Galton
institute

Genetics in Medicine

4. Epigenetics and Gene Regulation

By

Andrew P Read

Dian Donnai

Helen Middleton-Price

Genetics in Medicine 4. Epigenetics and Gene Regulation

By Andrew P Read, Dian Donnai and Helen Middleton-Price

Authors

Andrew P Read, Emeritus Professor of Human Genetics, The University of Manchester.

Dian Donnai, Professor of Medical Genetics, The University of Manchester.

Helen Middleton-Price, Visiting Fellow, The University of Manchester.

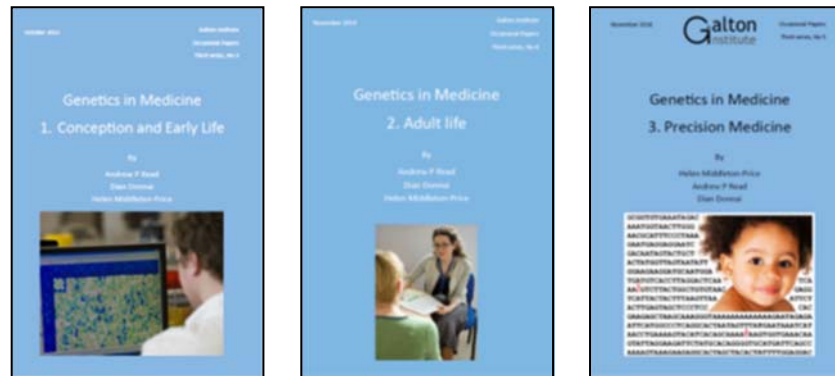# Genetics in Medicine 4. Epigenetics and Gene Regulation

By Andrew P Read, Dian Donnai and Helen Middleton-Price

Contents   Page

PMID numbers – in this booklet we are giving the unique identifying
number for some key publications in the PubMed database to assist
those who want to do further reading. See https://
www.ncbi.nlm.nih.gov/pubmed.

Introduction

GIM1   GIM2   GIM3

This is the fourth Galton Institute booklet in a series about genetics in modern medicine. In the first three (GIM1-3) we outlined the structure and function of **DNA** and the human **genome**, and how the **mutation** of DNA sequence that drives evolution can also give rise to disease. We described some of the more common **single-gene** variants that give rise to **Mendelian** disorders, some of which manifest from birth, others in later life. Further, we observed how significant loss or gain of genetic material can give rise to multi-system conditions such as **Down syndrome**. In addition, we looked at how genetic variants might influence some common illnesses.

We considered how the revolution in DNA sequencing has underpinned a rapid expansion in knowledge of the way genetic variation affects molecular mechanisms and pathways, and this in turn has led to new treatments for genetic disorders, cancer and **autoimmune** conditions. We also described how understanding the influence of genetic variation on the way we respond to drug treatments helps improve the safety and efficacy of medicines.

The scope of this booklet

Previous booklets in this series explained how humans have about 20,000 **genes** that encode all the **proteins** that make us and our cells function. These protein-**coding genes** make up a part of the DNA (deoxyribonucleic acid) that is

present in the nucleus of nearly every cell of our body. Each is a stretch of DNA that contains a recipe or blueprint for making a particular protein (see GIM1 pages 8-9 for a brief summary of how this works).

This booklet explores how genes are regulated and **expressed**. Genes don't simply keep constantly working: in different cells or at different times they are switched on or off. When a **gene** is switched on, it actively directs synthesis of its cognate protein; when it is switched off it is present but inactive. In this booklet we address how this is done. As we describe below, the behaviour and identity of a cell is determined by which of those 20,000 genes are switched on or off. Cells have many different mechanisms for regulating the way the DNA works. Some involve chemically modifying the DNA of the genome without changing the basic sequence of A, G, C and T **nucleotides**, while others affect the way the DNA is packaged within the cell nucleus. Epigenetics literally means 'above genetics' and is used to describe regulatory changes that are inherited from cell to daughter cell, but without changing the DNA sequence. Other regulatory processes work just within a given cell and are not directly passed on to daughter cells but can involve similar mechanisms.

The central role of gene regulation in making a human being

As a human develops from a single fertilised egg cell to a multi-cellular **embryo**, then to a **fetus**, a new-born baby and, eventually, a fully-formed adult, the ever-increasing diversity of cells is orchestrated by exceedingly complex and elaborate systems of gene regulation. Genes must be switched on and off with intricate precision and exact timing. One published estimate of the number of cells in an adult human body is 3.7×10^{13}. All these cells (with minor exceptions) contain the same DNA and, therefore, the same genes, yet they can look quite different from one another and function very differently (see figure overleaf). Muscle cells, skin cells, brain cells and every other of the several hundred cell types in the human body differ from one another because they use different subsets of the 20,000 genes they all share. The identity of a brain cell, skin cell, muscle cell etc is a function of the proteins that it contains, and this in turn depends on the subset of protein-coding genes that are switched on (expressed) in it. Thus, gene regulation is critical for cell identity and human development.
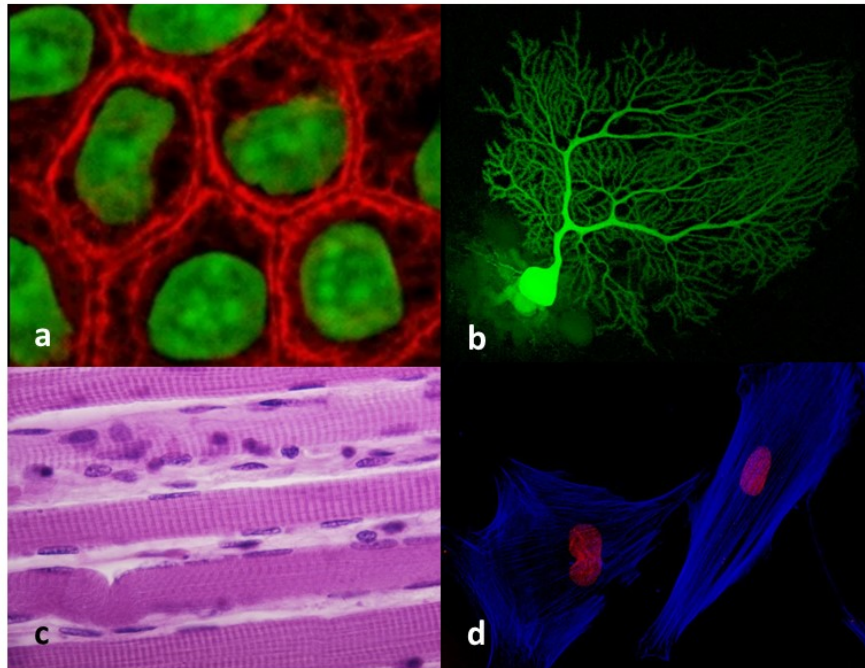
Figure: The diversity of cells. a) skin cells. b) a Purkinje cell (a type of brain cell). c) skeletal muscle cells. d) connective tissue cells (fibroblasts). All contain the same DNA. Colours are due to the different stains used in each case.

