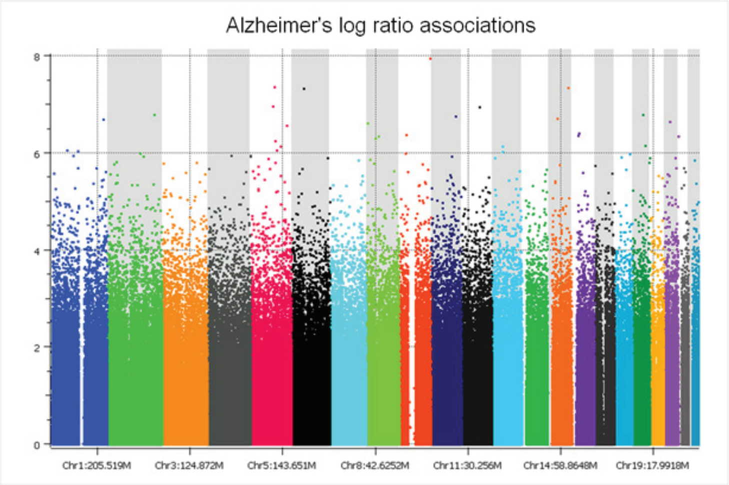
As I've traveled around the world and spoken with statistical analysts, I often feel I'm preaching to the choir. They readily agree and proceed to tell me how they are never asked to help with design before the experiment begins, only asked to clean up the mess after millions have been spent. I can only conclude, at least where the rubber hits the road with the wet work, many scientists are woefully inadequately trained in design of experiments. I also think that in some cases the problem stems from the division of labor in genotyping centers (the person that plates the samples is not the one that processes them, who is not the one that analyzes the data, and negative feedback never makes it back up the chain). I look back to my own education and cannot recall ever being formally taught DOE. I was just fortunate enough to be given on-the-job training while a graduate research assistant at a pharmaceutical company.

What does good design of experiments look like?

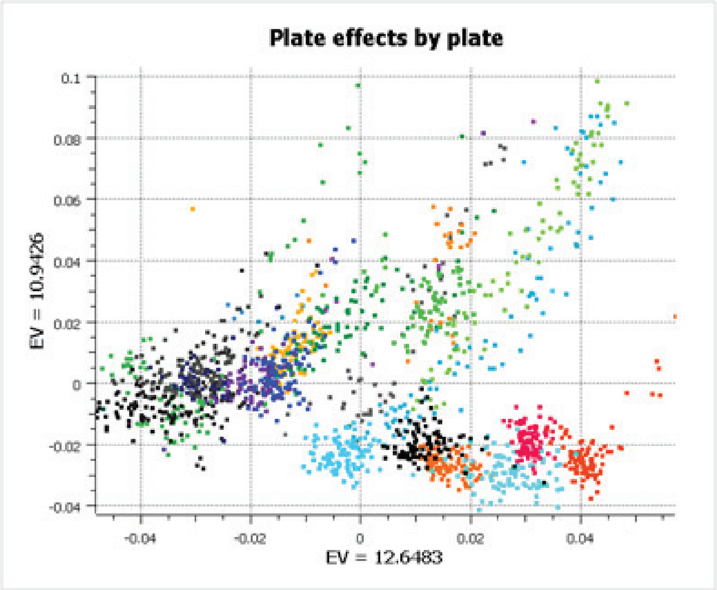
There are only two studies we have analyzed that did a proper randomized block design across plates – the GenADA Alzheimer's study from GlaxoSmithKline and nine medical centers from Canada, presumably done by clinical trial statisticians well versed in design of experiments; and a large Parkinson's study for which a Golden Helix statistician, Greta Peterson, designed the randomization scheme ("Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease", Nature Genetics 2010). In both studies, spurious associations due to plate effects were negligible. The following Manhattan plot for the Alzheimer's study was clean without filtering a single SNP!



