THE EPIGENETICS REVOLUTION

NESSA CAREY

How modern biology is rewriting our understanding of genetics, disease and inheritance

ICON BOOKS

Published in the UK in 2011 by
Icon Books Ltd, Omnibus Business Centre,
39–41 North Road, London N7 9DP
email: info@iconbooks.co.uk
www.iconbooks.co.uk

Sold in the UK, Europe, South Africa and Asia
by Faber & Faber Ltd, Bloomsbury House, 74–77 Great Russell Street,
London WC1B 3DA or their agents

Distributed in the UK, Europe, South Africa and Asia
by TBS Ltd, TBS Distribution Centre, Colchester Road
Frating Green, Colchester CO7 7DW

Published in Australia in 2011
by Allen & Unwin Pty Ltd, PO Box 8500,
83 Alexander Street, Crows Nest, NSW 2065

ISBN: 978-184831-292-0

Typeset in 12 on 16pt Times by Marie Doherty

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Chapter 3

Life As We Knew It

A poet can survive everything but a misprint.
Oscar Wilde

If we are going to understand epigenetics, we first need to understand a bit about genetics and genes. The basic code for pretty much all independent life on earth, from bacteria to elephants, from Japanese knotweed to humans, is DNA (deoxyribonucleic acid). The phrase 'DNA' has become an expression in its own right with increasingly vague meanings. Social commentators may refer to the DNA of a society or of a corporation, by which they mean the real core of values behind an organisation. There's even been a perfume called after it. The iconic scientific image of the mid-20th century was the atomic mushroom cloud. The double helix of DNA had similar cachet in the later part of the same century.

Science is just as prone to mood swings and fashions as any other human activity. There was a period when the prevailing orthodoxy seemed to be that the only thing that mattered was our DNA script, our genetic inheritance. Chapters 1 and 2 showed that this can't be the case, as the same script is used differently depending on its cellular context. The field is now possibly at risk of swinging a bit too far in the opposite direction, with hard-line epigeneticists almost minimizing the significance of the DNA code. The truth is, of course, somewhere in between.

In the Introduction, we described DNA as a script. In the theatre, if a script is lousy then even a wonderful director and a terrific cast won't be able to create a great production. On the other hand, we have probably all suffered through terrible productions of our favourite plays. Even if the script is perfect, the final outcome can

be awful if the interpretation is poor. In the same way, genetics and epigenetics work intimately together to create the miracles that are us and every organic thing around us.

DNA is the fundamental information source in our cells, their basic blueprint. DNA itself isn't the real business end of things, in the sense that it doesn't carry out all the thousands of activities required just to keep us alive. That job is mainly performed by the proteins. It's proteins that carry oxygen around our bloodstream, that turn chips and burgers into sugars and other nutrients that can be absorbed from our guts and used to power our brains, that contract our muscles so we can turn the pages of this book. But DNA is what carries the codes for all these proteins.

If DNA is a code, then it must contain symbols that can be read. It must act like a language. This is indeed exactly what the DNA code does. It might seem odd when we think how complicated we humans are, but our DNA is a language with only four letters. These letters are known as bases, and their full names are adenine, cytosine, guanine and thymine. They are abbreviated to A, C, G and T. It's worth remembering C, cytosine, in particular, because this is the most important of all the bases in epigenetics.

One of the easiest ways to visualise DNA mentally is as a zip. It's not a perfect analogy, but it will get us started. Of course, one of the most obvious things that we know about a zip is that it is formed of two strips facing each other. This is also true of DNA. The four bases of DNA are the teeth on the zip. The bases on each side of the zip can link up to each other chemically and hold the zip together. Two bases facing each other and joined up like this are known as a base-pair. The fabric strips that the teeth are stitched on to on a zip are the DNA backbones. There are always two backbones facing each other, like the two sides of the zip, and DNA is therefore referred to as double-stranded. The two sides of the zip are basically twisted around to form a spiral structure – the famous double helix. Figure 3.1 is a stylised representation of what the DNA double helix looks like.

Figure 3.1 A schematic representation of DNA. The two backbones are twisted around each other to form a double helix. The helix is held together by chemical bonds between the bases in the centre of the molecule.

The analogy will only get us so far, however, and that's because the teeth of the DNA zip aren't all equivalent. If one of the teeth is an A base, it can only link up with a T base on the opposite strand. Similarly, if there is a G base on one strand, it can only link up with a C on the other one. This is known as the base-pairing principle. If an A tried to link with a C on the opposite strand it would throw the whole shape of the DNA out of kilter, a bit like a faulty tooth on a zip.

