MARCO RAMONI: Today I'm going to talk to you about the-- about basic genetics, what geneticists do, and how genetics is moving into the genomic area by increasing the size, the scope, and the quality of genetic studies. We'll do--

[DOOR CREAKING]

You can close the door. So the origin of all of this is-- I will put myself at the particular time point, and the origin of all of this is the publication of the Human-- of the results of the Human Genome Project, a first draft of the human genome. And when they announced the publication of the human genome, the first item on the future agenda was a SNP map, and they said, this SNP map promises to revolutionize both mapping diseases and tracking human history. And I will try to show you how you do both using SNPs, and if you don't know what SNPs are, hold onto your breath. You will know it in a second.

[DOOR CREAKING]

Please. The SNPs, I will tell you in a second in more-- in detail. SNPs are things called single nucleotide polymorphisms, and there are subtle variations in the human genome across individuals. What this mean? It means that one of the most startling statistics out of the Genome Project is that we're all similar, very, very similar at genetic level. But there are some very, very subtle variations, one base pair swapped here and there, that make us different and, most of all, makes our genomes different.

And what you can say is that this is the real meat of the human genome. People paid billions of dollars not for the curiosity of a handful of scientists, but because the promise was that, once we know the human genome, it would be easier to track down the genetic causes of diseases and the genetic causes of-- and the-- possibly find some cure for them and get a better understanding. And as you will see, SNPs are the real meat of the human genome.

So I will start from the '80s, when people started thinking about-- seriously thinking about developing a human genome project. And the Human Genome Project-- I will tell you what genetic polymorphisms are, what type of polymorphisms exist, and then we give you some basic terminology of what are the terms that geneticists use to describe the properties of a genome. And then I will talk to you about the real interesting stuff, which are complex traits, or those traits, those diseases, those observable features of an individual that don't come from one single gene, but they are the result of an interaction of more genes or the interaction of genes with the environment, and I will tell you why these traits are complex.

And then I will tell you how people designed these kind of-- the experiments to identify these complex traits and also the simpler ones. And once they have designed it, I will tell you how do you analyze this data once you have them. And then I will tell you what's the word on the street on the latest fashion in analyzing the genome.

So as I was saying, the intuition behind genetics is that we can find causes for diseases without knowing the actual mechanism of the disease, just associating changing in the genome to observable traits, observable characters. Now, as you can imagine, this is not really an easy idea to sell, and when people started selling it in the early 20th century, people didn't really like it a lot, especially biologists, who say, what kind of science is this? There is no mechanism in this. There is no understanding of how the things work.

But then people started to deliver results, and how deliver results about this? Well, once we know that gene exists-- before, it was kind of difficult, but-- I will show you how you did it before, before knowing how to genotype somebody. But one of the-- the main intuition that changed everything is due to a guy called David Boston that, in 1997, showed that there are natural markers on the genome called polymorphisms that make us different and make this genome different, right?

So if I have a little change in my population, what I can try to do is collect people that have the disease and collect people that don't have the disease and see which marker shows up more than I would expect in one group or another. And I will tell you how you play these games. You can play them in several ways, but the way in which we rely-- we do-- we conduct the studies is not by tagging something, but leveraging our natural tags where exactly these polymorphisms.

This is just a summary of the central dogma of molecular biology. You all know this. You got a class from Dr. Butte, I think, about this. Some of you have a masters and PhDs in biology, so they know this better than I do.

The assumption is that there is DNA that is all identical in all our cells, and then this is translated in RNA. And RNA is then turned into proteins which determines whatever we can observe of somebody, the traits, the diseases susceptible to, his physiology, the metabolism, even drug resistance. So the idea that polymorphisms could be used as natural markers created exactly the background of the Human Genome Project because, at that point, we had an intellectual tool, a scientific tool to track down diseases using the code.

So the easiest thing we can do is to find out diseases that are caused by one single screw-up in one gene, and there are various ways in which this can actually be a screw-up or not. They can be dominant or recessive. Recessive means that you need to-- you know, you receive two chromosome, one from your father and one from your mother. To be recessive, you need to have both chromosomes that are screwed up, your gene in both chromosomes. To be dominant, one is going to be enough.

And we classify these diseases also by another dimension, which is, if it's on the X chromosome, which is-- has-- there's a symmetry for men-- and/or if it's any other chromosome. So if it's any other chromosome, it's going to be an autosomal and can be dominant or recessive, and examples are the Huntington's disease and cystic fibrosis. And if it's X-linked, it can be dominant or recessive, and I will give you right away a little example of this.

Today, there are about 400 single-gene diseases that have been identified. So for 400 diseases, we know that there is a little screw-up. Sometimes there is a very, very simple reason, because a base is turned from being a normal base, coding for amino acid, into stopping signals. So when that particular sequence is read, the protein does not code after that particular point and the protein is screwed up, and this creates disease.

OK, so I need-- this is the boring part. I need to give you-- I need to give you a little terminology. Now, an allele is a sequence of DNA bases. So for each SNP, for each gene, for each whatever, for each piece of your DNA, you have two alleles, one coming from your mother and one from your father.

Locus is a physical location on the chromosome. It's like [INAUDIBLE]. We know exactly where it is. We know it's there. We know on which chromosome is it, and we know the position of the chromosome.

Linkage is a proximity of two alleles on the chromosome, and it is kind of ambiguous because, as you will see, proximity may have several meanings in genetics. The marker is an allele with a particular position that we can track down somehow. The distance, the physical distance, is the number of bases between one point and another, but this is not the only distance we have, and actually is not really the most interesting distance we have.

The most interesting distance is the probability that two points on the chromosome will be recombined when you make your children. And the distance, this probabilistic distance, doesn't necessarily map on the chromosome with a constant number of bases, as you will see in a second. So there are-- this means that there are points on the chromosome that are easier to recombine and points on the chromosome that are harder to combine, less likely to recombine. The measure by which we identify this probabilistic distance is called centimorgan.

A phenotype is a observable character. It can be susceptibility to a disease. It can be diabetes, a disease itself. The genotype is the internally coded inherited information, so it's a piece of DNA. And the penetrance is, if you take the frequent interpretation seriously, is the probability that, given that you have the allele, you will develop the phenotype. OK?

So the physical distance between two alleles are a base pair, but the combination among them is not constant. Well, you all know Mendel's first law. We know that allele pair separate when the gamete is formed, and they are randomly reshuffled and create new pairs.

