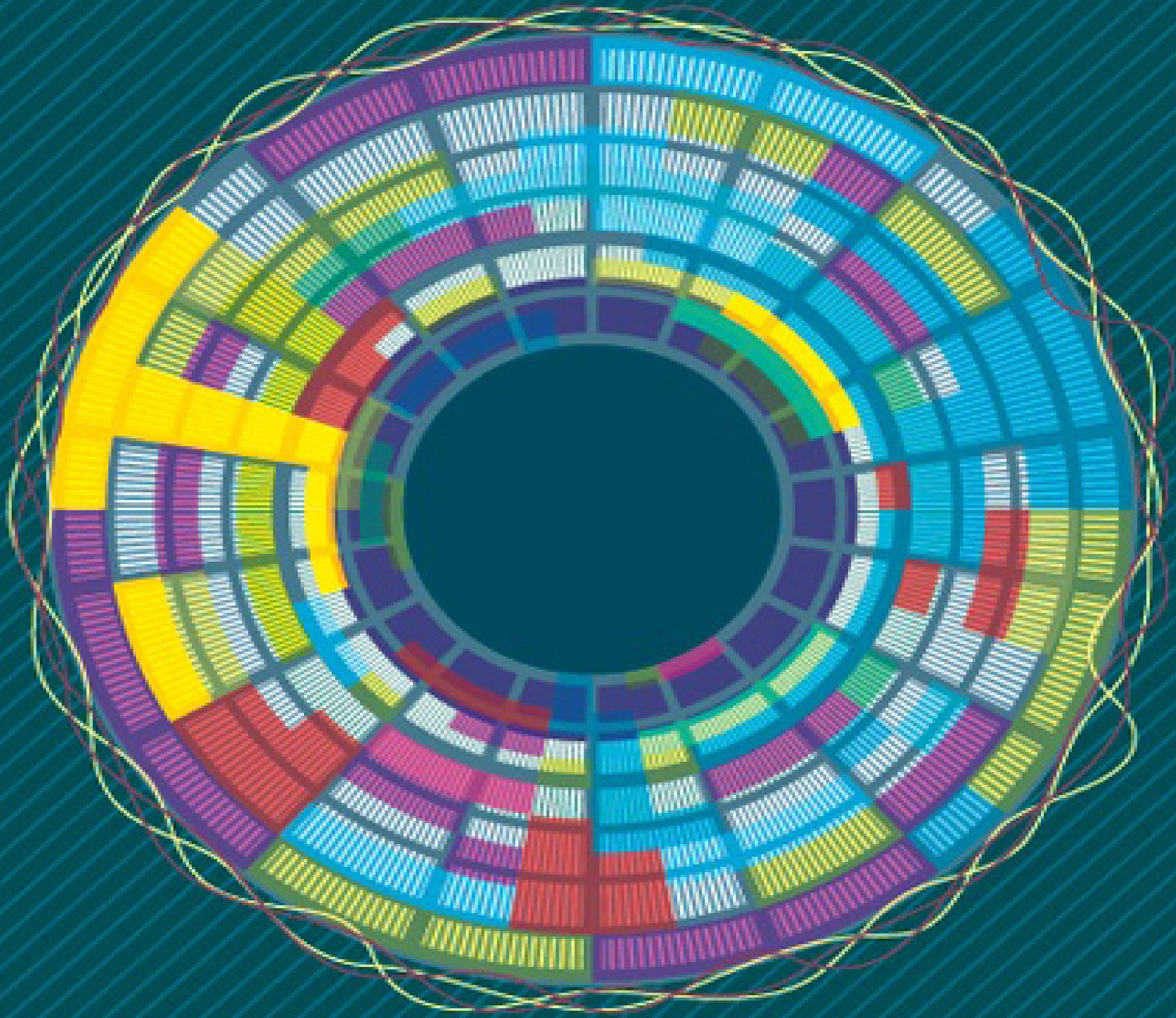
Author of *The Epigenetics Revolution*

NESSA CAREY



JUNK DNA

A Journey Through the Dark Matter of the Genome

JUNK
DNA

Also by Nessa Carey

The Epigenetics Revolution

JUNK DNA

A Journey Through the Dark Matter of the Genome

NESSA CAREY

ICON

Published in the UK in 2015 by
Icon Books Ltd, Omnibus Business Centre,
39–41 North Road, London N7 9DP
email: info@iconbooks.com
www.iconbooks.com

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

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

Distributed in Australia and New Zealand
by Allen & Unwin Pty Ltd,
PO Box 8500, 83 Alexander Street,
Crows Nest, NSW 2065

Distributed in South Africa by
Jonathan Ball, Office B4, The District,
41 Sir Lowry Road, Woodstock 7925