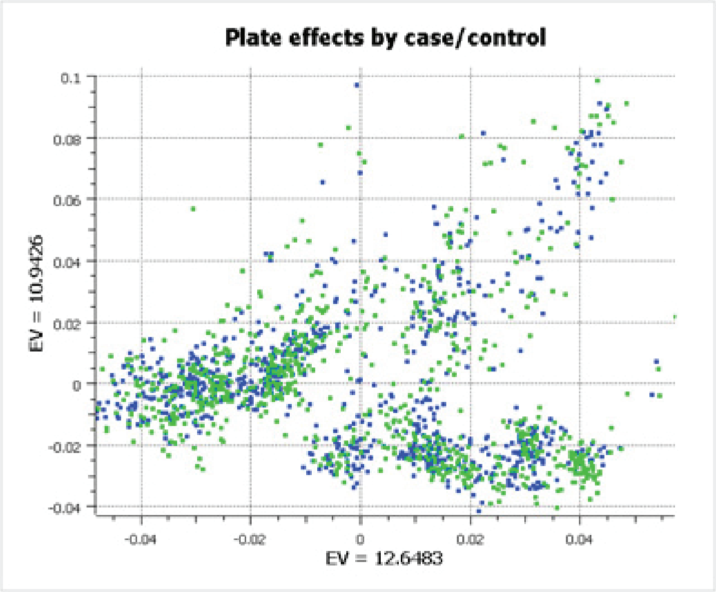
The copy number log ratio associations were nearly as clean, despite the potential confounding of DNA contributions from 10 different sites. To really understand the difference, compare the Y axis of the below plot with that of the previous T1D log R plot.



What is striking about this data is that there are large plate effects, but the point is that these plate effects are random with respect to case/control status and so don't show up as significant during analysis. If you look at the first two principal components of the data (below), color-coded by plate, you see very clear clusters, caused entirely by plate effects (not population stratification).



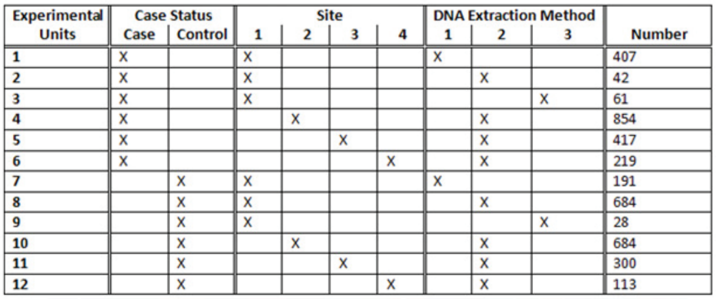
However, when you look at the same data color-coded by case/control status (page 4), you see that the data has been properly randomized. Each plate cluster contains equal proportions of cases and controls, eliminating spurious associations due to plate effects.



How to do it right

When plating samples, we strongly recommend a randomized block design. Below is an example illustrating the concepts by which we designed the plate layout for a 4,000-subject Parkinson's study run on an Illumina HumanOmni1-Quad_v1-0_B platform.

The study had 2,000 cases and 2,000 controls collected from four different sites using three DNA extraction methods. What could have been an experimental nightmare was actually not difficult to manage. We considered various design options, including randomizing on additional variables such as gender and age. The following table illustrates a block randomization involving case/control status, site, and DNA extraction method. Each row of the table contains counts of the various experimental units to be randomized across plates. The randomization ultimately employed in the study was somewhat more extensive, involving additional experimental units, but this illustrates the general concept of block randomization.



