

OK. Parts of a gene. We have our promoter, which is part of the untranscribed region of a gene, usually in the 5 prime end. Not always but for the genes we're talking about at the 5 prime end, the so-called 5 prime end of the gene, or so-called upstream of this transcribed region. And downstream of that there is more untranscribed region that interestingly can also contribute to the promoter, even though it's far away from this more upstream part of the promoter. But I'm going to call it just for now untranscribed, two flanking regions of untranscribed DNA sequence and one region of transcribed sequence. Now, I want to discuss with you very briefly a phenomenon called splicing. And this is a phenomenon that occurs within the RNA that is transcribed from a gene and, therefore, pertains to the transcribed region of the gene. It turns out that in this transcribed region there are two kinds of sequences. There are things called exons and there are regions called introns. The exons code for something, code for the final function of the RNA or for eventually a protein. So these are coding. The introns are noncoding. Both of them are transcribed. You'll see this definition is a little loose as we move on in today's lecture, but it's good enough. In the transcript that initially is made from the gene in this transcribed region, both introns and exons are present. So these are present in what's called the primary transcript or primary RNA. And primary refers to the first RNA that is transcribed from the gene. And subsequent to that, still in the nucleus, those introns and exons are subject to a process called splicing whereby the introns are removed -- -- or spliced out is the term, such that in your mature RNA only the exons are present. This process is likely a consequence. I'm going to put up a diagram that you had on your last time's handout. You can watch it now. And you can refer back to a previous lecture if you don't have it with you. This notion of introns and exons is probably a consequence of evolution whereby different parts of genes were combined and shuffled to give new kinds of genes and, therefore, new kinds of proteins. Here on my diagram I have exons in black and introns in blue, and they're all just DNA sequence, but when the RNA is transcribed in the first place is primary RNA. It's a copy of the gene. It has both exons and introns. And then a very complex enzymatic machinery comes on and it loops out and excises these introns. OK? So this is very interesting such that in your mature mRNA there are no introns. And the introns have been looped out and they form these little structures that are called lariats. And at this point your mRNA is mature -- -- and it moves to the cytoplasm. Now, this process what discovered by Professor Phillip Sharp here at MIT and he got the Nobel Prize for it in 1993. It's a very important process because it's absolutely required for maturation of RNAs. And also, and I'll come to this in a few lecture's time, it allows different proteins to be made from the same mRNA. So here's a rule. In this RNA there are what are called splice donor sites that I've put as a circle and splice acceptor sites that I've put as a square. Just watch this. Just watch this for now because we will come back to it. So watch what I'm saying rather than trying to madly write down. Any splice donor can join to any splice acceptor and remove the stuff between them. So in this top example I've got each introns being neatly removed because splice donors and inceptors interact. But look at the example below. I've got this splice donor next to exon one interacting with a spliced acceptor next to exon three. And when that happens you remove the hull of exon two. So you actually are going to make a different protein. Whereas, in the first case you'll have exons one, two and three and four. In the second case you'll have exons one, three and four. OK? So this process is very important for allowing different kinds of proteins to be made from the same gene. I want to make you aware of this now, and I will come back to it in the formation module when we talk about how different kinds of cells are generated. All right. So let's move onto the major topic of today's lecture -- -- which takes us back to the central dogma. And I want to introduce to you a term that is very important that you know and you understand. And this is the term gene expression. And really what we've been talking about is gene expression. Gene expression simply refers to the generation of the final product of a gene from the gene. So we're talking about the formation of a protein as directed by a particular gene. OK? So gene expression is, if you like, the readout. Here's another way of putting it. The readout, the final readout of a gene, or the generation of the final product of a gene. I'm going to come back to this term over and over again, and I will ask you to define it in your own way, but it's a term I want to throw out at you now because you do need to know it. It's very pervasive. Today I want to talk about the step in gene expression or translation whereby RNA is converted or is used to direct synthesis of a protein. So let's define translation because it is, I think, one of the most interesting questions in molecular biology. Certainly from a historical perspective that was true. And the notion in translation is that the base sequence of a mRNA somehow leads to the synthesis of a protein with a defined amino acid sequence. Now, if you think about DNA replication, transcription and translation, the relationship between them, there is a nice analogy that one can make. DNA uses the base code, four