Good afternoon. My name is David Valle. I'm in the Institute of Genetic Medicine at Johns Hopkins, and I'm here to talk to you today about the human genome and cardiovascular disease. So, the human heart, in its simplest form, starts off as a tube and, ultimately, develops into a four-chambered pump with four valves. It's comprised of muscle, wiring, and a blood supply. It beats about three billion times over the lifetime of the individual, and it is able to adjust its rate and stroke volume to meet physiologic demands over a wide range. And it also has the metabolic systems necessary to support these functions.

Now, from a genetic point of view, we want to know what genes are involved in the development and functioning of the heart and the cardiovascular system. What are the protein products of these genes, what are the structures and characteristics of the systems formed by these proteins, and what is the extent and consequences of normal variation of these genes and proteins? So, in particular, we want to know what, among these varients, which ones are normal, which ones are normal under certain circumstances, and pathologic under other circumstances, and which ones are pathologic under almost any circumstance? And, finally, we want to know what environmental variables interact with these variable gene products to produce either health or disease, depending on how things work out.

So, what are the approaches to answering these questions? Well, geneticists in the past have gone disease gene by disease gene and, certainly, the progress in this area has really been astounding. We know of at least five genes involved in hypertrophic cardiomyopathy, at least five genes involved in the long QT syndrome, roughly ten genes that contribute to coronary artery disease. But, increasingly we have, as a resource, increasing genomic knowledge that enables us to ask these questions in much greater detail and across a much broader horizon.

So, from a genomic point of view, we use comparative genomics, that is, comparing the genomes of one species versus another to help us identify genes and their function. We use expression profiling to look at the changes and levels of RNA transcripts of these genes in response to various physiologic, or iatrogenic perturbations. We look at proteomics to understand the protein products of these genes and how they interact and, increasingly, we're turning to systems biology to integrate these proteins into their normal physiologic systems and to understand how these systems behave in response to genetic variation and to stresses and strains from the environment.

So, we'll come back to these four approaches and how they can be used to understand cardiovascular function, both in health and disease, but first, let me spend a few minutes talking about the genome project and how it has provided us with a variety of new resources.

Before I do that, let me just make a distinction between genetics and genomics. First of all, genetics is the study of heritable variation and, for the last century or more, geneticists have been approaching usually gene by gene. With the advent of the genome project and an increasingly large number of whole genome sequences that are available, we have as a resource genomics, which is really directed at the constitution, function and evolution of genomes. The

study of genomics actually informs genetics and makes a variety of genetic approaches possible that we could not consider in the past.

So, let me just point out to you how recent and how young the science of genetics and genomics is. First of all, we really, although Mendel's laws began, were first postulated back in around 1860, we didn't even know what the heritable material was until the early 1940's when this man, Oswald T. Avery and his colleagues, published this astounding study entitled, "Studies on the Chemical Nature of the Substance-Inducing Transformation of Pneumococcal Types." And this was *the* paper published in 1944 that showed, unequivocally, that DNA is the heritable material. Before that, most people thought it was protein. And so, this really was a breakthrough and set the stage less than ten years later for Watson and Crick to, of course, publish their astounding one-page paper in Nature delineating the double-stranded anti-complementarity of DNA. And this DNA structure really underlies all of the molecular approaches that we can use today to identify genes and understand their function. So, that was in April 25, 1953, or about, just slightly over 50 years ago.

Now, the idea of the genome project got started in the mid-1980's and, initially, it was started as an idea that sprung out of physics as a grand project that was on the scale of projects like building giant telescopes to look out into space. And there was a good deal of debate about whether or not this was a wise way for money devoted to biological research to be spent and, ultimately, this led to the formation of a NSF, a national research council committee headed by Bruce Alberts, and the members are shown on this slide, two from Johns Hopkins, Victor McKusick and Dan Nathans, that ultimately issued in 1988 this very influential report which basically came to the conclusion that this would be a very useful and important thing to do. And so, two years later, in 1990, as shown on this sort of timeline slide, the genome project got underway. So, that's really only, at this point, roughly 14 years ago.

The first successes of the genome project were largely devoted to development of the technology to sequence DNA. However, in 1995, five years after its inception, the first whole genome sequence came on line, and that was of the prokaryotic organism, haemophilus influenza. That was followed shortly one year later by the first eukaryotes, the yeast saccharomyces cerevisiae . And then in 1998, the first multi-cellular organism, C-elegans , a roundworm.

The next milestone came in the year 2000 when two groups, the publicly-funded project headed up by Francis Collins, and a privately-funded project headed up by Craig Ventner, announced simultaneously the completion of a draft sequence of the human genome made front-page headlines in the New York Times, "The Genetic Code of Human Life is Cracked by Scientists." And there's Francis and Craig in the picture below the DNA molecule. And that was followed within the year by the publication of two landmark papers from the same two groups in Nature and in Science describing the draft sequence of the human genome enumerating the genes that were contained in that sequence and other features of the human genome.

The next real milestone was the completion of a whole genome sequence and the most important mammalian animal model, the mouse, which came along in 2002, less than two years ago, and there's the Nature paper describing the draft sequence of the mouse genome. And I'll refer to that several times during the course of my talk.

