PROFESSOR: So as you recall last time we talked about chromatin structure and chromatin regulation. And now we're going to move on to genetic analysis. But before we did that, I want us to touch on two points that we talked about briefly last time.

One was 5C analysis. Who was it that brought up-- who was the 5C expert here? Anybody? No? Nobody wants to own 5C. OK.

But as you recall, we talked about ChIA-PET as one way of analyzing any to any interactions in the way that the genome folds up and enhancers talk to promoters. And 5C is a very similar technique.

I just wanted to show you the flow chart for how the protocol goes. There is a cross linking. A digestion with a restriction enzyme step, followed by a proximity ligation step, which gives you molecules that had been brought together by an enhancer, promoter complex, or any other kind of distal protein-protein interaction.

And then, what happens is that you design specific timers to detect those ligation events. And you sequence the result of what is known as ligation mediated amplification. So those primers are only going to ligate if they're brought together at a particular junction, which is defined by the restriction sites lining up.

So, 5C is a method of looking at which regions of the genome interact and can produce these sorts of results, showing which parts of the genome interact with one another.

The key difference, I think, between chIA-PET and 5C is that you actually have to have these primers designed and pick the particular locations you want to query.

So the primers that you design represent query locations and you can then either apply the results to a microarray, or to high throughput sequencing to detect these interactions.

But the essential idea is the same. Where you do proximity based ligation to form molecules that contain components of two different pieces of the genome that have been brought together for some functional reason.

The next thing I want to touch upon was this idea of the CpG dinucleotides that are connected by a phosphate bond. And you recall that I talked about the idea that they were symmetric.

So you could have methyl groups on the cytosines in such a way that, because they could mirror one another, they could be transferred from one strand of DNA to the other strand of DNA, during cell replication by DNA methyltransferase.

So it forms a more stable kind of mark and as you recall, DNA methylation where something occurred in lowly expressed genes and typically in regions of the genome that are methylated. Other histone marks are not present and the genes are turned off.

OK. So those were the points I wanted to touch upon from last lecture. Now we're going to embark upon an adventure, looking for the answer to, wear is missing heritability found?

So it's a big open question now in genetics. In human genetics, which is that we really can't find all the heritability. And as a point of introduction, the narrative arc for today's lecture is that, generally speaking, you're more like your relatives than random people on the planet.

And why is this? Well obviously you contain components of your mom and dad's genomes. And they are providing you with components of your traits. And the heritability of a trait is defined by the fraction of phenotypic variance that can be explained by genetics.

And we're going to talk today about computational models that can predict phenotype from genotype. And this is very important, obviously, for understanding the sources of various traits and phenotypes. As well as fields such as pharmacogenomics that try and predict the best therapy for a disease based upon your genetic makeup.

So, individual loci in the genome that contribute to quantitative traits are called quantitative trait locis, or QTLs. So we're going to talked about how to discover them and how to build models of quantitative traits using QTLs.

And finally, as I said at the outset, our models are insufficient today. They really can't find all of the heritability. So we're going to go searching for this missing heritability and see where it might be found.

Computationally, we're going to apply a variety of techniques to these problems. A preview is, we're going to build linear models of phenotype and we're going to use stepwise regression to learn these models using a forward feature selection.

And I'll talk about what that is when we get to that point of the lecture. We're going to derive test statistics for discovering which QTLs are significant and which QTLs are not, to include in our model.

And finally, we're going to talk about how to measure narrow sense heritability and broad sense heritability in environmental variance.

OK. So, one great resource for traits that are fairly simple. That primarily are the result of a single gene mutation, or where a single gene mutation plays a dominant role, is something called Online Mendelian Inheritance in Man.

And it's a resource. It has about 21,000 genes in it right now. And it's a great way to explore what human genes function is in various diseases. And you could query by disease. You can query by gene. And it is a very carefully annotated and maintained collection that is worthy of study, if you're interested in particular disease genes.

We're going to be looking at more complex analyses today. The analyses we're going to look at are where there are many genes that influence a particular trait.

And we would like to come up with general methods for discovering how we can de novo from experimental data-- discover all the different genes that participate.

Now just as a quick review of statistics, I think that we've talked before about means in class and variances. We're also going to talk a little bit about covariances today. But these are terms that you should be familiar with as we're looking today at some of our metrics for understanding heritability.

Are there any question about any of the statistical metrics that are up here? OK.

So, a broad overview of genotype to phenotype. So, we're primarily going to be working with complete genome sequences today, which will reveal all of the variance that are present in the genome.

And it's also the case that you can subsample a genome and only observe certain variance. Typically that's done with microarrays that have probes that are specific to particular markers.

The way those arrays are manufactured is that whole genome sequencing is done at the outset, and then high prevalence variance, at least common variance, which typically are at a frequency of at least 5% in the population are queried by using a microarray. But today we'll talk about complete genome sequence.

An individual's phenotype, we'll say is defined by one or more traits. And a non-quantitative trait is something perhaps as simple as whether or not something is dead or alive. Or whether or not it can survive in a particular condition. Or its ability to produce a particular substance.

A quantitative trait, on the other hand, is a continuous variable. Height, for example, of an individual is a quantitative trait. As is growth rate, expression of a particular gene, and so forth.

So we'll be focusing today on estimating quantitative traits. And as I said, a quantitative trait or loci, is a marker that's associated with a quantitative trait and could be used to predict it. And you can sometimes hear about eQTLs, which are expression quantitative trait loci. And they're loci that are related to gene expression.

