The following content is provided by MIT OpenCourseWare under a Creative Commons license. Additional information about our license and MIT OpenCourseWare in general is available at ocw.mit.edu.

GEORGE CHURCH: OK. Welcome back. We're going to go through a very specific example of association studying, illustrating this extremely important statistic, the chi-square statistic in the simplest case. That simple case will be a single allele combination, two alleles and two possible phenotypic outcomes, HIV resistance and HIV non-resistance.

So to set up this association study, which would be computational, let's talk a little bit about the biology here. All viruses need to get into your cells somehow. HIV has a number of proteins to which it binds on the surface of your cell. These proteins are not-- they're not designed to bind HIV. They do something else.

This is in a family of chemokine receptors. This is involved in intracellular signaling, and so it's a receptor normally for these chemokines. But it is also a receptor for the virus.

And in the human population, as we alluded to earlier, there are at least two alleles, two common alleles. One of them, the top one, capital CCR5, has this long, open reading frame. And this you could translate by your pearl strip that you wrote for the problem set. And the DNA sequences in the middle here and the protein sequence derived from it is on the top strand. And that's the capital CCR5.

And the little delta or deletion Ccr5 is below it. You've locked out 32 base pairs, which is not an interval multiple of three. And as you know from last lecture, that means it's going to be a different genetic code. At every point downstream, this is going to be frame-shifted, and so you get a whole new set of amino acids for the entirety of the rest of the carboxy terminus.

So this in black here-- so we're showing here in the single-letter code a somewhat realistic folding of the-- schematic folding of the protein, showing here disulfide between two cysteines. Transmembrane region, we have these hydrophobic alpha helices. And finally, in the black amino acids, the C terminus, the end of the protein are all substituted in this deletion mutant.

Now that is sufficient to cause resistance to the virus. It probably has some effect on its ability to bind chemokines or react to them in some other way. And presumably, this is not-- there is some deleterious effect we do not know about.

But in any case, we can assess this. We can make a genetic assay for this. And we can look in human populations for their resistance to HIV and for the presence of either two alleles of the resistance or two alleles of the susceptible.

Now I'm kind of biasing you here in this. We haven't done the association study. I really shouldn't be referring to it as resistant or not, but I think it helps you visualize it. We'll be rigorous enough in the next slide.

So here is, if you will, the big allele, right? It's the original. It's the non-deleted allele. So when you do a PCR assay where you're amplifying with two primers across the region that could be deleted or not, you'll tend to get a large amplification product. You prime synthesis until you get something which has 403 base pairs. And there's enough of it that you can display it on electrophoretic gel, and it migrates slowly because it's large.

If you have a homozygote for the deletion allele-- we'll not call it the HIV resistance allele just yet-- then you'll get a 371-base-pair PCR product. And if you have the heterozygote, you'll get both, the large and the small allele represented on this electrophoretic assay. A very simple, very robust DNA PCR-based assay.

So now let's ask whether one allele or the other is more abundant in people which are observed to be seropositive or seronegative. Seropositive means that they have circulating antibodies in their serum which react positively to the HIV virus. And so that's not necessarily, but is commonly associated with being heavily exposed to the HIV virus, and therefore basically infected. Seronegative means that you are-- it's an indication that you're resistant.

Now we could do this as a two-by-two matrix here of alleles versus outcomes, or we could do it as a three-by-two where we have genotypes versus outcomes. The three genotypes are big-big, little-little, and big-little heterozygote.

But let's just do it-- just keep it simple in terms of alleles. If there is a selective advan-- or there's a perceivable association between the seropositive, seronegative, and the allele, you should be able to see it in both the genotype or just the allele. The allele is simpler. It's a two-by-two matrix.

And so these are the data, just these four boxes here. The big allele is 1,278 observed negative and 1,368 observed positive for a total of 2,646.

You can see that there are fewer total in the population surveyed of the deletion mutant. This is consistent with the claim that I made all along. It's about 9% in general human population, and it's 9% of this population. That's good.

But now you have to correct for that. You need to-- you can't just ask, is it more frequent here or not? You have to correct for the actual frequencies in the population.

And the way that's done is you calculate another table, which is the expected number of each of these combinations of big with seronegative, big with seropositive, and so forth, under the assumption that it's completely random. You know the allele frequencies in the population, but say that they just randomly associated with whether it's seropositive or seronegative.

