Chapter 1

Discovered!
A New Regulator of Gene Activity

The beauty of nature lies in detail.

From Wonderful Life: The Burgess Shale and the Nature of History*, by Stephen Jay Gould (Penguin Books, 1989, p. 13)*

Humans – as is all life when you really think about it – are a wonder. Imagine briefly the range of activities that your body performs on a typical day. Your heart pumping blood. The muscles that allow you to stand, walk, run and propel food through your gut. Your liver and kidneys, helping digest food, detoxifying your blood. Your immune system, providing constant surveillance of your inner health, capable of fighting known and new pathogens, recognizing what is 'you' and what is not. The electrochemistry of your brain, where over 80 billion cells are performing unfathomable computations, processing information from outside – such as sound, light, taste, smell, touch – and regulating every system inside the body. Enabling you to plan and execute, remember and forget, learn from mistakes, be creative (play a musical instrument), have empathy, feel love and sometimes act altruistically.

Our bodies perform an impossibly vast array of tasks and they do it extremely well and for a tiny amount of energy consumed, about 2,000 kilocalories a day. That's under 100 watts an hour. An electric kettle uses 30 times this amount of energy. That's a lot of 'bang for your buck', as they say. Your body does all of this through the collective actions of its basic building blocks: *cells*. Humans are made up of an estimated 30 trillion of these minuscule membrane-enclosed sacs of living chemistry. Cells work individually as well as together, teaming up to form larger structures: tissues, organs and entire systems such as the cardiovascular and nervous systems. Cells are specialised to carry out the functions of the body.

The abilities of cells to do all of this emerge from their inner biochemistry. The molecules they contain, the structures and bio-machines they form and the functions those machines carry out. In the Introduction to this book, I used the analogy of an orchestra to frame the central ideas of how precise tuning of this biological chemistry is essential for life. The sheets of music are our genes, the information needed to play the music correctly. But that is not enough for a top performance. A conductor is needed, to make adjustments to the molecular equivalent of sounds, volume and tempo, so that it all works perfectly. In between the sheets of music and the collective sound of the orchestra we have the *instruments*. The instruments we are talking about inside our cells are *proteins* and the conductor is a special form of *nucleic acid*. An information-carrying chemical cousin of DNA called ribonucleic acid (RNA).

Proteins perform vast numbers of functions in our cells. We have proteins that can read DNA. Proteins that form gates to control the movement of raw materials into our cells and let waste out. We have proteins that prompt our cells to grow and develop special features,

https://doi.org/10.1017/9781009466400.002 Published online by Cambridge University Press

turning simpler cell forms into sophisticated structures such as brain cells or any one of the hundreds of other cell types in our bodies. Proteins that form nano-machines to transport other proteins. Proteins that drive the motion of the hair-like projections called cilia on the surface of cells in our lungs that waft fluid and inhaled materials past the delicate air sacs. Other proteins form mechanical rod-like structures that give certain cells the ability to change shape, a motion that allows muscle cells to contract. We have proteins that carry out chemical reactions, converting glucose into the energy currency, called adenosine triphosphate (ATP), that all cells rely on. Sometimes the cell is more or less just protein. Red blood cells are packed with about 270 million copies of the haemoglobin protein, which can bind and release oxygen. Several of the proteins you will meet in this book are able to bind to RNA. These are called RNA binding proteins (RBPs). The human genome contains instructions to make about 3,000 different RBPs. They can bind to RNA of all different shapes and sizes. Some RBPs live out their lives in the nucleus close to DNA whereas others spend their days in the cytosol, the liquid outside the nucleus that makes up the remainder of the fluid in our cells. Some find their way to far reaches of the cell, making molecular journeys the equivalent of you travelling to the moon and back. Many RBPs work to keep RNA safe, protecting it from harm. And there are enzyme versions of RBPs that can cut, breaking the chemical bonds that hold together RNA. Molecular scissors that slice off and discard chunks of RNA akin to a sculptor chipping off pieces of marble to shape the statue beneath. Two of the RBPs that have this ability are fundamental to the lives of microRNA and we will meet them in Chapter 2.

All proteins are formed by the linking together of amino acids into long chains, assembled one at a time inside giant molecular machines called ribosomes. There are about twenty different amino acids that make up all the proteins in our cells. Amino acids are simple molecules made up of two main parts attached to a central carbon atom. One side has an amine group (one nitrogen and three hydrogen atoms; NH_3). The other end has a carboxyl group (one carbon, two oxygen and one hydrogen; COOH). Inside a cell, the two ends are charged, with the amine end carrying a positive charge and the carboxyl end a negative charge. This polarity provides them with the means to bond together in long chains, just as the positive end of one magnet is attracted to the negative end of another. Each amino acid differs in the nature of the atoms that form a branch off the central carbon atom called the *side chain*. Some carry a charge; others are neutral; others have highly reactive atoms that power chemical reactions. These differences are exploited by mixing together different combinations of amino acids to generate a protein that can do a particular job. Some amino acids are good at creating bulk and shape in a protein, while others sit at the centre of the reaction core where molecules are split apart or bonded together. The simplest amino acid is glycine. That contains a hydrogen (H) as the side chain. The amino acid cysteine contains a sulphur atom in its side chain and when paired with another cysteine can form a chemical bond called a di-sulphide bridge. This helps proteins fold and maintain their correct shape. Some amino acids, including alanine and leucine, are especially good at forming twists in proteins called α-helices.

The instructions for the order in which the amino acids are placed are carried by a molecule called *messenger RNA* (mRNA). This became a household name during the Covid-19 pandemic because some of the vaccines used the mRNA that codes for the spike protein on the virus to teach the immune system what to look out for. The instructions for the mRNA sequences to assemble your proteins are encoded in your DNA. These instructions for making you were inherited from your parents. Deoxyribonucleic acid (DNA) is

https://doi.org/10.1017/9781009466400.002 Published online by Cambridge University Press

a mega-molecule. Technically it is a polymer, a name given to molecules that are made up of fixed, repeating sub-units. A *gene* is commonly used to refer to a discrete section of DNA that has the instructions for making a protein. But there are many sites in the genome that this does not apply to. Indeed, less than 2 per cent of the human genome contains information to make proteins. The HUGO [Human Genome Organisation] Gene Nomenclature Committee (HGNC), which globally agrees a set of rules, defines a gene as 'a DNA segment that contributes to phenotype/function'. It is becoming increasingly clear that far more of our genome than we once thought meets this criterion. Most microRNAs are genes under the HUGO nomenclature, that is, distinct units that can be transcribed, that are book-ended by start and stop signals, that are heritable and that serve specific biological functions.