Now, use-- the probabilistic measure we want to use is the probability that two points will be recombined at the next passage to the next generation. So we can say that, on average, 1 centimorgan, which is the probability of 1% of being recombined, is about-- happens between 1 megabases. Every 1 megabases, you have one probability of being recombined.

Now, these are kind of little gossips. If human autosomal physical map is 3 billion bases, as you know, the linkage map in centimorgan-- the distance, probabilistic distance between points-- is different between humans and-- between male and female humans. There are 2,800 centimorgan for male and 4,200 centimorgan for females.

Why's that? Females have a bigger genetic code. Male have this little Y wimpy chromosome, and females had this second big, beautiful X chromosome, which is much bigger than ours, less stable, but bigger.

But if you compute the difference between these two things, you see that the probability of recombination of the correspondence between the probability of recombination for male and female is also different. One is a bigger code, and so the difference, you would say, is about 1-- on average, is 1 million pair-- 1 centimorgan for 1 million bases. And for female-- for male and female, it's about a bit more than a million for male and a bit less than a million for females. And one notion that would be absolutely important for our ability to use markers is of cosegregation, the fact that two alleles are transmitted together to the next generation.

AUDIENCE: I just have a quick question on that.

MARCO RAMONI: Sure.

AUDIENCE: So why--

MARCO RAMONI: You're upset because they said that women had better chromosomes than we have? No.

AUDIENCE: I agree. The question I had is-- was back in the definitions of a locus, and I just wondered, so the-- those things, can-- are those affected by, like-- so you said it's the physical location on a chromosome. Can the locus of a particular allele change if the packing into the chromosome structure is altered in any way? Like, is that something--

MARCO RAMONI: That's a very good question. So the--

AUDIENCE: And then follow-up to that is, if that is the case, does that affect the cosegregation as well?

MARCO RAMONI: Yeah. So this is a very good question. You can imagine a locus as a variable and an allele as the state of this variable. So you have-- you are given this locus, and at position 3,022 on chromosome 5, you have an A.

And if this is a polymorphism-- if it this is not a polymorphism, everybody has A. But if this is a polymorphism, we expect that something like above 1% of the population will have a different letter there, like a T or a G. Now, this is due to mutations, and there is nothing we can-- it's not affected by the locus. But the packaging of the genes, of the chromosomes, where they are next to each other, may affect the probability that it will be transmitted together.

So there are several reasons why this linkage disequilibrium exists. One is the physical recombination of this, which is very difficult to identify, and the second are historical reasons. As I will tell you in a second, Caucasians come from a handful of people who left Africa, and this handful of people were basically assimilated in this process.

So we all come from a very, very restricted number of people that were alive between 25,000 and 50,000 years ago, which is a blink of an eye from an evolutionary point of view. So if you look at the map of an African American population, or an African population and a European population, you will see that there is a staggering difference in the variability of those and our variability.

Now, if you look at our things, you may-- if you do a genotype study of Caucasians, you will see alleles that go together, but you will not be able to say if these alleles go together for actual physical recombination or simply because you live, we live-- we draw our genetic codes from a depleted pool in which only very few combinations are available. Does that answer your question?

AUDIENCE: Yeah.

MARCO RAMONI: I have a picture in a second that may be helpful in this. So let me go back to all this, a single disease works and how dramatic can be the effect that a single disease has. Hemophilia, as you know, is the X-linked recessive disease that is fatal for women, probably.

So if a man-- it's on the X chromosome. We men don't get an X chromosome. So we only get recessive. We only get the mother.

But-- and we-- so if the woman has it, she doesn't have any manifestation. She's just a carrier, and if a man has it, he has manifestations of the disease. We don't observe women with-- it's very rare to find women with both X chromosomes, male hemophilia.

Now, this is a major screw-up in the history of Europe, caused by a gene like that. Now, this is Queen Victoria's family tree, and Queen Victoria is the one-- is the second-- is the one right in the middle, half orange and half white. So if they're round, both are women, and square are men. And if it's a half colored, she's a carrier. If it's all colored, is an affected child.

So when Queen Victoria had her fifth child, but her first male-- we're talking about monarchy. Males have some kind of importance here. The first male, the male turned-- Leopold, the male, turned out to be hemophiliac. And she was very, very upset, and her declaration was that their blood was strong and there was no change in it, no weakness in it.

So if we take her seriously, what would you think had happened? I mean, this thing has incredible consequences. If you look at down here and you see this row of seven-- of five children with no descendants, that's the Russian family.

They were exterminated by the-- during the Russian Revolution, and people say that the wife of Nicholas II started seeing Rasputin-- was one of the major causes of the upsetting of the population-- because her first male born was hemophiliac, so a major screw-up. I mean, this led to the withdrawal of the Russians from the First World War and big, big, big changes. I mean we still pay taxes because they are communist, right?

So what do you think happened? Who is responsible for it? They start from the assumption that they're strong, their blood is strong.

You say, OK, so poor Victoria had a mutation. Would you believe that? Poor Leopold had the mutation? How would you compute that?

What I would do is look at the randomness of the distribution of those alleles in the second generation. We cannot genotype these people. We can infer the genotype only by looking at what they have, right? So you would expect that-- all of these guys on the third row, they are Victoria's offspring. You would expect that 50/50 distribution, right?

Now, for some of them, we cannot really evaluate. We have two-- we have three we have to remove out because they don't have the same, and we don't know anything about that. If you take the other, we have 1, 2-- we have 1, 2-- they are actually healthy, and we can almost say that.

And then we have three that are either affected or carriers. This is a good 50%, right? I mean, I would say that you cannot really get any better than that with this sample size. It's almost a perfect 50%. If you go down and look down streams, you will see that the ratio will be exactly the same.

For the women-- for the daughters of Nicholas II, we cannot know, but if you look at R in there, you will see, you have the same distribution. So I would say that from Victoria onward, everything is pretty consistent. So she didn't-- so, OK, she's might have a mutation, or she may have been cheating on her husband? No, because otherwise she would've been cheating with everybody, right, for all the children. We have a very pretty nice distribution of this.

So what happened? How about her mother? Well, her mother, we don't have really enough information. How about her father?

Her father married a woman 20 years younger than he was. He was rumored to be homosexual, and he died six months later Victoria was born. The postman maybe? If you have an hemophiliac postman, there would be a good explanation.