You use these totals here and the frequencies in the population to generate this expected end of the random assumption. And then you look for deviation of the real ones of the observed versus the expected.

So any deviation between the expected and the observed is going to be significant in this. So you want to do is take the difference between the observed and expected.

So we're working on a statistic here, a type of measure that will determine, how far from expectation are the observations on the left-hand side in this two-by-two? And so for every square in the two-by-two, you find a corresponding expected number and you subtract it. That's the starting point.

But you don't care whether it's negative or positive. You want to make it positive. So the trick we'll often use is to take the square of that. So you could take the absolute value.

You take the square for the chi-square, oddly enough. And then you want to put it on some kind of standard scale so that when you do a chi-square for any kind of phenomenon, you'll be able to compare.

And the way you do that is you divide by the expected number. Or as this puts it on, if these numbers were really huge, this would bring them down. Very small, it'll bring the same point.

And so you take the sum over all the squares in here, and that's 15.6. And that's saying that it deviates from the expectation by this amount normalized as a fraction of the expected.

Now to determine just a little sidelight on the chi-square, in order to determine the probability of 15.6 being significant, what does that mean, looking at this at 15.6? You want to turn this into a probability, because probabilities are the common language that we can all share our surprise at this being different from expected, the null hypothesis that this is the same as expected.

And so in order to evaluate that, you have to ask how much freedom there-- degrees of freedom. This is jargon for, how many different ways can these two numbers vary?

Well, since we know the total number, when we observe the number of CCR5, that fixes in the columns the number of deletions, right? Because it's just going to be the number of big alleles taken from the total.

So in a way, the degrees of freedom is just one. You fixed this one number, and the other one is now known. And the same thing's true for rows and columns. So the rows minus-- it's just one and the columns is one. And so the whole degrees of freedom is the product that rows and columns, which is one.

So you plug this in. If you look in your standard statistical-- or your favorite statistical book or software package, Excel or whatever, you plug in this chi-square value and this degrees of freedom, and you get this probability, which is very significant. Generally better than 5% means that you will only be wrong 20% of the time. Some people would prefer to-- this means you'll be wrong eight times in 10,000 or something like that. Very, very infrequently.

OK. Now this is great. This is a two-by-two matrix. We found an association. We have a plausible molecular mechanism in the previous slide. But how was it that we pulled CCR5 out of-- this rabbit out of the hat? I mean, why CCR5? There are 40,000-some genes in the human genome. Yes?

AUDIENCE: What is it? I was curious as to, what population is it? I mean, in which there's slightly more seropositive or seronegative people. This must be a particular--

GEORGE CHURCH: Oh, yeah. This is a case control study where you try to get a roughly equal number of negatives and positives that are matched for socioeconomic group and race and gender and things like that. So that's a setup.

But then the-- if there's not a huge risk ratio for the two alleles, then you'll expect them to be close to the population frequency, which means that you can't just look at it on inspection and say, oh, yeah. All the seropositive are the big allele. We wouldn't need a chi-square at that point. But this is pretty close to the population frequency, and so that's why we had to do a chi-square. OK.

But the question that we're taking now is, how did we pick CCR5? And I introduced this as being a putative receptor for the virus, but one of the first pieces of evidence that this was a putative receptor for the virus was this association study.

So how was it that it hit the chemokine receptor? Well, you could say, well, because some kind of hunch about chemokines being involved in immunology, and immunology being important in fighting viruses, but that wouldn't suggest that it's an actual receptor.

I think these are-- there's all kinds of inspired guesswork. There's biochemistry happening behind the scene, and so on. But let's just put that aside for the moment and take the more general case that you wanted to test not just this one hypothesis, that CCR5 is involved, but you've implicitly then tested or explicitly tested every gene.

You've gone through every gene and you've taken either the most common alleles or you've sequenced your own genome and you've found the alleles that you have, whether they're common or rare. You don't care. And you want to ask, what's the association?

Let's consider the upper right-hand panel here where we have some risk ratio, a fairly subtle one, 1.5. Remember, we were talking about risk ratios of 75 in the case of autism. This is just 1.5. Very subtle risk ratio, just like I think in the last one could have been a subtle risk ratio.

And the x-axis is going to be the number of alleles that you're hypothetically testing. This could be related to the number of genes that you're testing, or it could be more than the number of genes that you're testing, because you can have more than one allele per gene that you might want to test. Is this allele of this gene important in this disease, or is this other one?

