Hi, my name is Danni Fallin. I'm from the Department of Epidemiology here at Johns Hopkins, and I've been asked to talk to you about genetic epidemiology, some of the common designs that you might see, and some of the methodologies that you might encounter. So, I hope what you get out of today is just a little overview of some of the questions that we address in genetic epidemiologic research, some of the common sampling designs that are used in the analytic approaches, or methods, that we use to carry out these particular designs. And I'll try to give some relevant examples as we go through. Most of them will be cardiovascular disease, but many other disorders as they come up.

So, why would we want to look for genes that are associated with disease? And I think one of the biggest hopes when everybody started was that there would be this opportunity to design gene therapy, and once we found that a gene was going wrong in an individual, we could go in and change that particular gene, or modify that gene or its product in some way and, therefore, cure the disease. And, I think, well there's been a lot of hope for something like that; there are a lot of other things that we can do and other reasons that we should be looking for genes for human disease. And one common one is diagnosis and prognosis.

In the Alzheimer's field, for example, understanding whether or not someone has a particular genetic variant can help you to make a diagnosis of Alzheimer's disease versus some other kind of dementia. It may also help you in particular cancers to know the prognosis based on particular genotypes.

Probably the most important to me is that we can help to understand disease mechanisms better. The idea is that if we find genes that, from birth, have a particular mutation that will cause a disease at some time later in life, say at age 30 or even age 70, we don't, as clinicians, often get to follow somebody from the time they were born to the time they show up in a clinic. We do, however, get to model these kinds of DNA changes in other kinds of animals if we know about such DNA changes. So, for example, if we found a mutation that we know causes a disease later in life, we may be able to simulate that kind of change in an animal model, like a mouse model, we call this transgenic modeling and then follow the natural progression of pre-clinical and clinical disease in these animals and learn a lot about the mechanisms, or the pathogenesis, of the disease.

Finally targeted therapy, or pharmacogenomics, is probably an exciting area for a lot of folks. And this is the idea that, if we administer a drug, and I'll give you an example such as schizophrenia, where there are probably a multitude of drugs that work on some sub-set of individuals, but we're not good at knowing which drug works on which individual early on, and so we will try one, and if there's no response, we'll try the next, etc., until we find something that may work. If there were particular genotypes that predispose to doing well on one drug versus another, then we would want to target specific drugs early, and we could then shrink the amount of time until we found an efficacious drug in an individual and, hopefully, improve their outlook.

And then, finally, as an epidemiologist, I list this as the last reason that we'd want to know about genes that have to do with human disease, and that is prediction of disease. If we can identify folks that are at risk, much like other risk factors, then we can identify primary and secondary prevention strategies for those folks. This is also the idea behind genetic counseling that we can counsel individuals about their own risk or about the risk to their offspring for particular disorders. And, even in late onset diseases where there may not be known primary and secondary prevention right now, it certainly has implications toward retirement and care giving planning and social factors that would be quite important.

So, hopefully, that's convinced you that there's a reason to go out looking for genes that have to do with human disease.

So, let's start by what do we need to know if we're going to go through and do this? Well, the first thing I think is very important, and sometimes often overlooked, is to define what we call the phenotype, and this is the trait or measurable thing among individuals that we're interested in. And that can be something simple, like someone's hair color, or someone's height, or weight. It could be something biologic like what kind of protein does a particular person have in their serum in terms of isoforms or shapes of the protein, for example, or even what amount. Does a person have a few copies or a large number of that particular protein, or other kind of biologic measurement. And then often, in epidemiology, what we're interested in is diagnosis, or some event. So, for example, is someone hypertensive or not hypertensive?

So, once we've got the phenotype down, there are several questions we want to answer about that phenotype. So the first is, is there familial clustering of that phenotype. In other words, if we think there are genes responsible for hypertension, then it should be that hypertension clusters in families in some way. And I'll talk to you once we get through these questions about some of the methods that correspond to each of these questions.

So, if we've convinced ourselves that, yes, it looks like this particular trait is happening in some sort of familial clustering, or happens more often in families. Then, we may want to ask, well is it particularly a genetic effect? So, things could happen within families because families share the same household, or they eat the same diet, or they live in the same geographic community, right? So, what we want to do is parse that apart from truly genetic effects. And then, if we can answer that question, that it turns out that, yes, in fact, this familial clustering does appear to be genetic, or that some of the variation in this phenotype is due to genetic causes. Then, we might want to go a step further, and that's ask, is there evidence for a particular genetic model, and that is, can I ask questions about the mode of inheritance, and I'll talk to you a little bit more about what that means.

And, if I've convinced myself that all these things are true and, in fact, there is a gene, or several genes to be found that would influence my phenotype or trait of interest, then I want to go looking for the gene. And so, we'll talk about how to find where the disease, or several disease genes, are, mostly through two umbrellas of ideas called linkage analysis or association analysis.

And, finally, if we're lucky enough to find a gene, or several genes involved, we're not done because then we still have to go through this final step which is, how does this gene then contribute to disease in a general population, or to my phenotype, in a general population because, hopefully, we're all starting to realize that genes don't act in isolation. It depends on a particular frequency of the gene and particular environmental contacts. Many other things are involved and, so, once we find the gene, we need to then understand how important it is, how much of the variation is it contributing to for things that we see in the population or in a clinic.

So, I said I'd talk about the questions and then I'd go through some of the methods to answer those questions. And so, the first question was, was there a familial clustering, and one way to get at this answer is through familial aggregation studies, and we'll talk about those. And then we said, well what if there is evidence for familial clustering, is there evidence for particularly genetic effects. And we'll talk about heritability studies and how those can be used to answer that question. And then, I said if we're that lucky, we might want to answer questions about mode of inheritance, and this is what segregation analysis is for.