Further proof of the central importance of gene regulation comes from considering the nematode worm *Caenorhabditis elegans* (see figure opposite). This 1 mm long worm consists of only 959 cells and is widely studied as an example of one of the very simplest multicellular organisms. Yet it turns out to have just as many protein-coding genes as we have. Clearly the key to understanding our greater complexity must lie in more complex regulation of the same number of genes, not in having more genes. Thus, it should be no surprise that the epigenetic gene-regulatory mechanisms in humans and other higher organisms are subtle and exceedingly complex. There is much that we still don't understand, and the account given here is necessarily simplified.
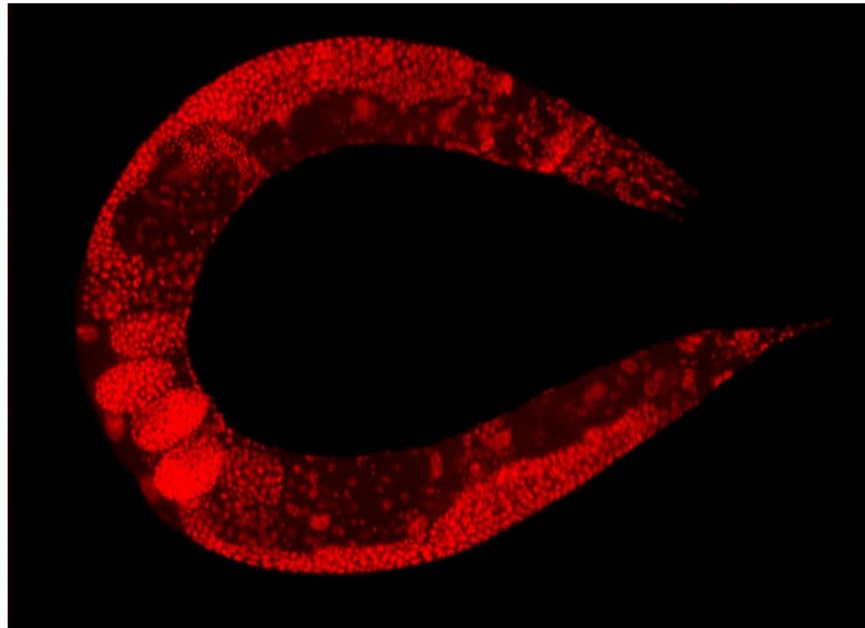
Figure: The nematode worm *Caenorhabditis elegans*. According to the *Ensembl* database, humans have 20,338 protein-coding genes while this 1 mm long worm consisting of just 959 cells has a similar number of genes.

Epigenetic programmes and reprogramming

As an embryo develops from a single fertilised egg cell, the various daughter cells specialise into all the different types necessary to make a multicellular organism (see figure overleaf). Cell identity among the differentiated progeny is determined by patterns of gene expression that are specific to each type of cell. When differentiated cells divide, the daughter cells exhibit the same type-specific patterns of gene expression. Stable cell identity is the result of epigenetic programmes that are applied at some point during embryonic development and retained thereafter. Under some circumstances cells can be reprogrammed (as described in GIM3 page 20).

A very early embryo should have an epigenetic clean slate to allow subsequent application of the different epigenetic programmes to progeny cells. Thus, while sperm and egg cells each carry their own specific epigenetic programme that

Figure: The steps from a fertilised egg via the **blastocyst** to stem cells which have the ability to develop into every different cell type.
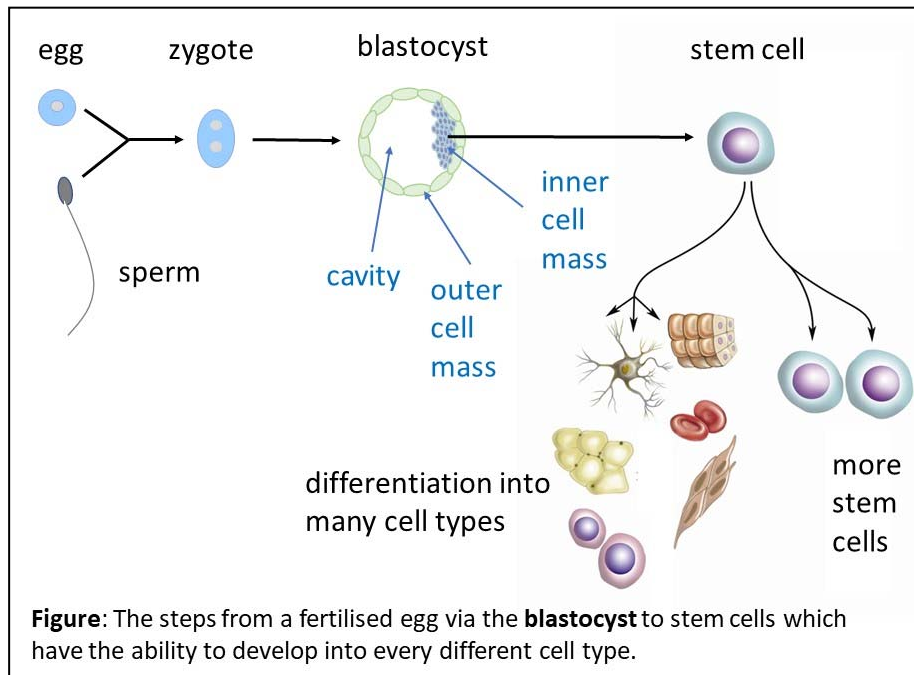
defines their identity, nearly all of these are erased during creation of an embryo. Hence, we would not, in general, expect epigenetic changes to be transmissible through pedigrees like conventional mutations. Inherited epigenetic disturbances are usually the downstream consequence of inherited conventional mutations (changes to the DNA sequence) – but later (page 29) we discuss some possible exceptions.

