Today, I'm going to discuss a new and very exciting area which is the genetic analysis of human disorders that display non-Mendelian patterns of inheritance. This has only been possible in the last ten years or so as genomics has become a much more expansive and inclusive discipline.

Non-Mendelian disorders have been known throughout the history of genetics and, to put it in view, one has to understand that at the level of traits or phenotypes, Mendelian or simple patterns of inheritance is observed only about five percent of the time. Overall, this is quite rare and, therefore, takes quite a bit of clinical acumen and expertise to try and tease out families that display Mendelian patterns of inheritance.

What this implies is that the majority of human traits, including those affecting us with disease, are decidedly non-Mendelian. Most traits are also environment-dependent, and this dependence is one of the biggest challenges that faces geneticists today.

Non-Mendelian traits are also known by many other names, and they're interchangeably called multifactorial, sometimes they're called polygenic but, almost invariably, they're referred to as complex genetic traits.

So, what is the evidence that human inheritance can be non-Mendelian? We have known for a very long period of time that even genetically identical twins are not always concordant for any given trait that we examine them for. This clearly implies that either chance or specific aspects of the environment are involved in the genesis of the trait that we are studying. Most disorders, however, shows familiality, meaning that relatives of an affected individual, or of those with a given trait, in fact, have an increased risk of having the disorder, or have an increased probability of displaying the trait over appropriate controls.

The second feature clearly implies the genes involved in the genesis of most human traits and diseases. The feature that makes the second familiality aspect genetic is that this risk of familiality decreases as the genetic relationship between a pair of individuals decreases. Families, however, do not show simple, meaning dominant or recessive patterns of inheritance, for the majority of traits as I alluded to before. And this, cumulatively, implies that most traits have familial and genetic tendencies and that genetic tendency has a complex pattern of inheritance.

So, what are the essential genetic features of such traits. Number one, they invariably arise because multiple genes are involved, and the pattern of inheritance is an amalgam of the patterns of inheritance of each of the genes involved. This also implies the susceptibility alleles, that is those alleles at any of the many genes that are involved, have low penetrance, meaning they're neither necessary nor sufficient for displaying the trait. Third, susceptibility alleles can also have high population frequency, an aspect that we're going to return to again and again in this lecture. And, finally, again as alluded to before, environmental, stochastic and epigenetic factors are important.

So, these four features are the hallmarks of a complex disorder or a multifactorial trait, and this is what we have discovered over many decades of studies of both experimental systems as well as some human diseases.

But in this molecular age, we need to know much more. What we really need to focus on in this genomics time of age are the identities of the genes involved in the multifactorial trait. We need to find the specific mutations and provide functional tests to show that these mutations are, in fact, the basis for the trait that we are studying. Third, we need to find appropriate molecular tasks that show interactions between the various mutations that are always suspected in such a complex trait and, most importantly, if multiple genes have, in fact, been inferred for any given disease, we need to recapitulate this trait in model systems, not by studying the genes one by one, but by studying them in a cumulative and total fashion.

The fact remains that, until today, very few complex disease genes had been identified. And two major impediments to this identification are that most traits and disorders that show these complex patterns are, in fact, often episodic, and they are age-dependent. So, the frequency of disease and, therefore, the frequency of familiality one can identify will depend on the age structure of the individuals that, in fact, have been enrolled into a study.

Secondly, these traits are also often gender, environment, and lifestyle-sensitive, so that having information on the environment and lifestyle is quite crucial to the analysis of the inference of genes. One might postulate, and there's beginning evidence that the mutations that lead to many of these complex traits are not like the simple, necessary and sufficient mutations of single gene disorders, but that these mutations may be regulatory, they, in fact, may be inducible and, therefore, the environment-dependence, and they might, in fact, be interactive, not only with one another but also with the environment. Consequently, the effects of these mutations are likely to be quite complex.

In order to find these genes that underlie complex diseases, it has turned out that we need to screen the entire genome, and we need to do the screen for two kinds of features — changes of genome structure in patients, and this could imply changes in the sequence itself, or changes in the dosage of genes, and a second kind of screen involving changes in the activity of genes, and this could be in the well-described changes in RNA expression, or gene expression, but could also include both proteins, or protein expression, as well as methylation.

