So, I want to thank the organizers of the Reynolds lecture in cardiovascular disease for having me here today. It's a pleasure to be here and to talk to you about SAGE analyses of human transcriptomes. And, of course, as we all know, the human genome project has progressed to a point where the determination of the human genome sequence, and the genes contained within it, has been completed. However, this really represents only one level of genetic complexity. A second, and equally important level of complexity, is the ordered and timely expression of these genes in the genome throughout the processes of development and differentiation.

The adult human body contains over several thousand different cell types, each of which presumably has its own pattern of gene expression specifically designed to carry out its physiologic function. And, of course, each of these cell types can be subjected to a wide variety of environmental states, thereby altering these expression patterns.

So, many years ago, in fact about a decade ago, we decided to develop a method that would allow us to look at gene expression patterns in a quantitative and global sense, and we wanted this methodology to allow us to compare both different physiologic states, but also pathologic states. And, finally, we wanted to develop a method that would allow us to also discover previously uncharacterized genes.

So, today I will be telling you about this method that we have developed for this purpose, which is called serial analysis of gene expression. So, in the first part of my talk, I will be describing the serial analysis of gene expression method, its technical details as well as its advantages and disadvantages compared to other methods of gene expression analyses. And, in the second part of the talk, I'll be describing an example of how we have used SAGE to analyze human cancer. And, finally, I'll talk about how we have used SAGE to look at the human genome in more detail and to identify previously undiscovered genes.

The SAGE method is based upon two basic principles, and that is that a short sequence tag of ten base pairs in length can distinguish among a million odd different transcripts, while the human genome contains no more than 100,000 different genes.

The second basic principle is that the short sequence tags can be linked together, or concatenated, to form long serial molecules which can be cloned and sequenced. Sequencing across one such concatemer can identify on the order of 30 or 40 tags in one sequencing reaction. And sequencing of multiple such concatemers can identify a very large number of sequence tags in a very short period of time.

The SAGE method essentially works by obtaining poly A RNA from a tissue or cell type of choice and obtaining short sequence tags from within these mRNA molecules. This can be done using two different types of enzymes — one a regular frequently cutting four base pair enzyme which we call an anchoring enzyme, and the second is a type 2S restriction enzyme which recognizes a non-palindromic restriction site and cuts into the DNA molecule.

Using these two enzymes, we're able to obtain short sequence tags from the transcript molecules, and these short sequence tags are then linked together, or concatenated, to form long serial stretches which, as I mentioned earlier, can be cloned and sequenced.

We have developed software that allows one to take the sequence information from these concatemers and to identify the tags that are contained within each concatemer . The anchoring enzyme sites that I mentioned earlier serve as punctuation marks which identify the beginning and end of each pair of tags within a concatenated sequence.

One can then use software to analyze these sequence tags against existing DNA databases to identify the transcript from where the tags originated. And, finally, one can simply count the number of times a particular tag is seen in any cell or tissue that has been analyzed. And one can see from the graphs below that tags that are present at higher numbers in one state versus lower numbers in another correspond to genes that are differentially expressed between those two tissues. If the number of tags remain similar between two tissues, that means that their gene expression pattern has not changed.

In this sense, SAGE is a digital technology. It allows one to quantify expression patterns in any cell or tissue type of choice simply by enumerating the number of times a particular transcript molecule has been observed.

Now, there are a number of ways in which one can look at gene expression profiles in a global way. There are a number of microarray technologies and other technologies which are becoming available and, certainly, these have a lot of advantages. One of them is that they can allow one to look at many different samples for low cost in a fairly rapid amount of time.

However, SAGE also has some advantages despite the fact that it is a little more laborious and expensive than these other approaches. The advantages that SAGE provides are that, first, it allows one to look for expression patterns in a more quantitative way. Because of the digital nature of the technology that I mentioned, the expression patterns that one obtains are absolute. They allow one to determine the expression level of any gene in a population of transcript molecules that one has analyzed. Of course, the genes under analysis do not need to be genes that have been previously identified as is the case with microarrays, for example, where one is limited to looking solely at the genes that are on a microarray. But, in this case, one can identify genes that have not been previously discovered and, although this might not seem like an important point where the genome has been fully sequenced, as I'll talk about later, there's actually quite a bit of evidence that there are many genes in the genome which have yet to be discovered.