So here, I sort of give you a summary of where we stand right now with whole genome sequences. Organisms from the three kingdom of all life have been sequenced. We now have more than 100 bacterial species that have been sequenced, at least two members of the third kingdom of life, the Archaea , a kingdom that was only recognized 20 or 30 years ago, and finally, of course, several members of the Eukaryotic kingdom, including our own species and several other animal species, as well as species representing other branches in the Eukaryotic kingdom. As of now, more than 150 organisms, the whole genome sequence of more than 150 organisms, is available.

In terms of relevance to cardiovascular disease, certainly the human sequence is very useful in understanding human genetic disease and making clinical correlations with disease phenotypes. The mouse is an important model for human cardiovascular disease, as is the rat, which has a wealth of data and studies available on cardiovascular physiology. The chicken sequence is just being completed, and that has been a very useful organism for understanding the later stages of cardiac development and, importantly, the zebra fish sequence is coming online now, and that's a very powerful animal model for understanding early cardiac development. So, many of these whole genome sequences are enormously important for understanding cardiac disease and normal cardiac development and function.

Now, it would be impossible to keep track of all this information if it were not for progress in the area known as bio-informatics. And any physician or biologist working nowadays needs to be familiar with the important databases that provide a wealth of bioinformatics data. There are three that one really needs to be knowledgeable of. The first is the National Center for Biotechnology Information (NCBI), which is a branch of the National Library of Medicine, and here's the home page of that website. Here is the website. I urge you to go look at all the tools and all the resources that are available here, including Entre and OMIM, and so forth. One of the things in terms of whole genome sequence that's often a question is, what is....you'll see reference to a build, and it turns out that the sequence is constantly being reevaluated and reassembled, and so, things may change a little bit in terms of the numerology from build to build, so you can go to the front page of the genome sequence statistics and find out what build number is currently up on line.

Another database put out by NCBI is, of course, online Mendelian Inheritance in Man, which comes from Victor McKusick and his colleagues here at Johns Hopkins. It's a very important database for clinicians and biomedical researchers alike. Here's the front page of that database. One can type in any condition that is likely to be genetic, let's say, for example, Marfan search, and you'll get a number of responses that you can then click on and go, let's say, for example, to Marfan syndrome and find out the gene responsible, its map position, a clinical

synopsis and, farther down the page, a list of mutant alleles that contribute to this phenotype, and so forth and so on. So, all physicians seeing patients should be familiar with this resource and use it to help them take better care of their patients.

The other databases that are useful from a genomic point of view include this one, which is the European counterpart to NCBI; it's called Ensemble, and there is its website. It takes the same genomic sequence but presents it in a different way so it's useful to be familiar with both. And the third that fits this category is the UC Santa Cruz database, and here is its genome; here is its URL as well.

So, I urge you to visit all of those websites and look up your favorite gene, or your favorite clinical phenotype, and see what's available for you at those websites.

If you find all of this daunting, there are useful guides. Here's a special issue of Nature that was published in 2002 that answers many questions that people have when they first try to understand these databases. Here's an example. How does one find a gene of interest and determine the gene structure, and so forth. And it goes from one simple question to the next and shows you how to find the answers in these publicly accessible databases.

There are also several excellent texts. My current favorite is called Bioinformatics and Functional Genomics by Jonathon Pevsner here at Hopkins, and it has many chapters dealing with the topics listed here in this slide, and it's quite current. And it's very accessible even if you are not knowledgeable about computers and so forth. So I urge you to take a look at that.

Okay. So those are the bioinformatics resources in the history of the genome project. What do we know about our own genome right now, and how does it compare to say a close relative, another mammal-mouse genome. So, here is the sort of summary of data that we now have. The human genome is roughly 2.9 gigabases in length. The mouse is slightly smaller at 2.5 gigabases. Both species have about 30,000 genes, probably a few less, but the number is still fluctuating as we're trying to sort out the differences between, trying to recognize all of the genes in the genome. We know of the gene products, the cDNA's that are produced, or the mRNA's that are then copied into DNA. We know about 60 percent of those; that is to say, 40 percent of the predicted gene products are new to us, and so we're still learning about what they do in terms of function and expression patterns, and so forth.

There are roughly, on average, 8.7 exons per gene across the genome. That means that the human genome has a total of roughly 200,000 exons to encode all the proteins that are necessary for normal human development and physiologic homeostasis. CPG islands is a genome feature that is frequently in and around promoters, so it's another way of counting genes, and that number, 27,000, is close to the 30,000 that's estimated by other mechanisms. And you can see the same similar data for the mouse genome as well.

Now, if you then ask what, of the total genome space, how much of it is taken up by genes, it turns out that about 40 percent of the human genome is taken up by genes and, by

that, I mean the combination of exons and introns. If you ask, what is the coding space; that is, what part of the genome is translated into protein, that is, just the exons, it's a very low percentage at about 1.5 percent, and that's the same in the mouse as well.

What fraction of our genome seems to be under selection pressure, that is, selection works on that part of the genome that's functional, and it's about 5 percent. So, that's about three or four-fold higher than the part of the genome that's actually coding. That means that many of these sequences that are not actually coding are involved in regulatory functions or, perhaps, structural functions that are important in terms of evolution of our species.