Now, somebody made some research and found out that a grandmother of Queen Victoria's had two affected siblings who died of hemophilia. So in this case, the blood is not really that strong because it's coming in from Saxe-Coburg, from that line, and becomes part of the monarchy of the European aristocracy. You don't like this kind of gossip? I mean, this is--

[LAUGHTER]

I mean, this was really the story of the day, of the century, in-- at that time. But you see how we did it. We had to wait 200-- a hundred and something years to find out what really happened because we had to observe-- we cannot genotype these people at that time, so we have to observe their, of course, characters of their things.

But if we were able to genotype them, what could we do? Well, we could look at this thing of these small variations in their chromosomes, and the oldest variation we have been able to use are called simple sequence repeats and microsatellites. So these are-- let me see if I have a slide explaining this. No.

So there are parts of the chromosomes of the genetic code, in which you have sequences like G-A-T-A that are repeated several times in your chromosome, in your genetic code. They are repeated 13 time in me and 15 times in you, two times in somebody else. And if we count these things, we can actually identify a region of the chromosome that by linkage-- remember, by being attached to it from evolutionary purposes-- we'll identify a part, a stretch of the genome that we can actually tag.

Now, the problem is that simple sequence repeats and microsatellites add this little drawback. They cannot occur in really interesting regions. You cannot change to match a genetic code by repeating the same thing 13 times or two times and hope to get coherent proteins out of it, right? So we are confined in interest in regions that, by design, are supposed to be kind of uninteresting from an-- from a functional point of view. There may be interest from an evolutionary point of view, but they're kind of leftover of something else.

Now, SNPs have this great property. They are one single base. And so it's something like this. You have your sequence, and you have that little base that in 90% of the population is T and in 10% of the population is G.

To be sure that-- to qualify a single change like this as a SNP and not just a simple mutation, it has to have a minimal appearance in the population. And we stipulated this 1%, but nobody takes the number seriously because that's below the genotype error-- the expected error. So when you genotype people, usually you have a higher cutoff to say, well, this is not really a SNP. This is something that is just due to genotype error.

They are also the most common type of variations. Not only they go in interesting places, but they can actually be in any place in the genome. And they have different functions and different roles in terms of protein coding.

So a cSNP is a SNP occurring in a coding region. An rSNP is a SNP occurring in a regulatory region. The sSNP is the gene that occurs in a coding region, but by changing the SNP, you don't change the amino acid. Right, we have all this redundant vocabulary of amino acids, and in this case, it's going to be functionally silent, but it's going still to be a marker.

Now, why actually we cannot do this-- why SNPs are important, and why we cannot have microsatellites in other regions? Well, this is a very old study, one of the very first. And what this Kreitman guy did was to sequence 11 alleles from a locus from a gene called alcohol dehydrogenase into drosophila.

Now, if you-- you have 11 coding regions, and you have 14 sites that have alternative bases. Now, if you simply imagine that these are random changes, you would expect that about 70%, 75% of them would change the amino acid. Now, when you actually look at them, you see that basically none of them does.

Why is that? Well, because this is a very important gene for this animal. They are born and they are nurtured in alcohol, like some of my friends in college. And they-- so the ability to detoxify alcohol is going to be a very important evolutionary point for them. So nature can make mistake, but it can give them so many changes in a critical region. But I want to challenge a little your Darwinian souls.

Mind, it's not that we observe them, we have the change, the random change, we kill off 75% of them, and then we see that the remaining 25% have survived. These things just don't happen. There is no natural selection, in terms of environment, that is killing the 75 people-- 75% of drosophila, OK? You with me?

So in a Darwinian scheme, what we would expect is that you have a random mutation. You go out. You don't run fast enough. You get killed.

In this case, you get a random mutation. Your random mutation is selected by the environment in some way, typically by killing you, and then-- before you can reproduce-- and that's it. But in this case, these things just don't happen, so there is no random change and then selection.

There is something there in the control mechanism that will prevent the animal to have the mutation to begin with. OK? This is the SNP map I was saying before. This is how you read a SNP map.

So you remember that you have two alleles for each locus, and in this case, what we have done is to genotype all these people-- they are the rows-- for these loci, that are the columns. So if an individual has a blue spot, the blue spot means that he is homozygous at the major allele, so he has both chromosomes with the most frequent allele, the most frequent in the population. If he's green, he's heterozygous, and if he's red, he's homozygous in the minor alleles. And his wife is a missing link.

Now, how would you do an association study? Suppose I draw that red line and I tell you, you know, down here, I have the phenotype. Up there, I don't have the phenotype.

Could you make some association? Well, you know, if you look here, there are some genes that, one side are all blue and one side are more colorful. You can do some statistics and say, well, this thing would allow me to distinguish between these two groups and predict which one is going to get the SNP-- the phenotype.

Now, I'll ask you a slightly different question. This is your SNP map. What is the phenotype? What makes these two groups different, something that is observable? Can you say that?

Come on. Buy a pizza. Guess. Why? Guess. No taker?

AUDIENCE: This is a human map or--

MARCO RAMONI: Yeah, it's a human map. These are two human samples.

AUDIENCE: I mean, [INAUDIBLE] African Americans.

MARCO RAMONI: You got the pizza.

[LAUGHTER]

Down there are African Americans. Up there are Caucasians. And so to go back to your question, if you look-- so the evolutionary pressure of the link between two elements in the European population-- take the fifth and the-- no, take the sixth and the seventh, two spots, right? There you-- no, let me do this this way.

Take these two points, OK? These two points, if you study them, they go together, if you look at the statistics, they will go together. Why-- if you do it in a European population. Why? Because you don't have alternatives.

You don't know if these things really go together because some physical reason, or they're really just recombining like crazy but we cannot observe because our population is a population that doesn't have that particular SNP and this SNP exists only in this other population. In this case, it's the African American people. Does this answer your question?

AUDIENCE: I'll have to think about it.

MARCO RAMONI: OK. So let me go back here. So the first quantitative law we-- the most important quantitative law we have in genetics dictates how many major, minor, and heterozygous people we'll have in a particular population. Let me-- come on. OK. So in this case, you take a single allele, and you can actually see that you have a-- you expect to have more people with the homozygous at the major allele, less people heterozygous, and less people that are homozygous at the minor allele.

Now, what rule exists there to make-- to distribute these proportions? So the law is the Hardy-Weinberg law. It states that the probability of having a major allele, the minor allele, and an heterozygous is this formula, is p squared plus 2pq plus p squared equals 1. We call this a situation of equilibrium.