And, finally, because the expression levels that one obtains by SAGE are absolute, the data can be stored in databases and allow one to analyze any new data that are obtained either in one's own laboratory or from other laboratories around the world to the existing SAGE datum. This allows the progressive accumulation of SAGE data to become an evermore useful

data source for individuals wanting to mine such data for expression analyses in particular tissues or cell types of interest.

Now, when we first developed SAGE, this method was not as high throughput as it is today. And, back in 1995, approaches for automated sequencing were just becoming widely available and, in fact, the first libraries that we generated were done using manual sequencing and, because of that, one could obtain only several thousand tags at one time. And, additionally, the SAGE technique involved a number of different steps which were technically challenging and required one to follow careful protocols.

Now, over the years, that process has become much easier with the availability of a commercially available kit for performing the SAGE technique and, additionally, methods for automated sequencing have made the process of sequencing SAGE tags much simpler. So, as you can see in this graph, over the years, the number of SAGE tags that have become available has been increasing almost exponentially. And we expect that this will continue as, using today's sequencing technology, one can easily obtain over 100,000 SAGE tags per day and, in the future, with massively parallel sequencing technologies, that many different companies and individuals have been talking about, it is conceivable that one will be able to look at on the order of a million different tags per day.

Now, these numbers are quite large and, in fact, with a million tags, one would be sampling a transcriptome of a cell, that is, the entire number of mRNA molecules inside of a cell several-fold over. A typical cell is thought to contain approximately 300,000 mRNA molecules and, by analyzing over a million tags, one would be analyzing this entire population of transcripts several times over.

Now, in addition to these technological advances that have made SAGE more facile and more high throughput, there are an increasing number of SAGE databases that allow one, as I mentioned earlier, to link one's own individual data for further analysis and comparison. These databases currently contain over 20 million tags from over a hundred libraries, and a major repository of these databases is the cancer genome anatomy project which currently contains over 15 million SAGE tags from cancers, as well as normal tissues from the brain, colorectal tissue, breast, ovarian, pancreatic, and prostate cancers. These are located at the website that's seen below and can also be accessed through the SAGE site that's listed below that.

Now, when considering a SAGE project, one should think about the fact that, in addition to the SAGE method itself, there are a number of up-front steps that must be performed as well as downstream steps that must be performed in any such analysis and, in many ways, the SAGE part of this, which is seen here in the middle of this slide and involves both library construction and sequencing of a SAGE library, both of these steps typically take less than two weeks to a month, in total, and one can perform multiple such libraries at the same time. We'll take up a much smaller fraction of the overall period of time that is required to both develop a test system, which I mentioned up front, and to both select the candidate

genes that are obtained by the SAGE analysis, as well as to evaluate the candidate genes further in a variety of biologic systems.

Now, let me talk a little bit about the development of a test system that essentially means either obtaining a cell type or tissue of interest or developing a model system with which to analyze, for example, a transcription factor, or a growth factor, or whatever one might want to analyze in more detail. Even in terms of obtaining a cell type or tissue of choice, one has to be extremely careful because the input that goes into a SAGE analysis for any gene expression analysis, for that matter, is really no better than the starting material. So, one needs to obtain material that one wants to analyze and is sure that that material faithfully represents the transcript analysis that one wants to obtain.

So, let me give you a brief overview of the types of analyses that have been done over the past decade using SAGE. Many of these have been performed on cancer. We have analyzed colorectal cancer, as I will talk a bit about later, but there have been a number of other cancer types that have been analyzed, both in terms of the tissue that's involved, as well as pathways that are thought to be important in tumorigenesis.

