

PRE-mRNA SPLICING: THE TIE THAT BINDS
Professor Joan Steitz
Yale University
24 and 26 May 2004

The title of my talk emphasises that pre-mRNA splicing is in fact a very central process in the pathway of gene expression. Let me go on to the next slide. I would like to start with a diagram that I have taken from a review article that was written by Tom Maniatis and Robin Reed several years ago. What it shows is all the molecular connections that had been uncovered at the time between splicing and other steps in gene expression. You see that splicing not only is connected to other steps in RNA processing like capping and polyadenylation, but it also talks to the various stages of transcription. We have known for a long time that splicing is necessary for export of the messenger RNA to the cytoplasm. That makes sense because, of course, you need to remove the introns before you send the mRNA out to be translated by the ribosome. A relatively recent realisation is that proteins that associate with the messenger RNA during the process or after the process of splicing are also very important for things that happen to it once it gets to the cytoplasm. Specifically, they play a role in surveying the message to see if it has any mistakes in it, and they appear to play a role in localising messages in particular parts of the cell for their translation. Just recently we have realised that even the efficiency of translation is linked somehow to the fact that introns have been removed from the messenger RNA in the nucleus. So since all of these things involve RNA-protein interactions, what we are really thinking about is interactions in the remodelling of RNA-protein interactions in a dynamic way, as we proceed through the pathway of gene expression.

I would like to start out with a couple of slides that illustrate the diversity of RNPs in the universe. Here is an RNP. These slides were taken in India a number of years ago, and this RNP is clearly very large and very utilitarian; you can think of it as being like a ribosome. There we also saw other RNPs. This one is smaller, much more exotic, but clearly the same sort of functional theme. We also saw small RNPs but more modern-looking and streamlined. In fact, this one dates from about the time that molecular biologists began to appreciate that there were a lot of different kinds of small RNPs in cells doing various things in gene expression.

My talk today is going to be divided into four parts. I would like to start out by indulging in history, and going back about 25 years to the time when we stumbled across the fact that there were these things called snRNPs (small nuclear ribonucleoproteins) in cells and that they were involved in the process of splicing. As you will see when I tell you the story, it is a very typical story in science in that there is a lot of serendipity involved. But it is also an atypical story in biomedical science in that it is not at all a bench to bedside story. It is more a bedside to bench story. It is a case, an unusual case, where tools provided by clinical medicine, namely autoantibodies from lupus patients, provided a way of beginning to dissect what was going on in the basic biology of higher cells.

After I talk about that, I want to go on and mention briefly two surprises connected with mRNA splicing that have happened subsequent to making the connection between snRNPs and splicing. At the end, I will tell you a current story from the lab about investigations that illustrate the previous slide, where you saw the many connections between splicing and other steps in gene expression.

Let me begin by saying a bit more about lupus. Systemic lupus erythematosus is perhaps the best known of a group of diseases that could be categorised as rheumatic diseases or autoimmune diseases. Other diseases in this group are things like scleroderma, mixed connective tissue disease, dermatomyositis, polymyositis, and some kinds of rheumatism. These diseases are not uncommon in the US. They afflict

about one in a thousand people, and I think the numbers are pretty much the same in the UK. They are more common in women than men, and they are more common in blacks than in whites. What all these diseases have in common is that in the sera of patients are circulating autoantibodies, antibodies against one's own cellular components. To start out here very simply, everybody knows that the immune system makes protein molecules called antibodies, which are designed to defend us against foreign invaders like bacteria, viruses and sometimes cancer cells—which can have on their surfaces things that appear foreign.

Our immune systems learn very early in life, and this is currently being worked out, how to discriminate between self and foreign. Sometimes this discrimination goes awry and people begin to make antibodies against their own cellular components, autoantibodies. If you have autoantibodies, the problems, the diseases, the pathogenesis are not caused by the autoantibodies getting into cells in high enough numbers to interfere with the normal functioning of their cellular targets, but rather because cells are lysing and dumping their contents into the blood stream. What then form are immune complexes; these build up and cause various problems. For instance, when they lodge in the fine capillaries they cause the red rash from which lupus got its name, or they can lodge in your hair follicles and make your hair fall out, or in your joints and give you joint problems. They affect all the internal organs and cause inflammatory responses. So these are really systemic diseases, and they are not nice diseases. One of the most interesting aspects of autoimmunity is that the cellular components that tend to be targeted, and this isn't understood at all, are components of cells which are very abundant and very highly conserved, namely the components which are involved in the central dogma of molecular biology. What we are looking at here is bacterial DNA being transcribed to make RNA, and ribosomes attaching to translate the RNA into protein. Many lupus patients, the majority, make antibodies against DNA, anti-DNA antibodies. Some make antibodies against ribosomes, and many make antibodies against snRNPs. But snRNPs, small nuclear ribonucleoproteins, aren't in this picture because this is a picture of the central dogma in bacteria, not in higher cells. SnRNPs are specific to higher cells.