A large fraction of our genome is made up of repetitive sequences, that is, short sequences that are repeated over and over again. Roughly 45 percent of the genome is comprised of this kind of sequence. And there's a special class of repetitive sequences called segmental duplications in which a region of the genome is duplicated and has diverged in terms of sequence by a very low extent, so you might find a region of 100 KB that's duplicated twice and the sequence is as close as 99.9 percent identical. So, we're just now learning what the function of these segmental duplications are, but current models suggest that, perhaps, they play a role in enabling our species to put certain genes aside into an area where they can undergo sequence change quite rapidly and that way contribute to the survival of our species.

Now, one interesting, now let me just talk about some other interesting features of the genome. The first has to do with gene number. And here I show you a plot of gene number versus genome size, and if you look across there at the orange columns, you'll see that the gene number for species as diverse as C-elegans , Drosophila , Ciona , which is a primitive chordate, and our own species, Homo Sapiens, is roughly around 20,000 genes for all those widely diverse (structurally and functionally) organisms. And this gene number has little relationship to genome size, as shown by the pale green column, so that you can see that, for example, we have a much larger genome than the genome of C-elegans, and yet the gene number is roughly the same.

Another way of looking at that is that is takes roughly twenty to, let's say, somewhere between 20 and 30 thousand genes for the development and function of a wide variety of organisms. Most of the organisms that exhibit bilateral similarity from very primitive organisms to our own species have roughly the same gene number.

Now, what about the discrepancy between gene number and genome size, and this comes down to a feature of our genome that was described by Robert Weinberg in a quote that states, "A gene appears as a small archipelago of information islands scattered amid a vast sea of drivel." And what the drivel is, is this repeated sequence. So, here you can see in a diagram looking at regions of the genome from humans, S.cerevisiae , bakers yeast, the fly, drosophila melanogaster , and e-coli, with the orange blocks representing coding sequences, and the blue, or blue/gray blocks representing repeated sequence that, depending on which species you're dealing with, that there's a good amount of repetitive DNA in and around the genes, or the orange coding regions. And so, one trick about studying genomes is to sort out the coding

regions and the related regulatory sequences from this vast sea of repetitive DNA sequence that appears to have relatively little functional significance.

Now, another feature in addition to the repetitive sequence in the gene number about the organization of our genome is that, if you look along a chromosome here shown on the left side of this diagram, you will see that the density of genes along the length of the chromosome is quite variable, so the blue plot shows the density of genes. And you'll see that there are areas in which there are a very large number, the gene density is very high, and other areas in which the gene density is extremely low. In general, the areas of low gene density correspond to the Giemsa-staining dark bands on chromosomes that we're used to seeing in stained metaphases of chromosomes.

The other point made by this slide is that there are some genes that are extremely large, more than 500 KB in length, and other areas of the genome in which there are almost no genes at all. These have come to be called gene deserts, and they may be areas as long as 500 KB to a megabase in which there's no recognizable gene in that stretch of DNA. So, you can see that, looking along the length of a chromosome, you have areas where there's a lot of genes and other areas where there are almost no genes at all, and that some of these genes are very big, and others are quite small indeed.

Now, it's interesting to extend this to ask, what is the variation in gene density by chromosome? And here I've plotted genes per megabase for each chromosome, and you can see that there's roughly a three-fold range, with chromosome 19 being the chromosome that has the highest gene density. It's interesting that the chromosomes with the lowest gene density, chromosomes 13, 18, and 21, are the three chromosomes in which live-born infants with trisomy are tolerated. So, those are the classic live-born human trisomic disorders, trisomy 21, 18, and 13 and, possibly, or probably, part of the reason why that chromosome elaboration is tolerated is because those chromosomes have a low gene density and are actually short in length, so relatively few genes are disturbed in terms of dosage number in those trisomies as compared to other chromosomes with a much higher gene density.

Now, what about the repetitive elements? Well, one could talk for hours about the repetitive elements in our genome, and I'm only going to say, make these few points, first of all, that there are a wide variety of kinds of repetitive sequences, some of which are tandemly arrayed , short tandem repeats, over and over again, tail to tail, head to tail, head to tail, head to tail, and then the classic sequences related to chromosome function, telomeres and centromeres .

And then there's another class of human repeats called interspersed repeats. Some of these are processed pseudo genes, transcripts in which the introns have been spliced out, gone to the cytoplasm as mRNA, and then by virtue of reverse transcriptase, being converted back to cDNA and then being reintegrated into the genome. And then there are a host of so-called mobile elements, the two major classes of which are called alu elements and line elements that can actually move around the genome and, in so moving, they may disrupt other genes and

serve to, at some extent, homogenize our genome and, perhaps, that plays an important role in the evolution of our species.

I mentioned at the outset that comparative genomics is very important and, certainly, here's a good example of how the mouse sequence can help us understand the human sequence. I mentioned that we're still trying to enumerate all the genes, and one of the ways that we can do this is to line up as best we can the human sequence and the mouse sequence. And what's shown here is that if you look at sequence identity, the regions of highest sequence identity are usually the exons, that is, those regions of genes that are actually translated into protein sequence. So, here's a particular gene, and you'll see the highest peaks of identity are in the first exon, the middle exons, and the last exons that corresponds to regions of the gene that are actually translated into protein. The five-prime UTR and the three-prime UTR have less identity than the coding sequences, and then the intragenic space has even less, on average, less identity. So, you can, by plotting the regions of high identity, you get a clue as to where the exons of genes are.