It is at a key step along the pathway from DNA to protein that microRNAs act. Indeed, while this book is about microRNAs, it is ultimately a book about proteins because the main job of microRNAs is to make sure that our cells have just the right amount of each protein. Is it important to have just the right amount of a protein in a cell? Yes. While most systems in the body can tolerate a degree of variation in the proteins that carry out specific functions, we know that without microRNAs, a key regulator of protein levels, you would not be alive to read this. If you're missing microRNAs at the start of life then you never get much further than being a ball of cells. If you remove microRNAs around the time of birth then you fail to develop much further. If you remove them when you reach adulthood, you can develop cancer or accelerated ageing. Stop them working in the brain and you develop seizures before the brain turns to mush. I am taking some liberty with the word 'you' here. This knowledge comes mainly from experiments in lab animals such as mice. But we are confident that the outcome would be more or less the same in humans. Indeed, people are born with errors in the machinery for making these molecules and this can have devastating impacts on their health. Some of the microRNA genes are so important that we never see people born without them because it is lethal at an early stage. So, this system for controlling protein levels in cells is essential for life. **Figure 1.1** provides a simple overview of the 'gene pathway'. Information flows from DNA to RNA and on to making a protein. I have highlighted the approximate position in this process where microRNAs act.

The Genetic Code

The human genome is a code running to three billion letters. Despite the extraordinary information it contains, including the 20,000 or more genes that code for proteins plus many other interesting parts that code for RNAs that do important things in our cells, it has oddly straightforward chemistry. The code is made from repeats of four simple chemicals called bases: adenine, cytosine, guanine and thymine or A, C, G and T for short. Like amino acids, they are simple molecules and about the size of one of the larger amino acids. The order of the DNA bases in a protein-coding gene determine which amino acid gets picked. Three bases (a trinucleotide) form the unit of information that codes for a particular amino acid. This is known as a *codon*. Because it is a triplet code, we have 4^3 possible combinations, giving us 64 different codons. Of these, 61 code for amino acids and 3 perform another function, signalling to terminate the making of the protein. Most amino acids can be coded for by more than one codon. For example, a glycine is placed in a protein if the mRNA sequence read contains the codons GGC, GGA or GGG.

https://doi.org/10.1017/9781009466400.002 Published online by Cambridge University Press

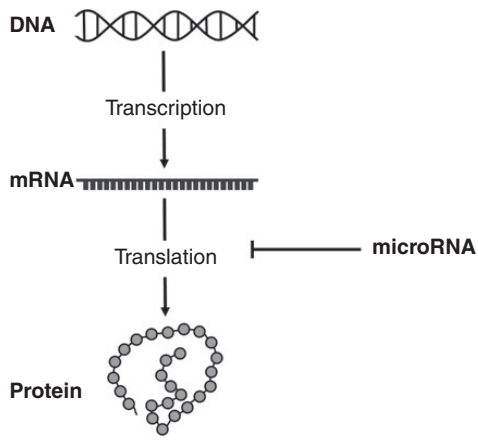

Figure 1.1 Overview of where microRNAs act on the pathway from gene to protein

In this simplified pathway from gene to protein, the first step is transcription, where an RNA copy is made from a gene encoded in the DNA. For protein-coding genes, the RNA formed is called mRNA. The nucleotide sequence of the mRNA (indicated by notches) is used as a template to generate a protein during translation. Proteins are formed by the sequential assembly of amino acids (round circles in the diagram). MicroRNAs act after the mRNA is formed but prior to the formation of the protein.

The structure of DNA slightly resembles a ladder or a spiral staircase, comprising two strands with the bases pairing to one another at regularly spaced intervals up the middle. The backbone (in the ladder analogy, the sides of the ladder you hold as you step up and down) is made up of two types of molecule. The first is a form of carbohydrate, closely related to the sugar you might add to sweeten something. Sugars are made up of carbon, oxygen and hydrogen. The one in the backbone of our DNA has five carbons and so is a pentose sugar; it is called deoxyribose. The deoxy refers to it missing a hydroxyl (OH) group. The other part is a phosphate group, a phosphorous atom with four oxygen atoms attached to it (PO_4). The deoxyribose and the phosphate are strongly bonded together and alternate as deoxyribose-phosphate-deoxyribose-phosphate and so on. The base attaches to the deoxyribose. So each rung of the ladder has a deoxyribose, a phosphate and one of the four bases. This unit – the base, the sugar and the phosphate – is called a *nucleotide*. Bases pair in a specific way. In DNA, an A is always paired across from a T and a C is always paired across from a G; Gs don't bind to Ts or As, and As don't bind to Gs or Cs. A gene sequence is simply an extremely long series of these four bases in different orders, for example A-C-T-G-C-G-T-A and so on. **Figure 1.2** provides an overview of the chemistry of DNA.