At this point I need to go back, to tell you about snRNPs, to the fact that as a graduate student, as a postdoc and as a beginning faculty member at Yale, I worked on RNA structure and function, but all in bacteria and its phages. When our first sabbatical leave from Yale was coming up in 1976/77, it was a good time to start thinking about something new. A lot of people at that time were turning their attention from working things out in bacteria to thinking about how they happened in higher cells. Everybody's presumption was that the basic features of gene expression were going to be the same, just more complicated in higher cells. As you all know, this has turned out not to be the case. I decided I would like to jump on the bandwagon, but I also didn't want to stray too far from RNA. The problem I decided to investigate is illustrated here. Again we are looking at DNA, this time from a eukaryotic organism. The RNA transcripts are decorated not with ribosomes, because this is all happening in the nucleus and ribosomes are in the cytoplasm, but rather with RNA-binding proteins. The curious fact had emerged that in higher cells there was a huge turnover, a huge wastage, of RNA. A lot of the RNA that was synthesised in the nucleus simply got decayed and maybe only in the order of 10% of it ever made it out to the cytoplasm to become messenger RNA. I thought if the nuclear RNA gets decorated with proteins and a little bit more was known about these proteins, maybe the proteins that bind are deciding which of the RNAs survive and which of the RNAs get decayed. I decided it would be awfully nice to have antibodies against these RNA-binding proteins. We were in the lab of Klaus Weber and Mary Osborn in Göttingen, where they are good at making antibodies. So I spent seven months isolating these proteins, injecting them into various animals trying to get antibodies that could be used as tools to study this phenomenon. But I completely failed

because these proteins are very highly conserved and very non-immunogenic. Thus, I ended up doing something else for the rest of my sabbatical. Of course, 1977 was the year that evidence came together from labs in all different parts of the world that told us that our genes are quite different from the genes in bacteria. Namely, they are interrupted by introns, bits of apparent junk in the genes. In order for expression to occur, the RNA that is made has to get spliced, removing these junk regions. So, that did a lot towards explaining the huge RNA wastage, because the introns are usually much larger than the exons. On the other hand, it raised the question of what could be the cellular machinery that would very precisely remove the introns and join the exons back together so that the message would be able to be read and read absolutely in frame so that it could be translated into proteins.

When I returned to Yale in the fall of 1977, since we were an RNA lab, everybody was very excited about working on splicing. However, I admit we were not too clear about what we should do and where we should go. Then, the first bit of serendipity came in January of 1978, when a new issue of the journal *Nature* arrived, and at the back was an article with an obscure title. I have underlined the salient sentence here that says: "Patients with MCTD have high titres of antibody to nuclear ribonucleoprotein (RNP) which also gives a nuclear speckled pattern on cell substrates in direct immunofluorescence." The reason that this caught my eye is because when I was trying to make antibodies and was failing, several people had said that they had heard of some diseases where patients made antibodies against something that was nuclear and had RNA and protein in it. But at that time I didn't know any clinicians to ask about how I could get such patient antibodies. When the article arrived, I had a new MD/PHD student in the lab named Michael Lerner. He had just been to all his medical school courses. I asked Michael, "Do you know anybody here at Yale that might have patients with MCTD?" He said, "Sure, I'll go and see Hardin." Hardin turned out to be John Hardin, head of the Rheumatology Section in the Department of Medicine. Michael went across the street and that very afternoon came back with a couple of vials of sera from patients with lupus and other related diseases. We began to work with them.

I want to inject here a rather sobering thought. What we did was possible 25 years ago. But if that happened today, before you can even think about working with human materials, you have to fill out all sorts of forms and go through all sorts of committees. We probably never would have proceeded, but we did at that time. We started working with the autoantibodies and very quickly found out they were not directed against the large RNPs that I had tried to make antibodies against previously. However, since the antibodies were targeted against something that was small, was very abundant, and was very highly conserved, we decided to keep working. It was very frustrating. For about a year Michael kept trying to fractionate the antigen and it kept disappearing. We later realised that RNase was chewing up the RNA component.