The publications that have involved SAGE also include a variety of different human diseases, both developmental disorders, as well as hereditary diseases, analysis of various molecular pathways, analysis of plant and a variety of different model organisms, use of SAGE to compare to other gene expression methods. A variety of different papers have used SAGE, as a sense of a gold standard, to look at their own expression technologies. And, finally, as I'll talk about at the end of my talk today, SAGE has been used to annotate the genome using the transcript information that is contained within it.

SAGE has also been used for analysis of a number of different cardiovascular diseases. These have included analyses of normal heart tissue, of cells that are under hypoxic conditions, of cells that differentiate into cardiomyocytes, vascular tissues, one of which I'll talk about later, which are analyses of tumor endothelial cells and normal endothelial cells, but also of human umbilical vein endothelial cells and also of other vascular diseases such as aneurysms. I will leave it to the viewer to look in more detail at these references for their various analyses, and I will now turn to talking about using SAGE to analyze colorectal cancer.

Now, colorectal cancer has been studied for many years, and one of the basic questions that still plagues this disease is trying to determine the molecular events that make the cancer tissue different from the normal colonic epithelial cells that surround it. And, of course, one of the major differences in cancer versus normal cells are alterations in specific genes at the level of the genome, and these include inactivation of certain tumor suppressor genes such as P53 and APC and mutational activations of certain oncogenes, such as beta catenin and K-ras .

But, while these genetic alterations serve as the basis for how we look at colorectal cancer today and serve as the canonical model for tumor progression in other cancers, they

really under-represent the number of changes that go on in cancer. And, indeed, one can imagine that, at these different stages of tumor progression, when one goes from normal colonic epithelium, to dysplasia, to adenomas , to carcinomas and, finally, to metastases, there are a whole host of other changes, including changes in gene expression patterns which may underlie these different states.

So, early on, we sought to use SAGE to better understand these pathologic changes that go on in cancer and to use the expression profiles to identify potential diagnostic markers, prognostic markers, and targets for therapeutic intervention either using small molecules or potentially with gene therapy.

Now, in analyzing cancer, as in any complex tissue, there are a number of challenges, and these include the fact that there are many changes that go on at the expression level, both throughout the course of tumorigenesis, but also within different parts of the tissue if the tissue is, in fact, quite complex, even when one thinks of the fact that cancer is a clonal disease, and I'll show you an example of that in a second.

And finally, there are a multitude of genes that can be involved in these processes, some of them characterized and many of them, perhaps, uncharacterized. This is an example of a cancer specimen and what one can see, although in many cases we think of cancer being a large number, a large collection, of cells that are all identical, one can see that that is typically not the case, as shown here by this slide. One is confounded with both necrotic tissue, which is on the left as well as stroma, and as well as normal surrounding tissue and inflammatory cells. And, in fact, the fraction of viable tumor cells, in many cases, is actually quite small.

So, several years ago, an M.D.Ph.D. student in our lab**,** Saurabh Saha**,** decided to use SAGE to further analyze human colorectal metastases. Now, he realized, based on the slide I just showed you previously, that one of the first things he would have to do would be to develop a method to isolate these metastatic cells from the contaminating normal and infiltrating cells. And this approach, which essentially is an immunoselection of these metastatic tumor cells, essentially involves two steps. One is a negative selection, where one uses antibodies to remove hematopoietic cells using magnetic beads and antibodies against those types of cells, and a positive selection which can allow one to use the BerEP4 antibody linked to magnetic beads to purify away the tumor epithelial cells.

The second step that needed to be done was to be able to modify the SAGE approach to analyze such small sub-populations of cells. And Brad St. Croix had, at the same time, been developing this for a different purpose, which I'll talk about a little bit later in my talk. And using microSAGE, one would now be able to analyze expression profiles from cell sub-populations, or microanatomic structures, of less than 50,000 cells or 5 micrograms of total RNA. That would normally have been impossible using the SAGE procedure or other gene expression profiles. So, this provided a whole new window into looking at gene expression profiles for such sub-populations of cells or tissue structures which would have been microdissecting.