Distributed in India by Penguin Books India,
7th Floor, Infinity Tower – C, DLF Cyber City,
Gurgaon 122002, Haryana

ISBN: 978-184831-826-7

Typeset in Janson Text by Marie Doherty

Printed and bound in the UK
by Clays Ltd, St Ives plc

For Abi Reynolds, who is always by my side
And for Sheldon – good to see you again

Contents

Acknowledgements

I am lucky that for my second book I continue to have the support of a great agent, Andrew Lownie, and of lovely publishers. At Icon Books I'd particularly like to thank Duncan Heath, Andrew Furlow and Robert Sharman, but not forgetting their former colleagues Simon Flynn and Henry Lord. At Columbia University Press I'm very grateful to Patrick Fitzgerald, Bridget Flannery-McCoy and Derek Warker.

As always, entertainment and enlightenment have been obtained from some unusual quarters. Conor Carey, Finn Carey and Gabriel Carey all played a role in this, and outside the genetic clan I'd also like to thank Iona Thomas-Wright. Endless support and lots of biscuits have been provided by my ever-patient, delightful mother-in-law, Lisa Doran.

I've had a blast delivering lots of science talks to non-specialist audiences since my first book was published. The various organisations that have invited me to speak are too many to namecheck but they know who they are and I've enjoyed the privilege immensely. It's been very inspiring. Thank you all.

And finally Abi. Who is mercifully forgiving of the fact that, despite my promises, I still haven't had that ballroom dancing lesson yet.

Notes on Nomenclature

There's a bit of a linguistic difficulty in writing a book on junk DNA, because it is a constantly shifting term. This is partly because new data change our perception all the time. Consequently, as soon as a piece of junk DNA is shown to have a function, some scientists will say (logically enough) that it's not junk. But that approach runs the risk of losing perspective on how radically our understanding of the genome has changed in recent years.

Rather than spend time trying to knit a sweater with this ball of fog, I have adopted the most hard-line approach. Anything that doesn't code for protein will be described as junk, as it originally was in the old days (second half of the twentieth century). Purists will scream, and that's OK. Ask three different scientists what they mean by the term 'junk', and we would probably get four different answers. So there's merit in starting with something straightforward.

I also start by using the term 'gene' to refer to a stretch of DNA that codes for a protein. This definition will evolve through the course of the book.

After my first book *The Epigenetics Revolution* was published, I realised the readership was quite binary with respect to gene names. Some people love knowing which gene is being discussed, but for other readers it disrupts the flow horribly. So this time I have only used specific gene names in the text where absolutely necessary. But if you want to know them, they are in the footnotes, and the citations for the original references are at the back of the book.

An Introduction to Genomic Dark Matter

Imagine a written script for a play, or film, or television programme. It is perfectly possible for someone to read a script just as they would a book. But the script becomes so much more powerful when it is used to produce something. It becomes more than just a string of words on a page when it is spoken aloud, or better yet, acted.

DNA is rather similar. It is the most extraordinary script. Using a tiny alphabet of just four letters it carries the code for organisms from bacteria to elephants, and from brewer’s yeast to blue whales. But DNA in a test tube is pretty boring. It does nothing. DNA becomes far more exciting when a cell or an organism uses it to stage a production. The DNA is used as the code for creating proteins and these proteins are vital for breathing, feeding, getting rid of waste, reproducing and all the other activities that characterise living organisms.

Proteins are so important that in the twentieth century scientists used them to define what they meant by a gene. A gene was described as a sequence of DNA that codes for a protein.

Let’s think about the most famous scriptwriter in history, William Shakespeare. It can take a while for us to tune in to Shakespeare’s writings because of the way the English language has changed in the centuries since his death. But even so, we are always confident that the bard only wrote the words he needed his actors to speak.

Shakespeare did not, for example, write the following:

> vjeqriugfrhbvruewhqoerahcxnqowhvgbutyunyhewqicx
> hjafvurytnpemxoqp[etjhnuvrwwwebcxewmoipzowqmr
> oseuiednrcvtycuxmqpzjmoimxdcnibyrwvytebanyhcux
> qimokzqoxkmdcifwrvjhentbubygdecftywerftxunihzxqw
> emiuqwjiqpodqeotherpowhdymrxnamehnfeicvbrgytrch
> guthhhhhhgcwouldupaizmjdpqsmellmjzufernnvgbyun
> asechuxhrtgcnionytuiongdjsioniodefnionihyhoniosdren
> iokikiniourvjcxoiqweopapqsweetwxmocviknoitrbiobeie
> rrrrrrruorytnihgfiwoswakxdcjdrfuhrqplwjkdhvmogmrfb
> vhncdjiwemxsklowe

Instead, he just wrote the words which are underlined:

> vjeqriugfrhbvruewhqoerahcxnqowhvgbutyunyhewqicx
> hjafvurytnpemxoqp[etjhnuvrwwwebcxewmoipzowqmr
> oseuiednrcvtycuxmqpzjmoimxdcnibyrwvytebanyhcux
> qimokzqoxkmdcifwrvjhentbubygdecftywerftxunihzxqw
> emiuqwjiqpodqeotherpowhdymrxnamehnfeicvbrgytrch
> guthhhhhhgcwouldupaizmjdpqsmellmjzufernnvgbyun
> asechuxhrtgcnionytuiongdjsioniodefnionihyhoniosdren
> iokikiniourvjcxoiqweopapqsweetwxmocviknoitrbiobeie
> rrrrrrruorytnihgfiwoswakxdcjdrfuhrqplwjkdhvmogmrfb
> vhncdjiwemxsklowe

That is, ‘A rose by any other name would smell as sweet’.

But if we look at our DNA script it is not sensible and compact, like Shakespeare’s line. Instead, each protein-coding region is like a single word adrift in a sea of gibberish.

For years, scientists had no explanation for why so much of our DNA doesn’t code for proteins. These non-coding parts were dismissed with the term ‘junk DNA’. But gradually this position has begun to look less tenable, for a whole host of reasons.

Perhaps the most fundamental reason for the shift in emphasis is the sheer volume of junk DNA that our cells contain. One of the biggest shocks when the human genome sequence was completed in 2001 was the discovery that over 98 per cent of the DNA in a human cell is junk. It doesn’t code for any proteins. The Shakespeare analogy used above is in fact a simplification. In genome terms, the ratio of gibberish to text is about four times as high as shown. There are over 50 letters of junk for every one letter of sense.

There are other ways of envisaging this. Let’s imagine we visit a car factory, perhaps for something high-end like a Ferrari. We would be pretty surprised if for every two people who were building a shiny red sports car, there were another 98 who were sitting around doing nothing. This would be ridiculous, so why would it be reasonable in our genomes? While it’s a very fair point that it’s the imperfections in organisms that are often the strongest evidence for descent from common ancestors – we humans really don’t need an appendix – this seems like taking imperfection rather too far.

A much more likely scenario in our car factory would be that for every two people assembling a car, there are 98 others doing all the things that keep a business moving. Raising finance, keeping accounts, publicising the product, processing the pensions, cleaning the toilets, selling the cars etc. This is probably a much better model for the role of junk in our genome. We can think of proteins as the final end points required for life, but they will never be properly produced and coordinated without the junk. Two people can build a car, but they can’t maintain a company selling it, and certainly can’t turn it into a powerful and financially successful brand. Similarly, there’s no point having 98 people mopping the floors and staffing the showrooms if there’s nothing to sell. The whole organisation only works when all the components are in place. And so it is with our genomes.

The other shock from the sequencing of the human genome was the realisation that the extraordinary complexities of human anatomy, physiology, intelligence and behaviour cannot be explained by referring to the classical model of genes. In terms of numbers of genes that code for proteins, humans contain pretty much the same quantity (around 20,000) as simple microscopic worms. Even more remarkably, most of the genes in the worms have directly equivalent genes in humans.

As researchers deepened their analyses of what differentiates humans from other organisms at the DNA level, it became apparent that genes could not provide the explanation. In fact, only one genetic factor generally scaled with complexity. The only genomic features that increased in number as animals became more complicated were the regions of junk DNA. The more sophisticated an organism, the higher the percentage of junk DNA it contains. Only now are scientists really exploring the controversial idea that junk DNA may hold the key to evolutionary complexity.

In some ways, the question raised by these data is pretty obvious. If junk DNA is so important, what is it actually doing? What is its role in a cell, if it isn’t coding for proteins? It’s becoming apparent that junk DNA actually has a multiplicity of different functions, perhaps unsurprisingly given how much of it there is.

Some of it forms specific structures in the chromosomes, the enormous molecules into which our DNA is packaged. This junk prevents our DNA from unravelling and becoming damaged. As we age, these regions decrease in size, finally declining below a critical minimum. After that, our genetic material becomes susceptible to potentially catastrophic rearrangements that can lead to cell death or cancers.

Other structural regions of junk DNA act as anchor points when chromosomes are shared equally between different daughter cells during cell division. (The term ‘daughter cell’ means any cell created by division of a parental cell. It doesn’t imply that the cell is female.) Yet others act as insulation regions, restricting gene expression to specific regions of chromosomes.

But a great deal of our junk DNA is not simply structural. It doesn’t code for proteins, but it does code for a different type of molecule, called RNA. A large class of this junk DNA forms factories in the cell, helping to produce proteins. Other types of RNA molecules transport the raw material for protein production to the factory sites.