The second piece of serendipity was when Joan Brugge came to Yale to give a seminar and talked about a new reagent called Pansorbin that had just come on the market. It was basically a preparation of *Staphylococcus aureus* cell walls, which has a protein in it called protein A, which binds to the constant region of antibodies. She was using Pansorbin with S35-labelled extracts of virus-infected cells to which there were antibodies to pull out the immune complexes and examine what proteins were there. So what we decided to do, as you see fractionated here, was to label HeLa cells with P32, which labels all small RNAs from tRNA size (about 70 nucleotides) up to the size of U2, less than 200 nucleotides, and use the same trick to try to pull out these complexes. In this lane what you see is Michael's own serum; happily he didn't have any autoantibodies which precipitated any RNA-protein complexes. But with the various patient sera that we had accumulated, we saw very distinct patterns. When we looked at the RNA molecules, we realised that one was a molecule called U1 and one was a molecule called U2, which had previously been characterised as small

nuclear RNAs in the labs of Busch and Weinberg and Penman. There were three other RNAs that we named U4, U5 and U6 in this lane. This is the RNP pattern, precipitating U1 RNA, that was mentioned in the *Nature* article. We knew from the medical literature that Sm was a related and overlapping autoantibody specificity. You can do the same experiment by labelling cells with S35 to look at the proteins. There was lots of evidence that the antibodies were directed not against the RNAs themselves but against proteins that bound to the RNAs. I am not going to show you the data but instead a cartoon that gives our conclusions. Namely, there were some proteins in common between the particles containing the U1 small nuclear RNA and the U2, and then there were some proteins that were specific to each of these particles. We then called these particles small nuclear ribonucleoproteins, snRNPs, or "snurps" for short.

If you possess antibodies against something and you want to try to figure out what its function might be, there are several things you can do. We immediately realised that the sequence at the five prime end of the U1 RNA was complementary to the five prime ends of introns, whose sequences were beginning to be determined. So we suspected that perhaps that at least this particle might be involved in splicing. You can also use antibodies to localise things in cells. Here we see an autoantibody against ribosomes, showing that they occupy the cytoplasm and the nucleolus, which is the locus of ribosome biogenesis in cells.

With the anti-RNP serum directed against the U1 particle, you see the converse pattern. It tells you that the U1 particles are in the nucleoplasm, which, of course, is where the chromatin is and where pre-mRNA is being made and spliced, but not in the cytoplasm nor in the nucleoli. We also did a number of experiments where we put pre-mRNA substrates into an extract and allowed snRNPs to bind and asked by protection methods where did they bind. Were able to deduce that U1 indeed bound to five prime splice sites and U2 bound to what was at that time becoming recognised as the branch site, a very important locus (I'll get to that in a moment) for the splicing reaction. One of the nice things we were able to rationalise even just at the time that *in vitro* splicing systems became available, was that the minimum size of introns was dictated purely by putting the different snRNPs on the pre-messenger RNA. This slide shows how much of an intron is protected from digestion by the binding of each of the snRNPs. If you add them up it comes to about 65 nucleotides, which is in fact the smallest size of our introns. Making an intron even smaller means you can't get all the snRNPs on. So, obviously, you can't splice out an intron that is smaller.

What I have told you so far is that gene expression in higher cells, in contrast to bacteria, involves a whole other class of RNAs, the so-called snRNAs. They are part of the snRNPs that assemble to make up the spliceosomes. They are, of course, necessary for the exact removal of introns from the pre-messenger RNA before you send the message out to the cytoplasm to be translated.

The next slide is a picture of the people in my lab at that time. This is Michael Lerner, whom I particularly wanted to point out. On the day this picture was taken, we were being visited by my colleague Sid Altman of catalytic RNA fame. That was because for a while we thought maybe the U2 snRNP might be his RNase P activity in higher cells. We had no idea initially what U2 was doing. It later turned out, of course, to be involved in splicing. So let me flash forward a bit.

Today we have beautiful pictures of the spliceosome in action. This is from Ann Beyer's lab in Virginia. We see here *Drosophila* chromatin being transcribed into RNA molecules and particles building up at the five prime and three prime splice sites. Here is an assembled spliceosome with the intron looped out. Thus, what one pictures from test tube experiments is in fact happening when visualised in the electron microscope. Work from many different labs achieved *in vitro* splicing reactions, from which we put together a picture of what is happening during the spliceosome cycle. It starts with recognition of the three prime splice site by protein factors. There are

probably on the order of 100 different protein factors involved in splicing, in addition to the proteins which are already tightly bound to the snRNAs. After the U1 and U2 particles bind, the U4/U5/U6 snRNP joins. It is a trisnRNP. The next step is the nucleophilic attack of the two prime hydroxyl of the branch-site A residue on the five prime splice site to form the lariat intermediate. Then, in the second step, the two exons are ligated, and the intron is degraded and the snRNPs recycled.