Keeping it pure

The base-pairing principle is incredibly important in terms of DNA function. During development, and even during a lot of adult life, the cells of our bodies divide. They do this so that organs can get bigger as a baby matures, for example. They also grow to replace cells that die off quite naturally. An example of this is the production by the bone marrow of white blood cells, produced to replace those that are lost in our bodies' constant battles with

infectious micro-organisms. The majority of cell types reproduce by first copying their entire DNA, and then dividing it equally between two daughter cells. This DNA replication is essential. Without it, daughter cells could end up with no DNA, which in most cases would render them completely useless, like a computer that's lost its operating software.

It's the copying of DNA before each cell division that shows why the base-pairing principle is so important. Hundreds of scientists have spent their entire careers working out the details of how DNA gets faithfully copied. Here's the gist of it. The two strands of DNA are pulled apart and then the huge number of proteins involved in the copying (known as the replication complex) get to work.

Figure 3.2 shows in principle what happens. The replication complex moves along each single strand of DNA, and builds up a new strand facing it. The complex recognises a specific base – base C for example – and always puts a G in the opposite position on the strand that it's building. That's why the base-pairing principle is so important. Because C has to pair up with G, and A has to pair up with T, the cells can use the existing DNA as a template to make the new strands. Each daughter cell ends up with a new perfect copy of the DNA, in which one of the strands came from the original DNA molecule and the other was newly synthesised.

Even in nature, in a system which has evolved over billions of years, nothing is perfect and occasionally the replication machinery makes a mistake. It might try to insert a T where a C should really go. When this happens the error is almost always repaired very quickly by another set of proteins that can recognise that this has happened, take out the wrong base and put in the right one. This is the DNA repair machinery, and one of the reasons it's able to act is because when the wrong bases pair up, it recognises that the DNA 'zip' isn't done up properly.

The cell puts a huge amount of energy into keeping the DNA copies completely faithful to the original template. This makes

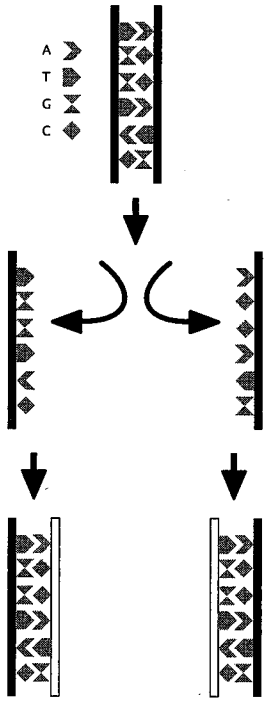

Figure 3.2 The first stage in replication of DNA is the separation of the two strands of the double helix. The bases on each separated backbone act as the template for the creation of a new strand. This ensures that the two new double-stranded DNA molecules have exactly the same base sequence as the parent molecule. Each new double helix of DNA has one backbone that was originally part of the parent molecule (in black) and one freshly synthesised backbone (in white).

sense if we go back to our model of DNA as a script. Consider one of the most famous lines in all of English literature:

O Romeo, Romeo! wherefore art thou Romeo?

If we insert just one extra letter, then no matter how well the line is delivered on stage, its effect is unlikely to be the one intended by the Bard:

O Romeo, Romeo! wherefore fart thou Romeo?

This puerile example illustrates why a script needs to be reproduced faithfully. It can be the same with our DNA – one inappropriate change (a mutation) can have devastating effects. This is particularly true if the mutation is present in an egg or a sperm, as this can ultimately lead to the birth of an individual in whom all the cells carry the mutation. Some mutations have devastating clinical effects. These range from children who age so prematurely that a ten-year-old has the body of a person of 70, to women who are pretty much predestined to develop aggressive and difficult to treat breast cancer before they are 40 years of age. Thankfully, these sorts of genetic mutations and conditions are relatively rare compared with the types of diseases that afflict most people.

The 50,000,000,000,000 or so cells in a human body are all the result of perfect replication of DNA, time after time after time, whenever cells divide after the formation of that single-cell zygote from Chapter 1. This is all the more impressive when we realise just how much DNA has to be reproduced each time one cell divides to form two daughter cells. Each cell contains six billion base-pairs of DNA (half originally came from your father and half from your mother). This sequence of six billion base-pairs is what we call the genome. So every single cell division in the human body was the result of copying 6,000,000,000 bases of DNA. Using the same type of calculation as in Chapter 1, if we count one base-pair every second without stopping, it would take a mere 190 years to count all the bases in the genome of a cell. When we consider that a baby is born just nine months after the creation of the single-celled zygote, we can see that our cells must be able to replicate DNA really fast.

The three billion base-pairs we inherit from each parent aren't formed of one long string of DNA. They are arranged into smaller bundles, which are the chromosomes. We'll delve deeper into these in Chapter 9.

Reading the script

Let's go back to the more fundamental question of what these six billion base-pairs of DNA actually do, and how the script works. More specifically how can a code that only has four letters (A, C, G and T) create the thousands and thousands of different proteins found in our cells? The answer is surprisingly elegant. It could be described as the modular paradigm of molecular biology but it's probably far more useful to think of it as Lego.