When everything is all right in the population, there hasn't been some major screw-up, this is the law that dictates the distribution. This is what we expect in a population in equilibrium. Now, in a hermaphroditic population we have an equilibrium in one generation by redistributing these things. We have no autosomal genes, so in our case, we need two generations to get an equilibrium.

Now, you can use this to make a lot of little games, like, for instance, how many Caucasians are carriers of cystic fibrosis? Well, we know that cystic fibrosis affects 1 of 2,500 Caucasians, so this is our q square, right? So 0.02 is going to be our q, and the number of nonaffected alleles is going to be 98-- 0.98, 98%, right? So we can plug p and q up there and compute the 2pq as 1 out of 25, which is a remarkably high number for a disease like this.

Now, we make all these calculations under some assumptions. First assumption we make is random mating, which is not a justification to have sex with anybody but is the mechanism of reproduction that we imagine that exists in our population. That is, the person you will mate with will not select you on the basis of a particular genotype.

If we are interested in that particular genotype, she would not ask you for your genetic map and say, no, you have a T in this locus, I don't want you. OK? If this is the case, all our calculations go down the drain. Well, we can make some adjustment then.

The consequence of this is that we have a lot of problems when the selection is based on things that are proxies, like being my relative. If somebody is my relative, I'm going to introduce a bias because I'm going to select her on the basis of her genetic code. I know I don't do this for the genetic code, but I'm going to have-- I'm going to do it anyway. And the other assumption we have is that we have an infinite population, which is kind of sensible for us, for 6 billion people.

So the other thing we-- so why do we need these assumptions? Well, because we know that the mechanism of spread of a character in the population-- it's called drift-- is the way in which a particular allele will get into the population and either conquer it all or just simply disappear, right? So at some point we will reach some kind of evolutionary stability in which these things will be around.

And the problem with this, the moment in which people in the-- mate locally, because, again, this creates some particular population-- some particular bias with the fact that your fishing pond is too small. So at the end, you are going to erase everything in that particular population. If everybody mates with each other in the particular-- in a very small population, whatever we can say about variation is going to disappear.

Now, the contrast of-- what is opposite of the drift is a mutation. So we have these random mutations that come up and show up in our genome, but as you can see, the funny thing between these mechanism-- that is, whatever we observe is drift and mutation-- mutation introduce changes, drift makes these change stable, so just get rid of them-- is in sharp contrast with what we have been taught in high school, that is, there is some kind of selection.

What is selection here? Where people get eaten, where people don't run fast enough? Is there a way to account for selection in this? Yes, there is.

So let's keep a quantitative representation of selection. A quantitative representation of selection is a function of fitness. That will tell you how good is somebody with that particular SNP, with particular allele, to survive in this particular environment, right?

Now, so suppose you have an allele that has a distribution of 0.6 and 0.4. So you apply the Hardy-Weinberg law, and you have that distribution of homozygous and the heterozygous in your population. Then you have a function.

We have a fitness function that will tell you that you have a-- your selection rate will be 0.2 because you're going to say that if you have the major allele or the heterozygous, you are all right. Your probability of survival is 1. If you are homozygous at the minor allele, you have a diminished capacity to survive. So we can compute out of this fitness function a selection function of 0.2 that will tell us how many people we lose in a particular category at each generation.

Now, if you look at the effect on the first generation, you see that there is an increase in the homozygous at the major allele, a decrease in the homozygous-- in the heterozygous, and sharper decreases in the homozygous at the minor allele. But there is something funny here, that because we used a differential there, the smaller is the number of individuals that are homozygous at the minor allele, the smaller will be their depletion, right?

So this mutation, that's a mutation doesn't really go away ever. It will simply stay there and slowly be taken out-- taking out individuals, but it will be lurking in the genetic code for the years to come, OK? We have no explanation in this way of why dinosaurs disappeared. This simply tells us that a particular mutation will keep floating in our population and we have no good mathematical models to explain how they will disappear.

Now, that's this-- everything is based on Hardy-Weinberg law, right? Does it really work? So this guy, this [INAUDIBLE] in 1975, in England, they made this experiment. They took blood samples. This is their blood groups, which are, again, governed by one single allele.

And they look at the population of this group, this random population, and they found that if you compute the expected values using Hardy-Weinberg and you compute the values, the observed values, these numbers are remarkably similar. You see, so we would expect 261.54 to be Mm from Hardy-Weinberg. We get 363. We get 636 for Mn, and we get 334.

Now, these numbers are so similar that somebody wrote a paper showing that actually they had been falsified, because you don't-- the precision of this number is above any reasonable statistical expectation of precision for a particular law, OK? But let's say they falsified the number because they didn't know statistics and they didn't know they could get a way with the better results. But let's buy this thing because we have seen in other population that it's not really that close but it really works.

But we've also seen some population in which it doesn't really work. So beta model being sickle cell anemia in West Africa, if you look-- if you compute the distribution using Hardy-Weinberg, and you compute and you observe the distribution in the population, you have this remarkable change. Look at the difference of those two numbers. We expect that, to have 254, and as a matter of fact, we have 64. Why?

AUDIENCE: [INAUDIBLE].

MARCO RAMONI: Say again.

AUDIENCE: [INAUDIBLE], so you can't observe.

MARCO RAMONI: OK. I think I-- no, I made a mistake. It's flipped. So we have-- sorry, we have 600 and something, not 64. This is 645, so your explanation is the opposite. So you have to explain the opposite phenomenon, not why it's smaller but because it's bigger.

[LAUGHTER]

As a matter of fact, you have a-- you see you have more homozygous-- heterozygous in the middle? Oh, no, no. The number is correct. So this is the thing.

[LAUGHTER]

The inconsistent thing is that you have too many heterozygous, given the homozygous and the minor alleles. You shouldn't have them. You see?

AUDIENCE: But isn't that evidence of a selection, evidence of gene selection?

MARCO RAMONI: You get two pizzas. Because the heterozygous get protection against malaria, so in doing so, you have a bias towards having a lot of heterozygous in the middle and keep the allele around in the population. And the other below there doesn't remain-- doesn't really change, but the selective advantage is for these 5,400 people who have the heterozygous. So in that particular case-- this is something that probably wouldn't be a good test to make here, because here malaria is not really a big issue. But in West Africa it's going to be a big problem.

AUDIENCE: If we did the-- did that test at birth or before birth, would it be consistent with the prediction?

MARCO RAMONI: Before birth?