Although screening the entire human genome for structure and activity is a new kind of feature in genetics, individual genes and variants in those genes that predispose to common diseases have been identified over the last several decades. It turns out that specific variants are sometimes susceptibility factors in human disease, whereas at other times they are protective factors with the disease resulting from the absence of such factors.

I'm going to point out from this slide two specific examples. This slide shows eight disorders, or classes of disorders, the specific genes in each of these disorders that have been implicated, and the specific name of a variant or variants that are susceptible or protective factors. In this list, gastric ulcers represent one of the earliest known human disease and

genetic marker associations in which the common blood type B appears to be a susceptible factor for gastric ulcers.

In much more recent times, even when family studies have not been possible, it has been shown that for AIDS, the chemokine receptor 5, CCR5, has a specific deletion variant termed delta 32 that is a protective factor, so that individuals who harbor one copy of this deletion variant, or two copies of the deletion variant, are provided some protection because the HIV virus cannot gain entry into T cells.

Common disease genes and variants then play a very important role because they show, as the list of these AIDS disorders exemplify, that despite knowing the specific variant in specific genes in these disorders that all of them display non-Mendelian inheritance.

A classic conundrum in genetics, then, has still remained which is, that if one considers the distribution of risk across the population at large, one might assume that it would have a normal distribution with individuals being clinically affected if they exceed some biological threshold, shown in red on the slide.

If one considers the genetic constitution, or genotypes, of the individuals marked in the red area under the curve, they might have genotypes under two different scenarios. Shown on the left is a scenario in which individual patients have genotypes that are mutant, at one gene and one gene only, such as an individual who carries two copies of a mutation, lower case a, but is wild type or plus at the remaining genes, or individuals who have mutations in gene b, but is spared from those at gene a and c, and so on and so forth.

Under this scenario, for a fixed incidence of the disease, each of these individual variants are going to turn out to be rare. On the other hand, it is likely that the individuals marked in red on this curve simultaneously carry genetic variants at multiple loci. Under this scenario for a fixed level of incidence, all of these variants are going to turn out to be common.

We still do not know for the majority of non-Mendelian disease where many genes are involved whether human disorders are largely due to the contribution of single genes with rare variants or whether they are due to the simultaneous contribution of mutations at many genes, thereby making the variants very common in the population at large, which clearly will have very different implications both for the identification of these individuals and for treatment and therapy.

We will come back to this issue later in a specific example. All of this, then, suggests that the human genome can be looked at in a somewhat different light than we have done before. I will first point out the structural and functional features of the human genome and then look at how the frequencies of variants in the human genome can lead to non-Mendelian disease.

The sequencing of the human genome has shown that the genome can be roughly divided into two equal parts — a unique or single copy DNA part, which is little over half the genome, and largely includes genes and control sequences for those genes. It's been clear for quite a long time that these elements, that is genes, are dispersed throughout the genome. About half the human genome is repetitive in the sense that DNA elements are present in more than one copy. Although it is considered by some to be largely junk, we know that this repetitive DNA contains functional features such as those that specify the centromere and the telomeres. These are also interspersed in the genome, and only future studies will identify the specific, if unknown, functional parts of this repetitive DNA.

Human genome sequencing has shown that the human genome, and most mammalian genomes, contain on the order of 20 to 25 thousand genes that are all encoded in the three gigabases of DNA units, or nucleotides.

Modern genetics has shown that the DNA sequence contains not only the coding sequences of genes, but largely non-coding DNA elements that also specify various aspects of gene function. I show on this slide some classical studies done by David Kingsley and colleagues at Stanford University where they have analyzed multiple mutations in a gene, here BMP5, or born morphogenetic protein 5, which lead to a specific mutation called short ear in the mouse. Short ear mutations have been known throughout mammalian genetics, and recent investigators such as the Kingsley group have identified specific mutations in the known exons of this gene as a standard in all genetic investigations.

The Kingsley group has also shown, by careful studies, that three prime to the gene are specific elements, as marked on the slide, that lead to functional deficits in many other tissues such as the lung. These elements are outside the coding sequence and not known to contain any coding potential. These studies have clearly shown, as have many others, that non-coding elements can have mutations, mutations that affect the function of nearby genes.