bases. Transcription RNA uses those same four bases as a code, but it's slightly different from DNA. So the synthesis of RNA using a DNA template is kind of like changing fonts in a document that you have. It's kind of like going from Times New Roman to Helvetica. You haven't really changed much. It just looks a bit different. Translation is very different. The use of mRNA to direct the synthesis of a protein is much more analogous to changing language where you've taken English and translated it into Chinese or Russian and translated it into French. OK? So this is a really different process. And it was clear from the outset, historically, that one had to think in a slightly different way about how this process was directed. And I want to talk about four things with respect to translation. Firstly, I want to talk about the genetic code that allows RNA to direct protein synthesis. I want to talk about something called the interpreter of that code. I'm going to talk about the factory in which the synthesis takes place. And then I'm going to get to a discussion of the molecule bases for genotype and phenotype. So let's think about the code. And thinking about this starts from a very simple logical place. And the place is this. One starts with four bases, A, G, C and T or A, G, C and U, depending if you're talking about DNA and RNA. And somehow those four bases have to be used in some kind of code to give you an outcome of 20 amino acids. And I am going to use the abbreviation AA for amino acids. So you can look at this and immediately understand there has to be some kind of combinatorial code in order to specify those 20 amino acids. So you can do combinations and you can save. OK. If two bases were used and you could have combinations of doublets. how

many combinations can you get to and would that be enough to specify those 20 amino acids? Well, no, because two base combinations would only give you 16 possible amino acid combinations, or the ability to specify 16 amino acids. OK? Four squared. How about three base combinations? Well, that's better. What you can get out of that is 64 different combinations. OK? And that is plenty to specify your 20 amino acids with some left over. And, in fact, this is what is used. Combinations of three bases. And these combinations of three bases are termed the triplet code. The discovery of the triplet code is really fascinating. I don't have time to go into it in this lecture, but your book is not too bad on the discovery. And I will post on your website, for those of you who really want to get into it, a reference to a very interesting historical account of the discovery of the triplet code and indeed of much of molecular biology. But it's a fascinating story. But I'm going to tell you the code is a triplet code. OK. So what does that mean? It means that three bases correspond to a particular amino acid. OK? So one triplet of bases correspond, I'm writing this out because it's really important that you know this, correspond to one amino acid. And this base triplet gets a special name. It's called a codon. And the thing that you will have noticed is that what I've told you is there are 64 possible combinations of triplets and only 20 amino acids. And so that leaves some over. What happens? Well, they're all used. And what happens is that although the code is universal, as far as we know it arose just once, all living organisms on our planet use this code, it is a redundant code. So I will write down it is redundant but not ambiguous, and tell you what that means. So what that means is that an amino acid can be specified by more than one triplet, and I'll show you that in a moment, but that any triplet of bases only corresponds to one amino acid. Let's look at some diagrams to show you what I mean. This is a table of your amino acid code. These letters in columns represent the bases. And next to them are written the amino acids that correspond to this particular code. Let's start with an easy one. This is methionine encoded by AUG. And that's one you should actually remember. OK? And for methionine there is only one possible codon. It is AUG and always AUG. But let's keep going here. And let's look at the amino acid lucine. lucine is encoded by six possible triplets, six possible codons, UAA, UAG, CUU, CUC, CUA and CUG. Any one of those in a mRNA can encode lucine. However, CUU only encodes lucine. It never encodes another amino acid. OK? And that's what I mean by redundant. More than one triplet can encode one amino acid, but any given triplet only corresponds to one particular amino acid. OK. You will have practice on this kind of thing as you go along. So let's get some basics down here. The template in the whole translation process is your mRNA. OK? It's the code. It contains the code. It is read to give a protein readout from 5 prime to 3 prime. And the readout of the protein, as I mentioned to you way back when, reads out from the amino to the carboxyl end. New amino acids are added onto the carboxyl end -- -- and the free amino group corresponds to the first amino acid polymerized. So it is read 5 prime to 3 prime, and that corresponds to the amino to the carboxy growth of the protein. All mRNAs start with the same amino acid, and that is methionine. And the start or initiation codon in all proteins is methionine, oops, is AUG which encodes methionine. Now, not all final proteins have got methionine at their amino ends because it can be cleaved off. OK? So you don't have to land up with a protein that has a methionine end, its amino end, but it starts off with methionine there. And then there are no gaps in the message. It is read without any