Lego used to have a great advertising slogan 'It's a new toy every day', and it was very accurate. A large box of Lego contains a limited number of designs, essentially a fairly small range of bricks of certain shapes, sizes and colours. Yet it's possible to use these bricks to create models of everything from ducks to houses, and from planes to hippos. Proteins are rather like that. The 'bricks' in proteins are quite small molecules called amino acids, and there are twenty standard amino acids (different Lego bricks) in our cells. But these twenty amino acids can be joined together in an incredible array of combinations of all sorts of diversity and length, to create an enormous number of proteins.

That still leaves the problem of how even as few as twenty amino acids can be encoded by just four bases in DNA. The way this works is that the cell machinery 'reads' DNA in blocks of three base-pairs at a time. Each block of three is known as a codon and may be AAA, or GCG or any other combination of A, C, G and T. From just four bases it's possible to create sixty-four different codons, more than enough for the twenty amino acids. Some amino acids are coded for by more than one codon. For example, the amino acid called lysine is coded for by AAA and AAG. A few codons don't code for amino acids at all. Instead they act as signals to tell the cellular machinery that it's at the end of a protein-coding sequence. These are referred to as stop codons.

How exactly does the DNA in our chromosomes act as a script for producing proteins? It does it through an intermediary

protein, a molecule called messenger RNA (mRNA). mRNA is very like DNA although it does differ in a few significant details. Its backbone is slightly different from DNA (hence RNA, which stands for ribonucleic acid rather than deoxyribonucleic acid); it is single-stranded (only one backbone); it replaces the T base with a very similar but slightly different one called U (we don't need to go into the reason it does this here). When a particular DNA stretch is 'read' so that a protein can be produced using that bit of script, a huge complex of proteins unzips the right piece of DNA and makes mRNA copies. The complex uses the base-pairing principle to make perfect mRNA copies. The mRNA molecules are then used as temporary templates at specialised structures in the cell that produce protein. These read the three letter codon code and stitch together the right amino acids to form the longer protein chains. There is of course a lot more to it than all this, but that's probably sufficient detail.

An analogy from everyday life may be useful here. The process of moving from DNA to mRNA to protein is a bit like controlling an image from a digital photograph. Let's say we take a photograph on a digital camera of the most amazing thing in the world. We want other people to have access to the image, but we don't want them to be able to change the original in any way. The raw data file from the camera is like the DNA blueprint. We copy it into another format, that can't be changed very much – a PDF maybe – and then we email out thousands of copies of this PDF, to everyone who asks for it. The PDF is the messenger RNA. If people want to, they can print paper copies from this PDF, as many as they want, and these paper copies are the proteins. So everyone in the world can print the image, but there is only one original file.

Why so complicated, why not just have a direct mechanism? There are a number of good reasons that evolution has favoured this indirect method. One of them is to prevent damage to the script, the original image file. When DNA is unzipped it is

relatively susceptible to damage and that's something that cells have evolved to avoid. The indirect way in which DNA codes for proteins minimises the period of time for which a particular stretch of DNA is open and vulnerable. The other reason this indirect method has been favoured by evolution is that it allows a lot of control over the amount of a specific protein that's produced, and this creates flexibility.

Consider the protein called alcohol dehydrogenase (ADH). This is produced in the liver and breaks down alcohol. If we drink a lot of alcohol, the cells of our livers will increase the amounts of ADH they produce. If we don't drink for a while, the liver will produce less of this protein. This is one of the reasons why people who drink frequently are better able to tolerate the immediate effects of alcohol than those who rarely drink, who will become tipsy very quickly on just a couple of glasses of wine. The more often we drink alcohol, the more ADH protein our livers produce (up to a limit). The cells of the liver don't do this by increasing the number of copies of the *ADH* gene. They do this by reading the *ADH* gene more efficiently, i.e. producing more mRNA copies and/or by using these mRNA copies more efficiently as protein templates.

As we shall see, epigenetics is one of the mechanisms a cell uses to control the amount of a particular protein that is produced, especially by controlling how many mRNA copies are made from the original template.

The last few paragraphs have all been about how genes encode proteins. How many genes are there in our cells? This seems like a simple question but oddly enough there is no agreed figure on this. This is because scientists can't agree on how to define a gene. It used to be quite straightforward – a gene was a stretch of DNA that encoded a protein. We now know that this is far too simplistic. However, it's certainly true to say that all proteins are encoded by genes, even if not all genes encode proteins. There are about 20,000 to 24,000 protein-encoding genes in our DNA, a

much lower estimate than the 100,000 that scientists thought was a good guess just ten years ago[1].