The human genome sequence, in fact, can be overlaid over such classical genetic studies, and what this shows is that these kinds of non-coding elements, just as the coding elements shown on the left-hand side of this graph, can be identified by sequence comparison. On panel B of this slide is an indication of the conservation in DNA sequence between the human and the mouse shown as tic marks along the X axis. The height of these bars on the Y axis, in fact, show the statistical significance of this conservation with greater height showing greater statistical significance.

If you look at this graph, it is clear that all of the exons of BMP5 can be identified by this sequence comparison, as well as all of the other non-coding elements that were identified by careful analysis of the three prime region of the gene.

These studies, then, give us a new view of the human and other mammalian genome, and it suggests that all the genes which occupy 1.5 percent of the genome are been the most

important functional elements that we have studied so far, that there are other parts to the genome that may contain function as well.

Most vertebrate genomes, at least their sequences suggest, have similar gene repertoires and the way in which one vertebrate genome differs from another vertebrate genome is by expansions and contractions of families of genes. Non-coding elements, as I suggested before, in fact, have turned out to be a very important part of the genome. They carry about 3 percent of the total conservation in genomes, and they are likely to be largely regulatory, but the function of many of these elements are still unknown.

It is increasingly becoming clear that mutations in both the coding and non-coding DNA are important in non-Mendelian disease, and the aim of this lecture will be to try and show you how modern human genome sequence and the sequence of other vertebrate genomes can be used to tease out the locations of genes, the locations of non-coding elements, and how to find mutations in each of these two segments of genomes.

In the next part of this lecture, I'm going to talk about a specific example to make concrete many of the features that I've discussed so far. Hirschsprung's disease, that goes by many names, but most appropriately congenital aganglionosis, is one of the most common genetic causes of intestinal obstruction, a functional intestinal obstruction, and is one of the model complex disorders that we know today. This is primarily because genetic studies by a number of groups over the last decade have clarified how multiple genes interact to produce this very common birth defect.

There are many features of Hirschsprung's disease that are typical of all non-Mendelian disorders. Pathologically, the disorder is identified by the absence of enteric ganglia and, on the right-hand side on the top is a cross-section of the GI tract that shows the specific absence of two kinds of ganglia that are of interest, and they are those of the Auerbach and the Meissner plexus. Because of absence of varying numbers of such cells along the gastrointestinal tract, the disorder can be roughly classified into a long-segment disease that about 20 percent of patients have and short-segment disease that 80 percent of patients have. Long and short-segment disease is classified by aganglionosis that either spreads into and beyond the splenic flexure or whether it's restricted to the most distal segments of the GI tract.

The phenotype has an incidence of about one in five thousand live births, affects four times as many male rather than female offspring, and has a heritability that's roughly a hundred percent. However, no single pedigree of Hirschsprung's disease displays Mendelian patterns, and this has been shown by studies by a number of groups that non-Mendelian inheritance is the explanation for the phenotypic resemblance that one observes in pedigrees.

The phenotype is also associated with other Mendelian features or non-Mendelian features such as Waardenburg syndrome, a pigmentary anomaly, trisomy 21, or Down's syndrome, and many kinds of distal limb and renal abnormalities.

The cumulative evidence from studies done at the genetic level, at the level of developmental biology in model systems, and by biochemical analysis of the effects of both wild type and mutations in many genes involved in Hirschsprung's disease has clearly shown that the phenotype results from a disrupted interaction between enteric neuroblast and gut mesenchymal cells.

What I've shown on this graph is a cartoon that shows secretion of very specific trophic factors from mesenchymal cells of the gut, such as GDNF or EDN3. GDNF, also known as glial cell neurotrophic factor, is a very potent trophic factor that activates a tyrosine kinase receptor called RET and requires a co-receptor called GDNF R- related factor alpha. These then signal to an adapter protein called GRB10 that leads to neuroblast differentiation.

A second pathway also required for enteric neuroblast differentiation is signaling through the G-protein couple receptor and the endothelial type B receptor called EDNRB that is stimulated by its physiological ligand, which is endothelin-3. This, then, also is important in normal enteric development and is mutant in Hirschsprung's disease.

This slide shows five specific proteins, or rather genes encoding those proteins, mutations of which have been found in Hirschsprung's disease, and these include GDNF RET, it includes EDN3 and EDNRB, as well as a number of transcription factors, chief among which is Sox10 .