Epigenetics and human disease

Epigenetic mechanisms have been implicated in a wide range of disorders:

- Fertilised eggs of laboratory mice can be manipulated so that they contain either just two paternal or two maternal genomes. Despite having the correct number of genes and **chromosomes**, these eggs never develop into normal embryos. Rare accidents at conception in humans show the same effect. Conceptuses with just two paternal genomes develop as hydatidiform moles, abnormal clumps of membranes with no embryo, while conceptuses with just two maternal genomes develop as ovarian teratomas, disorganised masses of embryonic parts with no supporting membranes. Similar parent-of-origin effects are seen in children with certain **syndromes**, as described later (page 24). This is because some parts of the genome carry an epigenetic 'imprint' of their parental origin and are functionally different depending on the sex of the transmitting parent, even though the DNA sequences may be identical**.**

- Mutations in DNA causing disruption of the molecular machinery responsible for epigenetic modification, described later (pages 20-21), have now been shown to cause a number of serious developmental conditions.

- All cancers are the result of cells running out of control, and epigenetic dysregulation is central to the way normal cells develop into **tumour** cells. Cancer cells usually have many genetic mutations that change the DNA sequence of genes, but the end effect of these mutations is very often to change the epigenetic regulation of cell behaviour.

- Changes in gene regulation are implicated in many common conditions, such as heart disease, asthma, **diabetes**, obesity, autoimmune conditions and disorders of mental health. Environmental factors such as diet and life events play a key role, but their effects are mediated through changes in gene expression.

All these aspects of gene regulation are discussed in more detail in following sections of this booklet.

> **X-inactivation: a classic example of epigenetics in action**
>
> We described the phenomenon of X-inactivation in the first of this series of booklets (see GIM1 pages 12-13). To recapitulate, this is a special mechanism that ensures each cell has only a single *functional* **X chromosome**, regardless of the actual number in a human cell (46,XY (male), 46,XX (female) or even conditions such as triple-X syndrome – 47,XXX). At one point very early in the life of an embryo each cell somehow counts the number of X chromosomes and then inactivates all except one. Hence, X inactivation persists in all cells in a female.
>
> Researchers have long been interested in X-inactivation as one of the best examples of an epigenetic mechanism. The inactivated X is still there, its DNA sequence unchanged, but it is inactive and nearly all genes on it are switched off by epigenetic mechanisms. Each cell with two X chromosomes chooses at random which X to inactivate: thus some 46,XX cells inactivate the X inherited from the mother, and some cells inactivate the one inherited from the father. When each cell replicates its DNA and then divides, it remembers which X was inactivated, and the daughter cells inactivate that same X. But when an XX female produces her eggs, the inactive X is *re*activated. It becomes indistinguishable from the hitherto active X, and either of the two is equally likely to end up in any particular egg cell.
>
> Thus X-inactivation is a perfect example of an epigenetic process: it involves switching genes on or off; the changes do not involve changes to the DNA sequence, they are remembered through **mitosis** when a cell divides into two daughter cells, but not during the process of **meiosis** that produces egg cells.

How gene regulation works

Every stage of gene function is regulated (see GIM1 pages 8-9 for a brief description of the stages). Short-term regulation allows cells to respond to a fluctuating environment; long-term patterns of selective gene expression govern the permanent **tissue**-specific identity of a cell. Epigenetic memory means that tissue identity can be transmitted from cell to daughter cell. Like the short-term changes, these epigenetic mechanisms do not involve any change to

the underlying DNA sequence. Although some of the mechanisms of short-term and long-term regulation are the same, by convention the label 'epigenetic' has been mostly restricted to describing the long-term changes that persist through cell division. Under normal circumstances the epigenetic programmes that drive development are irreversible, but the epigenetic slate is wiped more or less clean at the start of embryogenesis, so that the embryo can use the full range of epigenetic programming as it develops. Some more subtle epigenetic effects may be modified during a person's lifetime. At least some of the programming can also be reversed in the laboratory to convert fully differentiated cells into so-called induced pluripotent **stem** (iPS) **cells** (see GIM3 page 20).

Many of the epigenetic mechanisms operate at the level of **transcription** – DNA is transcribed to make an **RNA** copy of a segment of DNA. While the DNA in a cell consists of a small number of immensely long molecules, RNA molecules are copies of just short sections of the DNA. When these encompass a gene, the resulting RNA is called a **messenger RNA**. At each point in its life a cell must decide which bits of its DNA sequence should be transcribed. Transcription is controlled by regulatory sequences within the DNA that bind effector proteins or (sometimes) RNA molecules to exert their action. Two important regulatory elements stand out:

- *Promoters* are the sites on a DNA molecule where transcription starts. Epigenetic mechanisms control whether a promoter is active or inactive in any particular cell and at any particular time. Promoters are typically a few hundred **base-pairs** long and are located immediately upstream of gene sequences. Every gene must have at least one promoter; some have several alternative promoters. Initiating transcription requires the **enzyme** RNA polymerase, which synthesises the messenger RNA for each gene, together with an assortment of regulatory proteins. **Transcription factors** are primary among these: proteins that recognise short sequences in the exposed DNA of a promoter, attract the RNA polymerase to the correct sequence and then enable it to carry out its function.
- *Enhancers* are promoter-like sequences, but they are not immediately adjacent to the gene they regulate. They may be as much as a million

base-pairs away. Like promoters, they depend on epigenetic mechanisms for their activity. Enhancers attract sequence-specific DNA-binding proteins such as transcription factors. They are believed to exert their effect by the DNA looping round so that the enhancer and promoter come to lie together. The tissue-specific expression patterns of genes are largely the function of tissue-specific enhancers. Often one gene will have several enhancers, maybe each governing expression in a different cell type/tissue.