Editing the script

Most genes in human cells have quite a similar structure. There's a region at the beginning called the promoter, which binds the protein complexes that copy the DNA to form mRNA. The protein complexes move along through what's known as the body of the gene, making a long mRNA strand, until they finally fall off at the end of the gene.

Imagine a gene body that is 3,000 base-pairs long, a perfectly sensible length for a gene. The mRNA will also be 3,000 base-pairs long. Each amino acid is encoded by a codon composed of three bases, so we would predict that this mRNA will encode a protein that is 1,000 amino acids long. But, perhaps unexpectedly, what we find is that the protein is usually considerably shorter than this.

If the sequence of a gene is typed out it looks like a long string of combinations of the letters A, C, G and T. But if we analyse this with the right software, we find that we can divide that long string into two types of sequences. The first type is called an exon (for *ex*pressed sequence) and an exon can code for a run of amino acids. The second type is called an intron (for *in*expressed sequence). This doesn't code for a run of amino acids. Instead it contains lots of the 'stop' codons that signal that the protein should come to an end.

When the mRNA is first copied from the DNA it contains the whole run of exons and introns. Once this long RNA molecule has been created, another multi-sub-unit protein complex comes along. It removes all the intron sequences and then joins up the exons to create an mRNA that codes for a continuous run of amino acids. This editing process is called splicing.

This again seems extremely complicated, but there's a very good reason that this complex mechanism has been favoured by

evolution. It's because it enables a cell to use a relatively small number of genes to create a much bigger number of proteins. The way this works is shown in Figure 3.3.

The initial mRNA contains all the exons and all the introns. Then it's spliced to remove the introns. But during this splicing some of the exons may also be removed. Some exons will be retained in the final mRNA, others will be skipped over. The various proteins that this creates may have quite similar functions, or

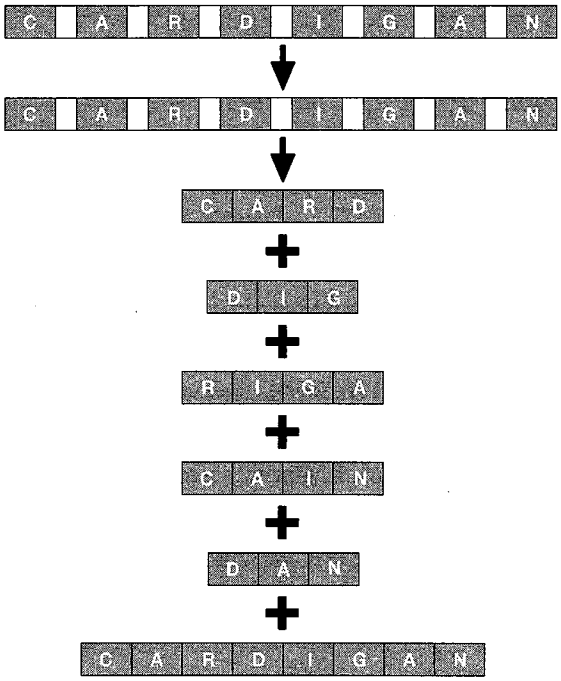

Figure 3.3 The DNA molecule is shown at the very top of this diagram. The exons, which code for stretches of amino acids, are shown in the dark boxes. The introns, which don't code for amino acid sequences, are represented by the white boxes. When the DNA is first copied into RNA, indicated by the first arrow, the RNA contains both the exons and the introns. The cellular machinery then removes some or all of the introns (the process known as splicing). The final messenger RNA molecules can thereby code for a variety of proteins from the same gene, as represented by the various words shown in the diagram. For simplicity, all the introns and exons have been drawn as the same size, but in reality they can vary widely.

they may differ dramatically. The cell can express different proteins depending on what that cell has to do at a particular time, or because of different signals that it receives. If we define a gene as something that encodes a protein, this mechanism means that just 20,000 or so genes can code for far more than just 20,000 proteins.

Whenever we describe the genome we talk about it in very two-dimensional terms, almost like a railway track. Peter Fraser's laboratory at the Babraham Institute outside Cambridge has published some extraordinary work showing it's probably nothing like this at all. He works on the genes that code for the proteins required to make haemoglobin, the pigment in red blood cells that carries oxygen all around the body. There are a number of different proteins needed to create the final pigment, and they lie on different chromosomes. Doctor Fraser has shown that in cells that produce large amounts of haemoglobin, these chromosome regions become floppy and loop out like tentacles sticking out of the body of an octopus. These floppy regions mingle together in a small area of the cell nucleus, waving about until they can find each other. By doing this, there is an increased chance that all the proteins needed to create the functional haemoglobin pigment will be expressed together at the same time[2].

Each cell in our body contains 6,000,000,000 base-pairs. About 120,000,000 of these code for proteins. One hundred and twenty million sounds like a lot, but it's actually only 2 per cent of the total amount. So although we think of proteins as being the most important things our cells produce, about 98 per cent of our genome doesn't code for protein.