punctuation marks, except for the fact that the codons are next to one another in a non-overlapping way. OK? So there are no gaps. And the only punctuation is the start codon and a series of stop codons which do not encode any amino acids. These are UAA, UAG and UGA. And you can remember them if you want, but we're not going to test that you do. OK? You can use your amino acid tables. OK. So your punctuation is the start and the end of the message. All right. So let's go on and talk about the interpreter and what I mean by the interpreter. In this diagram here I have got, look up here for a moment. This is quite a nice diagram not from your book. I've got your DNA strand, which is your template strand. Your corresponding RNA, your mRNA, and the readout of the RNA to the protein. And here are the codons, UGG, this is in the middle of the protein so that's why there's no methionine, UGG corresponding to tryptophan, UUU corresponding to phenylalanine. You can see how the codons are right next to each other, OK, but do not overlap. In fact, I'm going to write that on the board. So no gaps and no codon overlap. Very important that you understand that. So when people looked to this and figured out what the codons corresponded to in terms of amino acids there was the question of, well, how do you actually get those amino acids corresponding to those codons? And there was a sense that you needed some kind of adapter or interpreter molecule that both recognized the codon and recognized the amino acid. And that's the next thing that I'm going to tell you. And -- -- stop. Well, I apologize on behalf of our illustrious institute for the boards in this room. OK. So all right. So let's talk about interpreter. And I'll tell you that this is the class of RNA someone brought up earlier called tRNAs. So tRNA, as you may recall, are these very small RNAs. There are about 100 base pairs, 100 bases in length, and there are a lot of them. And there is a tRNA that corresponds to every codon. So tRNAs recognize both the amino acid and the specific codon. And they recognize, let's talk about the codon first. They recognize the codon by DNA complement, by RNA complementarity, by base pairing to a region on the tRNA called the anti-codon. So let's talk about methionine for a moment. The codon for methionine is AUG. That's the codon. Woops. Hold on one second here. 5 prime AUG, that's your codon. And what will be complementary to that on the tRNA from the 3 prime end is UAC. OK? So this anti-codon is on the tRNA. Anti-codons can either be written from the 3 prime end or you can switch them around and talk about 5 prime CAU. It's the same thing. OK? So that's one thing. I'll show you a picture in a moment. The other thing that a tRNA has to recognize is the amino acid. And that's more complicated. For different amino acids there are different parts of the tRNA molecule that recognizes specific amino acids. And it hasn't actually been figured out completely which part of which tRNA recognizes a particular amino acid, but the recognition is also on the tRNA -- -- and not really on the anti-codon. Or certainly not the anti-codon alone is probably fair to say. So let me show you a picture of a tRNA. tRNAs are single-stranded RNAs that fold up on themselves in a complex way. OK? Here's the representation of the three-dimensional structure of a tRNA. And these cross things are hydrogen bonds. So there's a lot of base-pairing within the tRNA. Represented more simply, the tRNA forms this kind of cloverleaf structure, and the anti-codon is at one end of the tRNA. OK? So this is the thing that's base pairing to the codon and the mRNA. The amino acid attaches to the very 3 prime end of

the tRNA at this site which is a CCA. OK? And there is a covalent attachment of the tRNA to the amino acid at this CCA region. All right. But the part that recognizes the amino acid can be somewhere in the rest of the tRNA molecule. It's very complex. OK. So let's move on now. Actually, let me tell you one more thing, though I'll tell it to you in a moment. OK. So let's move on now to the question of the factory. And by factory I mean the place where protein synthesis or translation takes place. And the factory here is the ribosome. We mentioned ribosomes right at the beginning of the course in the second lecture and haven't said a whole bunch about them since. Ribosomes are very large structures. They are not membrane bound, but they are very large. This is a representation of a ribosome from bacteria that has a small subunit and a large subunit. And, interestingly, ribosomes are an obligatory complex between the so-called rRNA, or ribosomal RNA, plus proteins. There is a small subunit, this is really bad. Let's try this one. Small subunit which consists of one ribosomal RNA of a particular kind and 33 proteins. And there is a large subunit. And I tell you this not because you need to remember this, but you need to appreciate that this is a very complex structure. It's a very cool and complex structure. The large subunit comprises of three RNAs and 45 proteins. You can represent the structure of the ribosome much more beautifully in this diagram, or in this representation, where the RNA is shown in gold, or the two RNAs are shown in gold, or the multiple RNAs are shown in gold, and some of the proteins are shown as these other structures and you can see the alpha helices of the proteins. OK? And what you should be able to see on this diagram, let me point to this one for a change, is this tunnel, this hole through the structure. And this is the tunnel through which the mRNAs thread as it is translated. So this is truly a factory. tRNAs come into