For example, sickle cell is important, but various other hemoglobin mutations are not. This chemokine deletion is important and maybe another one isn't. Many chemokine mutations will be neutral.

OK. So as you increase the number of genes and alleles from-- here's 10 to the fourth on up-- then the number of sib pairs-- so this is a simulation covering all kinds of experiments that you could do where you can use computational methods to help guide the design of experiments.

If you did an expensive experiment and you just happened to use too few patients and too many alleles, then you may have misused your resources. You should have maybe done fewer alleles and more patients or something like that. And this provides some guidance here.

But you can see that in order to get to a very large number of alleles, you need a fairly modest increase in the number of patients. And that's due to the exponential term that you have in these probability distributions.

So you actually-- but nevertheless, it's a big deal cost-wise going up from, say, 400 patients to 1,600. Nevertheless, it's only linear in patients while it's exponential in the number of alleles you can test, so that's the good news.

And you can see some of these other panels show the effect of varying other parameters. For example, here on the left you've got varying the population frequency of the allele. As you get to very, very rare alleles on the right hand of the horizontal axis here, then the number of sib pairs-- brothers-sisters, brother-brothers that you need-- starts to go through the roof. And now it's exponential on both axes, or that is to say that it's a direct relationship. Question?

AUDIENCE: What's that z value over there?

GEORGE CHURCH: Hmm?

AUDIENCE: What's that z value of over there?

GEORGE CHURCH: Oh, that's the population frequency that you're--

AUDIENCE: On the bottom?

GEORGE CHURCH: Hmm? Well, on-- on the one we were just talking about on the left, the population frequency is the horizontal axis, and it varies from near unity to 10 to the minus nine. It's a very rare allele frequency.

On the right, the one we started with, the top right, we just picked 0.5, which would be very-- be around here on the left-hand end of the left-hand quadrant. So you pick one of these allele frequencies, a fairly-- where it's equal frequency of the two alleles in the population as where you do the rest of the simulation.

You can think of all these panels could be some multi-dimensional display, but I wanted to take them out one at a time. And actually, I did all these in Excel using the equations that are present in this reference that's given in the slides here. A good sign of a well-written paper is for a sort of average individual to be able to reproduce it.

So we have how many-- so we've been talking about new polymorphisms. In fact, the question that came up during the break was, if selection and drift cause allele frequencies to fix fairly rapidly, drift in small populations, normal-sized populations will drift in fairly short evolutionary times and selection if you have a high selection coefficient of cost fixation. Why do we have any allele frequencies at all? Why isn't everything fixed at 100% for one particular allele?

And the answer is mutation. And where all these new mutations or polymorphisms come from-- actually, these should probably be called mutations, because most of them have frequencies less than 1%.

Well, here's a specific model which doesn't follow all the assumptions we had before, but we'll use a different-- we'll use a back-of-the-envelope calculation to give you a feeling for what is true for the-- what is likely to be the case for the human population.

So the human population, there's a little bit of unknown in some of these parameters, so you should take these all with a grain of salt. Actually, everything I say you should take.

But the number of populations that we've had since a bottleneck in the human population-- maybe there were as few as 10 to the fourth humans at some point or another. Maybe less.

But since that time, there have been about 5,000 generations. And during that time, our population has now grown to six billion people. That's the n prime nn. And the mutation rate, as mentioned earlier in response to a question, is around 10 to the minus eighth per base pair per genome per generation.

And so the genome size is about six billion. Coincidentally, the same as the population size. Total coincidence. Then 10 to the minus eight times that number of base pairs means you have about 60 mutations or so occurring in a generation. So you've got a steady flow of new mutations.

Now if you take that-- just roughly speaking, if you take that over 5,000 generations, you've got 60 times 5,000. You've got a very large number-- on the order of 10 to the fourth mutations that have accumulated in our population. Assuming relatively little drift because of this exponentially growing population and relatively little selection, then this is the number that might have accumulated.

You can do subtle corrections for this exponential growth. The number of mutations you got at the beginning will have higher frequency, but there'll be fewer of them, because population was smaller. But anyway, you get the picture, is that the total number of mutations in any one of this that are new since that new-- since 5,000 generations ago will be on the order of 10 to the fourth in your body, not all of them doing good for you.

And for each of those rare mutations-- they might have a frequency of about 10 to the minus five-- there will be about 10 to the fourth people on Earth that will share that very rare mutation with you. 10 to the minus 5 sounds very rare, but when you multiply it by six billion people, there are a lot of people who share that with you. It's a new mutation.