Until recently, the reason that we have so much DNA when so little of it leads to a protein was a complete mystery. In the last ten years we've finally started to get a grip on this, and once again it's connected with regulating gene expression through epigenetic mechanisms. It's now time to move on to the molecular biology of epigenetics.

Chapter 4

Life As We Know It Now

The important thing in science is not so much to obtain new facts as to discover new ways of thinking about them.
Sir William Bragg

So far this book has focused mainly on outcomes, the things that we can observe that tell us that epigenetic events happen. But every biological phenomenon has a physical basis and that's what this chapter is about. The epigenetic outcomes we've described are all a result of variations in expression of genes. The cells of the retina express a different set of genes from the cells in the bladder, for example. But how do the different cell types switch different sets of genes on or off?

The specialised cell types in the retina and in the bladder are each at the bottom of one of the troughs in Waddington's epigenetic landscape. The work of both John Gurdon and Shinya Yamanaka showed us that whatever mechanism cells use for staying in these troughs, it's not anything to do with changing the DNA blueprint of the cell. That remains intact and unchanged. Therefore keeping specific sets of genes turned on or off must happen through some other mechanism, one that can be maintained for a really long time. We know this must be the case because some cells, like the neurons in our brains, are remarkably long-lived. The neurons in the brain of an 85-year-old person, for example, are about 85 years of age. They formed when the individual was very young, and then stayed the same for the rest of their life.

But other cells are different. The top layer of skin cells, the epidermis, is replaced about every five weeks, from constantly dividing stem cells in the deeper layers of that tissue. These stem cells

always produce new skin cells, and not, for example, muscle cells. Therefore the system that keeps certain sets of genes switched on or off must also be a mechanism that can be passed on from parent cell to daughter cell every time there is a cell division.

This creates a paradox. Researchers have known since the work of Oswald Avery and colleagues in the mid-1940s that DNA is the material in cells that carries our genetic information. If the DNA stays the same in different cell types in one individual, how can the incredibly precise patterns of gene expression be transmitted down through the generations of cell division?

Our analogy of actors reading a script is again useful. Baz Luhrmann hands Leonardo DiCaprio Shakespeare's script for *Romeo and Juliet*, on which the director has written or typed various notes – directions, camera placements and lots of additional technical information. Whenever Leo's copy of the script is photocopied, Baz Luhrmann's additional information is copied along with it. Claire Danes also has the script for *Romeo and Juliet*. The notes on her copy are different from those on her co-star's, but will also survive photocopying. That's how epigenetic regulation of gene expression occurs – different cells have the same DNA blueprint (the original author's script) but carrying varied molecular modifications (the shooting script) which can be transmitted from mother cell to daughter cell during cell division.

These modifications to DNA don't change the essential nature of the A, C, G and T alphabet of our genetic script, our blueprint. When a gene is switched on and copied to make mRNA, that mRNA has exactly the same sequence, controlled by the base-pairing rules, irrespective of whether or not the gene is carrying an epigenetic addition. Similarly, when the DNA is copied to form new chromosomes for cell division, the same A, C, G and T sequences are copied.

Since epigenetic modifications don't change what a gene codes for, what do they do? Basically, they can dramatically change how well a gene is expressed, or if it is expressed at all. Epigenetic

modifications can also be passed on when a cell divides, so this provides a mechanism for how control of gene expression stays consistent from mother cell to daughter cell. That's why skin stem cells only give rise to more skin cells, not to any other cell type.

Sticking a grape on DNA

The first epigenetic modification to be identified was DNA methylation. Methylation means the addition of a methyl group to another chemical, in this case DNA. A methyl group is very small. It's just one carbon atom linked to three hydrogen atoms. Chemists describe atoms and molecules by their 'molecular weight', where the atom of each element has a different weight. The average molecular weight of a base-pair is around 600 Da (the Da stands for Daltons, the unit that is used for molecular weight). A methyl group only weighs 15 Da. By adding a methyl group the weight of the base-pair is only increased by 2.5 per cent. A bit like sticking a grape on a tennis ball.

Figure 4.1 shows what DNA methylation looks like chemically.