This slide makes the point that another area of highly conserved sequence are the regulatory elements, and here you see a region of relatively highly conserved sequence corresponding to exons in a particular gene, and then you see a region denoted in red of a short region of high sequence conservation, and that turns out to be a regulatory element, an enhancer for the gene shown on the left. So, we can use this comparative sequence analysis as a way to identify candidate regulatory elements, as well as candidate exons. It turns out to be very powerful in that regard.

Now, another aspect of this sequence that is important medically is the sort of rapidly expanding area that's come to be known as genomic disorders, and these are human disorders in which we observe recurrent DNA rearrangements involving unstable genomic regions. These are regions whose sequence actually predisposes them to some sort of rearrangement. Most result from non-allelic homologous recombination between regions of specific low copy repeat sequences. So you have two sequences that are very highly similar to one another and, when the DNA pairs, that sequences pair in the wrong way, and you end up with either a duplication or a deletion.

As I say, the rearrangements lead to loss or gain of dosage sensitivity of genes, or to disruption of the gene. And these low copy repeats are blocks of DNA sequence anywhere from 10 to 400 KB in length with greater than 97 percent sequence identity. So, depending on the size of the rearrangement, these genomic disorders may be either Mendelian disorders if they just disrupt a particular gene, and we now have more than 20 that are known. They may be contiguous gene syndromes in which several genes in sequence are disrupted by these rearrangements, or they may be gross chromosomal rearrangements involving hundreds of genes. Perhaps the best known of these genomic disorders involving the proximal part of the long arm of chromosome 22, a region known as 22Q11 shown here in aqua, and this is the region that's involved in the velocardiofacial syndrome and the DiGeorge syndrome, VCFS-

DiGeorge syndrome, also in deletion syndromes known as the cat eye syndrome, I mean duplication syndromes known as the cat eye syndrome.

So, here is a diagram of the long arm of chromosome 22 laid out horizontally and above three rectangles labeled low copy repeat 22 number 2, number 3A, and number 4. For the DiGeorge syndrome, the most common deletion comes from non-homologous recombination between low copy repeat 22,2 and low copy repeat 22,4. The end result is a three megabase deletion, accounts for about 90 percent of all patients with the DiGeorge syndrome. A smaller fraction, about 7 or 8 percent is the 1.5 megabase deletion shown just below it and, of course, the DiGeorge syndrome has many cardiac abnormalities, conotruncal abnormalities and facial abnormalities due to mal-development of certain of the pharyngeal arches, mal-development or often absence of the thymus leading to immunologic abnormalities and parathyroid developmental defects leading to defects in calcium homeostasis.

The milder form of this syndrome is known as the velocardiofacial syndrome, and here's a young woman with VCF, VCFS, a velocardiofacial syndrome. They have a characteristic facial appearance, they have learning disabilities, they have velopharyngeal insufficiency, and they have an increased incidence of cardiac defects, and an increased incidence, quite interestingly, of schizophrenia as adults. The frequency is about 25 percent, or about 25 times greater than the general population. And these defects, that is, the DiGeorge velocardiofacial syndrome, are quite common, about one in 4,000 newborn infants. Another genomic disorder, well-known genomic disorder, involving the cardiovascular system is William syndrome, which is another recurrent deletion.

Okay, now let's leave the genome for a minute and just say one brief word about the proteins that are encoded by the genes and our genome. So, we said there are roughly 30,000 genes and yet, by all extents, the number of proteins is vastly larger, perhaps greater than 100,000 proteins. This difference in number comes from the fact that many genes undergo alternative splicing to produce different forms of mRNA's that then lead to different proteins. If we look over evolutionary time and ask, how has the proteome changed as we go up the phylogenetic tree, one sees that there are a relatively limited number of protein motifs, or structures, but these are strung together in ever more complex arrangements to make more and more complicated proteins and, similarly, we also see expansion of certain protein families so that, for example, if we look at all proteins in the TGF beta family, there are 42 in our own species, only 2 in flies and worms. So, that this is a way that increasing complexity can be generated from relatively the same number of genes to start with.

Now, before we leave the genome, let's just ask one additional question and that is, how different are we within the same species, that is, within our own species? I talked a little bit about differences across species but, if you look around the room, you will see that every person looks differently, and we all know that different people are subject to different disorders. So, how much of this difference can we explain at the level of the DNA and, at first blush, it seems like there's really not that much difference at the level of DNA because, at a quick estimate, all humans are roughly 99.9 percent identical at the sequence level. But if you

put that into an evolutionary context, you would actually expect that sort of high degree of similarity of one person to the next.

Our species is a young species. We've only been around perhaps 100,000 years, or at most 200,000 years. We came from a small founding population, maybe as few as 10,000 individuals. And we have, if you look in terms of evolutionary time, we have a high degree of similarity with our relatives. So, if you look at coding sequence, we're 70 to 90 percent sequence identical with mice and, if you look at our closest primate relative, the chimp, we're about 98.5 percent identical to the chimp. So, that puts the 99.9 percent difference in context.