OK. So the take-home is-- and this is from this reference here. High genomic deleterious mutations accumulate over these 4,000-- some large number of generations. And they would confound linkage methods and they would confound assumptions that say that the common alleles-- common alleles are causative.

OK. So let's say we've done an association study. Either we've done it on one gene, picked one out of the hat like CCR5, or we've done it on a full genome survey, 40,000 genes, all the alleles we know of.

In order to do the latter, we had to have a big patient population. But let's say we've done that. Now we want to prove that that association is-- that great statistic that we got out, if we have a large enough patient population, is still just a statistic.

What will constitute proof is after we find the association, we make a copy of that mutation, isolated away from the three million other polymorphisms that are floating around in your body, and we do some kind of test. Ideally, we would make an isogenic pair of humans that only differ by this one mutation. That is not generally considered medically feasible or ethical.

But somehow or another, you can do it on human cells or you can do it in a mammalian model system. But you have to make something that's close to isogenic, a copy of this mutation, and show that it has a phenotype that makes sense. And then, just to make sure you haven't introduced other mutations, the really careful scientists will then revert that one polymorphism and show now you no longer have that phenotype.

OK. So let's walk through an example. Here's the third example in this lecture of a very specific allele where we've shown the molecular basis of it, and it's the second example where it's not coding, and it's the second example where it's going to repeat.

So this is just to-- I'm not giving you a random sample here. I'm specifically biasing this towards very interesting genes, very interesting traits that are associated with non-coding repetitive alleles.

In this case, the trait that got the attention of these researchers-- it isn't necessarily proven that this is the relationship. But the association that was studied was between anxiety-related traits, anxiety, and a polymorphism in the serotonin transporter. This is a science paper.

Now the next step in this is to-- OK, that was found associated. There were lots of other alleles randomly being moved around, because these are humans. We have no control over who mates with who, or very little. So you get what you're given. You can do the survey as best you can.

But now in order to move it towards a mechanistic basis and a proof of introducing just this one point mutation. Now what is the mutation found in this case of anxiety relationship in this serotonin transporter? It's upstream of the first RNA encoding axon one. It is a 44 base pair deletion in a repetitive region, in a region which might be responsible for promoting transcription.

You have the short and the long alleles, just like in the CCR5. In this case, it's in a putative promoter element. And when you make this construct and you hook it up to a luciferase enzymatic activity in vitro, in cultured cells so you don't actually have to construct a mutant human, you can now see the long allele always produces higher levels of expression than the short allele.

And these little error bars on top of each of the measures show your statistical measure of standard deviation, as we did in the first lecture. And if the black mean, which is the height of the bar, is different from the white mean by more than a couple of these standard deviations, which is the root mean square of the standard deviations of each of the measures, then you call it statistically significant.

And that's what these triple stars mean. It's the statistical shorthand for this is statistically significant at the cutoff that we're using. Say, 5%. Now in this case, it's kind of showing off because every one of them is statistically significant by two different-- completely different assays here.

But the point is this does not prove that anxiety is meaningfully associated with this. But it does prove that this repetitive polymorphism is causally capable of differential transcription levels for a reporter gene. So by introducing it into a clean cellular system, you can test at least part of the mechanism that might be involved in the association that originally got your attention.

Now as you can see, we're getting into a fairly mature and fairly new phase of human genetics. Some of the concepts that are used here will be useful in a variety of other systems where we have relatively little control over the genetics. But certainly in humans, there's a very large need.

And where this happened historically is with Mendelian linkage, which would involve very large families with a complete pedigree, multi-generation, great-grandfathers and mothers, all the way down, hundreds of people in a family. And that's the problem, because there aren't that many diseases for which you have large families with simple Mendelian inheritance. Common diseases tend to be much more complicated, involving multiple genes simultaneously and small families.

Then there's linkage disequilibrium, which depends on these common alleles where the population has gone through a particularly small bottleneck and where your common allele is a pretty good distance away from the causative one. But it allows you to map it into a ballpark, and then you hunker down and do the more expensive testing of hypotheses for all and sequencing and looking for potentially causative alleles.

Once you find a potentially causative allele, something that looks suspicious, maybe something that's in a coding region that it's conserved-- that's the first priority, despite all the counterexamples I've given. And then you go ahead.