So, let's begin then, with a very simple genetic model. It's going to be haploid, which means, of course, there's only one copy of each chromosome. Yeast is the model organism we're going to be talking about today. It's a haploid organism.

And we have mom and dad up there. Mom on the left, dad on the right in two different colors. And you can see that mom and dad in this particular example, have n different genes. They're going to contribute to the F1 generation, to junior.

And the relative color is white for mom, black for dad, are going to be used to describe the alleles, or the allelic variance that are inherited by the child, the F1 generation.

And as I said, a specific phenotype might be alive or dead in a specific environment. And note that I have drawn the chromosomes to be disconnected. Which means that each one of those genes is going to be independently inherited.

So the probability in the F1 generation that you're going to get one of those from mom or dad is going to be a coin flip. We're going to assume that they're far enough away that the probability of crossing over during meiosis is 0.5. And so we get a random assortment of alleles from mom and dad. OK?

So let us say that you go off and do an experiment. And you have 32 individuals that you produce out of a cross. And you test them, OK. And two of them are resistant to a particular substance.

How many genes do you think are involved in that resistance? Let's assume that mom is resistant and dad is not. OK. If you had two that were resistant out of 32, how many different genes do you think were involved? How do you estimate that? Any ideas? Yes?

AUDIENCE: If you had 32 individuals and say half of them got it?

PROFESSOR: Two, let's say. One out of 16 is resistant. And mom is resistant.

AUDIENCE: Because I was thinking that if it was half of them were resistant, then you would maybe guess one gene, or something like that.

PROFESSOR: Very good.

AUDIENCE: So then if only eight were resistant you might guess two genes, or something like that?

PROFESSOR: Yeah. What you say is, that if mom's resistant, then we're going to assume that you need to get the right number of genes from mom to be resistant. Right? And so, let's say that you had to get four genes from mom. What's the chance of getting four genes from mom?

AUDIENCE: Half to the power of four.

PROFESSOR: Yeah, which is one out of 16, right? So, if you, for example had two that were resistant out of 32, the chances are one in 16. Right? So you would naively think, and properly so, that you had to give four genes from mom to be resistant.

So the way to think about these sorts of non-quantitative traits is that you can estimate the number of genes involved. The simply is log base 2 over the number of F1s tested over the number of the F1s with the phenotype.

It tells you roughly how many genes are involved in providing a particular trait, assuming that the genes are unlinked. It's a coin flip, whether you get them or not.

Does everybody see that? Yes? Any questions at all about that? About the details? OK.

Let's talk now about quantitative traits then. We'll go back to our model and imagine that we have the same set-- actually it's going to a different set of n genes. We're going to have a coin flip as to whether or not you're getting a mom gene or a dad gene. OK. And each gene in dad has an effect size of 1 over n. Yes?

AUDIENCE: I just wanted to check. We're assuming that the parents are homozygous for the trait? Is that correct?

PROFESSOR: Remember these are haploid.

AUDIENCE: Oh, these are haploid.

PROFESSOR: Right. So they only have one copy of all these genes. All right. Yes?

AUDIENCE: [INAUDIBLE] resistant and they're [INAUDIBLE]. That could still mean that dad has three of the four genes in principle.

PROFESSOR: The previous slide? Is that where what you're talking about?

AUDIENCE: [INAUDIBLE] knew about it. So really what you mean is that dad does not have any of the genes that are involved with resistance.

PROFESSOR: The correct. I was saying that dad has to have all of gene-- that the child has to have all of the genes that are operative to create resistance. We're going to assume an AND model. He must have all the genes from mom. They're involved in the resistance pathway.

And since only one out of a 16 progeny has all those genes from mom, right, it appears that given the chance of inheriting something from mom is 1/2, that it's four genes you have to inherit from mom. Because the chance of inheriting all four is one out of 16.

AUDIENCE: [INAUDIBLE] in which case--

PROFESSOR: No, I'm assuming the dad doesn't have any of those. But here we're asking, what is the difference in the number of genes between mom and dad?

So you're right, that the number we're computing is the relative number of genes different between mom and dad you require. And so it might be that dad's a reference and we're asking how many additional genes mom brought to the table to provide with that resistance. But that's a good point. OK. OK.

So, now let's look at this quantitative model. Let's assume that mom has a bunch of genes that contribute zero to an effect size and dad-- each gene that dad has produces an effect of 1 over n. So the total effect size here for dad is 1.

So the effect of mom on this particular quantitative trait might be zero. It might be the amount of ethanol produced or some other quantitative value. And dad, on the other hand, since he has n genes, is going to produce one, because each gene contributes a little bit to this quantitative phenotype. Is everybody clear on that?

So, the child is going to inherit genes to our coin flip between mom and dad, right. So the first fundamental question is, how many different levels are there in our quantitative phenotype in our trait? How many different levels can you have?

AUDIENCE: N + 1?

PROFESSOR: N + 1, right, because you can either inherit zero, or up to n genes from dad. And it gets you n plus 1 different levels. OK.

So, what's the probability then-- well, I'll ask a different question. What's the expected value of the quantitative phenotype of a child? Just looking at this.

If dad's one and mom's zero, and you have a collection of genes and you do a coin flip each time, you're going to get half your genes from mom and half your genes from dad. Right.