Having made a messenger RNA, there are controls on **translation**, when the ribosomes use the sequence of **nucleotides** in the RNA to determine the sequence of **amino acids** in a protein that they are making. Further controls determine the target location of the protein inside or outside the cell and how actively it performs its function.

Packaging DNA in the nucleus

The first requirement for transcription is that the sequence in question should be accessible to the regulatory molecules. This is far from straightforward. The total length of DNA **double helix** in a human cell is an astonishing two metres (see box below). All this DNA has to fit into a nucleus that is typically 10 micrometres (10 millionths of a metre, 10 thousandths of a millimetre) in diameter – while still leaving room for all the other molecules in the nucleus. For the genome to function, the DNA must be precisely packaged and organised, not just squashed in. Many details of the packaging and three-dimensional organisation are still poorly understood. We do know, however, that they are crucial to regulating gene expression.

The total length of DNA in a human cell

You can work this out for yourself: in the Watson-Crick double helix successive base-pairs are 0.34 nanometres apart (1 nm is 10^{-9} metres or one millionth of a millimetre). The human genome comprises about 3,000,000,000 base-pairs, so its total length is about one metre. Each normal cell contains two genomes: one from mother and one from father.

The DNA in a cell nucleus does not exist as a naked double helix. The most basic level of packaging wraps the DNA double helix round little balls of proteins called **histones** to form bodies called **nucleosomes**. Electron micrographs of suitably prepared nuclear extracts show the nucleosomes along the DNA like beads on a string (see figure a) below). As well as being mostly complexed with histones in nucleosomes, a whole range of other DNA-binding proteins and some RNA molecules combine to organise it. **Chromatin** is a generic name for the packaged DNA. At the highest level, human DNA is organised into 46 chromosomes (in 23 pairs) that generally occupy distinct separate territories within the nucleus (see figure b) below). DNA that is near the nuclear periphery tends to be inactive; DNA in the centre may be actively transcribed. Between the levels of nucleosomes and chromosomes is a hierarchy of poorly understood loops that have vital roles in regulating transcription. The whole organisation is dynamic: sequences are accessible or occluded depending on the requirements of the cell.

Figure: a) Electron micrograph showing DNA packaged into nucleosomes like a string of beads. Each bead is a nucleosome consisting of 146 base-pairs of DNA double helix wrapped round eight molecules of histone proteins.
b) Chromosomes occupy distinct territories within a cell nucleus. Each chromosome has been stained a different colour. The right-hand panel identifies the individual chromosome numbers.

a) Courtesy of Jakob Waterborg, University of Missouri, Kansas City. b) Reproduced with permission from Nature Publishing Group. From Speicher MR & Carter NP (2005), The new cytogenetics: blurring the boundaries with molecular biology. *Nature Reviews Genet;* 6, 782-792.

Epigenetic marks

The accessibility of particular sequences is governed by the interplay of several regulators. The string of beads can be wrapped tightly (heterochromatin, which is 'closed' and mostly genetically inactive) or more loosely (euchromatin, 'open' and potentially active). The difference is largely a function of a variety of molecular tags ('epigenetic marks') attached either directly to the DNA or to the histone molecules in the nucleosomes.

One epigenetic mark, DNA **methylation**, is applied directly to the DNA. Certain **cytosine** nucleotides (letter C in the A-T-G-C DNA code) are marked by addition of a methyl group (see box overleaf for more detail). Many other marks are applied to the histone proteins in nucleosomes. How often these marks are strictly epigenetic (that is, passed directly from cell to cell) is arguable; at least in some cases this does happen. In any case they are crucial aspects of gene regulation, and even strictly epigenetic changes work through downstream histone modification. These histone marks are the work of three large sets of regulatory proteins:

- *Writers* are enzymes that add marks to histones. Some are activating marks whilst others are repressive. There are maybe a dozen different chemical tags that can be attached, and each can be attached to different specific amino acids of each of the four types of histone in a nucleosome. Thus, there is a large family of different specific writer enzymes.

- *Erasers* are enzymes that remove specific marks, either stimulating or repressing gene expression.

- *Readers* react to the marks and affect gene expression. The marks do not in general themselves directly affect gene expression; rather they act as signals to attract special effector proteins, or they lead to changes in the chromatin structure that influence whether those proteins can interact.

The figure opposite summarises these effects.