But the problem is that you're at the mercy of the recombination that might have occurred in that population, but you get to very interesting populations like people of African American descent where the population is very old, hasn't passed through a population bottleneck.

The useful linkage disequilibrium is on a couple of kilobases rather than hundreds of kilobases, so it actually is hard to find things that are linked. That is to say, insists but not causatives. Insists to the causative allele. Instead, you have to look directly for causative alleles. And that's where-- now you look for common causative alleles. This has the problem that we've been talking about, which is that maybe common alleles aren't often causative. They've been selected out.

So then you get into the scenario, what if we wanted to look at all alleles? Well, theoretically, there's nothing wrong with that. It'd be great. We could do the association studies. We'd have to have fairly large populations. But you saw it was linear with an exponential increase in the number of hypotheses, and you could rank those hypotheses by a priori likelihood and so forth, but it's expensive.

And so the discussion now has to include a little bit of discussion of the new technologies that might make this less expensive, and a discussion of how we got the first genome and how the new technologies might be a little different.

But we're going to do this in the context of computation in the sense of, how do we deal with random and systematic errors? As we plan our strategy for getting each of our genomes, personal genomes sequence for $1,000 or so, how do we choose which technology to pursue? Which ones have the lowest intrinsic, random, and systematic errors?

So we needed to study the first genome to see, what would we mean by random and systematic errors? How many people here know what the difference is already? A few. OK. But anyway, whether you do or not, it's good for the soul to go through some real examples of this.

A random mutation is something which occurs every time you do the experiment. You get a different error. And a systematic one is something where if you do it over and over the same way, you'll get the same error or a simple class of errors occurring more than once, over and over.

So for sequencing, the process involves picking something that you want to sequence-- we'll call it a clone or a template. Then actually doing the sequencing and then assembling this into a meaningful interpreted sequence.

And this is an example where we might have chosen these big clones. In other words, you might fragment the genome up randomly into large clones of 100 kilobases called bacterial artificial chromosomes, for example. And then that's all random, and so there's a shotgun of that scale.

And then we break it up randomly into smaller pieces, which provide us with little sequences. And then in the computer, we assemble these by methods that we'll discuss in the next lecture, which can take even fairly different sequences than assemble them. In this case, we're talking about very similar sequences.

And then you assemble the little sequences into big sequences, and you assemble the big sequences into even bigger sequences, and then you have the whole thing. But you can see there's a lot of opportunity for error here. We make it look simple in this slide, but we're going to talk about random and systematic errors in a moment.

Where did we get those sequences, those little sequences that we want to assemble? We're going to go through a few methods in a moment, but the most common one by far I think that generates over 90% of the human genome sequence in the last couple of years is capillary electrophoresis of fluorescently labeled polymerase terminated products.

When you electrophoresis things, you're separating DNA fragments, which are n nucleotides long, from a DNA fragment which is n plus one nucleotides long. That's a very subtle difference as n gets big. As n gets big, it gets harder and harder to separate n from n plus one.

As it does so, you start getting all kinds of errors. The total number of errors here in the lower left-hand corner is the number of insertions plus the number of deletions plus the number of substitutions plus n, which is an abbreviation for no call. It means the software felt that it was so close that it couldn't call it at all. It just said it's n. I don't know whether it's an A, C-- it doesn't know whether it's an A, C, G, or T. Calls it a no call.

And so if you look through this table, it shows on the vertical axis for each of these six bar charts is as you go up on the vertical axis, you go from very short reads where it's easy to separate n from n plus one electrophoretically to very long ones where you start accumulating insertions deletion errors, insertion deletion substitution, and n. They all go up with length.

And so you can think of that as a random error superimposed on the systematic error of if you always do the experiments, the same base pair is always at the end of your run, it's going to always have a higher random error rate. So this is kind of a combination of random and systematic error.

So let's go through some examples here. We have just isolating the template to prepare it for sequencing, there are systematic errors. If you have certain kinds of repeats, long, inverted repeats, or certain kinds of restriction elements that the bacteria doesn't like or likes to chew up, then you won't get the clone.

And that would plague the early part of the Genome Project, is you'd keep trying again and again the same way, and you just wouldn't get certain clones. It's as if there was a hole there. You know that there's something there, but you couldn't clone it.

Sequencing-- hairpins can form in the single strand of nucleic acids where you're trying to separate n from n plus one. And those hairpins then take out the gel electrophoresis and make it seem like it's much smaller than it actually is. Tandem repeats cause a problem in all three of these stages in sequencing. You get some polymerase stuttering and you get little artifacts.