In the cells of eukaryotes, organisms that include mushrooms, plants and animals, DNA is found in the nucleus, one of several organelles (another is mitochondria, which make most of our cells' ATP). The DNA in eukaryotic cells is found wrapped around the outside of proteins called histones. This provides a way to compact the genomic instructions and avoid DNA strands becoming tangled. The mix of DNA and histone proteins is called *chromatin*. This is a suitable place for a very short aside on the history of DNA. Key discoveries of what DNA is made of came from work in the early 1900s by Phoebus Levene, a Russian-American biochemist who identified the four DNA bases, the basic rules of their combinations and the presence of the ribose sugar. Most people are familiar with Watson and Crick. They are credited with solving the structure of DNA, which was

https://doi.org/10.1017/9781009466400.002 Published online by Cambridge University Press

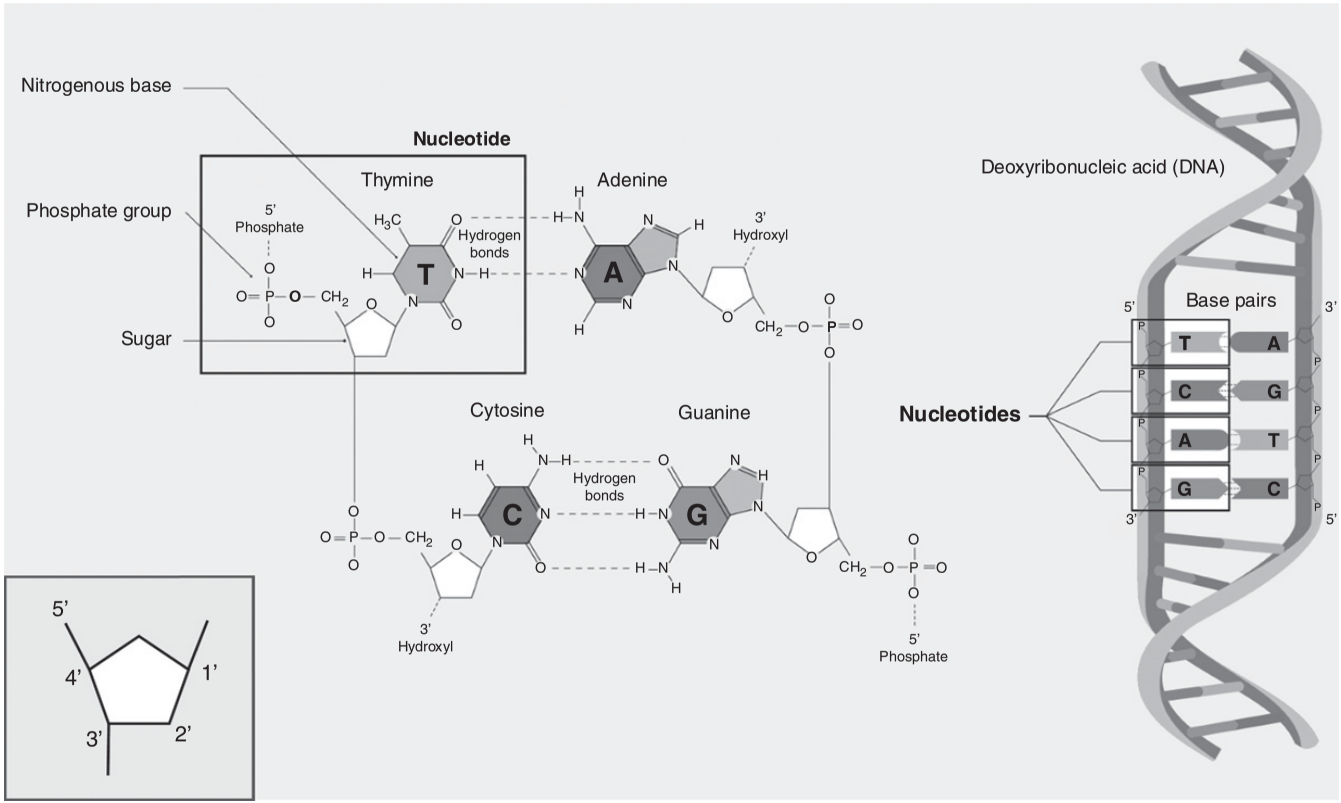

Figure 1.2 Chemical structure of DNA

This shows the basic chemical structure of the nucleotide, the basic building block of DNA, and the base-pairing rules. A nucleotide consists of a sugar molecule (deoxyribose in DNA) attached to a phosphate group and one of the four chemical bases. The bases in DNA are adenine (A), cytosine (C), guanine (G) and thymine (T). In RNA, the base uracil (U) takes the place of thymine. Molecules of DNA and RNA are polymers made up of long chains of nucleotides. The inset (bottom-left) shows the number-position labelling of carbons in the ribose sugar.

Source: Image courtesy of Darryl Leja and the National Human Genome Research Institute (www.genome.gov).

reported in the journal *Nature* in 1953, and this helped our understanding of how DNA worked (1953 is also famous for being the year that kainic acid was discovered, a key tool in neuroscience and epilepsy research – see Chapter 6). Key to confirming Watson and Crick's insights were X-ray images of a crystal of DNA. These were produced by Rosalind Franklin and shared without her knowledge. She was not initially credited with this work, but is now recognised as conducting the key experiments used in the discovery. Some accounts imply that Rosalind failed to grasp what her images revealed about the structure of DNA, but this is a misrepresentation. She was an equal contributor in the solution. She died from cancer owing to the radiation she was exposed to as part of her work; we owe much of modern scientific and medical discovery to this woman. Once the organisation of the DNA molecule was understood, with the sugar as the outside backbone and the bases pairing across from one another in the middle, how it coded could be understood. There was still a lot of work to do and Watson and Crick didn't determine what combination of bases coded to make proteins. That would come later. A couple of other key discoveries on the journey are worth a mention. Swiss scientist Friedrich Meischer gets credit for the 1869 discovery of DNA in the nucleus of cells. In 1944 we have Oswald Avery and colleagues, who proved that DNA is the hereditary material in our cells. They added DNA from a dangerous strain of bacteria to a normally harmless one, causing it to become deadly. In 1950, Erwin Chargaff reported that the amounts of Gs and Cs and the amounts of As and Ts were always the same, but the ratios were different in different species. Towards the end of that decade and in the first years of the 1960s, Marshal Nirenberg is credited with understanding how the code actually works in practice. That is, how sets of bases encode the amino acids in proteins. In the years since, the specific sequence of bases that code for each amino acid has been determined.