AUDIENCE: If you did if you did the same distribution of the alleles before they had this pressure of malaria?

MARCO RAMONI: Well, that's is the--

AUDIENCE: Is it the dying-off factor that's causing the difference or something--

MARCO RAMONI: No, there is no dying. Well, so let's put it this way. We actually do it before they're born, because the selection is actually on this parents. The pressure, as usual, is on the parents. So if you keep these things around, the chances that these people will go in older age and reproduce are higher.

So we-- the entire population has a global advantage in keeping around this particular variation, but only in the particular form, which is the heterozygous. The homozygous is going to be too bad for you, but the heterozygous is going to strike a nice balance between getting the disease and actually being protected by-- against malaria. If you do something like this here, you change the environmental condition, you're screwed. Here, I'm pretty sure, if you do the same test, this will not show up because the population as a whole doesn't have a pressure for this.

Now, how do we map these things? How do we use a marker as such? Well, first we want to find the genetic basis, a genotype for a particular disease that is our phenotype, and what's the mindset for somebody looking for this?

We are not really looking for the SNP or the mutation that is changing there. We are trying to find the marker that is in the proximity-- well, we'll be happy to find the marker that is in the proximity of the actual codes. So the mindset is more or less this one. I have a marker, and I observe in my data set a dependency between my phenotype and my marker. But in reality, the extent of my claim cannot be, this marker cause this phenotype.

The extent of my claim can only be, there-- this marker is linkaged to equilibrium with some real genotype that I haven't observed that is actually causing the phenotype. And you see where the complexities start emerging. The first complexity is that the phenotype and-- the genotype and the phenotype may have a complex form of causality. They may have an interaction with an environmental condition, or say they have incomplete penetrance.

And the other thing is that, our group is the linkaged equilibrium between the marker and genotype. I don't expect to have everybody getting exactly the same pair of genotypes, right? So this is where stochasticity comes in, and this is what is making our life slightly more complex but definitely more interesting.

Now, traits don't really follow single gene models. If you look at the list of diseases they gave, they are kind of minor diseases, minor. They have small incidence. They are horrible diseases, but they are not going to save the 50% of people who die of stroke or people who have diabetes or people that have this for more complicated alleles factors that cause more interesting diseases.

On the other hand, even some Mendelian traits are actually complex. So if you look at the-- so the-- and sickle cell anemia is a classical Mendelian disease, but-- so we know why you have it. But the phenotypic variability of the disease is immense. There are children that die at 13 by stroke, and there are people that live forever. And there are perfectly healthy people.

We don't know why this happens. We know these people get the disease, but we don't know what the-- why this happens. And so people are, for instance, studying the early mortality of kids with sickle cell anemia, try to find [INAUDIBLE] SNP, and they say, well, you know, these people get this particular disease that is genetically found. Maybe the difference between their long-term survival-- the chance of their long-term survival is really due exclusively to another SNP. Actually, they are almost finding it.

Another problem is incomplete penetrance when you have an interaction with another variable. This is a recent, very famous case called Bcr1. Bcr1 is a gene, is a particular locus that will be a predictor of the greater risk of developing breast cancer in a woman.

Now, we know that it's very difficult for a woman to develop breast cancer before menopause, so you expect in their 40s to have a very low incidence. But if look at-- if you average all this, the population in-- across people with BCR 1 and people without Bcr1, you will see that the evidence is not really that big. They get the slight increase in chances of developing the breast cancer.

But if you split them according to their age, so you use age as an environmental factor, you see how the changes increase along time. So when you're 40, it's 37 chances-- 37% chances. But when you're 80, it's 85% chances that it's going to make a change. But if you put them all together, eh, you have something that is not such a clear result.

On the other hand, the challenge here is actually to find out which is the environmental factor that is actually doing the deed. This way is to find out that it's age that is splitting these women in these different groups. So this is a necessary, sufficient problem, but we have also redundancy problem.

For instance, we have retinitis pigmentosa that may be caused by 14 independent SNPs, mutations, right? So any of these mutations will show up positive to a particular genetic test for association with the disease. And we have other diseases that are known to have polygenic causes.

If you have only one of these, you will not develop the disease, like the Hirschsprung disease, in which you'd need two different mutations on two different chromosomes, which, by the way, opens an interesting question on, what is the relationship, the evolutionary relationship, of objects that leave off loci that are on two different chromosomes? They are completely independent?

Well, we have very complex systems that are displaced around-- there are genes displaced around different chromosomes, and we are able to rebuild this system from one generation to another. How do we combine this with the idea that if you measure things by centimorgan, two things-- it doesn't make any sense to have-- to compare the distance of two loci on two different chromosomes?

So this is a notion, kind of radical notion, but proposed in the '30s by Sir Ronald Fisher, father of statistics and modern genetics, under the name of affinity. He didn't have genotypes at that time, but it was a paper on nature, talking about this very strange phenomenon and how likely it was that there was really no relationship on things that were very far on the genetic code, in this code he was imagining. And they were, at the same time, functionally tightly related.

So how do we-- how we try to dissect all this complex stuff and also the simple one? Well, you know, the traditional way is to do the same game we did for Victoria's secrets, right, find a large pedigree, wait, you know, those 150, 200 years, and see what happens across three generations. We can do this with the drosophila much faster, but if we want to approach human diseases, drosophila doesn't really always cut it.

So if we want to study the same thing in humans, what do we do? One is to find a large pedigree. One is to say, well, you know, I don't really need a large pedigree. What I can get are trios.

I get mother, father, and child, a very modern nuclear family. And then I imagine-- because I know the parents-- I imagine that I will get statistical evidence of transmission by repeating the same measurement in different spheres. Right? Like, they were a huge pedigree, and in this case, I can do everything in two generations.

I can have-- but sometimes, for some diseases, finding parents is very difficult. If you're looking for two complex traits that show up late in age or-- there are people that are actually studying this, the basis of longevity. And they recruit people in their 90s, late 90s, early first century-- early second century. It's kind of difficult to track down their parents, and most of all, it's difficult to genotype them.

So what do you do in this case? You get brothers. If you get the brother, you try to figure out what was the genotype. So you assume that there is some kind of random distribution of these things across the two children, and you try to figure out what was the original genotype of the parents. Or you can do a standard case control experiment in which you collect a bunch of people-- half of them or both of them have the disease, other then don't have the disease-- and compare the association between a SNP, a mutation, and if so, all three.