Other regions of junk DNA are genetic interlopers, derived from the genomes of viruses and other microorganisms that have integrated into human chromosomes, like genetic sleeper agents. These remnants of long-dead organisms carry potential dangers to the cell, the individual and sometimes even to wider populations. Mammalian cells have developed multiple mechanisms to keep these viral elements silent, but these systems can break down. When they do, the effects can range from relatively benign – changing the coat colour of a particular strain of mice – to much more dramatic, such as an increased risk of cancer.

A major role of junk DNA, only recognised in the main in the last few years, is to regulate gene expression. Sometimes this can have a huge and noticeable effect in an individual. One particular piece of junk DNA is absolutely vital for ensuring healthy gene expression patterns in female animals. Its effects are seen in a whole range of situations. A mundane example is the control of the colour patterns of tortoiseshell cats. At its most extreme, the same mechanism also explains why female identical twins may present with different symptoms of a genetically inherited disease. In some cases, this can be so extreme that one twin is severely affected with a life-threatening disorder while the other is completely healthy.

Thousands and thousands of regions of junk DNA are suspected to regulate networks of gene expression. They act like the stage directions for the genetic script, but directions of a complexity we could never envisage in the theatre. Forget about ‘Exit, pursued by a bear’. These would be more along the lines of ‘If performing *Hamlet* in Vancouver and *The Tempest* in Perth, then put the stress on the fourth syllable of this line of *Macbeth*. Unless there’s an amateur production of *Richard III* in Mombasa and it’s raining in Quito.’

Researchers are only just beginning to unravel the subtleties and interconnections in the vast networks of junk DNA. The field is controversial. At one extreme we have scientists claiming experimental proof is lacking to support sometimes sweeping claims. At the other are those who feel there is a whole generation of scientists (if not more) trapped in an outdated model and unable to see or understand the new world order.

Part of the problem is that the systems we can use to probe the functions of junk DNA are still relatively underdeveloped. This can sometimes make it hard for researchers to use experimental approaches to test their hypotheses. We have only been working on this for a relatively short space of time. But sometimes we need to remember to step back from the lab bench and the machines that go ping. Experiments surround us every day, because nature and evolution have had billions of years to try out all sorts of changes. Even the brief geological moment that represents the emergence and spread of our own species has been sufficient time to create a greater range of experiments than those of us who wear lab coats could ever dream of testing. Consequently, throughout much of this book we will explore the darkness by using the torch of human genetics.

There are many ways to begin shining a light on the dark matter of our genome, so let’s start with an odd but unassailable fact to anchor us. Some genetic diseases are caused by mutations in junk DNA, and there is probably no better starting point for our journey into the hidden genomic universe than this.

1. Why Dark Matter Matters

Sometimes life seems to be cruel in the troubles it piles onto a family. Consider this example. A baby boy was born; let’s call him Daniel. He was strangely floppy at birth, and had trouble breathing unassisted. With intensive medical care Daniel survived and his muscle tone improved, allowing him to breathe unaided and to develop mobility. But as he grew older it became apparent that Daniel had pronounced learning disabilities that would hold him back throughout life.

His mother Sarah loved Daniel and cared for him every day. As she entered her mid-30s this became more difficult because Sarah developed strange symptoms. Her muscles became very stiff, to the extent that she would have trouble releasing items after grasping them. She had to give up her highly skilled part-time job as a ceramics restorer. Her muscles also began to waste away noticeably. Yet she found ways to cope. But when she was only 42 years old Sarah died suddenly from a cardiac arrhythmia, a catastrophic disruption in the electrical signals that keep the heart beating in a coordinated way.

It fell to Sarah’s mother, Janet, to look after Daniel. This was challenging for her, and not just because of her grandson’s difficulties and the grief she was suffering over the early death of her daughter. Janet had developed cataracts in her early 50s and as a consequence her vision wasn't that great.

It seemed as if the family had suffered a very unfortunate combination of unrelated medical problems. But specialists began to notice something rather unusual. This pattern – cataracts in one individual, muscle stiffness and cardiac defects in their daughter and floppy muscles and learning disabilities in the grandchildren – occurred in multiple families. These individual families lived all over the world and none of them were related to each other.

Scientists realised they were looking at a genetic disease. They named it myotonic dystrophy (myotonic means muscle tone, dystrophy means wasting). The condition occurred in every generation of an affected family. On average there was a one in two chance of a child being affected if their parent had the condition. Males and females were equally at risk and either could pass it on to their children.[1]

These inheritance characteristics are very typical of diseases caused by mutations in a single gene. A mutation is simply a change from the normal DNA sequence. We typically inherit two copies of every gene in our cells, one from our mother and one from our father. The pattern of inheritance in myotonic dystrophy, where the disease appears in each generation, is referred to as dominant. In dominant disorders, only one of the two copies of a gene carries the mutation. It is the copy inherited from the affected parent. This mutated gene is able to cause the disease even though the cells also contain a normal copy. The mutated gene somehow 'dominates' the action of the normal gene.