I want to point out that by the time of the first step, already the U1 and U4 snRNPs are less tightly associated. Thus, the focus is on the U2, U5 and U6 particles as perhaps being part of the catalytic machinery of the spliceosome. Also, I want to emphasise that ATP hydrolysis is needed for both steps of splicing. This is not because the phosphates enter either the intermediates or the products, but because energy is needed for all the dynamic changes that take place during the process of splicing and probably also for the fidelity of the splicing process. What is going on catalytically hasn't been proven, but everyone in the field suspects that a spliceosome conducts RNA catalysis assisted by proteins. Part of that comes from looking in a number of different organisms at the structures of the snRNAs that are at the core of the spliceosome. Although the snRNAs can vary in length and in secondary structure, each one has a short region that is almost absolutely conserved from yeast to man. Those regions have been localised as being right where the action is in the spliceosome during either the first or the second step. For instance, I have already mentioned that the five prime end of the intron base pairs with the five prime end of U1 snRNA, which is conserved from yeast to man. There is a region in U2 that base pairs with the branch site, bulging out the branch-site A residue. That is also very highly conserved. A loop in U5 is also very highly conserved.

It has been shown in a number of experiments that it helps align the two exons for the second ligation step of the reaction. Finally, there is a sequence in U6 that is very highly conserved that replaces the five prime end of U1 at the five prime splice site before the catalytic steps occur. U6 is the most highly conserved snRNA, contributing to the belief that it is part of the catalytic core. U4, which you see paired to U6, is a sort of chaperone that brings U6 into the spliceosome. It is then released and U6 refolds, this time associating with U2. The remarkable thing about the spliceosome I have described to you is its versatility. It can splice introns of greatly different lengths—huge ones as well as small ones. It somehow manages to find the right splice sites in the pre-mRNA sequence. Also, it can splice premRNAs that have many different exons and introns. It manages to do this in an orderly fashion and also manages at different times of development or in different states of differentiation to combine the exons in different ways to give rise to alternative splicing.

This is one of the great current challenges in splicing— understanding how the spliceosome, under some conditions, ignores the existence of an exon and splices in a different way to give rise to a different protein isoform. Given this versatility, it turned out that it was a huge surprise—at least to me—that our cells and the cells of most other higher eukaryotes contain a second spliceosome. The second spliceosome is in much lower abundance. It is absolutely necessary to remove a subset of our introns. I want to tell you a little bit about it at this point.

The evidence that there might be another spliceosome first began to emerge in the early 1990s when databases were accumulating to the extent that people could line up the boundaries between exons and introns and look at the consensus sequences. What was realised that there were a few genes that instead of starting with the nearly invariant GU and ending with AG, like most introns do, appeared to start with AU and end with AC. At the DNA level this is ATAC, and so they were called ATAC introns for short. I list the first ones to be recognised here, and you can see that there is no particular theme holding their genes together. There would be just one of these introns in a gene; not very many of them, just a few. The ATAC intron wouldn't be either the first intron or the last intron, or the biggest intron or the smallest intron.

Just one of the introns would have these peculiar consensus sequences. It was Rick Padgett at the Cleveland Clinic who first pointed out that low abundance snRNPs of the same class as the splicing snRNPs might be involved. These had been discovered by a graduate student in my lab, Karen Montzka Wassarman, several years prior. She had called them U11 and U12, which turned out to be prescient because U11 is the U1 analogue and U12 is the U2 analogue. She could have named them the other way around, in which case it would have been hopelessly confusing. Padgett pointed out that like U1 base pairs to a five prime splice site and U2 base pairs to the branch site, there are sequences in U11 that could potentially base pair with the different ATAC five prime consensus sequence and in U12 that could potentially pair with the different branch point sequence. After that, a very talented postdoc in my lab, Woan-Yuh Tarn, managed to put together an *in vitro* splicing system that would in fact splice an intron containing the minor intron consensus sequences. She was able to show that U12 associates with the branch point by several means and that U11 is also part of the second spliceosome. At first our thinking was that since U6 was believed to be part of the catalytic machinery, and U4/5/6 were sort of a core particle, maybe you just used U11 and U12 to recognise the intron ends and then pulled in the core machinery. When Woan-Yuh tried to look for U4, U5 and U6 in the second spliceosome, she didn't find U4 or U6. It turned out that in fact there are low abundance counterparts. Here again we are talking about 1/100th of the amount. Two additional RNAs called U4atac and U6atac again can base pair with each other. U4atac brings U6atac into the spliceosome, thus contributing a corresponding function to the second spliceosome.