The base shown is C – cytosine. It's the only one of the four DNA bases that gets methylated, to form 5-methylcytosine. The '5' refers to the position on the ring where the methyl is added, not to the number of methyl groups; there's always only one of these. This methylation reaction is carried out in our cells, and

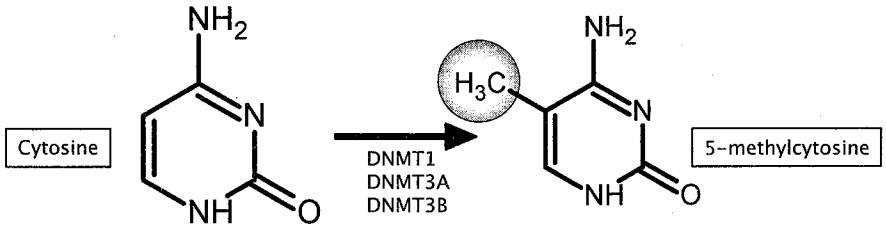

Figure 4.1 The chemical structures of the DNA base cytosine and its epigenetically modified form, 5-methylcytosine. C: carbon; H: hydrogen; N: nitrogen; O: oxygen. For simplicity, some carbon atoms have not been explicitly shown, but are present where there is a junction of two lines.

those of most other organisms, by one of three enzymes called DNMT1, DNMT3A or DNMT3B. DNMT stands for DNA methyltransferase. The DNMTs are examples of epigenetic 'writers' – enzymes that create the epigenetic code. Most of the time these enzymes will only add a methyl group to a C that is followed by a G. C followed by G is known as CpG.

This CpG methylation is an epigenetic modification, which is also known as an epigenetic mark. The chemical group is 'stuck onto' DNA but doesn't actually alter the underlying genetic sequence. The C has been decorated rather than changed. Given that the modification is so small, it's perhaps surprising that it will come up over and over again in this book, and in any discussion of epigenetics. This is because methylation of DNA has profound effects on how genes are expressed, and ultimately on cellular, tissue and whole-body functions.

In the early 1980s it was shown that if you injected DNA into mammalian cells, the amount of methylation on the injected DNA affected how well it was transcribed into RNA. The more methylated the injected DNA was, the less transcription that occurred[1]. In other words, high levels of DNA methylation were associated with genes that were switched off. However, it wasn't clear how significant this was for the genes normally found in the nuclei of cells, rather than ones that were injected into cells.

The key work in establishing the importance of methylation in mammalian cells came out of the laboratory of Adrian Bird, who has spent most of his scientific career in Edinburgh, Conrad Waddington's old stomping ground. Professor Bird is a Fellow of the Royal Society and a former Governor of the Wellcome Trust, the enormously influential independent funding agency in UK science. He is one of those traditional British scientific types – understated, soft-spoken, non-flashy and drily funny. His lack of self-promotion is in contrast to his stellar international reputation, where he is widely acknowledged as the godfather of DNA methylation and its role in controlling gene expression.

In 1985 Adrian Bird published a key paper in *Cell* showing that most CpG motifs were not randomly distributed throughout the genome. Instead the majority of CpG pairs were concentrated just upstream of certain genes, in the promoter region[2]. Promoters are the stretches of the genome where the DNA transcription complexes bind and start copying DNA to form RNA. Regions where there is a high concentration of CpG motifs are called CpG islands.

In about 60 per cent of the genes that code for proteins, the promoters lie within CpG islands. When these genes are active, the levels of methylation in the CpG island are low. The CpG islands tend to be highly methylated only when the genes are switched off. Different cell types express different genes, so unsurprisingly the patterns of CpG island methylation are also different across different cell types.

For quite some time there was considerable debate about what this association meant. It was the old cause or effect debate. One interpretation was that DNA methylation was essentially a historical modification – genes were repressed by some unknown mechanism and then the DNA became methylated. In this model, DNA methylation was just a downstream consequence of gene repression. The other interpretation was that the CpG island became methylated, and it was this methylation that switched the gene off. In this model the epigenetic modification actually causes the change in gene expression. Although there is still the occasional argument about this between competing labs, the vast majority of scientists in this field now believe that the data generated in the quarter of a century since Adrian Bird's paper are consistent with the second, causal model. Under most circumstances, methylation of the CpG island at the start of a gene turns that gene off.

Adrian Bird went on to investigate how DNA methylation switches genes off. He showed that when DNA is methylated, it binds a protein called MeCP2 (Methyl CpG binding protein 2)[3].

However, this protein won't bind to unmethylated CpG motifs, which is pretty amazing when we look back at Figure 4.1 and think how similar the methylated and unmethylated forms of cytosine really are. The enzymes that add the methyl group to DNA have been described as writers of the epigenetic code. MeCP2 doesn't add any modifications to DNA. Its role is to enable the cell to interpret the modifications on a DNA region. MeCP2 is an example of a 'reader' of the epigenetic code.

Once MeCP2 binds to 5-methylcytosine in a gene promoter it seems to do a number of things. It attracts other proteins that also help to switch the gene off[4]. It may also stop the DNA transcription machinery from binding to the gene promoter, and this prevents mRNA messenger molecule from being produced[5]. Where genes and their promoters are very heavily methylated, binding of MeCP2 seems to be part of a process where that region of a chromosome gets shut down almost permanently. The DNA becomes incredibly tightly coiled up and the gene transcription machinery can't get access to the base-pairs to make mRNA copies.