And then we move on. If we've convinced ourselves there's something to be found in the disease gene identification, and we'll talk about two concepts in terms of the disease gene identification. The first is, where are we going to look? Are we going to look across the entire human genome, at a particular chromosome or chromosome region, or are we going to look at candidate genes? And once we've made that decision, then we need to decide what kind of design and what kind of method we're going to use, and this is again where I'll talk to you about linkage analyses, which often are family-based designs versus association analyses, which may have family-based design applications, but are often what you think of more in terms of population sampling.

And just a little reminder if it hasn't been clear yet. This sort of process, both in terms of questions and in terms of the methods used to answer the questions, really occur at two stages. The top three pieces that I've spoken to you about, understanding that there's clustering or whether or not there are particular genetic effects, and whether or not there is a mode of inheritance that we can identify, we didn't need actual measured genotypes. So, we didn't actually have to collect DNA at those stages. What we're looking at in all three of those kinds of designs and analyses is whether or not the phenotypes that are measured have a particular pattern that would argue for a genetic cause. So, we haven't actually collected DNA, or not necessarily collected DNA at that level. But when we go, then, to looking for the particular gene, or set of genes involved, then we're going to ask for blood, or buccal, or some other kind of sample so that we can do genotyping.

Okay, so that first question, is there a familial effect? Well, one of the traditional ways to answer this kind of question was through migration studies. So this is the idea that if you follow a population in Japan, who then maybe moves to Hawaii or maybe moves to California where there's a drastic change in the environment and a diet, etc., that this person is exposed

to but, for at least a few generations, they're carrying genes from their ancestral geographic location which, in this case, would be Asia, we can see if they soon take on the disease rates of the new population or maintain the disease rates of their original population. And, if they were to maintain the ones of their original population, that argues for a stronger genetic effect, whereas if they quickly assimilated to the disease rates of the new population, that would argue more for environmental factors.

So, in addition to migration studies, what we call familial aggregation studies, are often used and, at its simplest, it's what you can imagine. We look at folks who have a particular disease, or phenotype, and folks who don't, so disease versus healthy, and we ask about their family history of that disease. Do they have first degree relatives, or even second degree relatives, who have also had that particular disease, and then we tally those for different kinds of categories, and we do normal cross-tabulation. In this case, you could estimate an odds ratio for disease where the risk factor is family history of that disease. And, any time this is greater than one, that's evidence of familial aggregation.

So, here's an example from a situation of MI, and you can see family history of MI versus being an MI case or control, and the odds ratio there is much greater than 1; it's, in fact, it's 1.7 in the ARIC Study with a significant confidence interval; it doesn't cross 1. So that would just argue that, at least, in this sample of individuals that there does seem to be familial aggregation of MI. And you could do the same idea for hypertension or diabetes, or several other disorders, and see this sort of evidence for familial aggregation in this way.

So, that's what I just showed you before. One other way that we answer the same familial aggregation question is to shift things slightly. And so what you'll see is, instead of looking at cases and controls and asking about their family history, what we do instead is collect relatives of cases and relatives of controls and treat them as the subjects, and then follow them for their disease rates or other kinds of measures of incidents for that particular disease. And the reason that this is nice is because then we can get what we call relative risks or, in this case, sibling relative risks, or I can look at the rate of a particular disease among siblings of a case versus the rate of that disease among people who are not siblings of cases. And, if I made that into a ratio, that's what we call the sibling relative risk, or Lambda S, and you might see this a lot in the literature because this is usually a pretty good measure of how genetic, or how familial a disease is and, therefore, how successful we might be at finding a particular gene for that disorder. So, it's much like the one I gave before, but now we're talking about using the relatives as the subjects and then following them for onset of disease.

So, this is a large list but, just to give you an example that this is now done for many disorders, this is an example from cancer, and you can see FRR is meaning Familial Relative Risks for children of folks with a particular kind of cancer, or Sibling Relative Risk for those who are siblings of individuals with a particular kind of cancer.

Okay, so if we may have passed that first hurdle, there tends to be familial clustering, we want to get past the next hurdle, which is, is there evidence for particularly a genetic effect

because, remember, just seeing familial aggregation means they could have shared environments or genes. So, on our method slide, this would be heritability studies.

So, heritability is essentially looking at familial correlations in a phenotype. And you can think of heritability as really the similarity in that phenotype that you see in individuals that is due to the genes they share. And, so if that's the way you think about it, it should be true that the more genes two individuals share, the higher correlation they should have in whatever phenotype you're looking at, if the phenotype is genetic. So, for example, identical twins who share 100 percent of their genes, should be more correlated than siblings who share, on average, 50 percent of their genes but siblings, in general, should be more correlated than cousins because they should share more genes than cousins should share, etc. So, if you keep with that thinking, you might end up seeing something like this. On the left is either the similarity, so, if we're looking at something like hypertension, if the two relatives that you're thinking about have exactly the same blood pressure value, then they would be considered very similar. If they have very disparate blood pressures, they would be considered very dissimilar, or no similarity, and then the whole range in between those two, and we sometimes call that the covariance between them and, if you look at distant relatives, you wouldn't expect them to lie very high on this metric, right? But, as you go to second cousins, maybe they're a little bit more similar. Cousins would be even more similar and then, as you get to these last two, siblings or dizygotic twins versus MZ twins, for a heritable trait, you would expect those to be more and more similar for that trait. But if it were non-heritable, you'd expect essentially to not be able to predict the covariance by the relationship.