But myotonic dystrophy also had characteristics that were very different from a typical dominant disorder. For a start, dominant disorders don't normally get worse as they are passed on from parent to child. There is no reason why they should, because the affected child inherits the same mutation as the affected parent. Patients with myotonic dystrophy also developed symptoms at earlier ages as the disorder was passed on down the generations, which again is unusual.

There was another way in which myotonic dystrophy was different from the normal genetic pattern. The severe congenital form of the disease, the one that affected Daniel, was only ever found in the children of affected mothers. Fathers never passed on this really severe form.

In the early 1990s a number of different research groups identified the genetic change that causes myotonic dystrophy. Fittingly for an unusual disease, it was a very unusual mutation. The myotonic dystrophy gene contains a small sequence of DNA that is repeated multiple times.[2] The small sequence is made from three of the four ‘letters’ that make up the genetic alphabet used by DNA. In the myotonic dystrophy gene, this repeated sequence is formed by the letters C, T and G (the other letter in the genetic alphabet is A).

In people without the myotonic dystrophy mutation, there can be anything from five to around 30 copies of this CTG motif, one after the other. Children inherit the same number of repeats as their parents. But when the number of repeats gets larger, greater than 35 or thereabouts, the sequence becomes a bit unstable and may change in number when it is passed on from parent to child. Once it gets above 50 copies of the motif, the sequence becomes really unstable. When this happens, parents can pass on much bigger repeats to their children than they themselves possess. As the repeat length increases, the symptoms become more severe and are obvious at an earlier age. That’s why the disease gets worse as it passes down the generations, such as in the family that opened this chapter. It also became apparent that usually only mothers passed on the really big repeats, the ones that led to the severe congenital phenotype.

This ongoing expansion of a repeated sequence of DNA was a very unusual mutation mechanism. But the identification of the expansion that causes myotonic dystrophy shone a light on something even more unusual.

Knitting with DNA

Until quite recently, mutations in gene sequences were thought to be important not because of the change in the DNA itself but because of their downstream consequences. It’s a little like a mistake in a knitting pattern. The mistake doesn’t matter when it’s just a notation on a piece of paper. The mistake only becomes a problem when you knit something and end up with a hole in your sweater or three sleeves on your cardigan because of the error in the knitting code.

A gene (the knitting pattern) ultimately codes for a protein (the sweater). It’s proteins that we think of as the molecules in our cells that do all the work. They carry out an enormous number of functions. These include the haemoglobin in our red blood cells that carries oxygen around our bodies. Another protein is insulin, which is released from the pancreas to encourage muscle cells to take in glucose. Thousands and thousands of other proteins carry out the dizzying range of functions that underlie life.

Proteins are made from building blocks called amino acids. Mutations generally change the sequence of these amino acids. Depending on the mutation and where it lies in the gene, this can lead to a number of consequences. The abnormal protein may carry out the wrong function in a cell, or may not be able to work at all.

But the myotonic dystrophy mutation doesn’t change the amino acid sequence. The mutated gene still codes for exactly the same protein. It was incredibly difficult to understand how the mutation led to a disease, when there was nothing wrong with the protein.

It would be tempting to write off the myotonic dystrophy mutation as some bizarre outlier with no impact for the majority of biological circumstances. That way we could put it to one side and forget about it. But it’s not alone.

Fragile X syndrome is the commonest form of inherited learning disability. Mothers don’t usually have any symptoms but they pass the condition on to their sons. The mothers carry the mutation but are not affected by it. Like myotonic dystrophy, this disorder is also caused by increases in the length of a three-letter sequence. In this case, the sequence is CCG. And just like myotonic dystrophy, this increase doesn’t change the sequence of the protein encoded by the Fragile X gene.

Friedreich’s ataxia is a form of progressive muscle wasting in which symptoms normally appear in late childhood or early adolescence. In contrast to myotonic dystrophy, the parents are usually unaffected by the disorder. Both the mother and father are carriers. Each parent possesses one normal and one abnormal copy of the relevant gene. But if a child inherits a mutated copy from each parent, the child develops the disease. Friedreich’s ataxia is also caused by an increase in a three-letter sequence, GAA in this case. And once again it doesn’t change the sequence of the protein encoded by the affected gene.[3]

These three genetic diseases, so different in their family histories, symptoms and inheritance patterns, nevertheless told scientists something quite consistent: there are mutations that can cause disease without changing the amino acid sequence of proteins.