Now, how can we have a lot of genetic differences from one person to the next if the overall sequence is 99.9? Well, if you just went out and picked two people on the street. Let's say, these two people shown here in the front of Time Magazine, Craig Ventner and Francis Collins, and you sequenced a particular chromosome from each of them, what you would find is that they differ only at one in 1,250 base pairs. That's what that 99.9 percent identical predicts.

On the other hand, there's no question that human variation is extreme as shown, perhaps, by these two individuals that certainly, in terms of size, represent different ends of the normal variation of our species. So, that translates into the fact that, if we're different at one in 1,250 base pairs, multiplying that times 3 billion base pairs, there are a lot of base pairs in which we are different one from another. So, there's a lot of room in our genome for genetic variation one person from the next.

Now, what are the sources of this genetic variation if we look in the genome? Well, the sequence variants really are in three classes. First is a sort of class in which there are small insertions and deletions. These have come to be called "indels," and they account for about 20 percent of the variation. Then there's a length polymorphism, and these insertions and deletions may be one base pair or two base pairs on that order. Then there's length polymorphisms, so-called short tandem repeat polymorphisms, or STRP's, and they account for about 10 percent of the genetic difference between one individual and the next. The largest class are what are known as single nucleotide polymorphisms, or SNPs, and they account for about 70 percent of the genetic variation from one individual to the next in our species.

The other factor that contributes to genetic variation is recombination. So this genetic variation is constantly being shuffled at each meiosis by recombination put into different combinations. So, let's just, before we leave this subject, talk a little bit more about single nucleotide polymorphisms because they turn out to be a very powerful marker for understanding human genetic variation and getting at genes involving human disease. So, single nucleotide polymorphisms and single-based pair variance with both possibilities relatively frequent, so you could imagine that one allele in which you have the sequence GATCA and another allele in which you have GAGCA, the T and the G then become a single nucleotide polymorphism.

In general, then, you have to ask well, how polymorphic are they? In other words, how common are the two alleles and, typically, these are categorized as the frequency of the minor allele, or the minor allele frequency, and the single nucleotide polymorphisms that are most useful are those in which the minor allele frequency is at least 0.05 or greater. That means that you can find, with relative ease, you can find both varients in the population. The single nucleotide polymorphisms are frequent, perhaps on the order of, depending on where you are in the genome, somewhere around one in 400 base pairs, or at least several million in the genome.

And actually, now, more than ten million have been identified. I show in this slide three million, but the number increases everyday, and we now know of at least ten million. And they're easily scored. They can be scored by sequence or a variety of other techniques. And they're binary; you either have one or the other, and so that makes them very useful markers, given that they're frequent and easily scored.

So, how common are they? Well, in a series of papers, one of which I quote here, people looked at a large number of genes by sequencing the same genes over and over again in different individuals. In this paper by Cargil et al., they looked at 106 genes, they sequenced a total of 196 KB in each individual, and it turns out to be about 135 KB of coding sequence in 57 and the SNP frequency that they came up with in that study, the frequency of non-coding SNPs was one in every 354 base pairs. The sequence of coding SNPs, that is, SNPs in the coding sequence, the part of the genes that actually is translated into protein, what was one in 346. Most of these were synonymous, that is, sequence changes that didn't change the amino acid sequence, so they were third base positions in codons, but an appreciable number were non-synonymous SNPs in which the amino acid sequence is actually changed with a frequency of one in 734 base pairs. And other studies have found similar numbers. So these numbers seem to apply across our population.

So, how different does this actually make us? Well, if there are three times ten to the sixth base pairs, or three million base pairs differences from one individual to another, that roughly calculates out into one coding SNP per KB, and that translates out into about 80 to 85 percent of our genes are polymorphic at the protein level. That means to say that, for 80 to 85 percent of our genes, there is a coding variant that changes amino acid sequence within a frequency of at least a 0.01, or 1 percent, in 80 to 85 percent of our genes. That translates out into 17 percent average heterozygosity. That is to say, if you looked at all of my genes, I would be heterozygous at least 17 percent of them, and we know that a little difference goes a long way. Recall that it only takes a single base pair difference to cause such things as achondroplasia or a greatly increased risk for colon cancer or retinitis pigmentosa or five-fold increased susceptibility to diabetes. So, this amount of variation is certainly sufficient to account for the genetic contribution to human disease.

Now, let me just say a word about recombination and how this shuffles this genetic variation. So, here we have two chromosomes pairing in meiosis. They've already duplicated their DNA as evidenced by the fact that both chromosomes are double stranded. We have a

breakage in recombination, and the strands of the two different chromosomes that leads to the four products as shown below, two products that are unchanged by the recombination, and two products actually have shuffled their sequence at the point of the break.