And so, those last two categories, sibs or DZ twins versus MZ twins, brings up a particular kind of study that we use to get at heritability of a disease, and this heritability again being this idea of the proportion of genetic influence on a trait. So, let me just remind you a little bit about the terminology, MZ versus DZ twins. MZ twins are monozygotic, or what we think of as identical twins, and they share 100 percent of their genome. DZ twins are what we call dizygotic twins, are like siblings in that they share, on average, 50 percent of their genome. And this is what you might hear as fraternal twins.

And so, one way to set a heritability is to consider, or exploit, this idea that, while both kinds of twins are born at the same time and raised at the same time and share essentially very similar environments, one set shares 100 percent of their genes whereas the other set only shares 50 percent of their genes. So, if a particular disease is genetically determined, you'd expect those that share 100 percent to be more alike than those that only share 50 percent, and you don't have to worry about the environmental contribution so much because, whether it's large or small, it should be equal, on average, between these two kinds of twins.

And so, if you're talking about a quantitative trait, that means that this row, or correlation, between MZ twins should be higher than between DZ twins or, if you're talking about a dichotomous phenotype such as yes or no for hypertension, you'd think that there would be more concord in MZ twins where, for example, both were hypertensive or both were not hypertensive versus DZ twins. And so this has really been useful in coming up with

what proportion of the overall variance in a population is due to genes versus environments in a very controlled way.

Okay. So, if we've been able to conduct a nice twin study where we've seen this higher concordance among MZ's versus DZ's, and we've convinced ourselves that there is a high heritability of the trait of interest to us, then, as I said, we might want to go one more step, which is, can we lay a particular genetic model onto this disease? So, what do I mean by that? And, I'll tell you that the way we're going to do that, if we're following our methods table, is through segregation analysis.

So, what do I mean by the different modes of inheritance that a gene may show? Well, hopefully, these two terms sound relatively familiar to you. There may be dominant inheritance or recessive inheritance. And these are pretty easy to understand. A dominant inheritance, in this case, is where you need only one copy of a particular mutation, or deleterious form of the gene, to show the phenotype of interest. So, in the example given, individuals who have a capital A, capital A, or only a capital A and a little A, as long as you have one copy of capital A, would show the phenotype of interest. But any individual who did not carry a capital A, such as the little A, little A type, would not show the disease.

Recessive is sort of the other side of that coin. In a recessive disorder, you would need two copies of the deleterious allele to become affected and, having only one copy would make you look phenotypically no different than somebody who is a non-carrier.

So, it's not quite as simple as that, in fact. We'll go through those again and then build a little bit. So, typically, what we do is we start by collecting a family of individuals. So, here I have a father, a mother, two sons, and two daughters represented by squares and circles for convention and, what I've shown you, is an underlying gene that reflects a trait, and so capital T would be the susceptibility allele, and little T would be a normal allele at this gene. And we start asking about other members of this family. So, the son may have married and has two children, this daughter has married and has four children, and maybe that granddaughter has also married and has children that are willing to participate in our study. So, we would end up sometimes with a multiplex family like this. And, if the gene were behaving as a dominant genetic disorder, then what you'd see is everyone in the family who carried at least one copy of that capital T would show the phenotype of interest. So, here, I've just shaded those. So, you can see anyone who has either one or two copies of that particular susceptibility allele ends up developing the disease. And this is an example of early onset Alzheimer's disease.

A recessive disorder would look quite different. Because, remember, in a recessive disorder, and now you need two copies of the deleterious allele before you would express a phenotype. So, now that son who has two capital T's shows a phenotype, and now the grandchild, or it looks like the great-grandchild, has two copies and also shows that phenotype. And, in general, when you need two copies like this to show a phenotype of interest, often this happens in highly isolated populations or where there's some situation of inbreeding and, so what you can see between the two grandchildren that create that recessive

child, is that these are actually cousin marriages. And that's indicated by a double line rather than a single line, and we call that consanguinity. And this often happens, like I said, when you see a recessive phenotype. And this is an example of cystic fibrosis.

It quickly gets more complicated than these two however. So, here all I've done is add a second gene that also is predisposing to a phenotype and have made the assumption that it's a quantitative trait, which means now that things act additively to create a phenotype. So, if you see someone who carries no copies of a capital allele, so if we look at the son, or the grandson, you can see there is a little t, little t, little g, little g for this person, and they are completely unshaded. So, they're the lowest you can be on that phenotype.

If you look at someone who has capital T, capital T, capital G, little g, they have three capital alleles, and those would have acted additively to have a pretty gray, but not completely, darkened phenotype, and then if you look at someone with all four capital alleles, they have the darkest shade that is shown in this family. And this would be something again like hypertension, or a quantitative phenotype, that could be measured where there'd be gradations and, if there were truly only two genes that reflected these gradations, you might see something like in this family. And you can start to imagine that, in fact, it's many genes, not just two genes, that would create a spectrum like this. And, if only that were true.

In fact, it gets a little bit worse, and most of the diseases that we think about today are complex traits where we have maybe several genes working additively, or interactively, and environments that add a particular context to risk. And so, here, you've seen just like the slide before except that, now, someone who should have only been slightly shaded because they only had one capital allele is now much darker in shade because they're circled with an orange. If I told you these represented folks who had had a high fat diet for the last twenty years, that might make some sense that they had increased their risk beyond that due to genetics alone, or maybe that environment had interacted with their genetic predisposition.