So, Saurabh used microSAGE and the immunopurification protocol that I described to analyze a total of about 100,000 SAGE tags from several different tissue specimens. He compared these data to SAGE data obtained from normal colorectal epithelial cells that had been previously obtained by Lin Zhang and Wei Zhou in our laboratory, as well as SAGE data obtained from colorectal carcinomas. And, using this analysis, he was able to hone in on a series of genes that were preferentially expressed in metastatic cells, but expressed at very low levels in normal cells or primary cancers.

And what you see here are the top 20 differentially expressed transcripts that were obtained by this analysis. And what you can see on the left is the SAGE tag, and next to it the number of times that tag was observed in normal colonic epithelium followed by the number of times it's observed in the primary cancers. And, finally, in the third column, by the number of occurrences observed in the metastatic cells. And for each tag on the far right column, you can see the description of the gene corresponding to that tag. This represents a typical output of a SAGE comparison.

And, while there were a number of genes that were differentially expressed in this way and seemed to be highly enriched in the metastatic cancer cells, Saurabh Saha decided to independently validate these using quantitative PCR, and what he found was actually quite surprising. While all of these top 20 genes that you see here were expressed at some frequency in metastatic cancer cells, and not expressed in normal cells or primary cancers, only one of these, the first one, PRL3, was expressed almost entirely in every metastatic sample that was analyzed.

This is the type of data that Saurabh saw when he did quantitative PCR on transcripts obtained from these different tissues. As you can see by the dark red peaks, these are the expression profiles of PRL3 that were obtained in purified metastases. To the right of those are unpurified metastases which also show expression of PRL3, albeit at a lower level. But again, remember that the unpurified metastatic lesions contain only a small number of metastatic cells. And to the left of those highly expressed columns, one can see the expression level of PRL3 in primary cancers and, even further to the left, in adenomas , as well as normal tissues. And you can see there, it's essentially not expressed.

Saurabh also looked at using quantitative PCR to study expression levels of PRL3 in colorectal samples that were obtained at different stages throughout tumor progression from the same patient. And what you can see in this slide are expression levels of PRL3 in normal epithelial cells, in primary cancers, and in metastases from six matched patients. And you can see in every case the metastatic lesion had increased levels of PRL3 compared to the normal cells. But interestingly, you can see in this case that the primary cancers did have a small increase in the amount of PRL3, suggesting that patients that already have metastases perhaps obtain expression of PRL3 prior to, or coincident with, invasion.

He then went on to use in-situ hybridization to specifically look at metastatic lesions and see if PRL3 was specific to the metastatic cells. And, indeed, as one can see from this analysis of a colorectal cancer metastasis to the brain, PRL3 is exclusively expressed in the epithelial cells of these metastatic lesions. An analysis of metastatic lesions from a number of different sites throughout the body shows that PRL3 is, indeed, over-expressed at high levels in all of these cases, and not in primary cancers or adenomas , or normal colonic tissue.

Now, this would have made PRL3 an interesting gene in its own right because it is expressed so uniformly in these metastatic lesions, but there are many changes in gene expression patterns that are not necessarily mediators of tumorigenesis but are, more or less, passenger changes. So, we wondered whether PRL3 might be linked more directly to tumorigenesis. And so, Saurabh undertook an analysis of gene content on the genomic loci where PRL3 was identified.

PRL3 is located a small distance away from the telomeric tip of chromosome 8Q. And what Saurabh found was that, in analysis of twelve metastatic lesions, three of those had an amplification of the region containing PRL3. Now this region is actually quite small, less than 100 KB in size, and only contained the PRL3 gene. This suggested that PRL3 is not only differentially expressed in metastatic lesions, but can be genetically altered and selected for during tumorigenesis, thereby implicating PRL3 in the tumorigenic process more directly.

Since this work was completed, there have been several studies looking at the function of PRL3, and I wanted to highlight one from a group by Guo et al. in Singapore, which essentially used cells expressing PRL3 or a mutant version of PRL3 that lacks a phosphatase domain to see the effect of these cells in mice. And what these mouse models have shown is that cells expressing active PRL3 lead to a very large and profound metastases in the lung of such mice, while cells expressing inactive PRL3 have essentially no effect.