There is another way to categorize these experiments. One is-- one are the experiments that are, like, double-sided, like case control experiments. I have two-- both sides of the story, the affected and the not affected.

Or I can have single-sided studies-- and I will show you how you analyze them-- in which you recruit parents-- they are healthy parents with an affected child-- and you try to see what is the transmission that goes from parents, child, that make the child affected. In this case, all your pool, all your recruited pool, would be made by people that are either affected or are related to an affected person. Which actually makes, usually, life much easier for the recruiter because, if I have to give-- to volunteer for study if I am affected or my child is affected, I am going to be more willing to volunteer for the study than if I'm just a random bozo that doesn't have the disease and by design I have to be independent of any subject with the disease.

So I will go into details for each of these studies. So the first thing we can do is to do linkage analysis. That is a traditional analysis of pedigrees. The second way we can do-- we can go is to-- is allele sharing.

Do you remember what I was telling you about, about brothers, and siblings in general? So if you have siblings, what you can do is, imagine that there is a random distribution for a particular SNP, for a particular mutation. And you would expect that if you have two brothers, these two brothers will have a particular probability of getting one allele or another.

Now, if you start deviating from this distribution, then you start becoming a bit suspicious. The association studies are standard case control studies in which you compute the likelihood that a particular SNP is causing, affecting your outcome. TDT, Transmission Disequilibrium Test, is the thing we actually use for the trios, and I will tell you in a second.

And then for complex traits, like quantitative traits, things that are not really binary diseases but are, for instance, your resistance to-- your inclination to start drinking, which may have different degrees, or your inclination to quit smoking, which, again, can have different degrees. The best way of doing them, of looking at these quantitative traits, is to use animal models, so crosses, which is not something we are allowed to do in humans.

Typically, these collections are hypothesis driven. So what these people do is, they wake up in the morning, they read a lot of papers the night before, and they take a shower and they say, oh, here are seven maybe related to my disease. So they go out, they ask money.

They recruit a bunch of people, use the money to genotype these people, run an experiment, and write a paper in which they either find something or they find nothing. Usually, if you don't find nothing, you can write a paper, and this induced misbehavior in scientists. The challenge here is that doing this is the old-fashioned way.

What is very precious for us today, or in this case, is the samples of people. Say, imagine you have a single drop of sample-- or of a-- from an individual and you can do a gazillions of these SNPs, but in this case, you don't have to read a lot of papers.

What you can do is-- or you can have your doubts. You can say, well, maybe it's not only this one. Maybe it's the second. Maybe the third. Maybe it's the fifth.

What if I collect 500 SNPs and then test all of them? What if I do an analysis that-- we don't collect any SNP? We collect all of them and try to look, to fish for dependencies in the data set.

OK, in this case, it becomes really, really difficult. This is how things used to be done. And this is a three-generation-- 3, 4-- four-generation linkage study. And in this case, you have a particular disease, and people that are affected, they are the red ones. People who are not affected are the bluish one. And you study a mechanism of transmission for a particular SNP, for a particular mutation.

And what you do is use a quantity called [INAUDIBLE] ratio or loss cortex. What is this? You develop a method-- a model of transmission that is accusing that particular-- the SNP to be responsible for the disease. OK? So the tracking down of the SNP will follow-- the transmission of the SNP, of the mutation, will follow the pattern of the observable characters in the individuals you see.

And then you make another one in which is-- in which the mutation is not responsible for it. And then using very simple statistical methods, you can compute that the data that you observe are generated by one model and the data that you observe are generated by the other model. And you can compare these two models, and you say, the model in which-- and you can say, the model in which this mutation is responsible for the disease, that is following the same pattern of inheritance, is n times more probable or less probable than the hypothesis in which the two patterns are not associated to each other.

Now, the problem here is that, if you have a large pedigree of-- you have many SNPs, you end up in the problem of multiple comparisons. So I have to tell you this because this is my job, but I don't believe in this. But if you one day want to publish a paper using standard statistical, classical things, you will end up in a thing called-- in a muddy area called multiple comparisons.

Now, multiple comparisons come from a classical superstition and from classical statistics. That is that you can infer-- you can say that two things are different if you can, by repeating the experiment a hundred times or a million times or some times, you will get less than 0.05 mistakes. You have heard of this p value. This is the p value. It is the probability that you will make an error by assuming that hypothesis more than n times. OK? In this case, it's five times out of a hundred.

Now, if I'm testing one hypothesis, this is my error level. Supposedly, I am testing two hypotheses. Well, the probability that I will do this by chance doubles, so it's going to be the product of these two probabilities, right, because these are two joint probability distributions.

Suppose I am observing 500, so it becomes 0.05 times 0.05-- smaller number. Suppose I have 500 SNPs. What should I do? That number will become so small that it would be impossible to prove anything, because the number would be-- because the probability that I pick up one hypothesis by chance would be so high that I will need a lot of evidence-- an unseemly amount of evidence to accept a particular hypothesis.

So frequentist people have this problem. Patient people don't. But it's a problem. If you want to publish in a journal, this is a-- this is something you will-- OK.

So this is nonparametric-- allele sharing is the thing I was telling you before, nonparametric method to assess linkage. And what you do is, you use siblings, and you assume that your particular SNP is identical-- the distribution of the SNP, of this mutation, is identical by descent. So I know what is the distribution I expect, and what I can do is to see, for a particular SNP, if there is a deviation from the distribution I would expect if these two guys were getting the way-- they were getting their mutation from the same pair of parents. Right?

So these are siblings. I look at the probability that we'll get this SNP by chance. And if they don't get the SNP by chance because they're all affected, what we find out is that these people that are actually affected have the greater chance of getting this particular SNP, and this will make my SNP suspicious.

Making it suspicious doesn't mean that you prove it. Now, if you want to prove something like this from a statistical point of view, using a non-parametric method-- because you make no assumption about distributional nature of your data-- you are going to have-- to need a huge amount of data. And this is not always feasible, but this is really, really feasible. So this is a weak test because it's simply telling you, the only thing you prove is that that particular allele is transmitted more than-- in a different way than random, well, more than its random population.

Now, association studies are typically done using parametric methods, and they test for association between a particular phenotype and the genotype in two sets, in two different populations, well, in two different samples. What is the problem? It's that sometimes you recruit your sample from two different populations.

I know it sounds absolutely silly, but this is one of the major concerns. Suppose the people in red are all people with asthma, and people with blue are all people without asthma. But then-- and they go and find out an association between the SNP and this-- and a mutation and then this phenotype. But then I discover that, for some reason, I have recruited everybody with asthma from Finland and everybody without asthma from Sicily.