So, it's not quite as simple as we first showed but, knowing which of these that I just showed you or, if it's something more complicated is at play, can help us when we statistically model particular genetic effects and, if we get that model right, we do a much better job at actually finding and identifying genes that are important. And so, that's why segregation analysis is important. Once you've found out that there should be a gene to be found, if you can follow these kinds of patterns and come up with something as simple as dominant, recessive, or additive, like those first few slides that I showed you, then you have a better chance of identifying that gene by modeling it that way.

Other things that segregation analysis will get you is some estimate of the disease allele frequency , maybe the magnitude or how strong that particular genetic effect could be and, again, you can use this when you model to look for the particular disease. And the way we do this is, usually through some sort of likelihood testing where we have a set of fitted explicit models, and we keep testing between them until we come up with the one that sort of fits the phenotype data for that family the best.

So, here's an example of doing that across about 115 families that had about 7, well, 676 members with HDL 3 cholesterol levels. And what you see shaded here is the overall distribution of cholesterol levels across these individuals. And you can see it makes a relatively simple shape but, if there were one major gene acting to create this particular distribution, you might expect what you see, but there's actually three different distributions underlying what you see in gray, and that is the red line which shows a first mean that's around 30, another line which is the blue line which shows another mean around 40 and then, finally, a group that's highly predisposed to high levels that has a mean around 50 and a little bit of spread. And it turns out, through doing something like segregation analysis, you can statistically prove that this model fits better than there just being one distribution across these family members, and this is evidence, in this case, of a co-dominant, or additive model, with there being a genotype that predisposes to that low mean group, another genotype that predisposes to a mean that's somewhere in between, and then a high susceptibility genotype that is that extremely high group.

Okay, so, if we can assign a model like that, that's great. Even if we can't but we believe that there is a genetic effect through heritability analysis and familial aggregation analysis, we're going to go on to the important part which is, what is the disease gene or set of genes, and so, we'll talk a little bit about why that's not as easy as it sounds at first and, as I said before, we're going to early on make some decisions about whether we're looking in the whole genome or a particular region or gene within the genome and what kind of design, family-based or population-based design, we're going to use.

So, how do we go about finding these genes responsible for disease? Well, unfortunately, it's difficult for at least two reasons. The first is there are many risk models. Like I said, you will do something like segregation analysis if you have the opportunity because you may be able to condense the number of risk models that you would try but, in general, we're talking about complex things where there may not be such a simple story to follow.

And, there are also many possible genes. There are approximately 30,000 human genes and, if we were thinking of these as different exposures in an epidemiologic study, that's 30,000 different exposures, all with several possible categories each. So, it's a large number of things to look through, and you have to make decisions early about whether or not you're going to look at all of them or you're going to pick particular ones of interest to you.

If we decided to look at all of them, we'd have to look across the sequence among all human chromosomes, including the X and Y sex chromosomes. So, that's a daunting task because, if you think about it in this sort of simple analogy, that's like thinking of an encyclopedia where each volume is like a human chromosome. So, if I'm looking for a particular mutation that's responsible for a disease, the first thing I have to do, and the analogy here would be that that's like finding a misspelling in this entire encyclopedia, do you think I just start at volume I and start reading until I found a misspelling all the way through.

That would take a long time and might not be very efficient and might not lead to the right answer since I'll probably get tired within the first volume and not be able to see things clearly.

So, instead, I'd want to be able to pick which volume to start with, and that's like saying which chromosome should I be looking at. And then I might want to get a little bit better. I might want to know which page I should be looking on. That's like saying which chromosomal region, or you might think of banding patterns in a chromosome. And that's great; I get to one page, but I really want to know which sentence, or which paragraph, right, and that's more like knowing which gene I should be looking at and then, finally, that's still a lot of letters and a lot of words, right? I need to figure out which is the particular misspelling that's causing the disease, or the mutation. And so, that's a lot of work to get done.

We might be able to circumvent it sometimes by jumping to particular pages, or particular genes in the genome, and looking at those. But, unfortunately, as we learn more and more about human disease, we learn more and more about potential proteins, or genes that code those proteins, that could be involved in that disease, and coming up with a list of 100, or even 200, things that we think could putatively be important, is pretty easy to do. And so, we're circumventing in a little bit when we do that, but maybe not that much. And we're also limited by our own imagination to come up with such lists.

So, thinking about those kinds of concepts, it's really too hard right now to read the entire volume, or genome, for every individual in a study. We hope to get there sometime in the future but, right now, we're really not there yet. So, we need some other strategy. Even looking at a large set of candidate genes while, I think is more and more feasible, could still lead us astray if we don't pick the right ones, or if we don't have the technologies available to do too many at one time.

So, we really need markers to represent sections of the genome, and then we need study designs and methods to find regions correlated with disease using these markers. And then, once we find markers that are associated, then we need to go back and understand more clearly which particular DNA variation is causing a problem in our families, or in our individuals. So, it's something like this, a very simple cartoon that, if we had looked at two markers for a cancer study, we might find that both of them seem to be associated with cancer in some way and, ultimately, if we looked between them at all the DNA variation, we might be able to come up with a particular gene that's responsible for cancer in that region.

So, this is just to remind you that, often times, we're talking about not being able to measure the genetic variant of interest but, instead, looking at markers of known location and using those markers to pinpoint areas of interest for us.

So, where are we in this term of methods? Remember, I said the top needs no DNA but, now, when we're looking for the genes, we do want DNA and, if we're going to use DNA, we need to know what to genotype. So, we need to know if we're looking across that

whole encyclopedia, or the entire human genome, or if we have a reason to believe that there's a gene for diabetes on chromosome 6, then we should only look at chromosome 6, or we think that calpain is the gene of interest for diabetes, and we're going to look at only one specific gene.