This is one of the reasons why DNA methylation is so important. Remember those 85 year old neurons in the brains of senior citizens? For over eight decades DNA methylation has kept certain regions of the genome incredibly tightly compacted and so the neuron has kept certain genes completely repressed. This is why our brain cells never produce haemoglobin, for example, or digestive enzymes.

But what about the other situation, the example of skin stem cells dividing very frequently but always just creating new skin cells, rather than some other cell type such as bone? In this situation, the pattern of DNA methylation is passed from mother cell to daughter cells. When the two strands of the DNA double helix separate, each gets copied using the base-pairing principle, as we saw in Chapter 3. Figure 4.2 illustrates what happens when this replication occurs in a region where the CpG is methylated on the C.

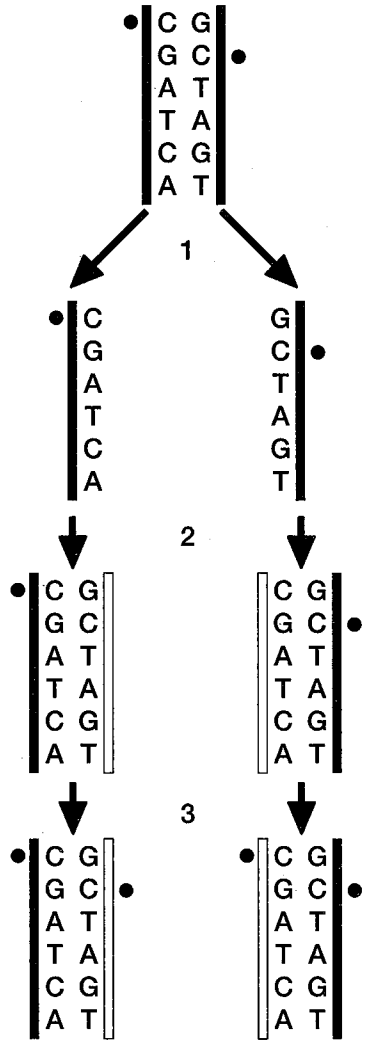

Figure 4.2 This schematic shows how DNA methylation patterns can be preserved when DNA is replicated. The methyl group is represented by the black circle. Following separation of the parent DNA double helix in step 1, and replication of the DNA strands in step 2, the new strands are 'checked' by the DNA methyltransferase 1 (DNMT1) enzyme. DNMT1 can recognise that a methyl group at a cytosine motif on one strand of a DNA molecule is not matched on the newly synthesised strand. DNMT1 transfers a methyl group to the cytosine on the new strand (step 3). This only occurs where a C and a G are next to each other in a CpG motif. This process ensures that the DNA methylation patterns are maintained following DNA replication and cell division.

DNMT1 can recognise if a CpG motif is only methylated on one strand. When DNMT1 detects this imbalance, it replaces the 'missing' methylation on the newly copied strand. The daughter cells will therefore end up with the same DNA methylation patterns as the parent cell. As a consequence, they will repress the same genes as the parent cell and the skin cells will stay as skin cells.

Miracle mice on *YouTube*

Epigenetics has a tendency to crop up in places where scientists really aren't expecting it. One of the most interesting examples of this in recent years has related to MeCP2, the protein that reads the DNA methylation mark. Several years ago, the now discredited theory of the MMR vaccine causing autism was at its height, and getting lots of coverage in the general media. One very respected UK broadsheet newspaper covered in depth the terribly sad story of a little girl. As a baby she initially met all the usual developmental milestones. Shortly after receiving an MMR jab not long before her first birthday she began to deteriorate rapidly, losing most of the skills she had gained. By the time the journalist wrote the article, the little girl was about four years old and was described as having the most severely autistic symptoms the author had ever seen. She had not developed language, appeared to have very severe learning difficulties and her actions were very limited and repetitive, with very few purposeful hand actions (she no longer reached out for food, for example). Development of this incredibly severe disability was undoubtedly a tragedy for her and for her family.

But if a reader with any sort of background in neurogenetics read this article, two things probably struck them immediately. The first was that it's very unusual – not unheard of but pretty uncommon – for girls to present with such severe autism. This is much more common in boys. The second thing that would have

struck them was that this case sounded exactly the same as a rare genetic disorder called Rett syndrome, right down to the normal early development and the timing and types of symptoms. It's just coincidence that the symptoms of Rett syndrome, and indeed of most types of autism, first start becoming obvious at around the same age as when infants are typically given the MMR vaccination.

But what does this have to do with epigenetics? In 1999, a group led by the eminent neurogeneticist Huda Zoghbi at the Howard Hughes Medical Institute in Maryland showed that the majority of cases of Rett syndrome are caused by mutations in *MeCP2*, the gene which encodes the reader of methylated DNA. The children with this disorder have a mutation in the *MeCP2* gene which means that they don't produce a functional MeCP2 protein. Although their cells are perfectly capable of methylating DNA correctly, the cells can't read this part of the epigenetic code properly.