this, the mRNA threads through, and as that takes place so the mRNA directs the synthesis of the protein. OK. This is a representation from your book. I don't like most of the diagrams from your book so I redrew most of them for you, but I left this one. This is a representation of translation. The mRNA is shown in green and the large subunit and small subunit of the ribosome come together, form the complete ribosome, and then the mRNA actually is threaded through the ribosome and the protein, well, here they've called it a polypeptide chain is threaded through. So let's explore this in a bit more detail. And in order to do so, I've got to conserve boards here because we are one board short. In order to do so I need to introduce you to the various parts of a mRNA. And this is on one of the diagrams that I handed out today. OK? So you don't need to redraw it. Just look at the diagram. In the mRNA, and this is crucial for translation, there are three parts that are really important. Two of them, excuse me. Two of them are actually added to the mRNA after it is transcribed. The thing at the very 5 prime end called the cap and something at the very 3 prime end, which is a long string of up to a couple of hundred A residues contiguous, which is called the poly A tail. And these parts of the mRNA are crucial for the first part of translation which is initiation. As in replication and transcription, you can divide up these synthetic processes into different steps. And initiation is the first step. And in order for initiation to occur one needs the parts of the RNA that are added on post-transcriptionally. You need the cap, this poly A tail -- -- and also a region that is just upstream or 5 prime of this AUG initiated codon in a region called the UTR, the 5 prime UTR which stands for untranslated region. And you also need the AUG codon. OK. And what happens is that the ribosome and various initiation proteins bind to the 5 prime cap and simultaneously to the poly A tail. So this is really cool. The mRNA is translated as a circle where this poly A tail, the very 3 prime end is brought all the way around to the 5 prime end, and you get a whole mess of proteins sitting on that part of the RNA and starting translation. So you get initiation proteins, which are called initiation factors, and you get ribosome assembly where the small subunit and the large subunit come together, and you get a tRNA carrying a methionine amino acid coming and sitting on the AUG. OK. Let me show you more. So here we have a cartoon, you have this in front of you but I'm going to show it to you in a step-wise fashion, of this ribosome recognition sequence. Actually, I'm not going to show you now but in your handout there are pictures of the circular RNAs being translated. OK? That's something new and it's something very interesting. I'm not going to dwell on it now. OK? Where the poly A tail comes all the way around to that 5 prime so-called cap region. I should just point out, again, I'm not going to dwell on it, the so-called 5 prime cap region is a modified guanine. OK? MEG stands for methyl guanine. You can call it the cap. It designates the very 5 prime end of the message. OK. So let us look at the sequence of translation. And what I'm going to tell you, before I go through the cartoon, is that in the elongation process sequential tRNAs carrying their particular amino acids are going to come in. And they're going to sit on these various codons. And peptide bonds are going to form between adjacent amino acids so you get the polypeptide chain growing. OK? So let's start off with the initiator. There's your tRNA that is joined to methionine. And I need to introduce you to a term now which is a charged, I didn't have space before. The term charged tRNA refers to the tRNA covalently linked to its amino acid. And then correspondingly the uncharged tRNA has no amino acid. The amino acid has fallen off or been used. OK. So there is a tRNA sitting on the first codon, the AUG, and that's the start of the sentence. That positions the beginning of the protein. Now, watch what happens. Here comes another tRNA that corresponds to leucine, and you're getting base pairing here. That first tRNA is base paired to the AUG codon through its anti-codon. The second tRNA is base paired to the second codon through its anti-codon. And now you've got a methionine tRNA sitting next to a leucine tRNA. OK. Everyone with me here? And what happens now is that a peptide bond forms between the methionine and the leucine. In particular, this methionine is going to move over to that leucine over there and lead to uncharging of that particular tRNA. Take a look. OK, so I've shown you that methionine is going to form a peptide bond with the leucine. Now, watch what happens next. Here's the methionine tRNA. It's lost its amino acid, OK, so it falls off the message. It's done its thing. It's no longer needed. Along comes, no, sitting there is this leucine tRNA which is now covalently attached to its peptide bond to the methionine. And there's a free amino end here which designates the first amino acid synthesized in a polypeptide chain. And here comes in the next tRNA that corresponds to a serine tRNA base paired by its codon, base paired by its anti-codon to the codon on the mRNA. OK? And the same thing is going to happen again. The leucine and the methionine is going to be transferred over and make a peptide bond with the serine, and so you get elongation of the polypeptide chain. So what I'm going to write under elongation is that adjacent amino acids join. Uncharged tRNAs leave. are released. and sequentially new tRNAs