So, if we decided to go the genome scan route, what would we have to consider, and what tools are available to us? And this is really where the human genome project so far has been a great success for genetic epidemiology and geneticists, per se, and that is that we now have what we call genetic and physical maps where we know, for particular kinds of variation in the human genome that occur commonly across individuals, we know the location of those spots that vary, and we can map that location with respect to other variants like that on a chromosome. So, this is a picture of chromosome I, and you can get this off the web through publicly available sites, and you can identify particular markers in a region of interest or, in this case, across the entire chromosome, and then across all the rest of the chromosomes, and genotype those in your study.

And if you did something like that, this is not for you to take home any specific message on this slide, but just to give you an overview of, if you did something like that, and you had markers going from chromosome I all the way across the chromosomes, you could start to see where there is a signal, or where you see in these plots, what we call a peak, where there's something that sort of lights up and looks like a mountain, in this case, that tells us that region may be harboring a particular gene, in this case, for prostate cancer. And we might want to go then and focus on that region, and now we've narrowed from looking across the entire encyclopedia to particular chapters within a volume, for example.

Like I said, the other side of that coin, is we don't want to look across the entire genome because we have a set of candidate genes that, we think, are most likely to be involved, then we might want to focus on those. And, for example, one thing that we commonly do is think about the pathway as we know it, or as we hypothesize it to be for a particular disease, spell that pathway out, and then think about what genes along that pathway could be modifying a cascade, or a risk, for a particular phenotype. And then look at variation in those genes for a relationship to phenotypes, or diseases of interest to us.

And, so that's really answered this genome-wide versus particular chromosome regions versus candidate genes but, once we've gotten to that and we know what to genotype, we need to know who we're going to collect, and how we're going to analyze those genotypes once we get them. And this is where we'll focus on linkage versus association studies, and you might hear a lot about these two terms, so I've separated them this way just to help be clear about the differences and similarities.

So, linkage analysis is a very traditional type of genetic analysis where we collect families and then we look for mitotic events as they go through families. So, from parent to offspring and then from that offspring to their offspring, etc. So the types of families that we usually collect are large families like I showed you in that previous pedigree slide, but even

sibling pairs or other family pairs, so, cousin pairs or parent offspring sometimes, or avuncular pairs, may be useful. And so, if we can't find huge families, we may be pretty good at doing a linkage analysis looking at sibling pairs who are all affected, for example. And this works very well for Mendelian diseases, especially in this idea where we had done segregation analysis and saw a clear pattern for dominant, recessive, or a simple additive model.

So, this is what we might be doing. Here's where I showed this sort of in between. It's not huge families, mostly nuclear, but maybe sometimes a third generation. And, often what we do is we genotype those markers on the map, as I showed you before, for chromosome I and, here, just for ease of explanation, I've made that each marker has only two forms, a capital and a lower case form. And so, for this particular region of the chromosome, if you've looked at the first family which, you'll notice, is that the person who's affected, that father, carries a capital A, a capital B, a capital C, and a capital D on one chromosome. And if you look at his children who are affected, they also carry a capital A, a capital B, a capital C, and a capital D. And, in fact, if you look at that affected son, he has an affected daughter who, again, carries those four alleles together in a set. And, if I saw this pattern often in a family, I would start to say that that segment is co-segregating with disease in that family. It only shows up when someone's affected, and it doesn't show up when someone is not affected, and it never changes. In other words, there's never a re-combination event. That whole segment from capital A to capital D tends to stay put, or stay together, from generation to generation.

If I look at the next family, I see a similar pattern. Now it looks like the affected individual has a little A, a capital B, a capital C, and a capital D, and her children who are affected also inherited a chromosome with exactly the same alleles as she had. So, again, there's been no changing of information between her generation and her children's generation and, when that particular segment was inherited, so was the disease. So, this would be the idea of linkage analysis, that you start to see large segments of a particular chromosome that go from generation to generation unchanged and co-segregate with disease.

And if you were to take markers like that and lay them on their side across a chromosome, and then plot areas where you see this, where it goes unchanged with disease, you might start to think that that area is linked to a disease locus, or a diseased gene. In other words, maybe a diseased gene lies somewhere between those four markers. And so, if we plotted it — this is, again, like the slide I showed you previously — you might see that, for that section represented by those particular markers that never changed but were always with disease, you would have a high Z score in this case, or high LOD score in other cases.

So, that was linkage analysis. What about association studies? Well, I put them together so that you can sort of compare and contrast. I showed you linkage analyses are looking at mitotic events through families, and you're really paying attention to what happened within any particular family at a time. Association studies are going to look at those kinds of genetic variance across families and so now we're going to not be just following what I showed you before as one family at a time, but rather looking at cross families to see what's in common with those who have disease. And, unlike linkage analyses where we're really

hoping to gather large families, or sibling pairs, or affected pairs, association studies have a larger spectrum of possible designs — case control designs, cohort designs; familial designs such as parent-affected trios could also be used. And there's been a lot of talk about how this may be more appropriate for complex diseases that don't have a simple pattern that you could have identified in segregation analysis.

So, what do I mean differently than the previous slide in terms of linkage versus association or, what I've called here, linkage versus linkage disequilibrium analyses. And, remember I showed you that there was co-segregation between capital A through capital D in the first family that went to the affected children, but if I showed you in the second family, it was little A, capital B, capital C, capital D that kept co-segregating with disease in that family. And, in fact, in the third family, it was capital A, capital B, capital C, little D. So, within families, you have this large segment of four markers that seem to be co-segregating without change through the family as disease is transmitted. But, across families, if you notice, the first family had a capital A among those four markers, the next family had a little A co-segregating, and then the third family had a capital A again.