One of the very interesting things about U6atac from human cells is that its sequence is more divergent from the regular U6 than the yeast and human U6s are from each other. This is quite divergent, but you can draw the same sorts of structures. What then turned out to be the case was that one snRNP is in fact utilised by both spliceosomes. That is the U5 snRNP, but somehow the minor class introns are recognised by four distinct snRNPs. Many of the protein components of the two spliceosomes turn out to be the same. This has been established in Reinhard Luhrmann's lab. It is still a mystery as to why the two systems don't get mixed up, with the minor spliceosome on the order of 1/100th the abundance of the major class snRNPs.

One of the most pleasing things about the minor spliceosome has been what it tells us about models for the catalytic core of the spliceosome. Here you see a picture, derived mostly from work in Christine Guthrie's lab—genetic suppression experiments—suggesting that U6 and U2 come together to form an elbow-like structure. Nearby are sequences that base pair with the five prime splice site and with the branch site. In this way, the branch-site A residue is juxtaposed for attack on the five prime splice site during the first step of splicing. Although only some of the sequences are conserved, the fact that you can draw the same types of structures lends credence to the idea that the model might be more or less correct.

We now know that minor class splicing is not determined by the terminal intron dinucleotides. More often they are GU and AG, just like major class introns. Instead, it is the longer sequences at the five prime end and the branch site that are determining. We also know that these introns and the second spliceosome must be at least a billion years old because they are found both in higher plants and in us. About one in three hundred of our introns are of this second sort. There are some current eukaryotic species, the yeasts and worms, that don't have these introns. The phylogenetic tree suggests that they might have had them at one point, and then lost them over evolutionary time. So the biggest question (I just want to leave you with this about the second spliceosome) is where did it come from.

Here Phil Sharp and his colleagues have proposed an idea that I actually quite like: maybe there was a progenitor spliceosome that then diverged in separate lineages

into what we now know as the major spliceosome and the minor spliceosome. At some later point the two came together, fused and mixed. Because the requirements for the minor spliceosome are more definitive, what has been happening since was that we have converted many minor class introns into major class introns to be used by a spliceosome that is more flexible. That then could account for what we see in current-day genomes. Next you must ask, why do minor class introns continue to exist with their corresponding spliceosome. What we do know, because of experiments done by Abhi Patel and an undergraduate in the lab, is that these introns are spliced more slowly. You can increase protein production by exchanging the splice-site sequences and thereby switching the spliceosome the cell uses. Therefore, potentially minor class introns could be regulatory for the splicing process and for the pre-mRNAs in which they occur. We don't yet know.

Let me go on at this point and talk very briefly about small nucleolar RNPs, or snoRNPs. This will be an introduction to the last little story I want to tell you about the connections between splicing and other steps in gene expression. We need to go on an excursion to the nucleolus. What you see here is just a part of the nucleus. The nucleolus is not membrane bound but rather where the repeated genes encoding ribosomal RNA are collected together and transcribed by RNA polymerase one into precursor rRNA molecules. Processing then happens in the fibrillar component, and in this larger granular component of nucleolus, the ribosomal proteins that have been made in the cytoplasm come in and assemble together with the newly synthesised ribosomal RNAs. The subunits then go back out to the cytoplasm to make proteins. One way of looking at the nucleolus is by where it's not. That is what you see here: again we are looking at where the splicing snRNPs are and you see them in the nucleoplasm and also the Cajal bodies that Angus Lamond has done so much lovely work on. They are not in the nucleoli. The same cells are here stained with anti-fibrillarin antibodies. Fibrillarin is an abundant protein component of the fibrillar parts of the nucleolus. As you shall see in a moment, fibrillarin turns out to be a nucleolar (sno)RNP protein. Thus, you see the nucleoli and also slightly the Cajal bodies. The main business in the nucleolus, as I have already mentioned, is to make a pre-ribosomal RNA. About half of the sequences in the pre-rRNA get thrown away in the process of releasing the mature RNAs, which are found in ribosomes. One of the things that happens very early to the pre-rRNA is that a number of nucleotide modifications are introduced. As you see here, they are introduced not into the spacer regions that get thrown away, but only into the parts of the prerRNA that are going to become mature molecules. As indicated, there are two sorts of modifications: two prime O-methyl groups and pseudo U groups. You'll see the latter in just a moment. It turns out that for each and every one of these modifications, the position is guided by a small RNA—a small nucleolar RNA—in the form of an RNP. Because there are so many of these, there is in fact a huge machinery in our nucleoli that is designed to put in these rRNA modifications. We are still struggling with what the modifications actually do for the ribosome, but the machinery is large.