An extra set of regulators are *chromatin remodelling complexes*. These are large multiprotein machines that can physically alter the state of chromatin by shuffling nucleosomes along, creating or abolishing the nucleosome-free stretches that are readily accessible to regulatory molecules. Furthermore, RNA molecules also play a role. Our genome encodes many so-called **non-coding**

Figure: Open and closed chromatin, and some of the epigenetic marks determining the difference.
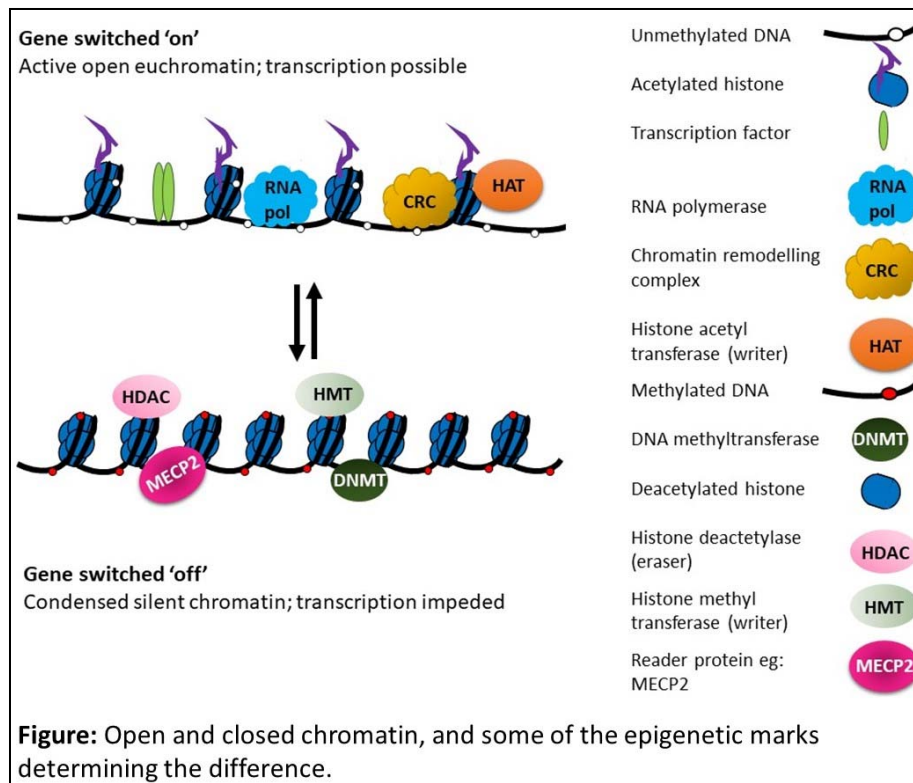
RNA molecules (see GIM1 page 9). These are not messenger RNAs; they do not code for protein synthesis, but they do other things in the cell. In fact, there are more genes for non-coding RNAs (22,521 according to the *Ensembl* database) than for proteins, and the resulting RNAs perform a great variety of functions in cells, often connected with gene regulation. RNA molecules can home in on matching DNA sequences, and so serve to bring enzymes or other proteins to those specific target DNA sequences.

It is tempting to ask which, among the various mechanisms, is the primary epigenetic signal and which are downstream epigenetic consequences. But this may be the wrong question. Gene regulation involves networks of effects; the various players form an interacting and mutually reinforcing system.

We have briefly sketched here just the rough outline of the very complex way

gene expression is regulated. If the complexity seems daunting, that is because it is. But remember, the system must somehow be able to direct construction of an entire functioning human, with all its different tissues and organs and all the functions that depend on those structures, using no more protein-coding genes than that 1mm long nematode worm. Not surprisingly, it is indeed both complex and remarkable, and our understanding of it is still at an early stage.

A bit of chemistry (for those not put off by chemical names or formulae).
The main epigenetic marks are methyl and acetyl groups.

- Methyl groups consist of one carbon and three hydrogen atoms.

- Acetyl groups consist of two carbon, one oxygen and three hydrogen atoms.
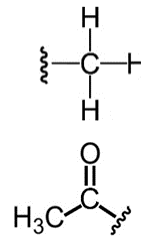
A small number of other chemical groups are sometimes also used as epigenetic marks.

Five main classes of enzymes add or remove these marks:

- DNA methyltransferases add methyl groups, predominantly to specific cytosine (C) bases in DNA. This addition of a methyl group is said not to change the DNA sequence because cytosine and 5-methyl cytosine both base-pair with G (**guanine**) in the double helix and function identically in the **genetic code**.
- Histone methyltransferases add methyl groups to various amino acids in the histone molecules in nucleosomes.
- Histone acetyltransferases add acetyl groups to histones.
- Histone demethylases remove the methyl marks from histones.
- Histone deacetylases remove the acetyl marks from histones.
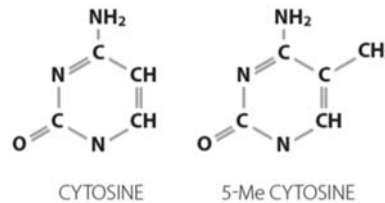
There are no DNA demethylase enzymes; methyl groups on DNA are removed by other mechanisms.

16

Studying epigenetics

Conventional techniques for studying DNA in the laboratory do not reveal epigenetic marks. In the first place, laboratories normally study isolated DNA, with all the histones and other cellular components purified away. Although the purified DNA does still carry the methyl marks attached to specific C nucleotides, most standard laboratory techniques do not study this DNA directly; instead they first synthesise copies of the DNA in the laboratory and study the copies. The manufactured copies will not carry those methyl marks. Thus, special techniques are needed to study epigenetic marks.

Certain alternative sequencing technologies do sequence the original DNA directly and, in principle, these systems can identify methylated cytosines. A widely used alternative method, compatible with standard sequencing methods, involves treating the original DNA with a reagent, sodium bisulphite, that alters normal cytosines so that they count as T (**thymine**) nucleotides, but leaves methylated cytosines unchanged. Sequencing samples of the same DNA twice, once with and once without the bisulphite treatment enables those cytosines that were methylated to be identified:

A sequence as determined on the untreated DNA: **GTGGAGCGGCCGCCGGAGAT**

Sequence as determined on bisulphite-treated DNA:

GTGGAGCGGTTGTCGGAGAT

The C nucleotides in red must have carried methyl groups in the original DNA; those in blue did not. **GTGGAGCGGCCGCCGGAGAT**

For studying histone modifications, **antibodies** are used. Specific antibodies will recognise and bind different specific modified histones in nucleosomes; the attached DNA can then be isolated and sequenced to reveal which DNA sequences were in nucleosomes carrying which marked histones.