The severe clinical symptoms of children with the *MeCP2* mutation tell us that reading the epigenetic code properly is very important. But they also tell us other things. Not all the tissues of girls with Rett syndrome are equally affected, so perhaps this particular epigenetic pathway is more important in some tissues than others. Because the girls develop severe mental retardation, we can deduce that having the right amount of normal MeCP2 protein is really important in the brain. Given that these children seem to be fairly unaffected in other tissues such as liver or kidney, perhaps MeCP2 activity isn't as important in these tissues. It could be that DNA methylation itself isn't so critical in these organs, or maybe these tissues contain other proteins in addition to MeCP2 that can read this part of the epigenetic code.

Long-term, scientists, physicians and families of children with Rett syndrome would dearly love to be able to use our increased understanding of the disease to help us find better treatments. This is a huge challenge, as we would be trying to intervene in a

condition that affects the brain as a result of a gene mutation that is present throughout development, and beyond.

One of the most debilitating aspects of Rett syndrome is the profound mental retardation that is an almost universal symptom. Nobody knew if it would be possible to reverse a neurodevelopmental problem such as mental retardation once it had become established, but the general feeling about this wasn't optimistic. Adrian Bird remains a major figure in our story. In 2007 he published an astonishing paper in *Science*, in which he and his colleagues showed that Rett syndrome could be reversed, in a mouse model of the disease.

Adrian Bird and his colleagues created a cloned strain of mice in which the *Mecp2* gene was inactivated. They used the types of technologies pioneered by Rudolf Jaenisch. These mice developed severe neurological symptoms, and as adults they exhibited hardly any normal mouse activities. If you put a normal mouse in the middle of a big white box, it will almost immediately begin to explore its surroundings. It will move around a lot, it will tend to follow the edges of the box just like a normal house mouse scurrying along by the skirting boards, and it will frequently rear up on its back legs to get a better view. A mouse with the *Mecp2* mutation does very few of these things – put it in the middle of a big white box and it will tend to stay there.

When Adrian Bird created his mouse strain with the *Mecp2* mutation, he also engineered it so that the mice would also be carrying a normal copy of *Mecp2*. However, this normal copy was silent – it wasn't switched on in the mouse cells. The really clever bit of this experiment was that if the mice were given a specific harmless chemical, the normal *Mecp2* gene became activated. This allowed the experimenters to let the mice develop and grow up with no Mecp2 in their cells, and then at a time of the scientists' choosing, the *Mecp2* gene could be switched on.

The results of switching on the *Mecp2* gene were extraordinary. Mice which previously just sat in the middle of the white

box suddenly turned into the curious explorers that mice should be[6]. You can find clips of this on *YouTube*, along with interviews with Adrian Bird where he basically concedes that he really never expected to see anything so dramatic[7].

The reason this experiment is so important is that it offers hope that we may be able to find new treatments for really complex neurological conditions. Prior to the publication of this *Science* paper, there had been an assumption that once a complex neurological condition has developed, it is impossible to reverse it. This was especially presumed to be the case for any condition that arises developmentally, i.e. in the womb or in early infancy. This is a critical period when the mammalian brain is making so many of the connections and structures that are used throughout the rest of life. The results from the *Mecp2* mutant mice suggest that in Rett syndrome, maybe all the bits of cellular machinery that are required for normal neurological function are still there in the brain – they just need to be activated properly. If this holds true for humans (and at a brain level we aren't really *that* different from mice) this offers hope that maybe we can start to develop therapies to reverse conditions as complex as mental retardation. We can't do this the way it was done in the mouse, as that was a genetic approach that can only be used in experimental animals and not in humans, but it suggests that it is worth trying to develop suitable drugs that have a similar effect.

DNA methylation is clearly really important. Defects in reading DNA methylation can lead to a complex and devastating neurological disorder that leaves children with Rett syndrome severely disabled throughout their lives. DNA methylation is also essential for maintaining the correct patterns of gene expression in different cell types, either for several decades in the case of our long-lived neurons, or in all daughters of a stem cell in a constantly-replaced tissue such as skin.

But we still have a conceptual problem. Neurons are very different from skin cells. If both cells types use DNA methylation to

switch off certain genes, and to keep them switched off, they must be using the methylation at different sets of genes. Otherwise they would all be expressing the same genes, to the same extent, and they would inevitably then be the same types of cells instead of being neurons and skin cells.

The solution to how two cell types can use the same mechanism to create such different outcomes lies in how DNA methylation gets targeted to different regions of the genome in different cell types. This takes us into the second great area of molecular epigenetics. Proteins.