There are two classes of small nucleolar RNAs (snoRNAs) and small RNPs (snoRNPs): one for introducing two prime O-methyl groups and one for the pseudo U groups. I am going to be telling you a bit more about the former class. These small RNAs have conserved sequence boxes called C and D, and internal copies that are similar called C prime and D prime. What happens is that the ribosomal RNA can base pair, usually perfectly, for 10–20 nucleotides upstream of either Box D or Box D prime. If you count 5 base pairs along this helix, that is where the methyl group is introduced into the ribosomal RNA. The two prime Omethylase turns out to be fibrillarin, which I already mentioned. It is an autoantigen that is often the target of autoantibodies in scleroderma patients. The most remarkable thing, however, about snoRNAs is how they are encoded in our genomes. It turns out that they are not in their own transcription units. Rather, they are encoded within introns, usually of protein-coding

genes, as you see here. Instead of introns being all junk, there are some fragments of introns that are in fact released, go to the nucleolus with bound proteins and have a second life. They do something in the nucleolus, namely guide the modification of ribosomal RNA. You see here that the intron is spliced out, is de-branched, and exonucleases chew from both ends as the snoRNA assembles together with proteins. At the time this slide was made (it is an old slide), we didn't know what the snoRNA function was. We now know that U15 is one of the RNAs that guides two prime O-methylation of a particular site in the large ribosomal RNA. Most of the snoRNA genes, as I mentioned, hide in the introns of protein-coding genes. Interestingly, they tend to be genes for ribosomal proteins, as you just saw, for translation factors, or for nucleolar proteins— all things that have something to do with protein synthesis or the biogenesis of the protein synthesis apparatus. That makes sense. But there are some very unusual genes in our genomes where almost every single intron has a snoRNA in it. When these genes are spliced, the spliced RNA doesn't have a long open-reading frame. It doesn't appear to be translated, but just very rapidly degraded. So these are sort of inside-out genes, where it's the intron pieces that are long lived, and splicing seems only to be a device for releasing the introns so that they can be processed into snoRNAs. As far as we know the spliced exons don't have any function in cells except to turn over.

With that I would like to move to the last part of my talk, where I want to talk about one particular connection between splicing and another step in gene expression that we have been working on recently. That is the assembly of snoRNAs with proteins into snoRNPs. Since the snoRNA sequences reside within introns, one might expect that this assembly process would be coupled to splicing. What I will be telling you about is the work of Tetsuro Hirose, who was a Human Frontiers-supported postdoc in the lab and recently went back to Japan, and Mei-Di Shu, a technician who worked with him. I will be telling you about the splicing-dependent assembly of intron-encoded Box C/D snoRNPs. There is also a small component of splicing-independent assembly that I shall mention.

First, I need to tell you a little bit more about the protein composition of the particles in the Box C/D snoRNP class. I already mentioned fibrillarin. Crosslinking experiments were done by a graduate student, Niamh Cahill, who introduced 4-thioU residues into each of the U positions in the conserved boxes, which were believed to bind proteins. From her work, we think there are probably two molecules of fibrillarin, which makes sense because there are potentially two sites that guide two prime O-methylation. There are two proteins, Nop56 and Nop58, which are related, but non-interchangeable, that bind to different distinct places. Both Niamh's work and the work of Lara Szewczak, a postdoc who used nucleotide analogue interference mapping to study RNA functional groups required for the assembly of snoRNP particles, came up with evidence that the 15.5 kDa protein is binding to the terminal stem. That conclusion was also reached by several other labs, notably Nick Watkins in Reinhard Luhrmann's lab. I also need to tell you a little bit about the terminal stem structure, which some of you will know a lot about as it has been worked on in David Lilley's lab. It is believed that the terminal stem forms a new RNA motif, a newly-recognised RNA motif, called a kink turn. We have crystallographic data on what this RNA motif looks like from structures of the ribosomal RNA in the ribosome and also from Luhrmann's lab, where the 15.5kDa protein has been co-crystallised with a piece of the U4 snRNA that forms this structure. What's seen in these structures are two base-paired regions and then two sheared GA base pairs, which extrude a nucleotide that makes very close contact with the 15.5 kDa protein. Although we don't have structural evidence for the Box C/D snoRNAs, they do have the potential for forming the sheared GA base pairs. Lara's evidence indicates that this particular U residue assumes a very unusual geometry within the structure.