I want to next turn to the methods that geneticists use, and have used, in order to pinpoint the specific genes that have been shown to be associated in a non-Mendelian disorder such as Hirschsprung's disease. The chief technology, and the chief method, that has found widespread use is called genome-wide linkage analysis. This method actually uses a very simple genetic argument, and the genetic argument is that, when mutations are rare, affected family members must share genetic material, meaning DNA, at specific sites in the genome where the disease gene is located. If there's only one such gene that contributes to the trait, such a site will be present only once in the genome and clearly, phenotypes that are due to multiple genes, will have multiple sites in the genome that will share such genetic material.

This has proved to be a powerful new method for identifying genes, known and unknown, suspected or unsuspected, in the etiology of any given genetic disorder. These kinds of methods have been used by a number of investigators to, in fact, identify the chief five genes involved in Hirschsprung's disease that I referred to earlier.

This slide, then, shows the five genes as alluded to before, the total number of mutations identified in each of these genes, the phenotype of the mutation homozygote and heterozygote in human patients, and the percent penetrance where it can be estimated because the number of mutations are large enough to get a reliable estimate.

What I show here is that marked in red are the two chief factors which are known to be mutant in Hirschsprung's disease, and they are the two receptors, the receptor tyrosine kinase RET and the G-protein couple receptor EDNRB. Each of them have multiple mutations, but the phenotypic effects are quite distinct.

Ret, in fact, has a nearly identical phenotype in all heterozygotes and in the single rare homozygote known. It has high penetrance; however, the disease is restricted primarily to heterozygotes. Mutation in its physiological ligand GDNF has the same feature as in the receptor. EDNRB, on the other hand, has Hirschsprung's disease as a phenotype in both mutant homozygotes and heterozygotes, again with high penetrance. However, mutant homozygotes also manifest other non-enteric features such as sensory neural deafness, often bilateral, and other pigmentary anomalies of the skin and eye. EBN3, which is the physiological ligand for EDNRB, also shows a pattern similar to the endothelial type B receptor.

Finally, the transcriptional regulator Sox10 is only known to be mutant in heterozygotes, but these mutant heterozygotes have features that are indistinguishable from the endothelial pathway mutations. These mutations also have high penetrance.

Taken in total, this shows that Hirschsprung's disease can be found in individuals who are mutant for only one gene, or individuals who are mutant for both alleles of a specific gene. They can carry non-enteric manifestations, both as homozygotes and heterozygotes. If these mutations turn up to be common in the population at large, you can imagine that the effects can be quite complex and not display any simple patterns of inheritance.

The previous slide showed Hirschsprung's disease genes identified primarily by looking at families that had patients with long-segment disease and patients that had syndromic associations. The vast and most common kind of Hirschsprung's disease is short-segment disease, and it turns out that linkage analysis of the type that I mentioned before leads to the identification of three different genes, or three sites of the genome, mutations which cumulatively lead to Hirschsprung's disease. It turns out that one of these three genes is RET, and the other two are currently unknown genes. Shown at each of these genomic sites are two quantitative quantities, the upper Q representing the estimated frequency of the mutations at that gene, and the lower lambda representing the risk of having one of those mutations.

If one looks at RET, it is clear that mutations occur with a fairly high frequency, one percent of the population, but the reason why these individuals might not be clinically affected is because they do not have mutations at the other genes in chromosome 3 and 19. However, RET mutations, when they occur, are powerful because they can increase risk about eight-fold over individuals who do not carry such a mutation.

Similar features of the two other genes in chromosome 19 and 3 show that they are also common in the population, having frequencies of four to five percent and that they increase risk four to five-fold over individuals who fail to carry such mutations.

A common feature and conclusion of all of the studies that are searched for mutations in RET in Hirschsprung's disease is that they represent loss of function mutations of RET, and that these mutations are necessary in all forms of Hirschsprung's disease whether they're syndromic or not, or whether they are long-segment or short-segment.

However, a second feature of RET mutations is that, even when linkage to RET, that is when the involvement of RET has been suspected, that mutations have not been uncovered. The mutations in RET that have been identified so far can affect any aspect of RET protein function, and they can occur anywhere within the gene and, therefore, affect any specific part of protein function in RET, such as those that impact on single peptide formation, or those that affect adherent binding in the extracellular portion of the protein or those that affect the catalytic function, that is, by having mutations in the tyrosine kinase domain itself.