corresponding to codons come in. OK. All right. So this whole process goes on until the mRNA, until the ribosome and all these tRNAs reach a place in the mRNA where there is a codon that doesn't correspond to an amino acid. A so-called stop codon. And at this point there is a process called termination where there is a stop codon that does not code for any amino acid and doesn't have a corresponding tRNA therefore. And at this point the protein polypeptide chain falls off the message. All right. You guys OK with that? OK. I'm going to refer you, I'm not going to go and watch this movie. Go and watch this movie. Go and watch the movie by yourselves. OK? I don't want to take the time to watch it now. It's an animation of what I've just told you. There are some diagrams in your book. You can look at them. They talk about things called A sites and P sites in the ribosome. To me that is less important than you understand the actual interactions between the tRNAs and the mRNAs. Here is a circular RNA with that poly A tail and the 5 prime cap of binding proteins to initiate translation as a circular RNA. All right. So, finally, let's move to this complicated, I think fantastic bringing together of mutation from genotype to phenotype. You've had a genetics module where you talked about mutations, you talked about the genotype, you talked about the phenotype. We've been throwing at you genotype has got something to do with the DNA base sequence. Phenotype has got something to do with the final product, particularly the protein sequence. Let's explore that in a bit more detail now and ask, what is the molecular basis for changes in genotype and how do these correspond to changes in phenotype? OK. So genotype to phenotype. And I want to emphasize again that phenotype is an outcome of a change in function of the final product of a gene. It isn't necessarily the same as the final product of a gene. OK? So, for example, a phenotype is gigantism. Someone who is very tall. The molecular basis for that could be multiple things. It could be production of too much of a hormone, a protein called growth hormone so that someone grows too tall or taller than normal. OK? So that is the phenotype is connected to the production of a particular protein that's not the same as. So here's another diagram, something for you to think about. Mutations, almost anywhere in a gene, can have an effect on the protein produced. And there are two ways the protein produced can be affected. One is in the amount of protein and the other is in the sequence of the protein produced. Now, if one gets a mutation in this promoter region or often in the introns, but particularly the promoter I've focused on, one can change the amount of RNA that is being transcribed from a particular gene. And that change in the amount of RNA will lead to a change in the amount of protein. And you may get a phenotype because you're making too little or too much protein. Conversely, changes in exons can lead to changes in the actual sequence, the amino acid sequence of the protein and, therefore, to its function. So those are two important distinctions to make. OK? So mutations can change the amount or the sequence of a protein. I'm going to go through some examples of mutations, and you will go through more in Section, and you are expected to know these changes and you are expected to know how the change in DNA sequence may lead or not to the change in protein sequence. So look carefully. I'm not going to get through all my examples today. You can go and you can do the examples that are posted on your website. You'll get more practice. And you really need to know this. OK, so here's a wild type gene. The top two strands are the DNA. The bottom of the strands is the template strand. This DNA is transcribed into a mRNA and that is translated into the protein indicated here. OK? Let's look at an example of what happens when there is a change in the DNA. So here I've got a change in the DNA, OK, such that this particular base pair has been changed. The mRNA, oh, this is the wild type again. Here's your wild type sequence, wild type mRNA, wild type protein. This is a class of change in the DNA that is called a nonsense mutation. Watch carefully. So at this position that I've underlined, watch this. Don't try to write anything down. OK? You'll have plenty of practice. This is all posted. Just watch. At this particular underlying position, instead of a GC base pair there is now an AT base pair. And that changes this codon UGG into UAG. And UAG happens to be a stop codon. So here's your gene, your mutant gene, here's your mRNA that comes from the mutant gene, and here's the protein. It starts OK with a methionine. But, look, the next codon is a stop. OK? So the protein is truncated. Now, there are a number of classes of mutation. I am going to write these on the board. I'm going to ask you to go and read your handout carefully. And you will cover these in section. Again, you need to know them so let me list the types of mutation. In the interest of time, I'm not going to go through them, but you will be able to work through these examples both in Section and on your own. So, to end off, the mutations in exons that you should know are silent mutations that don't change the sequence of the protein, nonsense mutations that I've just covered, something called missense mutations which change the amino acid sequence, and something called frameshift mutations which also are likely to change the sequence of the amino acid. OK? As I say, this will be covered. If you want to come and see me personally in office hours tomorrow or the next day, please do, and I'll go through these examples with you.