So, there's no consistency for that marker across families although, within a family, there was evidence for linkage. But if you look at what I've highlighted in blue, for the capital B, capital C markers, those, in fact, are the same alleles even across families as within families. And so, this is an idea of there being linkage disequilibrium, or an association with disease across families rather than just within the family. And that's really the fundamental difference between these two types of studies. And what we'll move into is that, with association studies, we can move to people being unrelated to each other. For example, if we just took a case from each of these families, we would still notice that capital B and capital C appear to be somehow associated with disease in the population.

So, that's linkage analysis and a little bit about association. You notice here I really pulled out two ideas of association — direct or LD, which means Linkage Disequilibrium. And I want to focus on that a little bit more.

So, I really think when we talk about genetic association studies, we need to focus on the idea of two different concepts — the direct method, which I've shown you here. We know a particular polymorphism, for example, in the ApoE gene, creates a different protein isoform, so this particular change reflects whether or not you have an ApoE protein versus a different kind of protein isoform. And, in that situation, since I know that that particular kind of variation reflects a protein change, that is, itself, of interest to me. So, it's like that I get to measure the exposure of interest, whether or not somebody has a C or a T on a chromosome, okay?

For an indirect concept, which we sometimes call linkage disequilibrium, or LD mapping, I may not have known that that CT existed and resulted in a protein isoform change. But I may have known that somewhere downstream in that same gene, there was a variant that was A or G that happened to occur with some frequency in a population. And so, I knew

where that was, so this is the marker idea, and I went and genotyped that in my individuals. Well, why would I do that? Why would that ever be useful? Well, to the extent that one of those alleles, either the A or the G, is correlated to the C or the T, I'd be picking up whether or not a person was an isoform T, I'm sorry, an isoform E4 or something else, because there's some correlation between what I got to measure, which is that AG polymorphism, and what I wanted to measure if I had known it existed, which was that CT. So, this correlation does, in fact, occur in most genes, and it's called linkage disequilibrium, and it's a population genetic term that has occurred over time, and I'll talk to you a little bit more about that.

But, what I really wanted to get across here is the difference between direct testing and indirect. In indirect, I'm measuring exactly what I want. I know its function, and I can just look at it in the population. In indirect, I'm using a marker idea, and I'm hoping that it's correlated with something that has a function. So, in the direct idea, I measure what I want, and I do something like a case control study, and I just run the analysis and see if I think it's involved. In direct, I'm not measuring what I want; I'm measuring something nearby, hopefully, and any results that I see there are a function of how good that marker was, or how good of a proxy that marker was for the unmeasured genotype.

So, here's an example of the direct method that I keep using, the apoE gene polymorphisms that reflect a particular protein isoform and, in fact, to get the common E3, E4, and E2 combinations, you would need two polymorphisms. I showed you the one that distinguishes E4 from the other two. If you also had the codon 158 polymorphism, you could distinguish all three of these common, and I've shown you the frequencies of these in most populations.

If I go ahead and just genotype those two polymorphisms in a data set, then I could ask questions about whether or not carriers of the E2 versus the E3 versus the E4 type seem to be more or less predisposed to a disease of interest.

And so, here's an example of doing just that for these three ApoE alleles for CHD. And what you can see is, across several studies, in fact, it does look like E4 carriers have an increased risk for CHD. And this is true for ApoE and many other diseases. Alzheimer's disease shows a consistent pattern similar to this one, and there's evidence for stroke and other disorders as well.

So, that's the direct method. We got to look at what we wanted and make some conclusion about whether or not we think it's associated. Well, the real thing that you have to think about when we're doing a direct method is we really have to know about that functional polymorphism, right? So, often we use what we call single nucleotide polymorphisms, or SNPs, which are just those C to T or A to G changes like I've been showing you, as candidate loci And, in fact, if we wanted to do a whole genome study, maybe we would just look at all of those kinds of SNPs, or what we call coding SNPs, that have a likely functional change. For example, they change the amino acid. And, if we did that across the genome, that might be a large number of direct tests that we could do. And that might be one way you'd want to go.

Unfortunately, we're nowhere near understanding how genes work and what kinds of variations in genes have particular phenotypic results. So, we can predict things like the coding SNPs that are non-synonymous may have an effect, but there are all sorts of regulatory features of genes that are within, or even distant from the gene, that may have an effect. And so, it's really not possible right now to come up with an exhaustive list of functional variance. And so, if you take this approach, you just say the caveat is that you may miss something.

And, just like we'll keep learning about, if you are going to take this approach at the genome-wide level, remember, about 30,000 genes, if you're trying to cover all the potentially functional changes in all of those genes, you have a large number of genotypings to do and to analyze.

And for some of those reasons that I just mentioned that we can't really focus on all of the potential functional differences in genes right now, we're left, if we want to do a very good job of understanding how variation in genes reflect variation in phenotypes; we're left with being able to query the gene overall in some way. And this is the idea of the indirect, or LD mapping, strategy where we're not focused on particular direct effects, but we're assuming that we have a set of markers in a gene, and we're hoping to understand the variation among those markers as a reflection of any variation that may or may not be important.

So, again, relying on that idea of correlation between a marker that we do get to measure and some functional change that we may or may not get to measure. And, remember, how good or bad that correlation is, is fundamental to whether or not the strategy works. So, the next two slides I'm going to hopefully convince you that you should expect that there is some correlation like that in most genes that we'll look at.