The Journey from Gene to Protein

Let us go briefly through the full process from start to finish: the start, the reading of the DNA code, through to the generation of mRNA that later gets read to form a growing amino acid chain that ultimately becomes a protein. Our depth of understanding of this process is now remarkable. I will cover enough detail to understand where and how microRNAs participate in this process. The first stage of the journey from DNA to protein is *transcription*. This is the process of reading a particular stretch of DNA, the gene, and making a copy of that sequence out of RNA. If we are making a protein, then the RNA made is an mRNA. It is common to refer to mRNAs as 'transcripts'. Other sections of DNA can also be read and an RNA copy made. But these other RNAs do not code for a protein and are *non-coding RNA*. There are many different types of non-coding RNA. Some, such as the RNA found inside the ribosome protein factories, are far more abundant in our cells than mRNAs. MicroRNAs are an example of such non-coding RNAs. **Figure 1.3** provides an overview of the basic chemistry of RNA and some of the different forms of RNA in cells.

Reading and copying DNA to make mRNA is a highly regulated, multi-step process. An enzyme called an RNA polymerase binds to a section of DNA. The two strands of DNA are unzipped from one another. The bonds holding the Cs to Gs and the As to Ts are broken and the base pairs are separated. The RNA polymerase now makes an RNA, with each base being inserted according to what it reads from the DNA strand. The strand of DNA that is read is called the *antisense* strand. The other ignored section of DNA is called the coding or *sense* strand because its sequence is the same as the mRNA that is generated. This reading process occurs in one direction only, analogous to how you read this sentence, from left to right.

https://doi.org/10.1017/9781009466400.002 Published online by Cambridge University Press

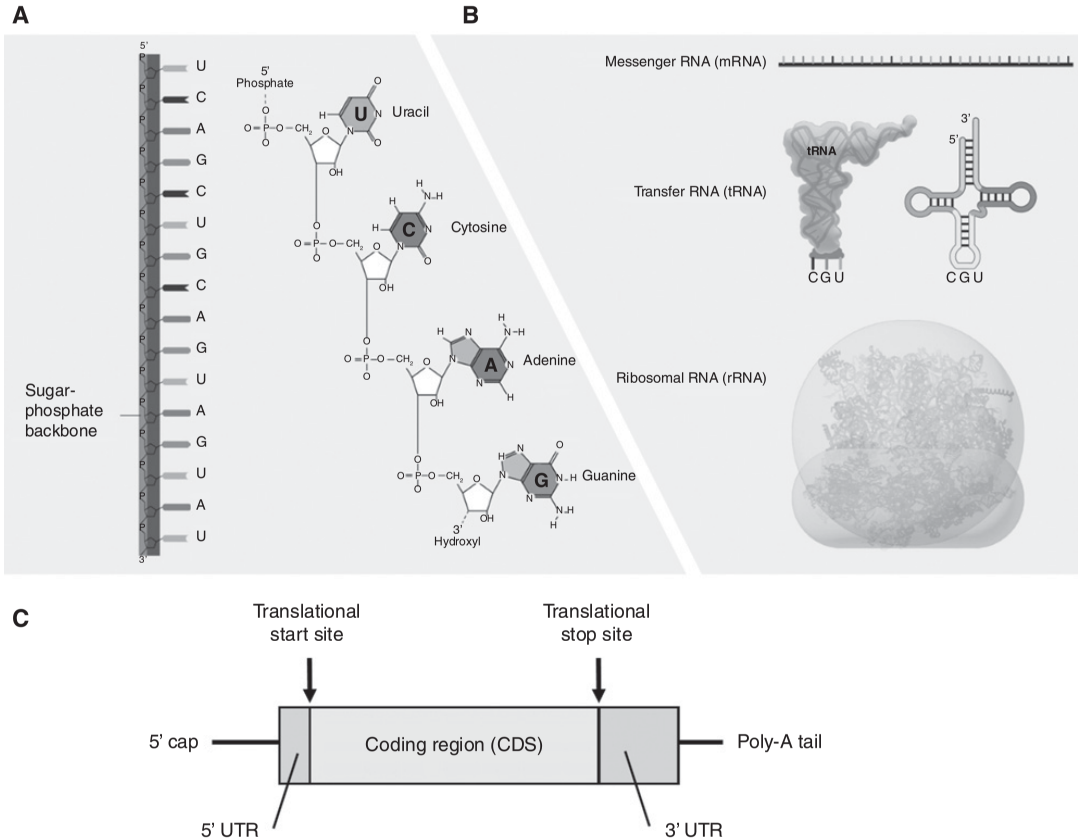

Figure 1.3 The base composition and different forms of RNA

Panel A shows the four different chemical bases present in RNA along the sugar–phosphate backbone (vertical column). Note that uracil (U) replaces thymine (T) in RNA. Within and between RNA molecules, base-pairing can occur between Us and As and between Gs and Cs. In the coding sequence of an mRNA, every three nucleotides comprise a codon for a specific transfer RNA to bring an amino acid during the building of a protein.

Panel B shows examples of some of the non-coding RNAs in a cell.

Panel C shows the basic subparts of an mRNA.

Key: U, uracil; A, adenine; C, cytosine; G, guanine; UTR, untranslated region; rRNA, ribosomal RNA; tRNA, transfer RNA.

Source: Images in Panels A and B courtesy of the National Human Genome Research Institute (www.genome.gov). Panel C is the author's own.

Each end of a DNA and an RNA molecule has a number. This is based on the position of the five carbon atoms in the ribose sugar (see **Figure 1.2**). The numbering works in a clockwise direction starting with one at the three o'clock position and running to position five. Carbon 1 is attached to a base and carbon 2 also faces in towards the centre. Carbon 4 faces out. This leaves carbons 3 and 5. Carbon 5 is at the top of the ribose and is the point of attachment for the phosphate group above it. Carbon 3 of the ribose bonds with the phosphate group below it. These are referred to as the 5 prime (5') and the 3 prime (3') positions. Both DNA and RNA use the same naming system.

During transcription, DNA is read by RNA polymerases in the 3' to 5' direction while the growing RNA strand is created in the opposite direction (5' to 3'). The RNA polymerase adds each new base onto the 3' end of the growing RNA strand. If the polymerase reads a G on the DNA strand, it places a C in the growing RNA molecule, while reading a T results in an A being placed in the RNA. One of the bases used by the polymerase is different from the standard four in DNA. The thymine base in DNA is not present in RNA. Instead, another base called uracil (U) is inserted into RNA. So wherever RNA polymerase reads an A base on the DNA strand it puts a U, not a T, into the growing RNA molecule. So, reading a stretch of DNA that had the sequence TAC GGA would generate an mRNA with the sequence AUG CCU. As the polymerase passes along reading the DNA sequence, the RNA strand passes out in one direction while the DNA strand is rejoined with the original strand it was separated from. The process is divided into four stages called initiation, promoter escape, elongation and termination.