The question that I posed is, when do the snoRNP proteins, and particularly the 15.5 kDa protein, initiate the assembly of the snoRNP particle? Does it occur before the intron is spliced? Does it occur at some particular step during the splicing process? Does it occur after the intron has been released and debranched? What we did know is that assembly has to occur before the exonucleases go to work because, at least for most of the snoRNP proteins, if they are not present the snoRNA disappears. It just gets completely chewed up. The first indication that Tetsuro Hirose had that the two processes of assembly and splicing might be mechanistically linked came when he examined about 60 sequences of introns containing snoRNAs from the human genome and plotted where the snoRNA sat within the intron relative to the five prime and three prime splice sites. What you see here is that they very much prefer to sit about 70-80 nucleotides upstream from the three prime splice site. The distance from the five prime splice site is much more variable. This suggested that there was something important in the positioning. What Tets first did were deletion experiments to look at snoRNA production in transfected cells. He concluded that there was no specific sequence required but that it was the distance between the snoRNA and the branchpoint A residue that was critical—not the distance between the snoRNA and the three prime splice site. That made sense in terms of what we know about splicing. If you try moving the snoRNA closer to the branch point, you simply don't see it ever being released and assembled. If you move it farther away, the efficiency of synthesis simply just drops off. On the other hand, the snoRNA doesn't have any effect on the splicing process. Again, all these data indicated some sort of synergy between the splicing process and snoRNP assembly.

To look at this further, Tets set up an *in vitro* system where he could use a splicing substrate with an intron and snoRNA within the intron. He got both the splicing, the snoRNA trimming process and assembly to work in a test tube. Here we see the pre-mRNA and the lariat product, the spliced intron product, and the intermediates—the two-thirds lariat and the excised five prime exon. Here we see that the snoRNA has extended sequences at its three prime end and gets trimmed down to the mature-sized snoRNA. Again the requirements for Boxes C and D and for the correct spacing were all apparent in the *in vitro* system, as in the *in vivo* system. One of the really nice things about lots of work having been done on *in vitro* splicing is that at this point we know much about the various stages of spliceosome assembly and function. We also have tools with which we can block the process at each of the various stages. As illustrated here, one of the ways of doing this is to use short RNAs, two prime O-methyl oligonucleotides, that interact with a snRNP RNA and block its action at a specific step. For instance, one oligonucleotide blocks the attachment of the U2 snRNP to the branch point. Another blocks the replacement of the U1 snRNP by the U6 snRNP. Yet another blocks the rearrangements that take place when the U2 and the U6 RNAs come together to form the catalytic core. Finally, it turns out that if you replace the AG at the three prime splice site, you block splicing after the first step and before the second step. The idea then was to block the spliceosome at these various stages and ask when do we see that the snoRNP proteins have assembled. Are they assembling at a particular step? I am going to show you just one piece of data in this slide from such an experiment.

What we are looking at here is immunoprecipitation with anti-fibrillarin antibodies. In an unblocked reaction, you can see association of fibrillarin with several different intermediates. But if you block between the first and second step of splicing, you see build-up of the two-thirds lariat intermediate. It is precipitated by anti-fibrillarin, antibodies against one of the snoRNP proteins. If the boxes in the snoRNA sequence are disrupted, you don't see that, so the results correspond to the features you would expect. Knowing that the two-thirds lariat is in fact already binding snoRNP proteins, you can ask whether something about the architecture of this particular intermediate is important. Or, does the lariat have to be generated during the splicing

reaction in order to get the snoRNP proteins assembled? What I am going to show you in my next slide is an experiment where Tets ran a splicing reaction, cut out the two-thirds lariat intermediate, and simply put it back into a splicing reaction. He asked whether it would pick up snoRNP proteins. Here we are looking at a different substrate, so things are running differently and we are using a tagged 15.5 kDa protein to do the immunoprecipitation. What you see is that if you throw the two-thirds lariat intermediate back into the reaction, it is not precipitated by association of the 15.5 kDa protein. However, if the lariat is generated during the course of the reaction, it is. This says something active is happening during the splicing process.