And why would that be? Well, here I've just shown you two individuals who have two chromosomes — capital A, capital B, capital C, capital A, capital B, little C for the first person, and then the other person has a slightly different arrangement. And, at some point in time, a disease mutation may have occurred between that capital B and capital C marker on the first chromosome of that individual.

Well, immediately after that occurs, any time you pull a chromosome out of the population that has the disease mutation on it, what are you going to see at those markers? A capital A, a capital B, and a capital C. It has to be true. That's the only one it's associated with, right? If you pull out a chromosome that does not have the disease mutation on it, will you be able to predict the status at the other markers? Well, not very well, because some of them have capital B, some of them have little, some of them have little C, some of them have capital C, etc. So, immediately, there's an induced association, or correlation, between the alleles that already occurred on a chromosome when a new mutation occurs.

And so, how does that help us when we're doing our tests? Well, if you follow that logic through, at the top of this slide, you'll see an ancestral chromosome where a new

mutation has occurred. So, we consider what's shaded as all of the alleles as they existed on that chromosome at the time that this new mutation, which is indicated as a plus, has occurred.

Now, when that person goes to mate gametes to have offspring, there will be some exchange of information between those two chromosomes, such that the children created by that person may have some portion of that chromosome that contained the mutation, but may have a little bit other sections of that particular chromosome that have been through exchange. So, you'll see one of those children could have the complete chromosome intact to the next generation; the other child had only the top part of that chromosome, but it had exchange, and the bottom half was from the other chromosome.

So, there's already been some shuffling going on. Well, if you follow that to the next generation, that kind of shuffling will occur again. And now, you'll see that there is a smaller segment from that ancestral chromosome that's still intact, meaning it has the same alleles. And, if you follow that through generation after generation, this kind of shuffling will occur so that only very small distances still reflect the ancestral pattern from this first mutation chromosome. But, right around where that plus has occurred, there will still be (you can see a little shading of that chromosome), there will still be a little bit of that ancestral chromosome existing, such that, if you're very, very near, if your marker is very, very near where that original mutation occurred, it will still show the alleles that existed at that marker when the mutation first arose. And so, we can borrow those markers to reflect mutation, even in a population now, which is like the bottom part of this slide. These folks may not even realize that they're related to each other because it was so many generations ago, yet they all, if they harbor the plus or the mutation allele, have similar marker alleles very close to them. So, that's what we're doing when we do this indirect LDG mapping.

So, what are the applications of this kind of approach? Well, one is localization, and I'll show you an example next. If we know of a region of interest because of linkage analysis, for example, we may want to see if we can narrow it through association or LD mapping. And then the other is, as I talked about, candidate gene studies. We could have taken the direct approach like I showed you with that particular polymorphism in ApoE or we could take this indirect approach and look for any variation in the gene.

So, if you remember, I showed you this slide before, and I showed you how the B and C markers seem to be the same allele across families, while the A through D seem to be linked within families. So, if these were four markers on a chromosome, A to D would be a longer distance than B to C, right? So, if you look at this next slide, it's like taking those markers and turning them on their side again, and what you see at the top is a linkage signal (that's what that line is where you see the mountain and the top peak), a linkage signal which you might think is across low side A through D and, on the bottom, is an association signal which would be more like this idea that, across families, you only see the signal between B and C. And so, although there's a whole region that shows linkage on the top, there's only a very small

distance that shows association across families on the bottom. And this is an example from IDDM1.

So, that's the way to localize. What about candidate gene LD studies? Well, remember I said, in the indirect context, you're measuring markers and hoping that there's some correlation between those markers and functional variation in that gene that you may or may not have known to genotype.

What kinds of studies might you want to take on to do this kind of test? Well, like I said, mostly it's unrelated individuals, cross-sectional data, case control studies, cohort studies. Even clinical trials have been used, both for pharmacogenetics questions and for etiologic questions, although you have to be very careful about the design if you're borrowing from a clinical trial because you're obviously not randomizing on genotype, so it's not actually a clinical trial when you do the study for genes.

The main difference when you're thinking about these designs is, what are you going to use for a comparison group? So, your options really are of two kinds. Either you could have unrelated individuals as controls, or folks with a low value on a quantitative measure versus folks with a high, for example, or you may have family designs.

Now, I've given you a little bit about the pros and cons of these two different strategies. Unrelated individuals may be nice because you don't have to collect parental genetic information or other kinds of family members, and sometimes that's a really difficult thing to do. It's nice to have allele frequency estimates that are representative of a case or a control group rather than inflated ones which often happens when you're looking at a family sample because, now, people's genes are correlated with each other. There is some evidence that this actually may be a more powerful approach than some family-based methods, but a lot of attention has been paid to this potential for compounding due to population stratification, and we'll talk just a little bit about that so that you can see where the controversy arises and make some of your own decisions about how to treat it in your studies.

Family-based designs are obviously the flip of that coin. You do need parental genotypes or some other family member. You don't have to worry about what the appropriate control group is. So, if you're doing a particular study where there isn't an appropriate control group at hand, family-based studies may seem more appropriate. And it avoids this confounding issue. So, unrelated samples, an example of something like a case control study. A family-based example, the most commonly used one right now, is something called a TDT designer, Transmission Disequilibrium Test design, and this is basically where you take affected individuals and collect genotypes on their parents. So, that's what's shown in this thing that looks a little bit like Mickey Mouse, where there's a parent, or two parents and a child, and there are three genotypes.