Transcription requires specific signals to be in place to mark the gene to be read. The key molecules that mark a gene as ready to read are proteins called *transcription factors*. There are also proteins called transcription repressors that do the opposite, having the effect of reducing any transcription of the targeted gene. Transcription factors receive their instructions from signals that originate outside or inside the cell. The transcription factor doesn't synthesise the mRNA; it simply docks onto the DNA. This happens a bit upstream of the section of DNA that codes for the protein. That site is called a *promoter*. It is a short stretch of DNA bases that acts as a 'come here and get started' signal for the enzyme complex that will make the mRNA copy. Alone, however, this doesn't result in much transcription. A second site needs to be activated; this is called an *enhancer*. This can be quite a distance from the gene and the promoter and a different set of proteins bind to this site. However, chromatin is flexible and a loop can be formed that brings the enhancer site close to the promoter site, resulting in a powerful 'on' switch for transcription.

Next, a complex of proteins called Mediator acts to communicate between these two sections of DNA and transcription now takes place. The RNA polymerase unwinds a stretch of fourteen base pairs of the DNA at what is called the transcriptional start site (TSS). An RNA strand starts to form (step 1 done). The polymerase now has to break free of the promoter to work properly, a process termed promoter escape (step 2). Now the polymerase can begin to read and form the full-length transcript, which grows at a rate of about 1,000 nucleotides per minute (elongation – step 3). Finally, the polymerase finishes and breaks loose (termination – step 4). By the end of transcription we have a long RNA molecule, although different genes vary in length and so do the RNAs transcribed from them.

We don't yet have our final mRNA. What we have by the end of transcription is really pre-mRNA. Synthesised mRNAs undergo a complex series of further adjustments in a process termed post-transcriptional regulation. These actions result in an mRNA that is prepped and ready to be sent to the ribosome for translation. The prepping processes

https://doi.org/10.1017/9781009466400.002 Published online by Cambridge University Press

include splicing, capping, tailing and editing. An estimated 100 different chemical modifications of mRNA have been discovered. We don't know what all of these do, but they generally affect the stability or the readability of the mRNA.

The parts of the pre-mRNA that code for the protein are called *exons*. When the pre-mRNA is made, it usually contains long sections of RNA that do not code for amino acids. These are called *introns* and they are snipped out. The sections of RNA either side are glued (spliced) back together. This leaves a series of exons joined together. The processing of the pre-mRNA may also result in inclusion of some but not other exons. For example, the pre-mRNA may include ten different exons but the final mRNA may omit one or more of these. That is called *alternative splicing*. It is carefully regulated by RBPs that bind onto the segments to be chosen for inclusion or exclusion; the enzymes that do the cutting are called nucleases. The proteins that result from these different splicing events display subtle differences in their properties. For example, one version might possess an extra signal that allows it to interact in a different way or directs the protein to a particular place in the cell. This is a way that the main repertoire of proteins encoded in the genome can be boosted even further. Thus, from around 20,000 protein-coding genes, cells can generate approximately 100,000 different proteins.

The part of the mRNA that is important for the sequence of amino acids is called the coding sequence (CDS) (see **Figure 1.3C**). Either side of this region are other stretches of RNA that serve regulatory functions. This includes the start codon at the 5' end of the CDS. This is usually the sequence AUG. Just in front of this start codon is a segment of RNA called the 5' untranslated region (5' UTR). This is only a few bases in length and is usually not a signal for an amino acid to be placed. Both far ends of the mRNA undergo modifications. At the 5' end of the mRNA, a structure called a cap is added. In most cases this involves the addition of a chemically modified guanine base. The cap serves several functions, protecting the molecule from being digested and identifying it for export to the ribosome. At the other end of the pre-mRNA, a stretch of adenines is added called a poly(A) tail. This can include more than 200 As in a row. Like the cap structure, this marks the mRNA for export and translation, and also affects its stability. The poly(A) tail doesn't last, at least at its full length. Specific enzymes shorten it over time and this ultimately dictates the lifespan of the mRNA. These modifications, the intron removal, the 5' cap and the 3' poly(A) tail, are coordinated and occur *co-transcriptionally*. That is, it all happens almost immediately as the transcript is generated.

Between the CDS and the poly(A) tail is a segment called the *3' untranslated region* (3' UTR). This contains important regulatory information which controls how much protein is made from the mRNA. It can vary substantially in length. Longer 3' UTRs tend to mean that the transcript is more heavily 'policed' during translation. This is the part of the mRNA where microRNAs usually exert their effects. It will feature heavily in later chapters. Once the modified and final mRNA reaches the cytoplasm, it is chaperoned to the ribosome. This is where the message is read and a protein is formed.

Translation

Proteins are made or synthesised inside giant molecular machines called ribosomes. These are conglomerates of a special form of RNA called ribosomal RNA or rRNA (see **Figure 1.3B**), intertwined with coils of various proteins in an approximately 60:40 mix. Ribosomes have one large sub-unit or part called the 60S and one smaller sub-unit, the 40S.

https://doi.org/10.1017/9781009466400.002 Published online by Cambridge University Press

It is an ancient structure in that its design and function and the sequences that code for the building blocks are highly similar or conserved in all eukaryotes. An mRNA arrives to a ribosome chaperoned by various RBPs and is then fed into a groove between the two sub-units. Next, another form of RNA called transfer RNA (tRNA) swoops in, carrying in its grasp a single amino acid that has been joined on prior to arrival. There are 40 to 60 different types of tRNA in human cells. Each tRNA carries its own amino acid. The tRNA structure slightly resembles a lower-case letter r. The 'bottom' of the r is called the anti-codon loop. This part contains a triplet sequence that is the complementary match for the codon, on the mRNA. For example, a tRNA with an anti-codon loop sequence CGU will pair to an GCA sequence on the mRNA. The amino acid is attached at the opposite end of the tRNA, the 'handle' of the r shape. As the mRNA is fed into the ribosome, the 5' end attaches to the small sub-unit. The ribosome moves along until it reaches a start codon. The large sub-unit is then brought in along with the first tRNA, followed by other tRNAs falling into place in sequence according to the codon:anti-codon pairing. Specific regions within the ribosome house and arrange the tRNA-amino acid molecules as they enter.