And so the expected trait value is 0.5. So for these added traits, you're going be at the midpoint between mom and dad. Right. And what is the probability that you inherit x copies of dad's genes?

Well, that's n choose x, times 1 minus .5 n to the minus x times 0.5 to the x. A simple binomial. Right. So if you look at this, the probability of the distribution for the children is going to look something like this, where this is the mean, 0.5.

And the number of distinct values is going to be n plus 1. Right. So the expected value of x is 0.5 and turns out that the expected value, or the variance of x minus 0.5, which is the mean squared, is going to be 0.25 over n.

So I can show you this on the next slide. So you can see, this could be ethanol production, it could be growth rate, what have you. And you can see that the number of genes that you're going to get from dad follows this binomial distribution and gives you a spread of different phenotypes in the child's generation, depending upon how many copies of dad's genes that you inherit.

But does this make sense to everybody? Now would be a great time to ask any questions about the details of this. Yes?

AUDIENCE: Can you clarify what x is? Is x the fraction of genes inherited--

PROFESSOR: The number of genes you inherit from dad. The number of genes. So it would zero, one, two, up to n.

AUDIENCE: Shouldn't the expectation of n [INAUDIBLE] x be n/2?

PROFESSOR: I'm sorry. It is supposed to be n/2. But the last two expectations are some of the number of genes you've inherited from dad. Right, that's correct. Yeah, this slide's wrong. Any other questions? OK.

So this is a very simple model but it tells us a couple of things, right. Which is that as n gets to be very large, the effect of each gene gets to be quite small.

So something could be completely heritable, but if it's spread over, say 1,000 genes, then it will be very difficult to detect, because the effect of each gene would be quite small. And furthermore, the variance that you see in the offspring will be quite small as well, right, in terms of the phenotype. Because it's going to be 0.25/n in terms of the expected value.

So as n gets larger, the number genes that contribute to that phenotype increase, the variance is going to go down linearly. OK. So we should just keep this in mind as we're looking at discovering these sort of traits and the underlying QTLs that can be used to predict them.

And finally, I'd like to point out one other detail which is that, if genes are linked, that is, if they're in close proximity to one another in the genome and it makes it very unlikely there's going to be crossing over between them, then they're going to act as a unit. And if they act as a unit, then we'll get marker correlation. And you can also see, effectively, that the effect size of those two genes is going to be larger.

And in more complicated models, we obviously wouldn't have the same effect size for each gene. The effect size might be quite large for some genes, might be quite small for some genes. And we'll see the effects of marker correlation in a little bit.

So the way we're going to model this is we're going to-- this is a definition of the variables that we're going to be talking about today. And the essential idea is quite simple.

So the phenotype of an individual-- so p sub i is the phenotype of an individual, is going to be equal to some function of their genotype plus an environmental component. This function is the critical thing that we want to discover.

This function, f, is mapping from the genotype of an individual to its phenotype. And the environmental component could be how well something is fed, how much sunlight it gets, things that can greatly influence things like growth but they're not described by genetics.

But this function is going to encapsulate what we know about how the genetics of a particular individual influences a trait. And thus, if we consider a population of individuals, the phenotypic variance is going to be equal to the genotypic variance plus the environmental variance plus two times the covariance between the genotype in the environment.

And we're going to assume, as most studies do, that there is no correlation between genotype and environment. So this term disappears. So what we're left with is that the observed phenotypic variance is equal to the genotypic variance plus the environmental variance.

And what we would like to do is to come up with a function f, that best predicts the genotypic component of this equation. There's nothing we can do about environmental variance. Right. But we can measure it. Does anybody have any ideas how we could measure environmental variance? Yes?

AUDIENCE: Study populations in which there's some kind of controlled environment. So you study populations that one population is one with a homogeneous. And another one was a completely different one.

PROFESSOR: Right. So what we could do is we could use controls. So typically what we could do is we could study in environments where we try and control the environment exactly to eliminate this as much as we possibly can, for example. As we'll see that we also can do things like study clones, where individuals have exactly the same genotype. And then, all of the variance that we observe-- if this term vanishes because the genotypes are identical, it is due to the environment.

So typically, if you're doing things like studying humans, since cloning humans isn't really a good idea to actually measure environmental variance, right, what you could do is you can look at identical twins. And identical twins give you a way to get at the question of how much environment variance there is for a particular phenotype.

So in sum, this is replicates what I have here on the left-hand side of the board. And note that today we'll be talking about the idea of discovering this function, f, and how well we can discover f, which is really important, right. It's fundamental to be able to predict phenotype from genotype. It's an extraordinarily central question in genetics. And when we do the prediction, there are two kinds of-- oh, there's a question?

AUDIENCE: Could you please explain again why the co-variance drops out or it goes away.

PROFESSOR: Yeah, the co-variance drops out because we're going to assume that genotype and environment are independent. Now if they're not independent, it won't drop out. But making that assumption-- and of course, for human studies you can't really make that assumption completely, right?

And one of the problems in doing these sorts of studies is that it's very, very easy to get confounded. Because when you're trying to decompose the observed variance and height, for example.

You know, there's what mom and dad provided to an individual in terms of their height, and there's also how much junior ate, right. And whether he went to McDonald's a lot, or you know, was going to Whole Foods a lot. You know, who knows, right?

But this component and this component, it's easy to get confounded between them and sometimes you can imagine that genotype is related to place of origin in the world. And that has a lot to do with environment. And so this term wouldn't necessarily disappear.