So, the conclusions to this point are that assembly of the snoRNP (and there are lots of data that I haven't shown you) does seem to occur at a particular step in the splicing reaction. If the snoRNA is too close to the branch site, assembly can't occur properly. Thus, you never get the snoRNP assembled and released, again arguing that there is synergy. How does it occur?

One possibility is that the spliceosome serves as a chaperone and helps the formation of the kink turn so that the 15.5 kDa protein can recognise it and seed the assembly of the rest of the snoRNP proteins. Other possibilities are that there are direct protein interactions between the 15.5 kDa protein and something in the spliceosome. There are actually two versions of this possibility. One is that something interacts and actually deposits the 15.5 kDa protein on the snoRNA at this particular stage of spliceosome function.

Another possibility is that only at a particular stage of splicing is the intron cleared of non-specific RNA-binding proteins that prevent the snoRNP proteins from getting on. Only at that point do the sequences at the termini become available for the binding of the 15.5 kDa protein. At the end, I want to quickly mention an alternative mode of snoRNP assembly. If I have convinced you that splicing is necessary for snoRNP assembly, some of you will have noticed in the graph I showed you earlier that there are some snoRNAs which sit very far away from the three prime splice site. So how do they get assembled? What Tets realised when he looked at one of our favourite multi-snoRNA host genes was that the snoRNAs that sit at the optimal distance have short terminal stems. In contrast, for the one that is far away, he could at least draw a longer stem in the vicinity. That also seemed to be the case for other snoRNAs which are located far from the three prime splice site of their host introns. In order to test this hypothesis, what one wants to do is to destabilise the stems. Various mutants were made that progressively destabilise the stem. One also wants to be able to move the snoRNA from its distant position to the optimal position and then ask what happens. Here are some *in vivo* experiments. What you see is that in the distant position, as you lower the stability of the stem, the efficiency of production of the snoRNA drops off. If the snoRNA is instead in the close position, the optimal position, the stability of the stem doesn't matter too much until it gets to be quite unstable. We think there are other reasons why this particular snoRNA is not well expressed. Finally, the "nail in the coffin" experiment to ask whether this idea is right or not, is to take a snoRNA that is in the optimal position and move it far away. You expect to see its efficiency of production drop, but then by adding a stem you expect to recover it. The final bit of data shows that is in fact what happens. If you move the snoRNA to a distant position, its efficiency of synthesis drops way down, whereas if you include a long stem in the flanking sequences, you can up the efficiency of production to a pretty good level.

What I have told you would suggest that for most snoRNAs that are located at an optimal distance, there really is a mechanistic link between splicing of the host intron and the assembly of the snoRNP. For those that are located far away, having an external stem that perhaps helps the kink turn to form so that the 15.5 kDa protein can bind, enables assembly. A somewhat more colourful version of the story is shown

here, with splicing-dependent assembly occurring at the C1 stage of the spliceosome reaction, and independent assembly occurring earlier. I would like to point out that at about the time we found this, there came a beautiful paper from Angus Lamond's lab which talked about the unusual trafficking of the 15.5 kDa protein to the nucleolus. Namely, it goes through the nucleus by transiting through speckles. Getting to the nucleolus moreover was dependent on RNA polymerase two transcription. Of course, this all fits very nicely with the idea that the 15.5 kDa protein is getting on to the snoRNA co-transcriptionally and co-splicing. It then moves from the speckles, where the snoRNP has been mostly assembled, to the nucleolus for the snoRNP to function. I am looking forward to discussing more of this with Angus tomorrow.

To give credit for the data that I have shown you, I again would like to thank Tetsuro Hirose and Mei-Di Shu and Human Frontiers, as well as HHMI and NIH. I would like again to return to the slide I started with, for there are lots more challenges here. There are many more interactions that need to be understood on the molecular level. A lot of fascinating biology is going to come out of such studies. What I find so remarkable is that a process that doesn't even exist in bacteria, the splicing process, has become so central to the whole gene expression pathway in our cells. I will leave you with that idea, and remind you that we started a long time ago with patients who had autoantibodies with various specificities against proteins that interact with small RNAs to form RNPs in cells. I want to end by saying thank you not just to the people that I mentioned specifically, but to all the wonderful students and postdocs that I have had in the lab over the years. They were the ones that made this story possible. Here are a few of them that turned up at a recent Halloween party. Each one of them is a different snRNP, and you see they are connected with exons. They have five prime caps and three prime polyA tails. The whole story is there.