In assembly repeats, you're assembling by sequence alignment. If you have a repeat, the repeat's going to align just as well because it's a repeat within the genome as a repeat due to experimental repeating, and so the alignment can be off.

The errors that you got from these earlier steps make it hard to assemble. Polymorphisms look like errors. Chimeric clones means you've got things that have been misassembled early on here, and so on.

OK. When we do this random selection of big clones and random selection of little clones for sequencing, we want to know when to quit. Now you could say, well, we're going to quit once we get it all assembled, but you need to accumulate a certain amount of data in advance before you try to assemble it.

And so there are various calculations as to when to quit. And this is related to the Poisson distribution, but it's not exactly. And in fact, one of the studies done in 1988 made some poor assumptions about-- remember, we mentioned in the first lecture the assumptions of the Poisson distribution that make it an approximation-- when it becomes an approximation to a more formal distribution like the binomial.

Anyway, that approximation, you should be getting, as you get higher and higher coverage, meaning more and more experimental repeats, you'll eventually fill all the gaps. And that means you'll get 100% complete coverage, and it should asymptotically approach that.

Well, if you use the Poisson incorrectly, as authors did in 1988, you basically go off to infinity. You get 200% coverage, which is physically impossible.

But both an earlier study, which they ignored, and a more recent study got it right. And these are slightly different measures, but they both converge on 100%. And I urge you to look at that if you're designing an experiment with simple formula.

On the other hand, if you want to design the experiment more explicit, you eventually get to a point where a simple analytic formula won't do, and you have to do Monte Carlo. We treated this in the first class of analytic versus numeric differentiable equations and then many other cases.

The simulation you want to do is beyond-- is too hard to do analytically. So Gene Myers set up-- just listed all the things he thought would be-- could affect the ability to assemble a genome sequence. The read length and types of repeats, and all this stuff was simulated. And he cranked the simulation on real human genome size projects and came to the conclusion that you could do a shotgun assembly of a mammalian genome.

A company called Celera was formed. Gene Myers was hired as the computer guru. They put together a large stable of computers and they started doing this on the human and the mouse genome.

The human genome was in the end not done by this method, but the mouse genome was, and it was a pretty good assembly. So the Drosophila genome missing the Drosophila repeats was also done by this method. So this is scaled very gracefully from the first shotgun sequence, which was on a four kV plasmid to mammalian size genomes.

No mammalian genome is completely sequenced, so we can't really declare victory. But the sort of simulation that he did here has played out very nicely.

Right here in Boston, there will be-- when we start thinking of the future of sequencing technology, we would all like-- this is almost uncontroversial as to how much we desire that the genome not cost $3 billion, but it costs $1,000. And now there are a number of people who are taking very definite steps to get us to that point.

To understand systematic and random errors that can occur in these steps, let's take this dideoxy gel electrophoresis we've been talking about here. You have four different colored terminators. These is where the polymerase can't go any further because there's a blocking group here, which is fluorescent labeled.

And if the template on the far left here is ready to accept an A, you'll get an A. And this n and n minus one will separate on this electrophoresis and the four colors will give you this four-colored pattern where the intensity reflects good termination at that position, and you can basically go along reading it. G, C, G, G, A, T.

Now this is in the well-behaved part. An example of a systematic error that you get in this extension occurs here in the upper right-hand corner of slide 40. Now you've got-- because of one of these hairpins that I mentioned before, you've got a pile-up of seven nucleotides all at the same position.

A completely alternative method that doesn't involve gel electrophoresis is called pirA sequencing, and this is an equivalent that can be done with fluorescent addition. Instead of separating them in time by their size in electrophoresis, in time you ask serially, do you want an A here? If yes, then you get a little green-- you get a little peek of representing the release of pyrophosphate or the incorporation of fluorescent A.

And then you ask, do you want a T, and so forth, and you go along. And each signal means yes to the answer that it's ready for that particular base. And you can see this doesn't have any problems with this hairpin region.

So when you have a systematic error, you have to change the method that you're using fairly radically. The opposite strand or a completely different method.

If you have-- early on, we had huge differences in intensity of the fluorophores and new enzymes and new fluorophores were developed by Tabor and Matthews and Glaser, and so forth. And this was a huge advance in making everything uniform and eliminating systematic errors.