An impossible disease

An even more startling discovery was made a few years later. There is another inherited wasting disorder in which the muscles of the face, shoulders, and upper arms gradually weaken and degenerate. The disease is named after this pattern – it’s called facioscapulohumeral muscular dystrophy. Perhaps unsurprisingly, this is usually shortened to FSHD. Symptoms are usually detectable by the time a patient is in their early 20s. Like myotonic dystrophy, the disease is dominant and passed from affected parent to child.[4]

Scientists spent years looking for the mutation that causes FSHD. Eventually, they tracked it down to a repeated DNA sequence. But in this case the mutation is very different from the three-letter repeats found in myotonic dystrophy, fragile X syndrome and Friedreich’s ataxia. It is a stretch of over 3,000 letters. We can call this a block. In people who don’t suffer from FSHD, there are from eleven to about 100 blocks, one after another. But patients with FSHD have a small number of blocks, ten at most. That was unexpected. But the real shock for the researchers was that they really struggled to find a gene near the mutation.

Genetic diseases have given us great new insights into biology over the last hundred years or so. It’s easy to underestimate how hard-won some of that knowledge was. The identification of the mutations described here usually represented over a decade of work for significant numbers of people. It was entirely dependent on access to families who were willing to give blood samples and trace their family histories to help scientists home in on the key individuals to analyse.

The reason this kind of analysis was so difficult was because researchers were normally looking for a very small change in a very large landscape, hunting for a single specific acorn in a forest. This all became much easier from 2001 onwards, after the release of the human genome sequence. The genome is the entire sequence of DNA in our cells.

Because of the Human Genome Project, we know where all the genes are positioned relative to one another, and their sequences. This, together with enormous improvements in the technologies used to sequence DNA, has made it much faster and cheaper to find the mutations underlying even very rare genetic diseases.

But the completion of the human genome sequence has had impact far beyond identifying the mutations that cause disease. It’s changing many of our ideas about some of the most fundamental ideas that have held sway in biology since we first understood that DNA was our genetic material.

When considering how our cells work, almost every scientist over the last six decades has been focused on the impacts of proteins. But from the moment the human genome was sequenced, scientists have had to face a rather puzzling dilemma. If proteins are so all-important, why is only 2 per cent of our DNA devoted to coding for amino acids, the building blocks of proteins? What on earth is the other 98 per cent doing?

2. When Dark Matter Turns Very Dark Indeed

The astonishing percentage of the genome that didn't code for proteins was a shock. But it was the scale of the phenomenon that was surprising, not the phenomenon itself. Scientists had known for many years that there were stretches of DNA that didn't code for proteins. In fact, this was one of the first big surprises after the structure of DNA itself was revealed. But hardly anyone anticipated how important these regions would prove to be, nor that they would provide the explanation for certain genetic diseases.

At this point it's worth looking in a little more detail at the building blocks of our genome. DNA is an alphabet, and a very simple one at that. It is formed of just four letters – A, C, G and T. These are also known as bases. But because our cells contain so much DNA, this simple alphabet carries an incredible amount of information. Humans inherit 3 billion of the bases that make up our genetic code from our mother, and a similar set from our father. Imagine DNA as a ladder, with each base representing a rung, and each rung being 25cm from the next. The ladder would stretch 75 million kilometres, roughly from earth to Mars (depending on the relative positions of their orbits on the day the ladder was put in place).

To think of it another way, the complete works of Shakespeare are reported to contain 3,695,990 letters.[1] This means we inherit the equivalent of just over 811 books the length of the Bard’s canon from mum and the same number from dad. That’s a lot of information.

If we extend our alphabet analogy a bit further, the DNA alphabet encodes words of just three letters each. Each three-letter word acts as the placeholder for a specific amino acid, the building blocks of proteins. A gene can be thought of as a sentence of three-letter words, which acts as the code for a sequence of amino acids forming a protein. This is summarised in Figure 2.1.

Each cell usually contains two copies of any given gene. One was inherited from the mother and one from the father. But although there are only two copies of each gene in a cell, that same cell can create thousands and thousands of the protein molecules encoded by a specific gene.

This is because there are two amplification mechanisms built into gene expression. The sequence of bases in the DNA doesn’t act as the direct template for the protein. Instead, the cell makes copies of the gene. These copies are very similar to the DNA gene itself, but not identical. They have a slightly different chemical composition and are known as RNA (ribonucleic acid, instead of the deoxyribonucleic acid in DNA). Another difference is that in RNA, the base T is replaced by the base U. DNA is formed of two strands joined together via pairs of bases. We could visualise this as looking a little like a railway track. The two rails are held together by a base on one rail linking to a base on the other, as if the bases were holding hands. They only link up in a set pattern. T holds hands with A, C holds hands with G. Because of this arrangement, we tend to refer to DNA in terms of base pairs. RNA is a single-stranded molecule, just one rail. The key differences between DNA and RNA are shown in Figure 2.2. A cell can make thousands of RNA copies of a DNA gene really quickly, and this is the first amplification step in gene expression.