OK. So there are two kinds of heritability I'd like to touch upon today. And it's important that you remember there are two kinds and one is extraordinarily difficult to recover and the other one is in some sense, a more constrained problem, because we're much better at building models for that kind of heritability estimate.

The first is broad-sense heritability, which describes the upper bound for phenotypic prediction given an arbitrary model. So it's the total contribution to phenotypic variance from genetic causes. And we can estimate that, right. And we'll see how we can estimate it in a moment.

And narrow-sense heritability is defined as, how much of the heritability can we describe when we restrict f to be a linear model. So when f is simply linear, as the sum of terms, that describes the maximum narrow-sense heritability we can recover in terms of the fraction of phenotypic variance we can capture in f.

And it's very useful because it turns out that we can compute both broad-sense and narrow-sense heritability from first principles-- I mean from experiment. And the difference between them is part of our quest today.

Our quest is, to answer the question, where is the missing heritability? Why can't we build an Oracle f that perfectly predicts phenotype from genotype?

So on that line-- I just want to give you some caveats. One is that we're always talking about populations when we're talking about heritability because it's how we're going to estimate it.

And when you hear people talk about heritability, oftentimes they won't qualify it in terms of whether it's broad-sense or narrow-sense. And so you should ask them if you're engaged in a scientific discussion with them.

And as we've already discussed, sometimes estimation is difficult because of matching environment and eliminating this term, the environmental term can be a challenge when you're out of the laboratory. Like when you're dealing with humans.

So, let's talk about broad-sense heritability. Imagine that we measure environmental variants simply by looking at environmental twins or clones, right.

Because if we, for example, take a bunch of yeast that are genotypically identical. And we grow them up separately, and we measure a trait like how well they respond to a particular chemical or their growth rate, then the variance we see from each individual to individual is simply environmental, because they're genetically identical. So

we can, in that particular case, exactly quantify the environmental variance given that every individual is genetically identical. We simply measure all the growth rates and we compute the variance. And that's the environmental variance. OK?

As I said for humans, the best we can do is identical twins. Monozygotic twins. You can go out and for pairs of twins that are identical, you can measure height or any other trait that you like and compute the variance. And then that is an estimate of the environmental component of that, because they should be genetically identical.

And big H squared-- broad-sense is always capital H squared and narrow-sense is always little h squared. Big H squared, which is broad-sense heritability is very simple then.

It's the phenotypic variance, minus the environmental variance, over the phenotypic variance. So it's the fraction of phenotypic experience that can be explained from genetic causes. Is that clear to everybody? Any questions at all about this? OK.

So, for example, on the right-hand hand side here, those three purplish squares have three different populations, which are genotypically identical. They have two genes, a little a, a little a, big A, a little A, and big A, big A. And each one is a variance of 1.0. out So since there are genetically identical, we know that the environmental variance has to be 1.0.

On the left-hand side, you see the genotypic variance. And that reminds us of where we started today. It depends on the number of alleles you get of big A, as to what the value is.

And when you put all of that together, you get a total variance of 3. And so big H squared is simply the genotypic variance, which is 2, over the total phenotypic variance, which is 3. So big H squared is 2/3. And so that is a way of computing broad-sense heritability.

Now, if we think about our models, we can see that narrow-sense heritability has some very nice properties. Right. That is, if we build and add a model of phenotype, to get at narrow-sense heritability.

So if we were to constraint f here to be linear, it's simply going to be a very simple linear model. For each particular QTL that we discover, we assign an effect size beta to it, or a coefficient that describes its deviation from the mean for that particular trait. And we have an offset, beta zero.

So our simple linear model is going to take all the discovery QTLs that we have--take each QTL and discover which allelic form it's in. Typically it's considered either in zero or one form. And then add a beta j, where j is the particular QTL deviation from mean value. Add them all together to compute the phenotype. OK.

So, this is a very simple additive model and a consequence of this model is that if you think about an F1 or a child of two parents, as we said earlier, a child is going to inherit roughly half of the alleles from mom and half of the alleles from dad.

And so for additive models like this, the expected value of the child's trait value is going to be the midpoint of mom and dad. And that can be derived directly from the equation above, because you're getting half of the QTLs from mom and half of the QTLs from dad.

So this was observed a long time ago, right, because if you did studies and you looked at the deviation from the midpoint of parents for human height.

You can see that the children fall pretty close to mid-parent line, where the y-axis here is the height in inches and that suggests that much of human height can be modeled by a narrow-sense based heritability model.

Now, once again, narrow-sense heritability is the fraction of phenotypic variance explained by an additive model. And we've talked before about the model itself. And little h squared is simply going to be the amount of variance explained by the additive model over the total phenotypic variance.

And the additive variance is shown on the right-hand side. That equation boils down to, you take the phenotypic variance and you subtract off the variance that's environmental and that cannot be explained by the additive variance, and what you're left with is the additive variance.

And once again, coming back to the question of missing heritability, if we observe that what we can estimate for little h squared is below what we expect, that gap has to be explained somehow. Some typical values for theoretical h squared.

So this is not measured h squared in terms of building a model and testing it like this. But what we can do is we can theoretically estimate what h squared should be, by looking at the fraction of identity between individuals.

Morphological traits tend to have higher h squared for the fitness traits. So human height has a little h square of about 0.8. And for those ranchers out there in the audience, you'll be happy to know that cattle yearly weight has heritability of about 0.35.