For each triplet sequence we get one additional amino acid. The tRNA that brought in the amino acid gets released and the ribosome bumps along to the next codon. The process ends when a stop codon is encountered in the mRNA. The common ones are UAG, UAA and UGA. The ribosome then releases the polypeptide. From these twenty amino acid building blocks we can generate every protein in the body. The growing chain of amino acids is created in a way that has a front and a back end. The beginning has an amine group (NH_3) and is referred to as the N-terminus. The end of the protein has a carboxyl group (COOH) and is called the C-terminus. When a protein is shown in a drawing, these ends are typically highlighted and the convention is to show the N-terminal on the left and the C-terminus on the right. The finished polypeptide is not yet a functional protein, however. As it is produced, it begins to fold and form more complex structures. These include the spiral α-helix mentioned earlier. These are called secondary structures. Further folding, tertiary structure creates the 3D shape of the protein. A higher level called quaternary structure refers to when groups of folded proteins form superstructures. And now we have reached the end of the DNA-to-protein pathway.

Asserting Control Over Proteins

The instructions for every protein are present inside every cell, with the exception of red blood cells. These originally contained DNA, but this is actively removed during the maturation process, creating more space for haemoglobin. Every other cell contains the instructions to make any and all proteins. But we know they aren't all made. A muscle cell is packed with structural proteins that allow the cell to contract and relax. A liver cell is packed with enzymes to break down ingested substances. If you are a neurone then you must be capable of transmitting rapid changes in electrochemical signals and producing, storing and releasing neurotransmitters. Clearly, cells pick and choose what proteins they make. This makes a simple but profoundly important point. The protein landscape, the collection of proteins that forms the structures of cells, enzymes, transporters, motors and everything else, is tailored to the job(s) the cells perform. And if a cell isn't making the right proteins or the correct amount of those proteins then problems quickly arise. At best, the cell doesn't do its job well enough. For example, it doesn't secrete what it should (insulin from the Beta islet cells of the pancreas of a patient with type 1 diabetes), contract how it should or signal

https://doi.org/10.1017/9781009466400.002 Published online by Cambridge University Press

how it should. At worst, a cell can become cancerous. Dividing, invading and ultimately killing its host. How much of a given gene's encoded protein a cell makes and when and where its product is generated must be subject to precise control. But how is this control achieved?

The correct amount of a protein is determined by regulatory steps at multiple points in the pathway. Researchers continue to apply new and more sophisticated ways of measuring global protein output relative to mRNA levels. A key influence on protein levels is how much actual transcription occurs in the first place. Put simply, the more mRNA, the more protein. But, by several estimates, mRNA abundance accounts for only about half the levels of a protein. Accounting for differences in the mRNA:protein ratio are various steps along the way. These include factors acting after transcription but before translation, including on mRNA stability and other post-transcriptional controls, and after translation, including on protein stability and turnover. MicroRNAs act post-transcriptionally. They latch onto the mRNA and alter how much is translated into protein. So how were they discovered? How many are there? What proteins do they control? What are the rules they live by?

Gene, RNA and Protein Naming Conventions

We are about to dive into a section that shifts between talking about genes among our DNA, transcribed mRNA, and the translated product of the mRNA, the protein. There are conventions for naming each of these. A quick overview will make it easier to keep track of which part of the 'molecular soup' is being referred to. Let's imagine a hypothetical gene called 'read every day'. Because there is more than one version of this gene, let us give it a number – 1. In humans, the full gene name would be written as 'read every day 1'. When referred to in its abbreviated form, the gene is *RED1*. All capital letters in italic. The mRNA for this gene would also be *RED1*. The protein is not italicised, though, and is written as RED1. Now we come to other species. The full name of the mouse version of the gene would also be 'read every day 1'. But the abbreviated form of the mouse gene only has the first letter capitalised. The rest are lowercase. The gene would be *Red1*, the mRNA would also be *Red1*, but the protein would be uppercase RED1. Many of the early discoveries on microRNA were made in an animal called *Caenorhabditis elegans* or *C. elegans* for short. The language convention for naming organisms is to use Latin. The *C. elegans* gene naming convention includes a hyphen when there is a number and the first letter is not capitalised. So we get the gene *red-1*, the mRNA *red-1* and the protein RED-1.

A Worm's Tale

The discovery of microRNAs in animals can be traced back to two landmark papers in the same 1993 issue of the journal *Cell*. The discovery of microRNAs was not made in humans but rather in *C. elegans*, a roundworm. What were the researchers looking for in the worm and why were they using worms? The *C. elegans* is what is called a model organism. Model organisms are an essential tool in biology, allowing complex problems to be understood in a simpler context. Another advantage of model organisms is how quickly they develop, shortening the time it takes to run experiments: *C. elegans* reaches adulthood three days after hatching from an egg. Adult *C. elegans* worms measure about 1 mm and so they are cheap to house and feed. Further, *C. elegans* can survive being deep-frozen, making them easy to store. Model organisms also allow experimentation including genetic mutation that might

https://doi.org/10.1017/9781009466400.002 Published online by Cambridge University Press

not be otherwise possible for ethical or technical reasons. Other examples of model organisms include the bacterium *E. coli*, the fruit fly *D. melanogaster* and a fish called *D. rerio.*