And, simply what we're asking is, if the child in this case had a capital A, capital A and developed the disease, if that capital A, capital A predisposes to disease, then more often

than not, individuals should be given a capital A from a parent if they're affected. And, so what do I mean by that? Well, think of it as the controls here being the two parents' genotypes, the particular pieces of them that were not transmitted to an affected child. So, in this case, the father was a capital A, little A, the child is a capital A, capital A, so it had to be that the father gave the capital A to the affected child and didn't give the little A, and the same is true with the mother in this case. And the way I would tally this, on this table shown, is that for the father-child pair, a capital A was transmitted, a little A was not transmitted, so that would be one count in the cell shown. For the mother, a capital was transmitted and a little was not transmitted. That would be another count in the table as shown. So, now we have two counts in that cell and zeroes in the rest.

If I collected several trios in this way and filled a table like this, ultimately, if the capital A allele were associated with disease, I would see that parents more often than not transmitted a capital but didn't transmit a little. And this cell, which I'll call the B cell, if you're labeling A, B, C, and D, would be highly inflated compared to the C cell, which would be if they over-transmitted the little A. And, so this is a nice matched pair design that's been highly useful for family-based association studies.

And I won't say much about this slide, other than I really want to make sure that folks who are thinking about these kinds of studies consider a larger context of things that may have occurred. Any time you do a study, you want to be able to explain your findings and make sure that the explanation that you give is that there's an actual genetic effect and not that something else about your design, or your population, created bias, right?

So, several of the other things that could go wrong, or that should be considered, when we're doing these kinds of indirect tests, are population stratification, which I'll talk one more slide about. Recent ad-mixture can create problems. Things like genetic drift and selection, these are population genetics that are used that could also create structure that you'd want to account for. And that's beyond the scope of what I'm talking about today, but I encourage you if you're going to think about doing these kinds of studies to get an expert on board who can talk to you in more detail about this to make sure you're not going down the wrong path when you have an exciting finding.

So, what about this population stratification issue? Well, this is really simply a confounding issue. The classic example is, you have a population shown here that has a disease prevalence of 30 percent. Another population has a disease prevalence only of 5 percent. And because they have different genetic backgrounds — one's an American Indian population and the other is probably European Caucasian population. They have slightly different marker frequencies. So, if you look at markers 1, 2 and 3, especially that third marker, 68 percent of the American Indians have a capital A, but only about 30 percent of the Caucasians have a capital A.

Well, what do you think would happen if I do a case control study in a population where both American Indians and Caucasians reside? By sampling my cases, don't you think a

lot of them are going to come from the American Indian population since it's so much more prevalent in that group, and then most of my controls will probably come from the other population, since 95 percent of that population are not affected with this disorder. Well, if I do that, what else happens? Well, if most of my cases came from the American Indian group, then they probably look like marker 3, being a high frequency, whereas if my controls are mostly from the Caucasian group, then their marker 3 alleles are probably closer to that 30 percent, right? So, if I then do a study with marker 3, using these cases and controls, I think, Bam, I've got a great result that has to be involved in this particular disorder. I get an odds ratio of almost 5 but, unfortunately, that's not at all true. It's simply confounding. And so, if you remember the traditional confounding triangle. If you can have an association between your exposure in the confounder, and between the confounder in the disease, that pathway can explain what you observe as an association between the marker and disease.

So, it's debatable how often this occurs, but lots of folks are a little bit worried, and this is why things like matching on gene pool have been proposed, such as the family-based studies that I just showed you because, there, you obviously are controlling for this problem by their controls being from the same gene pool.

56:00
Self reporting ethnicity may get at this a little bit but be cautious that self-reporting ethnicity may not always correspond to gene pool origin and certainly for things like add-mixed populations, self-reporting ethnicity itself as a report for being add-mixed. So instead of just matching, what else could we do in these large scale studies that we don't know for sure? I would recommend two different approaches. One is termed genomic control and it's by some folks at Pittsburg and their idea is to correct for the over-dispersion in a test statistic when you have this particular type of problem occurring and so you can actually measure other markers or other DNA polymorphisms in your dataset and use those to adjust your analysis. Another similar idea is what is called Structured Association Analysis and instead of correcting a test statistic in this case you study something that is probably more palatable to epidemiologists, which is we use additional sets of genes to predict membership of an ancestral population. So say I do 50 genes and then I try to predict just based on those genotypes whether someone is American Indian or Caucasian. And then I would use that predicted status as a covariant that I adjust for to do the genetic analysis of interest.

So, whether or not this is a huge concern is debatable but you can certainly either match a design or do these kinds of adjustments in the analysis by adjusting just a few more things to address that potential problem. And the reason that I brought up the rest of these, as I said, is just to make sure that you know your own population. You understand how the genetics have played a role in your own population before making decisions about the design and analysis of your particular study.

And then finally, we are lucky enough to go through those kinds of studies and find a gene or a set of genes that are responsible for a disease or phenotype, remember, I said that we are really not done. We have to then answer the more important question in my mind which is,

how does this gene then work in the context of the rest of what we know about that disorder? What is its frequency, what is its risk magnitude, in other words, does it only play a risk by 1.5 or does it play a risk by 10 fold, what is the attributable risk, in other words, what proportion of those that are affected in a population are affected because of that genetic risk and how do environmental factors play a role? How important is it to also know the environmental context when predicting how this gene influences the disease? So here's an example for hypertension and other kinds of cardiovascular outcomes and phenotypes it's not a simple story. Once you find genes involved in one particular phenotype you need to understand better its context in terms of other things that are playing a role in disease for that individual or that population.

So hopefully these have given you a very quick overview the questions a genetic epidemiologist can address, what kinds of designs you might undertake to get at those questions, a little bit about the analytic approaches that are used for each of those designs and not in terms of any detailed analysis but just more that you would you know, Oh, I've heard of segregation analysis now I know why you do it and what kinds of information you get out of it.