Now, things like life history which are fitness traits are less heritable. Which would suggest that looking at how long your parents lived and trying to estimate how long you're going to live is not as productive as looking at how tall you are compared to your parents. And there's a complete table that I've included in the slides for you to look at, but it's too small to read on the screen.

OK, so now we're going to turn to computational models and how we can discover a model that figures out where the QTLs are, and then assigns that function f to them so we can predict phenotype from genotype. And we're going to be taking our example from this paper by Bloom, et al, which I posted on the Stellar site. And it came out last year and it's wonderful study in QTL analysis.

And the setup for this study is quite simple. What they did was, is they took two different strains of yeast, RM and BY, and they crossed them and produced roughly 1,000 F1s. And RM and BY are very similar. They are about, I think it's about 35,000 snips between them.

Only about 0.5% of their genomes are different. So they're really close. Just for point of reference, you know, the distance between me and you is something like one base for every thousand? Something like that. And then they assayed all those F1s. They genotyped them all.

So to genotype them, what you do is you know what the parental genotypes are because they sequence both parents. The mom and dad, so to speak, at 50x coverage. So they knew the genome sequence is completely for both mom and dad.

And then for each one of the 1,000 F1s they put them on a microarray and what is shown on the very bottom left is a result of genotype in an individual where they can see each chromosome and whether it came from mom or from dad.

And you can't see it here, but there are 16 different chromosomes and the alternating purple and yellow colors show whether that particular part of the genome came from mom or from dad. So they know for each individual, its source. From the left or the right strain. OK.

And they have a thousand different genetic makeups. And then they asked, for each one of those individuals, how well could they grow in 46 different conditions? So they exposed them to different sugars, to different unfavorable environments and so forth.

And they measured growth rate as shown on the right-hand side. Or right in the middle, that little thing that looks like a bunch of little dots of various sizes. By measuring colony size, they could measure how well the yeast were growing. And so they had two different things, right.

They had the exact genotype of each individual, and they also had how well it was growing in a particular condition. And so for each condition, they wanted to associate the genotype of the individual to how well it was growing. To its phenotype.

Now, one fair question is, of these different conditions, how many of them were really independent? And so to analyze that, they looked at the correlation between growth rates across conditions to try and figure out whether or not they actually had 46 different traits they were measuring.

So this is a correlation matrix that is too small to read on the screen. The colors are somewhat visible, where the blue colors are perfect correlation and the red colors are perfect anti-correlation.

And you can see that in certain areas of this grid, things are more correlated, like what sugars the yeast liked to eat. But suffice to say, they had a large collection of traits they wanted to estimate.

So, now we want to build a computational model. So our next step is figuring out how to find those places in the genome that allows us to predict, how well, given a trait, the yeast would grow. The actual growth rate.

So the key idea is this-- you have genetic markers, which are snips down the genome and you're going to test a particular marker. And if this is a particular trait, one possibility is that-- let's say that this marker could be either 0 or 1. Without loss of generality, it could be that here are all the individuals where the marker is zero.

And here are all the markers where the marker is 1. And really, fundamentally, whether an individual has a 0 or a 1 marker, it doesn't really change its growth rate very much. OK? It's more or less identical. It's also possible that this is best modeled by two different means for a given trait.

That when the marker is 1, you're growing-- actually this is going to be the growth rate on the x-axis. The y-axis is the density. That you're growing much better when you have a 1 in that marker position than a zero.

And so we need to distinguish between these two cases when the marker is predictive of growth rate and when the marker is not predictive of growth rate.

And we've talked about lod likelihood tests before and you can see one on the very top. And you can see there's an additional degree of freedom that we have in the top prediction versus the bottom because we're using two different means that are conditioned upon the genotypic value at a particular marker.

So we have a lot of different markers indeed. So we have-- let's see here, the exact number. I think it's about 13,000 markers they had in this study. No. 11,623 different unique markers they found. That they could discover, that weren't linked together. We talked about linkage earlier on.

So you've got over 11,000 markers. You're going to do a lod likelihood test to compute this lod odds score. Do we have to worry about multiple hypothesis correction here? Because you're testing over 11,000 markers to see whether or not they're significant for one trait. Right.

So one thing that we could do is imagine that what we did was we scrambled the association between phenotypes and individuals. So we just randomized it and we did that a thousand times. And each time we did it, we computed the distribution of these lod scores.

Because we have broken the association between phenotype and genotype, the lod scores which we should be seeing if we did this randomization, should correspond to essentially noise. But we would see it random. So it's a null distribution we can look at. And so what we'll see is a distribution of lod scores.

This is the lod. This is the probability from a null, a permutation test. And since we actually have done the randomization over all 11,000 markers, we can directly draw a line and ask what are the chances that a lod score would be greater than or equal to a particular value at random?

And we can pick an area inside this tail, let's say 0.05, because that's what the authors of this particular paper used and ask what value of a lod score would be very unlikely to have by chance? It turns out in their first iteration, it was 2.63. That a lod score over 2.63 had a 0.05 chance or less of occurring in randomly permuted data.

And since a permuted data contained all of the markers, we don't have to do any multiple hypothesis correction. So you can directly compare the statistic that you compute against a threshold and accept any marker or QTL that has a lod score greater, in this case then 2.63 and put it in your model. And everything else you can reject.

And so you start by building a model out of all of the markers that are significant at this particular level. You then assemble the model and you can now predict phenotype from genotype. But of course, you're going to make errors, right. For each individual, there's going to be an error.