The roundworm *C. elegans* has been helping biologists understand the genetic programmes that control how animals develop for more than fifty years. A key figure is Sydney Brenner, a Nobel prize winner who passed away in 2019. One of the most brilliant minds of the twentieth century, he entered university in his home country of South Africa at the age of fifteen and five of the people he trained went on to win Nobel prizes. He worked for a time with Francis Crick at the University of Cambridge, contributing to our understanding of the triplet code for amino acids. He later headed the university's laboratory of molecular biology. He was interested in how sets of genes work to give rise to the complex organisation of organisms, including their nervous systems. He recognised that simple or lower organisms share much of the same biology as humans. For example, the *C. elegans* genome has similar versions to 60 per cent to 80 per cent of human protein-coding genes, despite being separated by 500 million to 600 million years of evolutionary time. Our last shared common ancestor was in the pre-Cambrian period, slightly before the fossilisation of the animals that are the focus of the history of life story featured at the start of this chapter. Brenner's true passion was always to understand how the brain worked. Writing in an article in the journal *Genetics* in the early 1970s, he laid out the case for how, by understanding the complete structure of a simple organism's nervous system, the timing of its development, the underlying genetic programmes that control behaviours, you could gain insight into much more complex systems (i.e., human systems). His work on *C. elegans* had begun in the 1960s. He had pondered using *D. melanogaster* but decided the numbers of neurones were still unmanageable. Adding to the argument to use *C. elegans*, the laboratory conditions for maintaining were simple and already established. The worms are also hermaphrodites, meaning they can self-fertilise and their offspring are clones and therefore genetically identical. At the time he started, *C. elegans* was thought to have about 200 neurones. These numbers have been slightly revised; a typical adult worm has ~300 neurones of its total of ~1,000 cells. While it has a nervous system, a reproductive system and a digestive system, it lacks a respiratory system or a circulatory system. Over many years, Brenner's and other teams transformed our understanding of how genes control the developing nervous and other systems. Indeed, *C. elegans* was the first animal to have its complete nervous system 'connectome' mapped, in a wiring diagram in which every connection between every neurone has been documented.

After hatching from an egg, *C. elegans* goes through four larval stages termed L1–L4, before becoming an adult worm. The genetics of its growth and development have been studied in such depth that we know precisely what genes switch on when, what processes they control, the number and whereabouts of every cell made, what tissues they form and even how many cells are killed off. Indeed, Brenner's Nobel prize was shared with others who discovered genetically encoded pathways that eliminate excess cells generated during development; some of the same genes in humans get inactivated in cancers. A common trick to learn what a gene did was to expose worms to known mutagenic chemicals that randomly inserted errors in their genomes. The position of the errors could be traced, allowing researchers to understand a bit about the function of the gene that contained the mistake. By repeating this enough, researchers learnt which genes affected worm development. Researchers now have the ability to engineer the *C. elegans* genome to precisely change sequences as well as introduce molecules that are fluorescent under a microscope, allowing them to track the growth, movement and final resting place of any cell in the worm.

https://doi.org/10.1017/9781009466400.002 Published online by Cambridge University Press

MicroRNAs were discovered during studies on the development of *C. elegans*. The scientists at the centre – Victor Ambros and Gary Ruvkun – were both Boston area–based. At the time, a number of genes called *heterochronic* genes were known to control the timing of development. Ambros had performed experiments that showed that mutations in these genes resulted in changes to the timing of specific events during worm development, for example the reappearance of early developmental changes. One of these heterochronic genes was *lin-14*. The 'lin' stands for abnormal cell lineage. The *lin-14* gene encoded a transcription factor that appeared early during larval development, in the embryo and the L1 stage. Experiments had already shown that the LIN-14 protein, normally present in the nucleus as befits its job of controlling gene activity, later disappeared and was not detectable at later larval stages and in mature worms. That is, a time-dependent, decreasing gradient of LIN-14 was required for normal development. But researchers noted that the mRNA for *lin-14* didn't disappear. In fact, it remained present the whole time. The message was still being made, but something was now preventing translation. Something was acting *post-transcriptionally* to lower production of LIN-14 protein and control the timing of worm development.

At the time, transcription of another gene, called *lin-4*, was known to affect the timing of larval development. Mutations that caused loss of *lin-4* resulted in animals going back into earlier larval phases. They would re-start previous rounds of development and would be missing certain adult features. *Lin-4* was necessary for maturation of the worm. Loss-of-function mutations in *lin-14* that prevented the protein being made produced the opposite features. Worms skipped steps in early larval development, jumped ahead, and the result was the inappropriate appearance of adult-like features in the young worms (see **Figure 1.4A**). In contrast, overproduction of *lin-14* produced worms that looked similar to the faulty *lin-4* mutants, displaying re-engagement of early development features. This was consistent with *lin-4* acting as a negative regulator of *lin-14*.

By studying the mutant *lin-14* worm, Ruvkun's team had narrowed down the genetic error that was stopping the decline in LIN-14 protein during later development. The mutant worm that couldn't 'switch off' LIN-14 was missing a section of the *lin-14* mRNA. The missing piece was not, however, in the protein-coding sequence. It was found in the untranslated region of the mRNA, the 3' UTR. The 3' UTR of *lin-14* contained information in the form of a nucleotide sequence that was important for blocking production of LIN-14 protein. Was this linked somehow to whatever *lin-4* was doing that seemed to control translation of the LIN-14 protein?

Ambros' team had been working since the early 1980s on *lin-14* and other genes including *lin-4* that controlled the timing of *C. elegans* development. His team, in parallel with Ruvkin's, were about to make a major advance in molecular biology. By uncovering the mechanism by which *lin-4* regulated translation of LIN-14 protein, they would discover a fundamental mechanism by which gene activity in all animals, not just worms, was controlled. Ambros' work in the seminal *Cell* paper[1] began by a detailed analysis of the *lin-4* locus, the specific position on a chromosome where the gene information resides. He and his team noted that the sequence lay within an intron of another gene and analysis of the genetic sequence revealed that it did not code for a protein. Indeed, introducing changes to the sequence of *lin-4* that would have scrambled or blocked translation of a protein-coding mRNA had no effect on the ability of *lin-4* RNA to affect developmental timing. When they searched for evidence of expression of RNA from the locus, they detected two transcripts. One was sixty-one nucleotides in length and the other was twenty-two nucleotides. The longer one was

https://doi.org/10.1017/9781009466400.002 Published online by Cambridge University Press