So the sort of things that are on the horizon that hopefully will be discussed tomorrow in Boston, we've talked about this short number of base extensions like power sequencing, the longer extensions in capillary arrays, which, capillary arrays are changing from low-end capillaries into microfabricated chips. I'll show an example of a mass spec in just a moment. Sequencing by hybridization on arrays will illustrate as a prelude, as an example of Affymetrix technology in this slide.

OK. So the idea here, this is mainly for re-sequencing. Some of the ways of reducing the cost of sequencing the human genome will not apply to sequencing brand new genomes. But still, we need to pursue them, because this may be the way that we get the $1,000 human genome, even if we don't get other ones.

And here, you know the sequence, except there's a possibility that there could be a polymorphism at any base. Could be any single base substitution at any base. And so you don't necessarily know in advance what the common ones are or the rare ones, but you do know the canonical sequence.

And so at every position, you'll make a 25-mer oligonucleotide which will bind to a fluorescently labeled version of your genome, or a piece of your genome. And this is actually developed for HIV re-sequencing by Affymetrix.

And at this middle position, you'll put in all four possible substitutions, T, G, C, or A. And you'll consider each possible template. So if you have the template, which is-- let's say these are the two alleles that occur in a human population or in your sample.

You can have this sequence and then all the variations on it, stepping along, changing T, G, C, A for the first base, the second base, the third base, and so forth, till you hit this one. And this is where the real polymorphism occurs.

And you could have either this as the context or this as the context. And this is schematic And this is the real data. These are real data down here where you have the context of the A allele or the context of the C allele. You can have the homozygotes or the heterozygotes.

And in the homozygous A, you can see it lights up the best, the best hybridization when you have the A in the middle position. The middle position is most sensitive to hybridization changes. And for the C, you have it in the C row here.

And remember, you're changing all the bases in every position. One, two, three, four. This is the one where the middle base of the middle position is in the right context. So this is in the A context and the C context. And in the heterozygote, you get both the A and the C.

And so this has been done on HIV. It's been done on BRCA1, on mitochondria. And now they've applied it with whole wafers to the entire human genome. And this probably costs on the order of $3 million or so.

Mass spectrometry is another way that's used, probably not for sequencing the whole genome, but it's used for single nucleotide polymorphisms. This costs on the order of $0.50 per polymorphism. If there's three million of them in your genome, that's a lot of $0.50.

[LAUGHTER]

And here's what it looks like when you read it out. It's really just like electrophoresis. You can now separate an addition of an A from an addition of a G. And the difference in mass between an A and a G, this is even more subtle than the difference of an n and an n plus one.

This is just a difference between an A and a C. It's detectable in this. In fact, it's detectable and quantitative enough that you can pool samples. This is a bit of a stunt, but it's an important stunt to show that this is really a very precise method, albeit still fairly expensive.

Now just in closing, I want to give you the simplest possible example of how we can search through sequences. Next week, we'll give you a much more rigorous way that you can look through the sequences with very extensive differences between the sequences.

But here, the theme of today's talk is the subtle polymorphism differences that occur between you and me. And here, one different-- so generally, you're looking for exact matches. And a good way to look for exact matches is-- good ways are hashing, suffix arrays, and suffix trees where basically in each of these, you're looking for using a word, a word either that's built up and stored from the end, the suffix, one letter at a time, or it's a chunk that you might have as a hash.

And you make up a lookup table. And the size of that lookup table-- it's a trade-off between speed of searching and the size of the table. The size is going to be-- if the word is n nucleotides long, it's going to be four to the n. It's going to be the storage space you have to put on disk or RAM, RAM if you want it to be a fast search.

And so 16 is the magic number, sort of in the ballpark for a human genome, because four to the 16th is four billion sequences you can represent. But it's a huge table. You have to have a table of four billion times however many bytes you need to store the positions. Typically, about four bytes of storage.

If you cut back on this a little bit, you'll end up with collisions where you'll have two things that have the same hash or suffix. If you make it-- and that's if you make it smaller. It'll take less space. If you make it bigger, it'll take a ridiculous amount of RAM.

And then here's a kind of whimsical-- another example of pearl where you not only want to find all the mutations here at a very high density-- a ridiculously high density, not one particular base for every few base pairs. You not only want to find them, but you want to correct them. And here, the pearl does the substitution. And after gene therapy, everybody walks out happy. OK.

[LAUGHTER]