You're going to have a residual for each individual that is going to be the phenotype minus the genotype of the individual. So this is the error that you're making.

So what these folks did was that you first look at predicting the phenotype directly, and you pick all the QTLs that are significant at that level. And then you compute the residuals and you try and predict the residuals.

And you try and find additional QTLs that are significant after you have picked the original ones. OK.

So why might this produce more QTLs then the original pass? What do you think? Why is it that trying to predict the residuals is a good idea after you've tried to predict the phenotype directly? Any ideas about that?

Well, what this is telling us, is that these QTLs we're going to predict now were not significant enough in the original pass, but when we're looking at what's left over, after we subtract off the effect of all the other QTLs, other things might pop up. But in some sense, we're obscured by the original QTLs. Once we subtract off their influence, we can see things that we didn't see before.

And we start gathering up these additional QTLs to predict the residual components. And so they do this three times. So they predict the original set of QTLs and then they iterate three time on the residuals to find and fit a linear model that predicts a given trait from a collection of QTLs that they discover. Yes?

AUDIENCE: Sorry. I'm still confused. The second round? [INAUDIBLE] done three additional times? Is that right? So the--

PROFESSOR: Yes.

AUDIENCE: Is it done on the remainder of QTL or on the original list of every--

PROFESSOR: Each time you expand your model to include all the QTLs you've discovered up to that point. So initially, you discover a set of QTLs, call that set one. You then compute a model using set one and you discover the residuals.

AUDIENCE: [INAUDIBLE].

PROFESSOR: Correct. Well, residual [INAUDIBLE] so you use set one to build a model, a phenotype. So set one is used here to compute this, right. And so set one is used. And then you compute what's left over after you've discovered the first set of QTLs.

Now you say, we still have this left to go. Let's discover some more QTLs. And now you discover set two of QTLs. OK. And that set two then is used to build a model that has set one and set two in it. Right.

And that residual is used to discover set three and so forth. So each time you're expanding the set of QTLs by what you've discovered in the residuals. Sort of in the trash bin so to speak. Yes?

AUDIENCE: Each time you're doing this randomization to determine lod cutoff?

PROFESSOR: That's correct. Each time you have to redo the randomization and get to the lod cutoff.

AUDIENCE: But does that method actually work the way you expect it on the second pass, given that you have some false positives from the pass that you've now subtracted from your data?

PROFESSOR: I'm not sure I understand the question.

AUDIENCE: So the second time you do this randomization, and you again come up with a threshold, you say, oh, above here there are 5% false positives.

PROFESSOR: Right.

AUDIENCE: But could it be that that estimate is actually significantly wrong based the fact that you've subtracted off false positives before you do that process?

PROFESSOR: I mean, in some sense, what's your definition of a false positive? Right. I mean it gets down to that because we've discovered there's an association between that QTL and predicting phenotype. And in this particular world it's useful for doing that.

So it's hard to call something a false positive in that sense, right. But you're right, you actually have to reset your threshold every time that you go through this iteration. Good question. Other questions? OK.

So, let's see what happens when you do this. What happens is that if you look down the genome, you discover a collection. For example, this is growth in E6 berbamine.

And you can see the significant locations in the genome, the numbers 1 through 16 of the chromosomes and the little red asterisks above the peaks indicate that that was a significant lod score. The y-axis is a lod score.

And you can see the locations in the genome where we have found places that were associated with growth rate in that particular chemical. OK.

Now, why is it, do you think, that in many of those places you see sort of a rise and fall that is somewhat gentle as opposed to having an impulse function right at that particular spot?

AUDIENCE: Nearby snips are linked?

PROFESSOR: Yeah, nearby snips are linked. That as you come up to a place that is causal, you get a lot of other things are linked to that. And the closer you get, the higher the correlation is.

So that is for 1,000 segregants in the top. And what was discovered for that particular trait, was 15 different loci that explained 78% of the phenotypic variance. And in the bottom, the same procedure was used, but was only used on 100 segregants.

And what you can see is that, in this particular case, only two loci were discovered that explain 21% of the variance. So the bottom study was grossly under powered.

Remember we talked about the problem of finding QTLs that had small effect sizes. And if you don't have enough individuals you're going to be under-powered and you can't actually identify all of the QTLs.

So this is a comparison of this. And of course, one of the things that you don't know is the environmental variance that you're fighting against. Because the number of individuals you need, depends both on the number of potential loci that you have.

The more loci you have, the more individuals you need to fight against the multiple hypotheses problem, which is taken care of by this permutation implicitly. And the more QTLs that contribute to a particular trait, the smaller they might be. And there you need more individuals to provide adequate power for your test.

And out of this model, however, if you look at for all the different traits, the predictive insight versus the observed phenotype, you can see that the model does a reasonably good job.

So the interesting things that came out of the study were that, first of all, it was possible to look at the effect sizes of each QTL. Now, the effect size in terms of fraction of variance explained of a particular marker, is the square of its coefficient. It's the beta squared.

So you can see here the histogram of effect sizes, and you can see that most QTLs have very small effects on phenotype where phenotype is scaled between 0 and 1 for this study.

So, most traits as described here have between 5 and 29 different QTL loci in the genome. They're used to describe them with a median of 12.

Now, the question the authors asked, was if they looked at the theoretical h squared that they computed for the F1s, how well did their model do? And you can see that their model does very well. That, in terms of looking at narrow sense heritability, they can recover almost all of it, all the time.