So one, on average, at each meiosis, there's one recombination per chromosome arm in each generation. So, what does this mean in terms of the distribution in Nature of the single nucleotide polymorphism? So, imagine if we had one population which was completely homogeneous. So, let's say we had a chromosome pair in this population and we consider three SNPs whose genotype is indicated here by capital letters. The A gene, the B gene, and the C gene. So, if you had, and we need to define a term, so a haplotype is the genotype of a set of markers linked together at a segment of the same chromosome. So here, the haplotype would be capital A, capital B, capital C on both of these chromosomes.

Now, suppose then, we have some recombination between these chromosomes as they are passed from one generation the next. If the population in question is completely homogeneous, that is, it had no genetic variation, then we wouldn't be able to score the recombination because the switching of the different segments of the chromosomes would not be scorable markers, so you'd still have the same haplotypes in the next generation.

But, let's suppose a second population moves in, here represented by a smaller number of blue sequences, and the blue chromosome for these three markers is exactly different so, at the haplotype on the blue chromosome is little a and little b, little c. Now, suppose we have a recombination, and now we've generated two different haplotypes. In addition to the two parental haplotypes, we now have the haplotype big A, little b, little c, and little a, big B, little c. So, recombination shuffles the SNP genotypes into different combinations as shown here.

Now, let's suppose again that we had this population and then, after many, many generations of mating between individuals who, as these two populations became homogenized, we would find all different kinds of chromosomes with all different segments of the original blue and pink chromosomes in all different possible haplotypes.

Now, if the frequency of these haplotypes calculated out this way. If we measure the frequency of the various haplotypes, and we said the frequency of the little a allele was 0.05, and the frequency of the little c allele was 0.05, and then the frequency of the haplotype, little a, little c, should be the product of those two frequencies — .05 x .05 should be .0025. If that .0025 is the measured frequency of the little a, little c haplotype, then this population is in linkage equilibrium. That is to say, there's been so much recombination, that the original arrangement of these haplotypes has been completely homogenized, and the association of the little a with the little c allele is predicted exactly by the frequency of each of them separately.

But, let's suppose the frequencies are the same — little a, .05, little c, .05, and when we measure the frequency of the haplotype, that is, little a, little c, we actually find something other than predicted by the frequency of the two alleles, namely .05. That means to say that

recombination has not had a chance to homogenize the combination of those two markers, and so those two markers are in what is called linkage disequilibrium, and you'll see a lot in the literature about using linkage disequilibrium mark mapping to identify transmission of segments of chromosome containing, perhaps, disease genes down through the generations.

Now, the other thing to realize about single nucleotide polymorphisms, and I already mentioned that you have a frequency. The frequency of the two alleles may differ, one being much more common than the other. And, as a rough measure, usually the allele that has the lowest frequency is usually the youngest allele. That is to say, it happened relatively few generations ago, and it has not had a chance to expand in the population.

So, here in this diagram, I indicate the frequency of the different alleles by the size of the diameter, and you can see that, and time is going down here, and so you can see that the alleles with the lowest frequency, that is, the littlest circles, are the ones that happened most recently, as indicated across the genome here. And I refer you to this paper by Aravinda Chakravarti if you're interested in this to learn more about it.

Okay. So, we've covered a lot of ground. Let me just hit some terminology that you should have taken away and, actually, terminology that's frequently misused. So, let's talk first of all about mutation, and I think you should just consider a mutation a heritable change in the DNA sequence. Notice that it says nothing about frequency, and it says nothing about functional consequences. It says simply that a heritable change in the DNA sequence is a mutation.

Now, what's a polymorphism? That's a sequence change in which the frequency of the most common allele is less than .99. That is to say that there are other alleles that are easily identified. Sometimes in the case of a single nucleotide polymorphism where there are only two alleles, the minor allele of frequency must be .01 or greater.

So, what's a neutral mutation? Then that's a mutation that has no functional consequence. Frequently people refer to neutral mutations as a polymorphism. That's because they think that alleles that are frequent are likely not to be under selection so, therefore, they must be neutral. Your thinking will be much clearer if you separate your thinking about frequency and function.

So, what's a pathological mutation? That's a mutation that has a functionally significant change that, in some way, disrupts the function of the normal protein. So, you're much better off using the terms mutation just to refer to heritable change in DNA sequence and, if you want to mean a pathological mutation, then say that.

So, another way to get around all this is just to refer to sequence variance as variance and then specify whether these variances are pathologically significant or pathologically not significant.

Now, a mutation by its very Nature produces linkage disequilibrium, and so here is a region of a chromosome in which there is a single nucleotide polymorphism, a G or an A at the indicated position. You have a mutation on one chromosome that introduces a T, and so now that T is always together with the G, since that's the chromosome that it appeared, until recombination occurs enough times between the G and the T to homogenize those two chromosomal segments.

So, at least in the initial generation, and for many, many generations thereafter, particularly if the mutation is close to the G, so that it's harder for recombination to occur in that stretch, the G and the T will be in linkage disequilibrium.

So, let's suppose then we have a larger region of the genome, let's say 10 KB, and there are nine single nucleotide polymorphisms, or SNPs, spread out across this 10 KB, as shown here on this slide at variable intervals and we could guess that, since there are two possibilities at each one of these positions, that the theoretical number of haplotypes would be two to the "nth" or, in this case, 512 different possibilities. When we actually look, we often find that the actual number is far, far less but, perhaps, maybe just two or three different haplotypes covering this region of 10 KB and that's because most of the SNPs in this particular region are in linkage disequilibrium, and so if you know what the SNP genotype is at position one, then you have a good idea of what the genotype will be at the other SNPs in this small region.