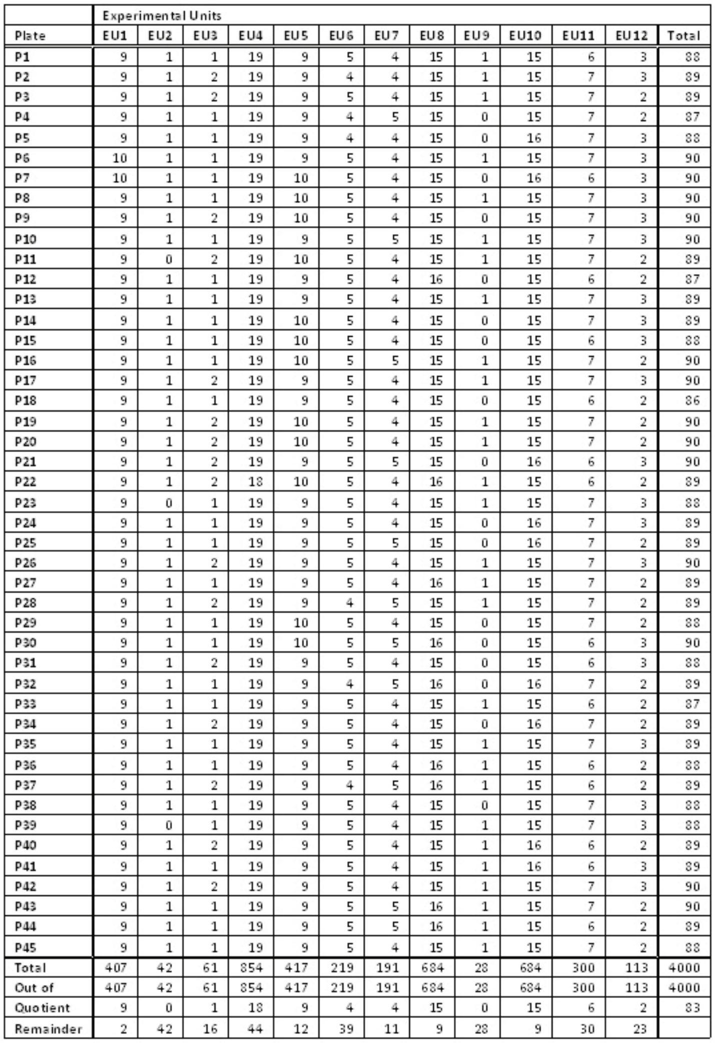
Experimental	Case Status		Site				DNA Extraction Method			
Units	Case	Control	1	2	3	4	1	2	3	Number
1	X		X				X			407
2	X		X					X		42
3	X		X						X	61
4	X			X				X		854
5	X				X			X		417
6	X					X		X		219
7		X	X				X			191
8		X	X					X		684
9		X	X						X	28
10		X		X				X		684
11		X			X			X		300
12		X				X		X		113

The case/control status is the most important variable to randomize. With quantitative traits, some form of discretization into experimental units can be employed. While plating will not remove the confounding due to the non-ideal data collection of different numbers of cases and controls by site and DNA extraction method, we can at least ensure these will not be further confounded with plate artifacts. That is, if we need to correct for data distortions due to site or DNA kit later, those can be done with a handful of dummy variables in a logistic regression, rather than an untenable cross-product of 45 plate variables with those dummy variables. (In an ideal world we would be able to block design the data collection as well, but often an experimenter must take the data as it comes.)

How does this table translate to plating? The 407 samples from Experimental Unit 1 containing cases from Site 1 with DNA Extraction Method 1 would be evenly divided at random among the 45 plates, resulting in either nine or ten samples per plate (never off by more than one, see the "EU1" column in the below table). We similarly divide Experimental Units 2 through 12 to construct 45 plates with 90 samples or less (see figure on page 5). The remaining six wells on each plate were reserved for a male and female control sample and four other replicates.

Note that it is not sufficient for good experimental design to randomly place cases and controls on plates — it is key that the randomization is controlled so that the experimental units are almost perfectly balanced. We have seen studies that placed the samples at random, with many plates by chance alone being quite unbalanced and thus creating spurious plate-driven associations.

As mentioned above, the results of block randomization were clearly evident once the samples were processed. The Manhattan and Q-Q plots were excellent, and the inevitable plate-based experimental artifacts did not confound the results. Again, without filtering a single SNP for low call rate or departures from HWE, there were essentially no spurious associations (data not shown). The publication actually did apply standard SNP quality filters; but they were really unnecessary other than to follow convention. With proper DOE one can go back to manually inspecting and applying thoughtful scientific inquiry to the handful of extreme values to see if they are real or artifacts, rather than resorting to automated filters. We would like to note that it is entirely possible to have your genotyping center work with you on getting a good plate layout; CIDR was extremely cooperative and helpful in running the experiment with our final design. We'd also like to acknowledge Haydeh Payami for her vision in allocating the resources up front on the experimental design to improve the outcome of the study.