So, in summary, what these studies using SAGE have shown is that PRL3 is consistently, dramatically and specifically expressed in metastases derived from colorectal cancers, that PRL3 can be amplified in a sub-set of cases, and that over-expression of PRL3 promotes tumor metastases in animal models. Now, the exciting aspect of PRL3, of course, is that it encodes a protein tyrosine phosphatase, an enzyme and, as we know from several recent examples in human cancer treatment, especially those by Gleevec and Iressa that target various tyrosine kinases, kinases and phosphatases represent attractive targets for therapeutic intervention. And a lot of work is underway trying to develop new therapeutic molecules that will specifically inhibit PRL3 in the hope of one day using this as a therapeutic modality for colorectal metastases.

Let me now move on to another example of how we have used SAGE to look at a different aspect of tumorigenesis, and this involves not the cancer, per se, but the surrounding tissues in which the cancer develops. And one important tissue type that has been recognized over the years to be essential to tumor formation are vessels. And, in fact, in the 1970's, Judah Folkman hypothesized that tumors, in order to grow beyond a millimeter, or two millimeters,

in size needed angiogenesis to occur, and that this angiogenesis would be responsible for the formation of larger lesions.

This turns out to be the case and, as one can see in this colorectal cancer shown on the left panel, not only do vessels occur in these primary colorectal cancers, but one can see that there's a ring of viable tissue just surrounding the vessel but, that further away, there is necrosis, suggesting that the cancer cells are dependent on the vessel for both micronutrients and for oxygen for continued tumor growth.

So, one way to use the process of angiogenesis for therapeutic modalities is to target existing genes that are expressed in angiogenic cells. But Brad St. Croix hypothesized that perhaps a better way to do this is to identify genes that are specifically expressed in cells that are involved in tumor neoangiogenesis as opposed to cells that exist in normal vessels of the colon epithelium, or other vessels throughout the body. So, he devised a method to purify endothelial cells that's very similar to the method I described earlier for purifying epithelial cells that essentially involves both the negative selection and a positive selection. The negative selection in this case removes the epithelial cells, as well as other hematopoietic cells, and the positive selection uses anti-epithelial antibodies to purify away just the endothelial cells.

And he used this analysis to look at 100,000 transcripts from colon cancer, endothelial cells, as well as approximately 100,000 transcripts from normal mucosa endothelial cells, again using the SAGE approach. And what this study showed, depending on the way one performs this analysis, is that a number of different markers involved in endothelial cells emerge. One are the panendothelial markers, which are genes that are highly expressed both in tumor endothelium as well as normal endothelium, but expressed at very low levels in a number of different cell lines that are tumor derived and should not contain any endothelial cells within them.

He identified 93 such panendothelial markers. He also identified markers that were highly expressed in normal endothelial cells but expressed at low levels in tumor endothelium or cell lines grown in vitro. And finally, he identified 46 tumor endothelial markers, that is genes that are highly expressed in just tumor endothelium, but expressed at low levels in normal colonic epithelium, as well as in normal human cancer cells grown in vitro.

While these data can be useful for a number of different reasons, Brad was specifically interested in looking further at the tumor endothelial markers as these potentially could be useful to target tumorigenesis. And he looked at the most highly differentially expressed such TEMs as they're called, using *in situ* hybridization and found that, in all cases, these were specifically expressed only in endothelial cells derived from colon tumors, but not from normal colonic epithelium. This was performed using RTPCR.

Using insight to hybridization, Brad was able to also show that these tumor endothelial markers are present not just in primary cancers, but also in tumors of other origins such as lung tumors and brain tumors, in liver metastases obtained from colorectal cancers and, finally,

it appears that these tumor endothelial markers are expressed at higher levels in normal processes of neoangiogenesis, such as wound healing and corpus luteum formation.

This is an example of a tumor endothelial marker as seen by in-situ hybridization in a colorectal cancer. And you can see that the endothelial cells nicely ring this small vessel, and the same is the case in this breast cancer and lung cancer that shows in-situ hybridization analyses of TEM 7.