A second way of finding genes that the first linkage method often fails to find is by doing genome-wide association studies. Genome-wide association studies are recently becoming popular, and the reason for this is that this are relevant for finding common mutations and not the rare ones that are family-specific. Many new genes can be identified by using this principle, and the main feature of this method is the assumption that affected individuals, not only within families, but across all families, share some genetic material at the disease gene. This is, of course, only possible if the mutation occurred somewhere in the history of humans and has since spread to many humans who are now no longer closely related to one another.

Almost all of the examples of common disease variants that I mentioned earlier, such as the B blood type, leading to gastric ulcers, or the deletion mutation in CCR5 that leads to HIV resistance are, in fact, mutations of this sort.

Genome-wide association studies have now become possible because the human genome project has led to the identification of many, many genetic variants across the human genome and the genome of many other organisms. The simplest kind of association study that one can perform is, of course, a case control study, but what I show here for the example of Hirschsprung's disease is a family-based association study in which affected individuals can be searched for genetic variants in comparison to the variants that this individual received from his or her parents.

The box here shows a search of the human genome of 566 STRP's, or short tandem repeat polymorphic sites, 9,000 SNP's, or "snips", or single nucleotide polymorphisms in 35 trios . The trio here represents an affected individual, termed here TT, who gets a transmitted chromosome T from each of his parents, and an untransmitted chromosome U from the other chromosome in each of the two parents.

By comparing the frequency of alleles on the transmitted and untransmitted chromosomes, one can perform an association test because, clearly, mutations that are common will be enriched on the transmitted chromosomes and will be depleted on the untransmitted chromosomes.

This association study was done in Hirschsprung's disease in the Mennonites because the Mennonites are known to have a ten-fold increase in the incidence of Hirschsprung's disease, and we considered, given their isolated status, that they would, by virtue of their common ancestry, have a common mutation.

A search of the human genome, then, showed that both the endothelial type B receptor and the RET receptor, in fact, had mutations of variants, genetic variants that predisposed to Hirschsprung's disease in this population. What I show in this panel are specific variants of mutations in each of these two genes which have a background frequency on untransmitted chromosomes of 8 percent for the endothelial type B receptor, but a frequency that is nine times increased on transmitted chromosomes.

For the RET receptor, there are two haplotypes, or regions, that contain genetic variants of the RET receptor, a five-prime and a three-prime haplotype, each of which have fairly high frequency or non-transmitted chromosome, 30 percent on untransmitted chromosome, nearly doubled untransmitted chromosome for the five-prime haplotype, and a four-fold increase from 6 percent on the three-prime haplotype. There is some initial evidence that a similar factor could also lie on human chromosome 16, but its status is still uncertain and needs to be tested in a larger sample.

A second feature of these data, and that can now be routinely studied in non-Mendelian disorders is that, even though the genetic variants at the endothelial type B receptor and RET are in two different human chromosomes, chromosome 13 and chromosome 10, respectively, and the fact that these two chromosomes segregate independently, that we still find evidence of non-random transmission of the two sets of genetic variants on these two independent chromosomes. This is denoted by a high level of statistical significance as shown at the bottom of the slide showing that these two sets of genetic variants interact with one another in order to produce the disease phenotype. This interaction is also clear if one classifies all of the Mennonite patients, as I show on this slide, by both the genotype at RET, as well as the genotype at the endothelial type B receptor.

I show two sets of figures, a percentage penetrance, as well as, in parenthesis, the number of patients who are clinically affected over the total number of patients who have the indicated genotype. And it is clear that, as the RET susceptibility factor, denoted A in this graph, denoted A in this slide, increases, and the dosage of the endothelial type B receptor mutation increases, shown as C on this slide, that the penetrance of this combined genotype also increases, meaning those that carry more mutations at both loci, show higher rates of affection than those that carry fewer mutations at one or the other, or both, loci.

So, this is the hallmark of non-Mendelian inheritance, or the inheritance of different numbers of mutations in different genes leads to different degrees, or probabilities, of affection by Mendel's laws.