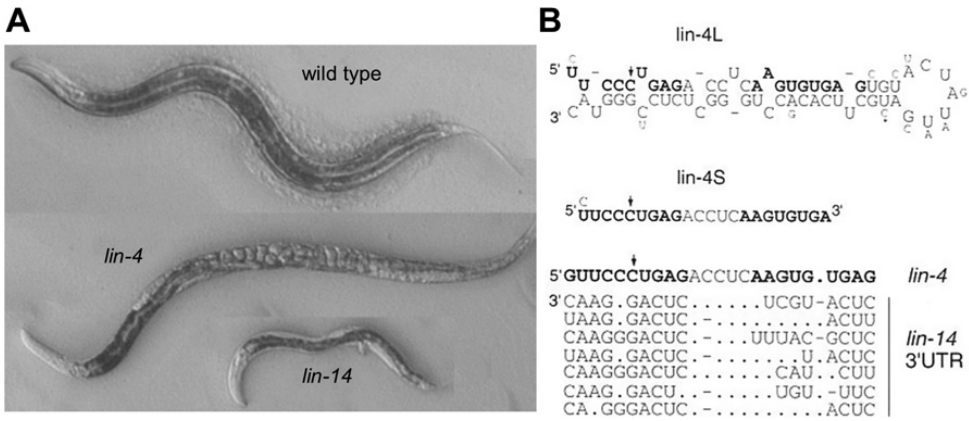

Figure 1.4 *C. elegans* and the discovery of the first microRNA

Panel A shows three *C. elegans* worms. Compared to the wild-type worm, adult *lin-4* loss-of-function mutants lack many adult structures and they are unable to lay eggs on account of a failure to develop a vulva (note the spherical eggs accumulating within their bodies). The bottom worm is a *lin-14* loss-of-function mutant. This has developed certain adult features precociously at larval stages, resulting in smaller, poorly formed adults.

Panel B shows the sequences deduced for (top) the precursor *lin-4L* including the hairpin loop and (middle) the shorter *lin-4S*, the mature microRNA. Sequences in bold are those complementary to the 3' UTR of *lin-14*. The smaller type letters are the annotation of sequences that differ in other species of *Caenorhabditis*. Other markings indicate positions of introduced mutations. The lower panel shows complementary sequence alignment of *lin-4S* with seven sites in the 3' UTR of the *lin-14* transcript. At the time, the RNA sequences were deduced from the DNA sequence and transcript mapping and not from direct sequence analysis of RNA.

Source: Panel A reprinted with permission from Ambros, *Nature Medicine*, vol. 14, pp. 1036–40 (2008).[2] Panel B adapted with permission from Lee et al., *Cell*, vol. 75, pp. 843–54 (1993).[1]

predicted to have a region that folded over on itself, creating a loop structure. The team deduced the sequences of both, discovering that the short form, *lin-4S*, was identical to the 5' end of the longer *lin-4L*. This suggested the *lin-4S* had originally been part of the longer transcript, so perhaps *lin-4L* was a precursor of *lin-4S*. Both sequences were also found in three other species of *Caenorhabditis*, whereas the parts of *lin-4* outside of this region were poorly conserved between different worm species (see **Figure 1.4B**).

Now came the eureka moment. They aligned the *lin-4S* sequence with the 3' UTR section of the *lin-14* mRNA known to be critical for blocking translation. There was a series of seven repeating sections of complementarity between the two RNAs: stretches of nucleotides on the *lin-4* RNA that were perfectly complementary to *lin-14*. One region of the 3' UTR of *lin-14* had the sequence CAAGGGACUC, which would be able to base-pair to a region of *lin-4* contained within the twenty-two nucleotide shorter RNA that ran GUUCCCUGAG. The final piece of the puzzle was in place. *Lin-4* was reducing the amount of LIN-14 by direct RNA-to-RNA interaction within the 3' UTR of *lin-14*. In their conclusions, they offered that *lin-4* RNA may bind to *lin-14* mRNA in the cytoplasm and inhibit its translation. *Lin-4* was acting like molecular Velcro, sticking onto the *lin-14* RNA, forming a duplex and preventing it from being translated into a protein. This type of inhibitory mechanism had a name – *antisense*. We previously came across antisense in the context of transcription. When the DNA helix is unwound, it is the antisense strand that is the template that the RNA polymerase uses. The generated mRNA is a sense strand and is also identical except for

https://doi.org/10.1017/9781009466400.002 Published online by Cambridge University Press

the swapping of a T for a U to the other sense DNA strand. Here, an RNA from one part of the genome was being made that could control the effects of another RNA by sticking to it. This was not completely unheard of because an RNA called *XIST* had been identified that was involved in shutting down one of the two copies of the X chromosome in the cells of females. But *XIST* is 16,500 bases long. *Lin-4* was 0.1 per cent of this length. The authors realised that *lin-4* might be the first of a new class of *small regulatory* RNA. **Figure 1.4** shows examples of the *lin-4* and *lin-14* mutant worms and the sequence of *lin-4*, as well as the sequence alignment to the 3' UTR of *lin-14*.

Ruvkun's team performed similar experiments on the levels of *lin-4* and *lin-14*/LIN-14, confirming that it was a post-transcriptional event and mapping the seven complementary interaction sites.[3] They also performed some clever experiments that further confirmed this new gene control mechanism. They took the section of the *lin-14* 3' UTR that contained the regulatory information and inserted it into another protein-coding mRNA that was not sensitive to developmental timing. The transplant was a success. The protein product of the recipient gene now showed developmental down-regulation, even though the host gene had nothing to do with the functions of LIN-14. As with the original *lin-14* observations, the amount of RNA of the hybrid containing the transplanted 3' UTR remained unchanged, with only the protein decreased. Finally, they found that the acquired timing effect of the transplanted 3' UTR only worked if *lin-4* was also present. The tampered gene returned to its original unregulated state if *lin-4* was mutated. The 3' UTR of *lin-14* contained the complete instruction for the translational block by *lin-4*.

A major new regulatory step had been discovered in the pathway from gene to protein with *lin-4* the founding member. But the significance of the findings were rather overlooked at the time. That would all change at the turn of the millennium. The year 2000 would be a big year. What could have been an idiosyncratic finding in a worm turned out to be something fundamental to complex life on Earth.

https://doi.org/10.1017/9781009466400.002 Published online by Cambridge University Press