Now, when a mutation occurs on this segment of the chromosome, here indicated by the orange star, then originally the mutation is exactly identifiable by the SNP genotype across this haplotype, this region of the genome, the haplotype of this region of the genome, because the mutation occurs on a particular chromosome with a particular haplotype in the area flanking the mutation. As this mutation is passed down through the generations, then recombination takes place and, gradually, the block of sequence containing the mutation becomes smaller and smaller in terms of having its neighboring SNP still in linkage disequilibrium without a mutation. If you're using those neighboring SNPs to find that mutation, then the longer ago the mutation occurred, the smaller the block, and the smaller the number of markers that will be in linkage disequilibrium with a mutation, the harder it will be to find the mutation.

So, one can think of the genome as a mosaic of blocks of linkage disequilibrium separated by regions, where there's been a lot of recombination, and in which there are relatively low linkage disequilibrium. Here, the aqua indicates the regions of linkage disequilibrium, and they're separated by red regions where there's a lot of recombination and low linkage disequilibrium. In the blocks, there may be two or three different haplotypes that are common, so I've indicated here....in block one, for example, there are three haplotypes that are common and, in block two, four haplotypes that are common, and so forth and so on.

And there's a lot of work now on the haplotype block structure of the human genome. I refer you to the references on this slide describing the idea of haplotype blocks and the

number of haplotypes for each block. The size varies depending on how recently the population has gone through a population bottleneck. The oldest populations that have existed without strong bottlenecks are found in Africa, whereas in other regions of the world where the populations have more recently undergone severe population bottlenecks, the blocks have not had a chance to be reduced by recombination and, on average, they are larger.

And there is a project ongoing right now that will be, I think, of great help to understanding the genetic origin of human disease, and this so-called haplotype map project, it's aim is to define the haplotype block structure across the entire genome and then compare the block structure in three populations from around the world — Northern Europeans, Africans, and Asians — and define the frequent haplotypes for each block in each population. If an investigator studying, let's say, a disorder in one of these populations, the investigator will know something about this population, about the region around the gene of interest, and so forth.

The haplotype map project also plans to study other populations in a sub-set of genomic regions to make sure that, to determine whether or not these other populations, whether the population the investigator wishes to study are similar to one of the reference populations or whether you need to sort of do some additional study of this population before one embarks on trying to identify genes that contribute to a particular disease in a particular population. And I urge you to see this reference for, if you want to learn more about the haplotype map project.

Now, this website, the hap-map that defines where we stand in understanding the haplotype structure and the SNPs in these different populations. This information will be of tremendous use for association studies where one looks for the frequency of markers in large populations of individuals with the disease, and a carefully matched control population without the disease in a so-called case control study.

Now, let's come back to, with that sort of information about the structure of our genome, the amount of variation in our genome, and some information about the frequent genetic markers in our genome, how this might be applied to understanding cardiovascular disease. So, the areas that one can anticipate are likely to be particularly powerful in identifying the genes and genetic variance that contribute to cardiovascular disease are shown here: comparative genomics for identifying genes and understanding their function; expression profiling for understanding how these genes are expressed and how the expression changes in response to various physiologic perturbations or non-physiologic perturbations; proteomics in terms of understanding how the changes in expression in RNA are translated into changes in expression at the protein level; and then systems biology - how these proteins interact and how did the systems behave in response to various injuries and insults.

So let's, in the last couple of minutes here, just take a couple of examples to see how we'll use this. So, first of all, what about comparative genomics? Well, this turns out, as I've already indicated, to be a very powerful way to find genes and to understand their function,

and a wonderful example of this comes from the study of Pennacchio et al. that first appeared in Science in 2001. These individuals were looking at a region of the genome where apolipoprotein genes were known to be found, and they noticed a region of conserved sequence in comparing the human and the mouse sequence. It turned out, then, that this region of high sequence comparison encoded a previously unknown apolipoprotein, so-called apoA5 that was present both in mice and in human species.

They went ahead and used mouse models to understand the function of this gene and the protein that it encoded and, by making transgenic mice and then looking at the lipoprotein profile of these animals, they could show that when they over-expressed apoA5, that what changed in the lipoproteins of these animals were the triglyceride levels, and they came down when you over-expressed this protein and, conversely, if you knock this gene out, what changed was, again, the level of triglycerides. In this case, the triglyceride level increased.

So, these data immediately pointed to an important role in apoA5 in regulating the level of triglycerides in blood plasma. And subsequent studies, and I don't have time to go into them, but subsequent studies have shown that this gene, in humans, has a wealth of genetic variation, and this genetic variation contributes to triglyceride levels and adds to atherosclerotic cardiovascular disease in humans, all found in whose function was understood by virtue of comparative genomics.