Seeing that there are two groups of genes, or two groups of variants in individual genes that leads to a non-Mendelian disorder is, in fact, straightforward, but in order to prove the joint action of multiple genes, one clearly needs experimental models.

I'm now going to discuss two such models, one for showing that the disease phenotype can be recapitulated, and this will use the RET and EDNRB mutations as examples. But animal models and experimental models can also be used for identifying and testing of new interactions in disease, and I'm going to show one example of how the copper zinc superoxide dismutase , called SOD 1, interacts with RET to lead to Hirschsprung's disease.

RET and EDNRB not only interact in the Mennonites, but also interact in a mouse model for Hirschsprung's disease. In order to prove this, we created mice with the appropriate genotype and, shown here in this two by two slide, are the genotypes of mice that individually could be wild type or plus at RET, or carry one copy of a heterozygous known mutation. At the endothelial type B receptor, mice could either be heterozygous for an allele named piebald or S or, alternatively, a second allele called piebald lethal at EDNRB, or they could be homozygous, or transheterozygous, for piebald and piebald lethal alleles.

What this slide shows is that mice of the individual genotype, that is, mice that have mutants at either RET, or EDNRB, do not show signs of Hirschsprung's disease, but that mice that are heterozygous for RET, and are homozygous, or compound heterozygous for two mutations at EDNRB, in fact, are 100 percent affected. This, then, completely recapitulates what we observed in the Mennonites, and it's quite important to show that segregation at more than one gene is simultaneously required in order to lead to the disease phenotype.

This is, of course, seen phenotypically in the mouse where the mouse here in the picture at the bottom is the wild type, or control mouse, but the one on the top that is heterozygous for RET and homozygous for piebald lethal clearly shows the abdominal distention seen in human patients as well.

Upon dissection, it is clear that that mouse that has heterozygous genotype at RET and homozygous for the piebald genotype also has megacolon as well as, upon detailed dissection, shows aganglionosis of the GI tract. Intriguingly, these mice of compound genotype also show a sex difference at least in the onset of expression in that male mice show patterns of clinical expression that are much earlier than female mice.

A second kind of a screen, genetic screen, that is becoming very important to the study of non-Mendelian disorder now uses drosophilia , or the fly, as a model. RET, in fact, also

contains gena-function mutations, and this gena-function mutations lead to a series of neural endocrine tumors of multiple endocrine neuroplasia. What's shown here in the center is a photograph of a drosophilia eye in which expression of RET gena-function MEN2B allele has been forced. This leads to a phenotype in the drosophila eye that's been called a rough eye. This fly can be mated to existing stocks that either have a deletion, or mutation, in different parts of the fly genome, and these resulting offspring can then be screened with respect to the eye phenotype to find flies that either have a worse, or enhanced, phenotype shown on the left, or one in which there's a milder phenotype, or a suppressor, that's on the right.

By observing now from the drosophila sequence genes that are absent in these deficiencies, or deletions, across the fly genome can lead to the identification of specific genes that, in conjunction with the known gene such as RET, can lead to Hirschsprung's disease. This has now been done by a variety of studies by us that have led to the identification of the copper zinc superoxide dismutase as a modifier of RET expression.

In the similar experiment on the left, you see a wild type drosophilia eye in which every individual facet has been clearly demarcated. On the right-hand side are two sets of flies, in the middle, an MEN2A kind of mutation in which the rough eye phenotype is evidenced as is a dimunition of the size of the eye. When a fly is now made transgenic for a SOD1, what one observes is still the rough eye so that the phenotype is not completely lost but, clearly, the size of the eye is much greater, showing the proliferation induced apoptosis that is seen in the MEN2A transgenic allele is abrogated in the MEN2A SOD1 transgene, suggesting that SOD1 interacts with RET in a very direct way.

These kinds of models, then, are going to turn out to be very important in order to prove that the different genes that we identify in Hirschsprung's disease, or in any other non-Mendelian disorders are, in fact, the true culprits.