Gene expression can be studied directly by isolating messenger RNA (see GIM1 page 8). RNA itself cannot be sequenced directly routinely (although new technical developments are changing this), but it can be sequenced indirectly using a special enzyme (reverse transcriptase, so-called as it acts in the opposite direction of transcription in the cell) that can make DNA copies of RNA (called

complementary DNA or cDNA). A typical experiment would isolate all the messenger RNA from a collection of cells, create cDNA using the reverse transcriptase enzyme, and then bulk-sequence the resulting cDNA copies. Only those genes that are currently being expressed will have cDNA copies, and the relative numbers of copies of the different cDNAs gives an indication of how intensively different genes are being transcribed. Typically, one would compare the results of such an experiment with results from a different tissue, or the same tissue under different conditions and relate these to the different epigenetic marks present in the two cases.

Gene regulation in disease

There are many ways in which disruption of gene regulation can cause illness. The writers, readers and erasers of epigenetic marks are all proteins. Similarly, the chromatin remodelling complexes are multi-protein machines. Like any other proteins, these are all encoded by specific genes and like any other genes, these may suffer mutations that cause the protein to malfunction (see GIM1 page 17). Thus, there is a whole range of single-gene disorders that directly affect the application or removal of epigenetic marks, or in some other way disrupt the chromatin configuration in the region. The ultimate causes of these conditions are DNA sequence changes that are heritable in the conventional way, but the pathogenic mechanism is epigenetic dysregulation, and so they are collectively known as epigenetic diseases. Some typical examples are described below.

Genetic conditions can also occur where epigenetic marks are disrupted locally as a secondary effect of a primary sequence variant. For example, Fragile-X syndrome, described below, is the result of silencing of the *FMR1* gene through methylation. This happens when the local chromatin structure is changed because of an abnormal DNA sequence near the gene.

Dysregulation of gene expression is the essential hallmark of cancer, and if the dysregulated **phenotype** were not passed from cell to daughter cells they would never form a tumour. The transmission may be of DNA mutations that cause dysregulation or of epigenetic changes such as abnormal DNA methylation.

Many pathogenic regulatory changes occur independently of any DNA sequence change. There are many common **complex conditions** where environmental factors play significant roles in the causation. Famine, diet, trauma and toxins have all been implicated as factors that cause regulatory changes. Some of these are discussed in more detail below. In most cases it is not clear whether the changes in gene regulation are directly passed on from cell to daughter cell (and hence qualify as epigenetic) – but where the pathogenic effects appear long after the environmental trigger has vanished, there must be a suspicion that this is happening.

As explained above, we would not, in general, expect epigenetic changes to be transmissible through pedigrees like conventional mutations (unless caused by just such a conventional DNA mutation as in the case of the single-gene conditions below). Nevertheless, there is evidence that some environmentally-induced effects can indeed affect future generations (so-called 'transgenerational effects'). Most probably this occurs through epigenetic changes to the genomes of eggs and sperm. Thus, in some circumstances an organism can pass on to its offspring characteristics that it had acquired during its lifetime – a hallmark of so-called Lamarckian inheritance (see page 39).

Single-gene conditions affecting the epigenome
Disorders influencing chromatin structure

Many different epigenetic diseases are caused by conventional mutations: DNA sequence variants in genes encoding proteins essential for chromatin structure, illustrating how important chromatin is for normal development. A gene which encodes an enzyme that modifies chromatin structure in turn regulates expression of many other genes. This could be likened to a conductor directing an orchestra – mutation in the main (conductor) gene leads to dysregulation of many genes (the orchestral players) and hence clinical manifestation in many body systems. Examples among many include Kabuki, Rett and Coffin-Siris syndromes, caused by mutations in a histone writer, a reader and a part of a chromatin remodelling complex respectively.

Kabuki syndrome is a rare condition that affects many parts of the body,

19

although specific symptoms and severity can vary. Features include a characteristic facial appearance, reminiscent of the makeup used in the classical Japanese Kabuki theatre, including interrupted eyebrows. Other features may include prominent finger-tip pads, skeletal abnormalities, short stature, heart defects and intellectual disability, microcephaly, poor muscle tone, eye problems and cleft palate.
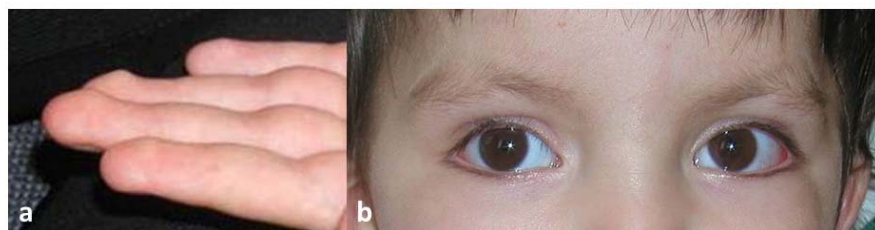
Figure: Individuals with Kabuki syndrome showing a) prominent finger-tip pads b) interrupted eyebrows.

Kabuki syndrome is most often caused by mutations in the *KMT2D* and *KDM6A* genes. Both these genes code for enzyme proteins responsible for transferring methyl groups on and off histones.

Rett syndrome is a rare severe genetic disorder, mainly seen in females; it affects the way the brain develops, causing a progressive neurological impairment, seizures (epileptic fits), stereotypic movements and lack of speech. Babies with Rett syndrome appear normal at birth, so it usually remains undetected until later in the first year of life, when infants can lose many acquired skills such as crawling or use of hands and become withdrawn.

Figure: A young girl showing the stereotypic hand wringing typical of Rett syndrome.

Rett syndrome is an X-linked *dominant* condition and is rarely seen in males; a mutation on a male baby's single X chromosome is most often fatal at an

early stage of development; in females the second normal X chromosome ameliorates the effect of the mutated X chromosome.

Rett syndrome is caused by mutations in the X-chromosome gene *MECP2*, which codes for a reader protein that recognises and binds methylated bases in DNA, regulating transcription and chromatin organisation. This protein is particularly abundant in the brain, in which it is important for the function of nerve cells, or **neurones**. The MECP2 protein most likely plays a role in maintaining connections (**synapses**) between neurones, where cell-to-cell communication occurs.