However, the problem comes here. Remember we talked about how to compute broad-sense heritability by looking at clones and computing environmental variance directly.

And so they were able to compute broad-sense heritability and compare that the narrow-sense heritability that they were able to actually achieve in the study. And you can see there are substantial gaps. So what could be making up those gaps? Why is it that this additive model can't explain growth rate in a particular condition?

So, the next thing that we're going to discover are some of the sources of this so-called missing heritability. But before I give you some of the stock answers that people in the field give, since this is part of our quest today to actually look into missing heritability, I'll put it to you, my panel of experts.

What could be causing this heritability to go missing? Why can't this additive model predict growth rate accurately, given it knows the genotype exactly? Yes.

AUDIENCE: [INAUDIBLE] that you wouldn't detect from looking at the DNA sequence.

PROFESSOR: So epidemic factors-- are you talking about protein factors or are you talking about epigenetic effects?

AUDIENCE: More of the epigenetic marks.

PROFESSOR: Epigenetic marks, OK. So it might be now, yeast doesn't have DNA methylation. It does have chromatin modifications in the form of histone marks. So it might be that there's some histone marks that are copied from generation to generation that are not counted for in our model. right? OK, that's one possibility. Great. Yes.

AUDIENCE: There could be more complex effects so two separate genes may come out, other than just adding. One could turn the other off. So it one's on, it could [INAUDIBLE].

PROFESSOR: Right. So those are called epistatic effects, or they're non-linear effects. They're gene-gene interaction effects. That's actually thought to be one of the major issues in missing heritability. What else could there be? Yes.

AUDIENCE: [INAUDIBLE].

PROFESSOR: Right. So you're saying that there could be inherent noise that would cause there to be fluctuations in colony size that are unrelated to the genotype. And, in fact, that's a good point. And that's something that we're going to take care of with the environmental variance.

So we're going to measure how well individuals grow with exactly the same genotype in a given condition. And so that kind of fluctuation would appear in that variance term. And we're going to get rid of that.

But that's a good thought and I think it's important and not appreciated that there can be random fluctuations in that term. Any other ideas? So we have epistasis. We have epigenetics. We've got two E's so far. Anything else?

How about if there are a lot of different loci that are influencing a particular trait, but the effect sizes are very small. That we've captured, sort of the cream. We've skimmed off the cream.

So we get 70% of the variance explained, but the rest of the QTLs are small, right, and we can't see them. We can't see them because we don't have enough individuals. We're underpowered, right. We just-- more individuals more sequencing, right.

And that would be the only way to break through this and be able to see these very small effects. Because if the effects are small, in some sense, we're hosed. Right?

You just can't see them through the noise. All those effects are going to show up down here and we're going to reject them. Anything else, people can think about? Yes?

AUDIENCE: Could you content maybe the sum of some areas that are-- sorry, the addition sum of those guys that have low effects. Or is that not detectable by any [INAUDIBLE]?

PROFESSOR: Well, that's certainly what we're trying to do with residuals, right? This multi-round round thing is that we take all the things we can detect that have an effect with a conservative cut off and we get rid of them.

And then we say, oh, is there anything left? You know, that's hiding, sort of behind that forest, right. If we cut through the first line of trees, can we get to another collection of informative QTLs? Yeah.

AUDIENCE: I was wondering if this could be an overestimate also. Like, for example, if, when you throw out the variance for environmental conditions, the environmental conditions aren't as exact as we thought they were between two yeast growing in the same set, setup.

PROFESSOR: Right.

AUDIENCE: Then maybe you would inappropriately assign a variance to the environmental condition whereas some that could be, in fact-- something that wouldn't be explained by.

PROFESSOR: And probably the other way around. The other way around would be that you thought you had the conditions exactly duplicated, right. But when you actually did something else, they weren't exactly duplicated so you see bigger variance in another experiment. And it appears to be heritable in some sense. But, in fact, it would just be that you misestimated the environmental component.

So, there are a variety of things that we can think about, right. Incorrect heritability estimates. We can think about rare variance. Now in this particular study we're looking at everything, right. Nothing is hiding. We've got 50x sequencing. There are no variants hiding behind the bushes. They are all there for us to look at.

Structural variants-- well in this particular case, we know structural variants aren't present, but as you know, many kinds of mammalian cells exhibit structural variance and other kinds of bizarre behaviors with their chromosomes. Many common variants of low effect. We just talked about that. And epistasis was brought up. And this does not include epigenetics, I'll have to add that to the listen. It's a good point. OK.

And then we talked about this idea that epistasis is the case where we have nonlinear effects. So a very simple example of this is when you have little a and big B, and big A and big B together, they both had an effect. But little a, little b, have no effect. And big A and big B have no effect by themselves. So you have a pairwise interaction between these terms. Right.

So this is sort of the exclusive OR of two terms and that non-linear effect can never be captured when you're looking at terms one at a time. OK. Because looking one at a time looks like it has no effect whatsoever. And these effects, of course, could be more than pairwise, if you have a complicated network or pathway.

Now, what the authors did to examine this, is they looked at pairwise effects. So they considered all pairs of markers and asked whether or not, taken two at a time now, they could predict a difference in trait need. But what's the problem with this? How many markers did I say there were? 13,000, something like that.