So, what this study has shown is that, using SAGE, one can analyze specifically genes that are present in small sub-populations of cells within the endothelium. Obviously, these would have been difficult to analyze if one had looked at the bulk tissue and, of course, have been missed all these years in analyses of various EST libraries which have also been obtained from bulk tissues.

In addition, the studies have shown that normal and tumor endothelium are highly related. There are many genes that are highly expressed in both of these types of endothelial cells and not expressed essentially anywhere else in the human body. But, despite that, tumor derived endothelial cells are quantitatively different from normal endothelial cells, or normal cells, and these genes that are tumor specific, can be used not only to characterize these endothelial cells but, perhaps, to target them.

So, the immediate goals of this study are to identify promising tumor endothelial markers but, in the future, one can imagine using this sort of differential expression patterns to better understand neoangiogenesis, to detect angiogenesis and, finally, to target it in human cancer. And, of course, one of the nice things about these processes is that they are not limited to just colorectal cancer, but may be applicable to a large number of different cancers.

So, for the last part of this presentation today, I will talk about using SAGE to further explore the genome. And the basis for this study came out of the fact that, as more and more SAGE data accumulated in the existing transcript databases, and this is an example of one such description of the cancer genome anatomy project database which housed, at the time, about seven million transcript tags from 27 tissues represented by 171 libraries and showing over 104,000 distinct transcripts, that there were many such transcripts that were not matching genes that were known to be annotated in the GenBank databases.

In other words, these transcripts likely represent novel genes. It turns out that these look like transcripts that are expressed at low levels, and this might be what one would expect from the fact that most of the genes that have been annotated have been annotated using EST databases, which have a limited depth of analysis. What one can see here is that a significant fraction of the genes detected, of the transcripts detected by this analysis, did not match known genes in GenBank. These turned out to be genes that are expressed at lower levels less than five transcript copies per cell, and this would be consistent with the fact that many of the genes that have been annotated in GenBank depend upon the genes having been highly expressed in order for investigators to have found them over the years.

Now, what is the status of the human genome project and the genes that are contained within it? The initial draft analyses of the human genome show that there were approximately 30,000 genes. These were comprised by 15,000 well-characterized genes, as well as 10 to 20 thousand gene predictions, depending on the prediction models that have been used. However, when this number came out, it surprised a number of biologists because there had been and, after the studies came out, continued to be a number of studies that suggested that there were additional genes in the genome, and these included EST studies, full-length cDNA studies, micro-array studies, and the SAGE studies that I mentioned.

So, we wondered whether we could use SAGE, not only to characterize gene expression patterns of known genes, but also to identify previously undiscovered genes. And, what one can imagine doing is taking SAGE tags from individual transcripts and matching them back to the genome, as shown here in this figure. When one does that, one can match both known genes, novel introns of known genes, that is, introns that may not have been appreciated in alternative splice transcripts or, finally, the tags may match completely undetected genes that are present in the genome.

And so, we wondered whether we could use existing SAGE approach to do that, but it turns out that the length of the SAGE tags is simply not sufficient to allow one to match the tag back to the genome in a specific manner. So, we simulated, as shown in this slide, that one would need a tag of 19 to 21 base pairs in length to specifically identify the location in the genome from where these tags were derived.

Now, remember, that even the shorter SAGE tags are specific to a particular transcript molecule if one is simply looking at the cDNA molecules inside of a cell or cDNA databases but, as the genome is a much larger sequence, one needs longer tags in order to match these tags specifically to the genome.

So, Saurabh Saha thought of a way to do this using SAGE, that is, to obtain tags that were longer than the conventional SAGE tags, and he did this by using a difference type 2S restriction enzyme, called MNE1, and this provided a 21-base pair tag. And, initially, we use this to analyze a colorectal cancer cell line and to show that, similar to our simulations, these long SAGE tags can, indeed, be matched to specific locations in the genome.