Figure: Children with Coffin-Siris syndrome showing a) sparse hair and b) unusual shaped feet and underdeveloped toenails.

Coffin-Siris syndrome is a rare condition that causes variable degrees of learning disability, speech and motor developmental delay, distinct facial features and sparse scalp hair amongst other features. It can be caused by a DNA mutation in any one of several genes in different chromatin remodelling complexes.

Fragile-X syndrome (FXS) is an X-linked condition that is one of the most common causes of moderate to severe inherited learning disability in males. The mutation in the gene responsible, *FMR1*, also affects chromatin structure, but by an indirect mechanism: the sequence variant is an unstable expansion of the non-coding CGG repeat sequence – $(CGG)_n$ – in the section at the beginning of the gene that is not translated into protein (the so-called 'untranslated region'). In normal individuals, there are about 6-60 CGG repeats, but in affected boys (and girls, although they tend to be less severely affected as they have a second X chromosome producing a normal FMR1 protein) the expansion inflates to >200 repeats (see figure overleaf). This massive expansion affects the chromatin architecture in the region to the

extent that the X chromosome – when seen under the microscope – displays a break at the position of the FMR1 gene. This break is known as a 'fragile site', hence the name Fragile-X syndrome. The large DNA expansion at the beginning of the gene becomes methylated, and the histones deacetylated, possibly through the hybridisation of small RNA molecules produced during early development when the mutated *FMR1* is still transcribed. Eventually transcription of the gene is switched completely off. The exact mechanism for this process is not yet fully understood.

Figure: Transcription of *FMR1* proceeds when the repeat sequences are limited to 6-60 repeats. Transcription is prevented when the expansion is >200 repeats.
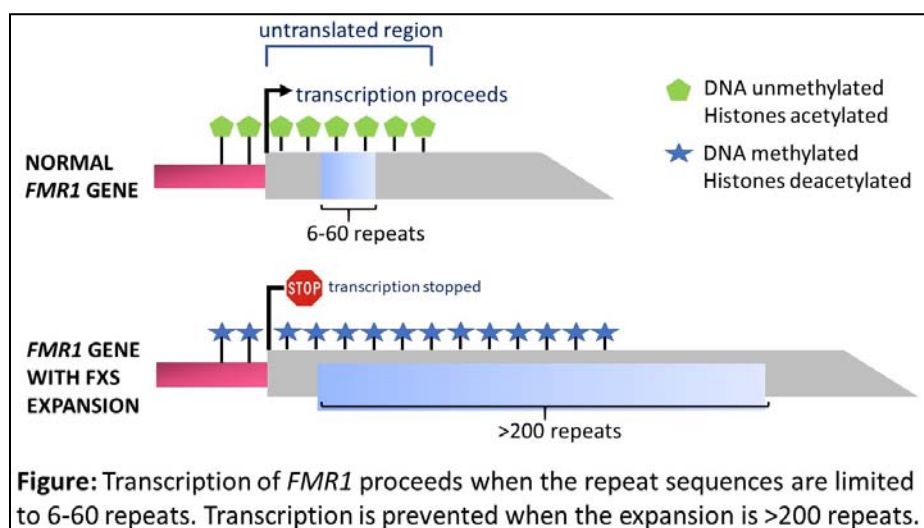
A further line of enquiry relates to how the 'silencing' of the *FMR1* gene leads to the symptoms of Fragile-X syndrome. The protein (called FMRP) binds RNA and is thought to inhibit translation of many genes involved in the development of the nervous system; it may be through the disruption of these finely controlled neurodevelopmental processes that the loss of FMRP has its effect.

Parent-of-origin effects and disorders of genomic imprinting
Parent-of-origin effects occur when the consequences of a gene on a person's phenotype depend on whether the gene is inherited from an individual's mother or father. About one hundred of the 20,000 protein-coding genes in our genomes carry different epigenetic 'imprints' depending on the parent of

origin: they are switched 'on' on the chromosome from one of our parents but switched 'off' on the other chromosome.

Like X-inactivation, imprinting is epigenetic and reversible and does not change the DNA sequence. For example, a man will have inherited some imprinted sequences from his mother. In all his cells those sequences carry a maternal imprint (he will also have second copies of those sequences, inherited from his father, which do not carry the imprint). Yet when he passes any imprinted maternal sequence on to his child it must then carry a paternal imprint (see figure below). So, the imprint is erased and then re-set during the creation of the **gametes** (egg and sperm cells).

Most, *but not all*, parent-specific epigenetic marks are erased in the early embryo. Marks in imprinted regions remain.

fertilisation

blastocyst

New epigenetic marks are applied during early stages of embryonic development. Parental imprints remain.

Imprint persists in the body's cells.

Parent-specific epigenetic marks

or

imposed later on gametes (eggs and sperm).

Epigenetic marks (including imprints) erased in newly formed eggs and sperm.

Figure: There are two rounds of genome-wide epigenetic reprogramming, one in the gametes and one in the early embryo. Parent specific imprints resist erasure and reprogramming in the second round.