All pairs of markers is a lot of pairs of markers. Right. And what happens to your statistical power when you get to that many markers? You have a serious problem. It goes right through the floor. So you really are very under-powered to detect these interactions.

The other thing they did was to try to get things a little bit better as they said, how about this. If we know that a given QTL is always important for a trait because we discovered it in our additive model. Well consider its pairwise interaction with all the other possible variants.

So instead of now 13,000 squared, it's only going to be like 22 different QTLs for a given trait times 13,000 to reduce the space of search. Obviously I got this explanation not completely clear. So let me try one more time. OK.

The naive way to go at looking at pairwise interactions is consider all pairs and ask whether or not all pairs have an influence on a particular trait value. Right. We've got that much? OK.

Now let's suppose we don't want to look at all pairs. How could we pick one element of the pair to be interesting, but smaller in number? Right. So what we'll do is, for a given trait, we already know which QTLs are important for it because we've built our model already.

So let's just say, for purpose of discussion, there are 20 QTLs that are important for this trait. We'll take each one of those 20 QTLs and we'll examine whether or not it has a pairwise interaction with all of the other variance. And that will reduce our search base. Is that better? OK, good.

So, when they did that, they did find some pairwise interactions. In 24 of their 46 traits had pairwise interactions and here is an example. And you can see the dot plot, or the upper right-hand part of this slide, how when you BYBY. You have a lower phenotypic value then when you have just any RM component on the right-hand side.

So those were two different snips on chromosome 7 and chromosome 11 and showing how they interact with one another in a non-linear way. If they were linear, then as you added either a chromosome at 7 or a chromosome 11 contribution it would go up a little bit.

Here, as soon as you add either contribution from RM, it goes all way up to have a mean of zero or higher. In this particular case, 71% of the gap between broad-sense and narrow-sense was explained by this one pair interaction.

So it is the case that pairwise interactions can explain some of the missing heritability. Can anybody think of anything else they can explain missing heritability? OK.

What's inherited? Let's make a list of everything that's inherited from the parental line to the F1s. OK. Yes.

AUDIENCE: I mean, because there's a lot more things inherited. The protein levels are inherited.

PROFESSOR: OK.

AUDIENCE: [INAUDIBLE] are inherited as well.

PROFESSOR: Good. I like this line of thinking.

AUDIENCE: [INAUDIBLE].

PROFESSOR: There are a lot of things that are inherited, right? So what's inherited? Some proteins are probably inherited, right? What is replicable through generation to generation as a genetic material that's inherited?

Let's just talk about that for a moment. Proteins are interesting, don't get me wrong. I mean, prions and other things are very interesting. But what else is inherited? OK, yes?

AUDIENCE: [INAUDIBLE].

PROFESSOR: So there are other genetic molecules. Let's just take a really simple one-- mitochondria. OK. Mitochondria are inherited. And it turns out that these two strains have can have different mitochondria. What else can be inherited?

Well, we were doing these experiments with our colleagues over at the Whitehead and for a long time we couldn't figure out what was going on. Because we would do the experiments on day one and they come out a particular way and on day two they come out a different way. Right.

And we're doing some very controlled conditions. Until we figured out that everybody uses S288C which is the genetic nomenclature for the lab trained yeast, right. It's lab train because it's very well behaved. It's a very nice yeast. It grows very well. It's been selected for that, right.

And people always do genetic studies by taking S288C, which is the lab yeast, which has being completely sequenced and so you want to use it because you can download the genome with a wild strain. And wild strains come from the wild, right.

And they come either off of people who have yeast infections. I mean, human beings, or they come off of grape vines or God knows where, right. But they are not well behaved. And why are they not well behaved?

What makes these yeast particularly rude? Well, the thing that makes them particularly rude is that they have things like viruses in them. Oh, no. OK. Because what happens is that when you take a yeast that has a virus in it, and you cross it with a lab yeast, right. All of the kids got the virus. Yuck. OK.

And it turns out that the so-called killer virus in yeast interacts with various chromosomal changes. And so now you have interactions-- genetic interactions between a viral element and the chromosome.

And so the phenotype you get out of particular deletions in the yeast genome has to do with whether or not it's infected with a particular virus. It also has to do with which mitochondrial content it has.

And people didn't appreciate this until recently because most of the past yeast studies for QTLs were busy crossing lab strains with wild strains and whether it was ethanol tolerance or growth and heat, a lot of the strains came up with a gene as a significant QTL, which was MKT1.

And people couldn't understand why MKT1 was so popular, right. MKT1, maintenance of killer toxin one. Yeah. That's the viral thing that enables-- the chromosomal thing that enables a viral competence.

So, it turns out that if you look at this-- in this particular case, we're looking at yeast that don't have the virus in the bottom little photograph there. You can see they're all sort of, you know, they're growing similarly.

And the yeast with the same genotype above-- those are all in tetrads. Two out of the four are growing, the other two are not, because the other two have a particular deletion.

And if you look at the model-- a deletion only model, the deletion only, only looks at the chromosomal compliment doesn't predict the variance very well. And if you look at the deletion and whether or not you have the virus, you do better.

But you do even better, if you allow for there to be a nonlinear interaction between the chromosomal modification and whether or not you have a virus. And then you recover almost all of missing heritability.

So I'll leave you with this thought, which is that genetics is complicated and QTLs are great, but don't forget that there are all sorts of genetic elements.

And on that note, next time we'll talk about human genetics. Have a great weekend until then. We'll see you. Take care.