Now, what about expression profiling? Well, there are now many examples of using expression profiling. Generally, these are either chip base assays or so-called serial analysis of gene expression, sage assays, to look at the RNA levels in various physiological states in cardiovascular tissues — myocardium, myocardium in response to myocardial infarction, in response to congestive heart failure, and so forth. One of the ones that I found particularly fascinating, and I think really will, in the long run, have important ramifications in the way we treat patients, is this one by Storch and his colleagues that appeared in Nature in 2002. And what they did is looked at changes in gene expression over the 24-hour period of the day. And they identified about 10 percent of genes expressed in cardiac tissue that undergo significant changes over the 24-hour period of the day. That is to say that they were cycling in a circadian fashion.

The interesting thing is, when they looked at the same genes in liver, they found that a different set of genes undergoes circadian changes in liver. So, from this and other subsequent studies, we have a picture that, in every tissue, there are a set of genes that undergo circadian changes, and that if we really want to understand the physiology of a particular organ, or tissue, we have to understand how its physiology changes in a circadian way over time and how, when there is an insult to this tissue, be it due to a genetic defect or due to some environmental stress or strain, how this changes the circadian expression and how that affects the overall well-being of the individual.

Now, what about proteomics in terms of understanding the cardiovascular system? And again, these are technologies in which one finds a way of trapping all of the proteins that

are present in a particular tissue at a particular physiologic state, isolating these proteins, separating them typically by 2D gel, or by some sort of elution from a solid support mechanism, and then using mass spec to identify these proteins. And this technology takes advantage of the tremendous power of mass spectrometry to identify small fragments of proteins and the increasing database that we have that enables identification of many, many different proteins.

So, here's a method in which one added a biotinylated aspect to all the proteins expressed in cardiac tissue and then used avidin beads to isolate those proteins and then run them out on 2D gels and identify them by mass spec. You can also use GST glutathione S-transferase tags to identify and then run them over columns, glutathione, and then eluted from these columns with glutathione and then run the samples out on some sort of gel matrix and isolate those proteins and identify them by a mass spec.

So, these technologies, either in human tissue or in model organism tissue, enable us to identify the relative amounts of proteins in different tissues under different physiological times and understand how the organ responds to disease states.

Now, what about, in closing, what about the systems biology? Well, we've talked a lot about genes and their protein products, but if you actually look at living systems, it turns out that this reductionist sort of approach is good for finding genes and for, perhaps, understanding their individual function but, virtually, all gene products work together in some sort of complicated systems or, here, and a seminar paper, actually, by Lee Hartwell and his colleagues in modules. These modules are integrated into larger modules, or systems, and then eventually integrated into organisms.

So, if we really want to understand the organism and understand how the organism responds to various stresses, we need to know something about these systems and how they behave. And this work is really in its early stages, but I think it's going to be extremely important in terms of understanding disease and how we treat disease. I cite a simple example here that, I think, makes the point, and this is work by L. Barabasi and his colleagues, and they took the entire yeast proteome; saccharomyces has only 6,000 genes, and it's possible to look at the interaction of the protein products of these genes one by one, and the yeast research community has done this and asked which of these proteins interact with one another and, in this case, have scored the interaction for proteins that interact physically. And then they mutagenized the gene for each of these proteins, so you can ask, what is the phenotype if you knock out one of these proteins.

So, Barabasi looked at the phenotype identified by the yeast genomics community and where the protein sat in a particular system, and he characterized in this case the proteins by virtue of the number of links with other proteins. What he found was that most proteins had less than five links, that is, less than five interactions with other proteins that made up about 93 percent of the total and, of these, 21 percent were essential. That is to say, when the gene encoding the protein was knocked out, 21 percent of them turned out to be lethals. However, a

small fraction of genes and their protein products were highly linked. That is to say they had more than 15 links with other gene products. This made up less than 1 percent of the total but, of these, 62 percent were essentials. So that gives you a rough estimate, a rough indication, that the number of interactions that a protein has is, in a way, predictive of how the system will be perturbed if this particular gene product is disrupted.

So think, for example, of the cardiac system here of the proteins involved and dilated cardiomyopathy, the system we're understanding in greater and greater detail as we dissect out the genetic causes for dilated cardiomyopathy. I refer you to this paper if you want to learn more about all the gene products and how they interact with one another. But you can see it's a highly interactive and integrative system. And it turns out that all systems have a number of parameters that define their robustness, that is, how the systems behave in response to perturbation. So, we want to ask, how does this system adapt to perturbation, how sensitive is it to changes in the individual components of a system — that's called parameter insensitivity — and if the system is going to fail, will this failure occur gradually, that is to say, gracefully, or will it be catastrophic as shown here in a little patient of mine who has an inborn errors of metabolism and has had a catastrophic failure of one of his metabolic systems.

So, by understanding systems biology and how these systems behave, how robust they are to genetic change, we'll do a much better job at understanding disease pathophysiology and, ultimately, what we can do to shore up that system and reduce the phenotypic consequences of a particular genetic disorder.

So, I will close at this point. We've covered a lot of ground, and I hope that you'll be able, since this is recorded, be able to go back over it in areas where I've gone very quickly. And I've listed a lot of references. I urge you to read some of those papers in areas that are of interest to you, and I think we're at a very exciting time as we understand much more about genetics and genomics and how this is responsible for normal variation and for the variation that contributes to problems that bring our patients to the hospital.

Thank you very much.