Now, is this SNP going to account for asthma, or is this SNP going to be account-- going to account for whatever this variance is between these two populations. These are called population admixtures, and I know it sounds silly for people to recruit things across different countries or buried, segregated populations. On the other hand, you may have-- you no guarantee that you don't have stratifications in your population. Even if you recruit in beautifully multicultural Massachusetts, the probability that you will find out something about the-- something about your sample that you don't really like, like a stratification, is not negligible.

So what is really, really trendy today is to use transmission disequilibrium tests. Transmission disequilibrium tests are based on the assumption that you're actually using the parents of an individual as his controls, so you're sure you are not going to get any stratification, right? So I'm going to recruit a bunch of triads, father, mother, and child, and the child is affected and the parents are not affected, OK?

And then what I'm going to do is to compute for each particular SNP, for a particular SNP, for a particular mutation, if there is a dependency, if there is a pattern, so to speak, between my distribution of transmission and the fact that all these kids are actually affected. And in this way. I have a very powerful test and non-certified controls. The only problem is that it's not always easy, as I was saying, to find parents for this.

And the other problem is that you may find phenotypic stratification with that. Suppose you're looking for a particular disease that is related to weakness. An example I have is cholera. There are-- people believe that there is a SNP and people are looking for a SNP that is more-- making people more susceptible to have a bad outcome from cholera.

And what they do is to go and recruit the people in households, so when somebody shows up in the hospital, they recruit the entire household and use the parents as control. Now, at that point, the problem is, because the phenotype is not really that easy to identify, we don't know if these people, for instance, got an immunity 20 years before because they got cholera. These are parents of a child who has-- who is 15 years old. In places where they have serious cholera, usually these are still young people, but from the societal structure, they're kind of old people. They have been there around for, like, 35 years. And this means that, at that point, you have this particular population that by design has a very different-- a phenotype that is very difficult to characterize.

So QTL-- traits that, like in this case, variability in intensity. I'm not looking for something that is black and white. I'm looking for something with different degrees of severity. And until last year, there was really no way of doing these things. Well, nobody has proven ways of doing this.

Now, people are working very hard because QTL are actually one of the most interesting things. People are looking really hard to find ways to characterize at least some type of QTL. And you can-- and complex QTL, like with census data. So one of the studies we have been doing is about the development of-- breast development in women.

You say "why?" Well, because breast development has been associated to breast cancer, so we know that late breast development is protective of breast cancer, at least one type of breast cancer. Now this is a QTL. Why? Because you're looking at age, which is not a binary variable.

And it's a very complicated QTL because it is-- you may have sensor data. You may have girls that at some point decide for one reason or another to drop out of the study. So the way which you can organize for this particular QTL-- not for every QTL-- is to imagine, to represent the structure as a survival study, in which you imagine that your SNP, a particular SNP, is going to be a treatment and the other is going to be the control.

You draw your Kaplan-Meier curves, in this case, not for survival but for breast development, which is a kind of a deadline point. And at the end, you can compute the difference between these two samples and get evidence that actually there is a factor there that is putting them together. And you can use other multivariate structures that allow you to study the interaction of SNPs in QTL. But typically, if you want to have a general model for QTL, you have to resort to animals, animal models.

Now there is another interesting thing that is happening on this street mostly, and has been happening for the past four or five years. There is-- we have two or three large phenotypic studies run by Harvard and affiliated hospitals. The oldest one is known as the Framingham Heart Study.

Framingham Heart Study collects, I don't know, 50,000 people. They have been followed, and now they are the third generation. They have been followed. Their family members have been followed. We know about these people basically everything.

They're-- these are selected for heart disease, but there are also studies-- there are cohort studies that have not been selected for anything, like the Nurses' Study. Nurses are wonderful individuals, incredibly compliant to doctor's order. So even if you harass them for 30 years, asking them every other year a complete report of what they have done, diseases they have, what is their diet, how much they weigh, they will comply.

Nurses' Study has 150,000 women that, every other year, go and fill a questionnaire just for the sake of it. So about these women-- and this is now in the second generation-- we know absolutely everything. We know how much carrots they eat every year.

And you probably see one-- the reason why you see these articles in the New York Times, they come out of chance, or the School of Public Health, but usually they mean chance, in this case. There are people who own this data.

And these articles are like, you know, red hair are going to make you live longer. Or, you know, eating five carrots a day has been associated to having brighter eyes. Where those findings come out from? From the nurses study. There are these people that mine the data set.

Again, they don't read in this case. They take a shower in the morning and say, what I can go and look into the nurses study? And they say, well, how about the association between mascara and blindness?

We have 150,000 women that use mascara every day. Well, actually, you probably have a good proportion that uses mascara and a good proportion that doesn't use the mascara. And you know it.

So they go there. And the joke on the street is that they don't publish in *New England Journal.* They publish directory on *The New York Times* because these are usually very high impact questions that people get right away. But imagine if you can take all those things, all those phenotypes, and you could genotype those women-- we have blood for about 90,000 of them-- and run full genetic scan for these women, what could you do?

You could find an association between mascara and SNPs, or blindness and SNPs, or whatever and SNPs. Can we do this? Why don't we do it? Well, OK, genotype is net cost about $0.45, just to let you know where you are. To snip an individual map is going to cost-- so sorry, an individual map is going to be about 90,000, OK? So 1,000 individuals is going to cost you 90 million.

And that money, they're very difficult to find in that structure. If you go to NIH and ask for this kind of money, you are going to get them. Work is not the problem. Postdocs at Harvard are cheap. And they work very hard. So we could actually have 1,000 of complete SNP maps in about 7,000 days, now, which is about two years for 10 postdocs, nothing.

How can we solve this problem? Well, you remember when we said that these things go together? Maybe we don't need to genotype all of them. This is an example of a gene. I remember we-- I don't remember what gene. But this is a map of a gene that tells you the distance in terms of r-squared, which is the measure I was telling you before, the distance between-- the pairwise distance between two points.

And as you can see, you can identify blocks there, the blue areas in which the correlation among the r-squared in this area is so high that you don't really need to genotype all of them. You can just decide where you go. And you say, OK, pick any of them. Problem is to know this, you first have to genotype them. So at that point, there is not much purpose for it.