Brach Peters, then, used long SAGE to look in more depth at one specific human tissue to answer this question as to whether there might be more genes that could be detected using SAGE. And he chose human brain, and he chose human brain tissue because this is a highly complex tissue in terms of the transcripts that are expressed. We have seen this both by SAGE and, as well as other investigators have shown this. And he decided to analyze RNA from human brain that was highly purified; it was doubly poly A selected and was DNAase treated in order to avoid any possible contamination with genomic material.

And, finally, he used long SAGE to analyze over 660,000 total transcripts in order to identify the vast majority of transcripts, even those that would be expressed at a single transcript copy per cell. And analysis of this number of transcript tags using long SAGE identified the majority of the annotated genes that were present in the genome. And there were over 17,000 of these annotated genes that were detected out of the approximately 30,000 genes that are known to be present in the genome.

These known transcripts were distributed, of course, throughout the genome, and their expression ranged from less than one to over a thousand transcript molecules per cell and, in some cases, this can be seen in chromosome 19. The expression pattern was localized to a specific region within the chromosome. These are areas that have been recently termed ridges for the fact that they have increased levels of expression at those regions.

But, surprisingly, what Brach found was that, in addition to those tags matching annotated genes, a very large number of tags, 28,000, matched regions in the genome that did not correspond to known genes. Approximately half of these matched intronic regions within genes, suggesting that they encoded for unannotated exons of known genes while the other half, that is, about 15,800, corresponded to regions that were far away from known genes. This was surprising because these transcripts were located very far away from known genes. In fact, their median distance was 40KB and their average distance was over 180KB.

And Brach wondered whether these do, indeed, represent new genes, and so he independently measured their expression in several ways. One was by comparing them to EST databases, and he found that approximately half of these transcripts matched ESTs that were not included in the current gene annotations. These provided one level of evidence that these transcript tags are, indeed, expressed.

Together with Saurabh Saha, they next evaluated 129 of these novel transcripts by RTPCR and found that 123 of them were, indeed, expressed in an RT dependent fashion. And, finally, he looked at 15 of these novel transcripts in a variety of different tissues and found that 13 of these 15 were differentially expressed among different tissues. He suggested that these genes are not only expressed but may play different physiologic roles in these different tissues.

So these data were surprising to us because of the extent of transcripts that appeared to correspond to new genes. So, at least two questions emerged from these data. One is, how many genes are in the human genome? Now, of the known genes that have been annotated in the human genome, this SAGE analysis detected 17,000 of these known genes. In terms of the novel genes, over 15,000 distinct transcript tags were detected. If one looks at a window size of 15 kilobases in size, that is, the average size of a gene, to try to cluster tags that might be coming from the same gene, one still obtains over 11,000 different such clusters. Therefore, this study suggests that, rather than the 30,000 or so genes that the human genome is thought to contain, there are probably at least 45,000 to 53,000 different transcripts in the genome.

The second question from this study is, what do these genes do? And, of course, many years of additional work will be necessary to fully answer that question, but one thing that is already clear is that these genes are structurally different from known genes. For one, a small fraction of them are detected by gene prediction programs such as Genscan. Less than 20 percent of these transcript tags were found to be within a gene prediction program, while over 70 percent of known genes are routinely detected by these programs.

The second structural aspect that is different about these genes is that the GC content of the transcript tags is lower than for the known genes. It's 42 percent for the identified novel genes versus 47 percent for annotated genes. Interestingly, both of these features, the fact that these genes are now detected by programs and the lower GC content is very similar to features of non-protein coding *ab initio* transcripts that have been recently described in the literature.

So, in summary, what these long SAGE analyses have shown is that the draft genome analyses substantially underestimated the number of genes in the human genome. These studies have also confirmed the fact that experimental approaches must be used to identify new genes in the genome and also to convert gene predictions to *bona fide* genes. And, finally, these studies suggest that the human genome contains on the order of twice as many genes as have currently been annotated, and that these transcripts are structurally different from known genes and may predominantly encode non-protein coding transcripts. Of course, much work will be needed over the next decade or so to fully evaluate the function of these genes in both normal physiology, as well as disease states.

And finally, the paragraph at the bottom is a disclosure that we receive funding support from Genzyme for these SAGE studies.