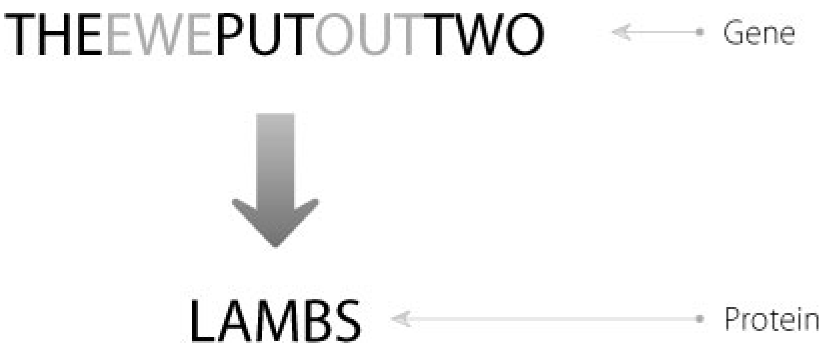


Figure 2.1 The relationship between a gene and a protein. Each three-letter sequence in the gene codes for one building block in the protein.

The RNA copies of a gene are transported away from the DNA to a different part of the cell, called the cytoplasm. In this distinct region of the cell, the RNA molecules act as the placeholders for the amino acids that form a protein. Each RNA molecule can act as a template multiple times, and this introduces the second amplification step in gene expression. This is shown diagrammatically in Figure 2.3.

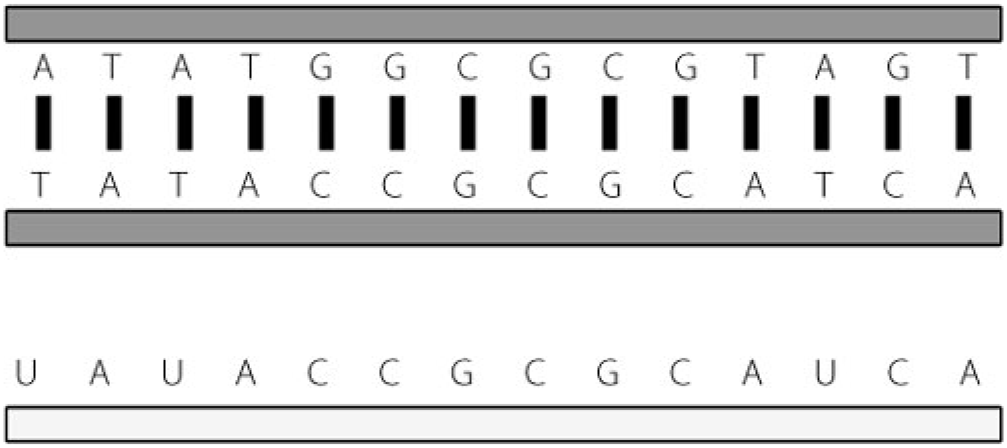

A T A T G G C G C G T A G T
T A T A C C G C G C A T C A
U A U A C C G C G C A U C A

Figure 2.2 The upper panel represents DNA, which is double-stranded. The bases – A, C, G and T – hold the two strands together by pairing up. A always pairs with T, and C always pairs with G. The lower panel represents RNA, which is single-stranded. The backbone of the strand has a slightly different composition from DNA, as indicated by the different shading. In RNA, the base T is replaced by the base U.

We can visualise this using the analogy of the knitting pattern from Chapter 1. The DNA gene is the original knitting pattern. This pattern can be photocopied multiple times, akin to producing the RNA. The copies can be sent to lots of people who can each knit the same pattern multiple times, just like creating the protein. It's a simple but efficient operating model and it works – one original pattern resulted in lots of soldiers with warm feet in the Second World War.