And the other problem is once you have this-- so here it's easy, right? When I have a region-- that region of recombination, there are four or five of lower recombination. I can say, OK, I pick up one, two, three, four. Now, the question is, which one I pick? Now, suppose they don't have five. I have in this case, 59, which are not that many. Regions like this there are 229 long.

So well, your guess. How many SNPs do I need to genotype? Not which ones, of course. It would require a little. But how many do I need to genotype here to account for all the variations? So these are the SNPs. These are all the combination I get in the population, OK? These SNPs go together. So I don't expect to see all possible combinations. How many of them I will need to account for the 14 different alternative variations of [INAUDIBLE] over these 59 SNPs?

I want the number. Jose, you're up to your third pizza. 15?

AUDIENCE: 10.

MARCO RAMONI: 10? OK, we have a 10 here. Anybody getting closer? That's a good guess. You will buy 10? Buy 10? Five. These five SNPs allow you to reconstruct all the variations without any information loss. I have a guy, Channing, a good friend of mine. And when we develop the method to identify this thing, bought me dinner because he has spent two years staring at these pictures and trying to figure out which were the SNPs that came [INAUDIBLE].

And actually, when he was wrong, getting a lot of shit from people because, oh, you got me the wrong thing. Now I mentioned it's a mind numbing task to stare at this picture and try to find out which are the optimal one to genotype. So what they have been doing is to play with the structure called haplotypes. Haplotypes are the things you get from your mother and from your father. And there is stretch of the genome that is consistent one chromosome on another.

Now, the problem with haplotypes is that they're very difficult to identify for us. When you sequence somebody, you are going to have maps that look like the one I showed before. I can tell you if you're heterozygous, homozygous, or majority all homozygous, and of the minority of. But I cannot give you two bases. So if that particular SNP is A and T, as two alternative bases, I can tell if you are A-T, T-T, A-A. But I cannot tell you if you are T-A. I cannot decide if two things are coming both from your mother or both from your father, which is exactly the kind of information we need to build these kind of maps to reduce this information.

So these structures are called haplotypes. And there are molecular methods. They are horribly expensive to do this. So they go into the chromosomes. And they tear apart the chromosome and return to separate chromosomes. And you know that one is coming from your mother and one is coming from your father. There are stochastic methods to do this in which you can try to figure out by making some assumption how these things are transmitted on a population. What is the distribution of an entire haplotype?

But the best way of doing this is to use trios and try to guess what is the SNP that is coming from the mother and which is the one coming from the father. And so what people are doing now is developing a thing called HapMap in which we will have a map-- you can go to hapmap.org. And you will find all the data that's currently available.

It's systematic. Now the resolution is 30,000 bases, 30 kilobases, a systematic genotyping of all SNPs on a relatively small population. There are 33 years in which we can actually find out which are the phases-- the fact that one is coming from the father and the mother is called the phase of a genotype. We can identify the phases from the parents. And why you do this? Well, because next time we want to design a genetic study, I don't have to run the genetic study and then say, oh, I could just genotype this one.

I can actually go there, see I am interested in this particular gene, find out which are the SNPs that I really need to genotype. But there is some even more good news that apparently, all these things actually go together in stretches. So this is a paper from a couple of years ago in which they show that they made a high resolution map of a region of chromosome four associated with Crohn's disease.

And what they found aside from the clinical things, what they found out is that this region is actually from an evolutionary point of view broken in 11 subregions. These regions are stable. They are transmitted together, so to speak. And they are interrupted once in a while by [INAUDIBLE] regional recombination. And the [INAUDIBLE] is they found out that if you look at all these variations of the alternative common haplotypes you have in the population, they actually come-- you see the different colors?

They actually come from four ancestral haplotypes that are recombined. They've been recombined over the generations. So again, this gets back to your question. I'm not sure that there is, from a physical point of view, there is no-- you see the big block there in the middle, 92 kilobases? Well, I'm not entirely sure that there is not a very high recombination spot there. But because they all come from four haplotypes, with not other variation, I cannot see any recombination there because the probability of recombining requires a little pool to fish from.

And in this case, they pool to fish from is very, very poor. Needless to say, these are all Caucasians. And they all come from this handful of individuals. So apparently, these people all come-- there's 129 SNP pairs-- not SNP pairs, trios. They all come from four ancestral haplotypes. The way you do is to use-- the way you identify this block is to use the Markov models.

I will talk about the Markov models next week for the machine learning method. So I will not bother you with this. And at this point, what we can do once we have these blocks is to identify the SNPs that are tagging those particular blocks and then run them. And the same thing we've shown is an exponential saving with length. So it's about an average 10%. But if your block is very big, the number of SNPs you will need will be much smaller than 10%. And if it's a small block, the number of SNP you will need will be a bigger number.

So the take home message, the fundamental take home message for a medical class is that all this technology has these beautiful properties of finding out stuff. But the only way to find out this stuff is to have good phenotypes.

One day, running a SNP will cost $0.01. Maybe it will cost a tenth of a cent-- near nothing. But what makes very precious, what makes these studies possible, is the fact that you have something like the [INAUDIBLE] study that you have critically annotated, carefully recorded phenotypes and nicely characterized. And this is where the other side of medical knowledge comes in, not to find out what is really the cause, but just to identify precisely, as crisply as possible, a particular phenotype.

At that point, you can actually map the phenotype. If your phenotype is slippery, yeah, well, it's very hard you will find any association. And if you find it, it probably will not be association with the thing you want because that's going to be a different phenotype, the one you're thinking. So the critical thing here is not, to go back to the very first slide, is not really to do bean bag genetics.

The real critical thing here is to have good medicine that is able to characterize phenotypes that in turn allowed these studies to happen and to create good explanations for these phenotypes. And the other big take home message is that hypothesis-driven is out of fashion. We don't need hypothesis anymore. With enough money, we can have skills. We have the intellect. And we have phenotypes. We have enough information. We have enough juice to go and get our answer without thinking them in the shower, but looking systematically at the genome and taking advantage of the fact that something is more likely with something else.

So if I tell you that something is associated with something else and this is my only study, we will need 10 years before people can consolidate the studies in a meta analysis study and say, oh, the evidence is not that big. But if I start telling you, I analyzed 500, or 5,000, or 50,000 SNPs and here it comes. This SNP is a million times more probable than any other SNP to be associated with this phenotype. That's a measure that you cannot get from an hypothesis-driven test.

So OK, so when you will be reviewers of papers and grants, usually back people because their studies are not hypothesis-driven, remember this. See you next--