	Experimental Units												
Plate	EU1	EU2	EU3	EU4	EU5	EU6	EU7	EU8	EU9	EU10	EU11	EU12	Total
P1	9	1	1	19	9	5	4	15	1	15	6	3	88
P2	9	1	2	19	9	4	4	15	1	15	7	3	89
P3	9	1	2	19	9	5	4	15	1	15	7	2	89
P4	9	1	1	19	9	4	5	15	0	15	7	2	87
P5	9	1	1	19	9	4	4	15	0	16	7	3	88
P6	10	1	1	19	9	5	4	15	1	15	7	3	90
P7	10	1	1	19	10	5	4	15	0	16	6	3	90
P8	9	1	1	19	10	5	4	15	1	15	7	3	90
P9	9	1	2	19	10	5	4	15	0	15	7	3	90
P10	9	1	1	19	9	5	5	15	1	15	7	3	90
P11	9	0	2	19	10	5	4	15	1	15	7	2	89
P12	9	1	1	19	9	5	4	16	0	15	6	2	87
P13	9	1	1	19	9	5	4	15	1	15	7	3	89
P14	9	1	1	19	10	5	4	15	0	15	7	3	89
P15	9	1	1	19	10	5	4	15	0	15	6	3	88
P16	9	1	1	19	10	5	5	15	1	15	7	2	90
P17	9	1	2	19	9	5	4	15	1	15	7	3	90
P18	9	1	1	19	9	5	4	15	0	15	6	2	86
P19	9	1	2	19	10	5	4	15	1	15	7	2	90
P20	9	1	2	19	10	5	4	15	1	15	7	2	90
P21	9	1	2	19	9	5	5	15	0	16	6	3	90
P22	9	1	2	18	10	5	4	16	1	15	6	2	89
P23	9	0	1	19	9	5	4	15	1	15	7	3	88
P24	9	1	1	19	9	5	4	15	0	16	7	3	89
P25	9	1	1	19	9	5	5	15	0	16	7	2	89
P26	9	1	2	19	9	5	4	15	1	15	7	3	90
P27	9	1	1	19	9	5	4	16	1	15	7	2	89
P28	9	1	2	19	9	4	5	15	1	15	7	2	89
P29	9	1	1	19	10	5	4	15	0	15	7	2	88
P30	9	1	1	19	10	5	5	16	0	15	6	3	90
P31	9	1	2	19	9	5	4	15	0	15	6	3	88
P32	9	1	1	19	9	4	5	16	0	16	7	2	89
P33	9	1	1	19	9	5	4	15	1	15	6	2	87
P34	9	1	2	19	9	5	4	15	0	16	7	2	89
P35	9	1	1	19	9	5	4	15	1	15	7	3	89
P36	9	1	1	19	9	5	4	16	1	15	6	2	88
P37	9	1	2	19	9	4	5	16	1	15	6	2	89
P38	9	1	1	19	9	5	4	15	0	15	7	3	88
P39	9	0	1	19	9	5	4	15	1	15	7	3	88
P40	9	1	2	19	9	5	4	15	1	16	6	2	89
P41	9	1	1	19	9	5	4	15	1	16	6	3	89
P42	9	1	2	19	9	5	4	15	1	15	7	3	90
P43	9	1	1	19	9	5	5	16	1	15	7	2	90
P44	9	1	1	19	9	5	5	16	1	15	6	2	89
P45	9	1	1	19	9	5	4	15	1	15	7	2	88
Total	407	42	61	854	417	219	191	684	28	684	300	113	4000
Out of	407	42	61	854	417	219	191	684	28	684	300	113	4000
Quotient	9	0	1	18	9	4	4	15	0	15	6	2	83
Remainder	2	42	16	44	12	39	11	9	28	9	30	23	

Closing thoughts

The bottom line is we need to stop the inefficient usage of taxpayer money inherent in running large-scale genetic studies without proper DOE. Perhaps it is time for the NIH to set forth policy on this before further money is wasted. Perhaps grant reviewers should insist on experts in experimental design being involved before the experiments are run. It most certainly is time for every field that calls itself a science to devote teaching time to the theory and extensive hands-on practice of design of experiments. And then my head will not explode!

...And that's my two SNPs.

About Christophe Lambert

Dr. Christophe Lambert is the President and CEO of Golden Helix, Inc., a bioinformatics software and services company he founded in Bozeman, MT, USA in 1998. Dr. Lambert graduated with his Bachelors in Computer Science from Montana State University in 1992 and received his Ph.D. in Computer Science from Duke University in 1997. He has performed interdisciplinary research in the life sciences for over twenty years. Dr. Lambert is also currently the co-chair of the Genome Wide Copy Number Variation Data Analysis Team of the Microarray Quality Control Consortium.

About Golden Helix

We are inspired by significance. Not only statistical, but technological, scientific, and personal significance. It's embodied in everything we and our customers do. And we believe the only way to achieve significance is by transcending the status quo. Every day we strive for extraordinary analytic and technological advancements that empower scientists around the world to pursue that which is truly significant: from uncovering the genetic causes of disease and transforming drug discovery to developing genetic diagnostics and advancing the quest for personalized medicine. To learn more about Golden Helix, visit www.goldenhelix.com.