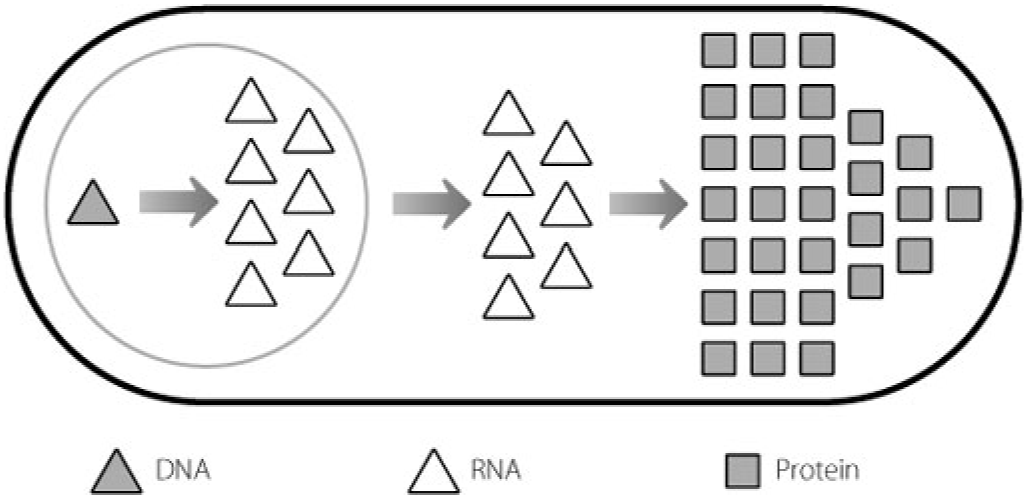


Figure 2.3 A single copy of a DNA gene in the nucleus is used as the template to create multiple copies of a messenger RNA molecule. These multiple RNA molecules are exported out of the nucleus. Each can then act as the instructions for production of a protein. Multiple copies of the same protein can be produced from each messenger RNA molecule. There are therefore two amplification steps in generating protein from a DNA code. For simplicity, only one copy of the gene is shown, although usually there will be two – one inherited from each parent.

The RNA molecule acts as a messenger molecule, carrying a gene sequence from the DNA to the protein assembly factory. Rather logically it is therefore known as messenger RNA.

Taking out the nonsense

So far, things might seem very straightforward but scientists discovered quite some time ago that there is a strange complication. Most genes are split up into bits that code for the amino acids in a protein and intervening bits that don't. The bits that don't are like gobbledegook in the middle of a string of sensible words. These intervening bits of nonsense are known as introns.

When the cell makes RNA, it originally copies all of the DNA letters in a gene, including the bits that don't code for amino acids. But then the cell removes all the bits that don't code for protein, so that the final messenger RNA is a good instruction set for the final protein. This process is known as splicing, and Figure 2.4 shows diagrammatically how this happens.

As Figure 2.4 shows, a protein is encoded from modular blocks of information. This modularity gives the cell a lot of flexibility in how it processes the RNA. It can vary the modules which it joins together from a messenger RNA molecule, creating a range of final messengers that code for related but non-identical proteins. This is shown in Figure 2.5.

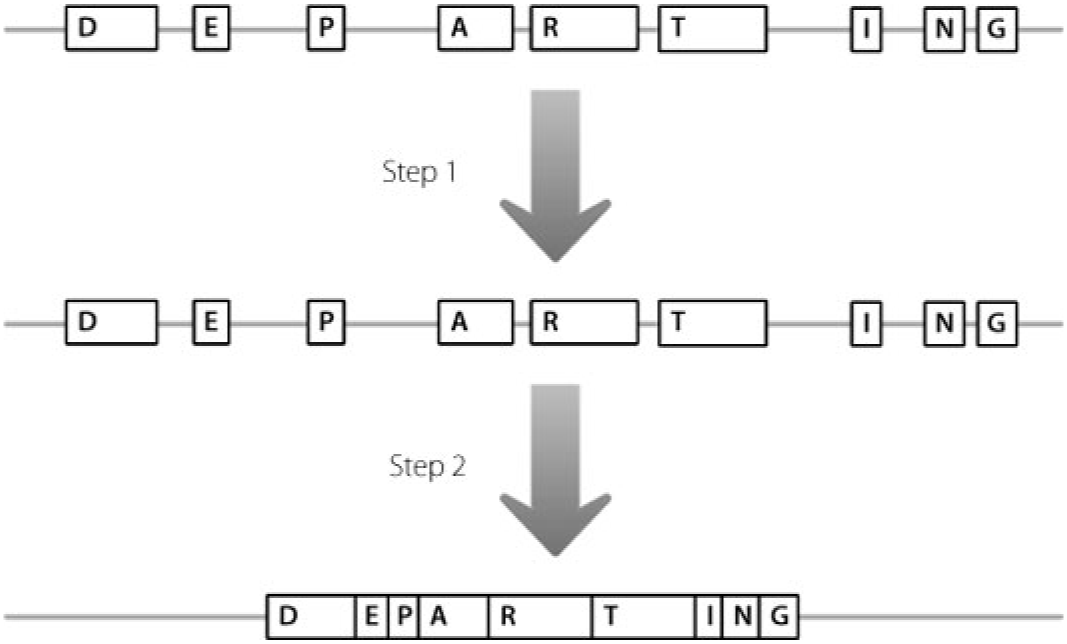


Figure 2.4 In step 1, DNA is copied into RNA. In step 2, the RNA is processed so that only the amino acid-coding regions, denoted by boxes containing letters, are joined together. The intervening junk regions are removed from the mature messenger RNA molecule.