I have mentioned earlier that RET has lots of function mutations that lead to Hirschsprung's disease. These mutations generally are rare and often family-specific. However, I'm going to point out now and show some evidence that a common loss of function mutation is found in most human populations, especially populations that reside outside Africa. The finding of this mutation is quite significant because I have alluded to the major problem in Hirschsprung's disease and RET being that, even though RET appears to be involved by analysis of linkage, we can find mutations in only about half of familial cases. This suggests that the remaining mutations either lie in the non-coding sequence in RET, which is clearly likely given the features of the genome that we know, or it's due to a tightly linked gene, a gene immediately in the neighborhood of RET, which linkage analysis cannot distinguish from genetic segregation at RET.

In order to distinguish between these possibilities, modern genomics offers us the possibility of comparing the sequences of multiple genomes to narrow down the regions of the human sequence that have been conserved for very long periods of evolution time.

I'm showing here 350 kilobases of sequence surrounding the human RET gene in comparison to the sequence of 12 other vertebrates going all the way back to three distant fish — such as zebra fish, the Fugu and Tetradon . What one sees from these comparisons are regions of very high sequence identity, particularly among the primates and among the other mammalian vertebrates and regions of somewhat less, but still statistically significant sequence identity shown in green of 50 percent or greater. These kinds of studies done in detail show that not only are the coding sequences that comprise the RET gene conserved throughout evolution but, about twice as much sequence outside the coding region is conserved as well.

A small segment of that sequence is shown in much greater detail in this diagram where these kinds of coding elements are called MCS's, or multiple conserved sequences. This shows the intron one of the receptor of the gene encoding the receptor RET where, in blue, the coding sequences are clearly demarcated. What I show here are the exons 1 through 3. Beyond the conservation of these exons, which are shown by increasing height on the Y axis, are also shown three other elements that have been marked and that are MCS's. These MCS's marked in red represent the fact that these are not known to code for coding sequences. These MCS's, then, are thought to provide some regulatory function and, therefore, are candidates in which DNA sequencing can uncover variants as well.

We have used a family-based association study to demonstrate that the highest degree of association that we see, in fact, maps on top of one of these associated sites, in fact, on the third such conserved element that's been marked as MCS plus 9.7. When one compares the sequence of the site to all of the other known vertebrates, it's clear that a single mutation distinguishes all human patients with Hirschsprung's disease from those that are considered to be the wild type in not only humans, but in other species as well. This element is now known to function as an enhancer. We've been able to demonstrate this by a series of control experiments, as well as a series of transfections into a specific neuroblast tumor cell line called neuro 2A.

A piece of DNA that contains both the variant sequence, called CS + 9.7V, as well as the non-variant, or wild-type sequence, CS + 9.7NV, clearly shows that the variant site reduces transcription of the RET gene by six-fold or greater. If one includes the remaining, or the secondary site of conservation at position 5.1, here denoted as CS + 5.1, then that also shows a dimunition of transcription of the RET gene.

The previous studies have clearly shown that the mutation we have discovered is a likely loss of function mutation, although not a complete loss of function mutation, of an enhancer of RET. It is quite surprising, then, that this mutation has a worldwide distribution as shown on this graph. This shows a study we've done in which 51 populations had been sampled across the world and, in each, the frequency of the wild-type change, shown in chocolate, and the mutation in blue have been plotted for each of the populations that have been sampled in the indicated portions of the world.

It is clear that this mutation is very rare within Africa. It increases to a frequency of somewhere between 15 to 30 percent throughout most of Europe and, in fact, Southeast Asia, whereas in Eastern Asia, as well as in the northern parts of Asia, it shows a frequency of 50 percent, or even greater.

This shows clearly a mutation that has increased in frequency despite having a strong deleterious effect in human patients, or currently in human patients, and suggests that the susceptibility allele is very common because it was likely positively selected for some time in human evolution.

This, of course, is not completely new and has been known in human genetics as a central feature why some disease alleles are found to be so common. The chemokine mutation, delta 32, that's involved in HIV protection, is thought to play similar role. Its protective effect in HIV transmission is clearly known, and it's possible that, in the past, that the same mutation provided protection to perhaps some other kind of viral agent.

A second good example of such protection is seen in another chemokine receptor that leads to the duffy blood group in which a known mutation is known to protect individuals in Africa where this mutation is found in nearly 100 percent frequency from plasmodium vivax infection. So, it is quite likely that the reason why this RET allele that loses function and is still so common because it provides protection to heterozygotes for RET for some kind of selective agent, although exactly what that agent is currently unknown.