Angelman syndrome and Prader-Willi syndrome
Disorders of genomic imprinting were identified in the late 1980s, when two quite distinct conditions, Angelman syndrome (AS) and Prader-Willi syndrome (PWS) were observed to be caused apparently by the same **deletion** – loss –
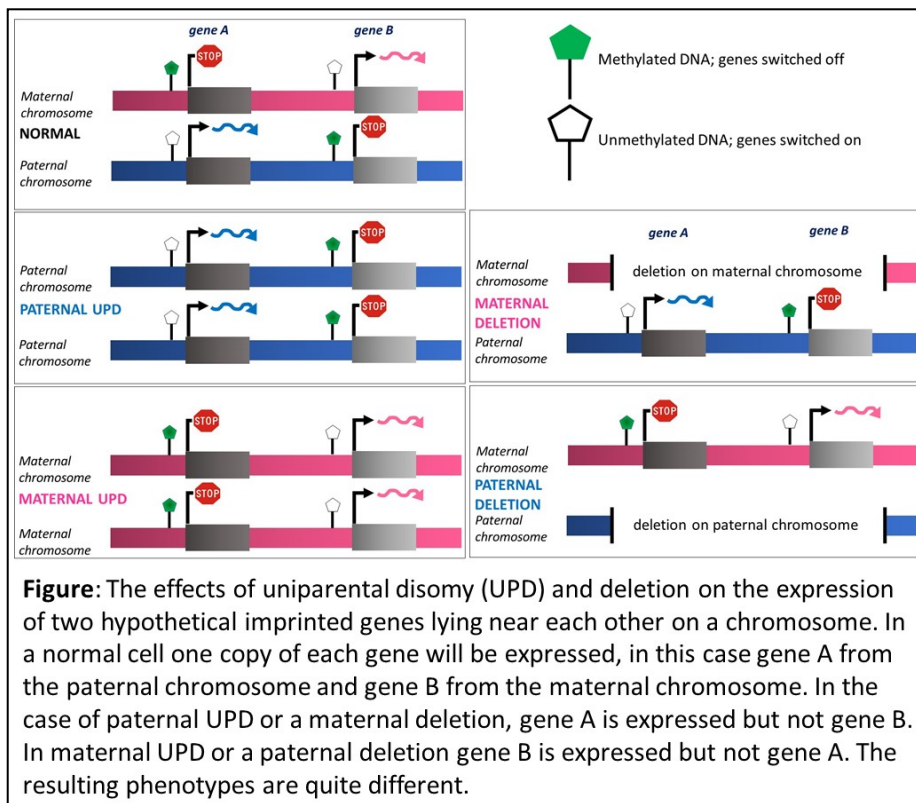
of part of chromosome 15.

AS patients have delayed development, intellectual disability, severe speech impairment, problems with movement and balance and an apparently 'happy' disposition. Individuals with PWS are born with weak muscle tone, delayed development and some behavioural features in common with autism, feeding difficulties and poor growth, later developing an insatiable appetite leading to obesity.

It was soon discovered that AS was caused by deletions of the chromosome 15 inherited from the mother (the *maternal* chromosome) and PWS was caused by deletions of the chromosome 15 from the father (the *paternal* chromosome), leading to the proposition that there are differentially imprinted genes that are essential for normal development. It was also observed that these conditions could be caused by uniparental disomy (UPD), when the individual inherits *both* copies of chromosome 15 from a single parent, caused by errors at the time of egg, sperm, **zygote** or early embryo formation. Paternal UPD (resulting from loss of the maternal chromosome and gain of a second paternal chromosome) produces the AS phenotype, while *maternal* UPD (resulting from loss of the paternal chromosome and gain of a second maternal chromosome) produces PWS. In both cases there are losses of (different) genes that are, normally, only expressed from either the *maternal* or *paternal* chromosome 15, as they are silenced on the other chromosome 15 by the imprints made by the epigenetic processes described earlier. The clinical disorder arises when the normally expressed gene(s) is lost (see figure opposite).

AS and PWS can also – more unusually – be caused by small deletions or mutations in individual specific genes/DNA sequences. Critically for parents who have a child with one of these serious conditions, the risk for families with large deletions or UPD of having a second child with the condition is very low as they are caused by *de novo* – chance, new mutation – events, whilst the risk of recurrence for small deletions or other mutations in key genes is much higher as they can be passed down from the relevant asymptomatic parent (mother in AS and father in PWS).

Several other rare conditions including Beckwith-Wiedemann and Silver-Russell syndromes similarly involve abnormal complements of imprinted genes. The

Figure: The effects of uniparental disomy (UPD) and deletion on the expression of two hypothetical imprinted genes lying near each other on a chromosome. In a normal cell one copy of each gene will be expressed, in this case gene A from the paternal chromosome and gene B from the maternal chromosome. In the case of paternal UPD or a maternal deletion, gene A is expressed but not gene B. In maternal UPD or a paternal deletion gene B is expressed but not gene A. The resulting phenotypes are quite different.

detailed molecular mechanisms are quite complex, and one might reasonably ask, what could be the purpose of all this? One theory relates it to a conflict of interest between maternal and paternal genes in a fetus. The theory suggests the evolutionary interest of the father is best served by having the fetus as well-nourished as possible, even if that is to the detriment of the mother's health. The mother's interest is to restrict the growth of a 'parasitic' fetus so that it can traverse the birth canal, and she can survive to have future offspring. Thus, imprinted genes that are paternally expressed are in general growth-promoting, and maternally expressed ones are growth-restricting in the fetus. Hydatidiform moles, with two paternal genomes, consist of just the membranes that would normally extract nutrients from the mother, while ovarian teratomas, with two maternal genomes, lack all those membranes. The phenotypes of many imprinting disorders support this interpretation.

25

Responding to the environment

When environmental factors such as infections, toxins, poor diet or famine, neglect or abuse assail an individual, there are obvious immediate pathological responses. However, in addition, epidemiological data show that there can be long-term consequences for health. The increased **prevalence** of many common diseases – such as asthma, diabetes and obesity – over the last century cannot be due to genetic changes as the speed of change is too high to be accounted for by genetic mutation; it may be that changes in environment and life-style could combine with a genetic predisposition to change gene expression resulting in long-term effects.

Life events can change patterns of gene expression through a variety of mechanisms, not all of which are epigenetic. An external signal may cause a transcription factor that had been sequestered in the **cytoplasm** of the cell, away from the DNA in the nucleus, to move into the nucleus and start switching on genes. Pre-existing messenger RNAs in the cytoplasm may be stabilised or destabilised in response to environmental changes. Thus, there is much room for debate about how far any particular response to environmental factors is mediated by epigenetic changes.

Many studies have investigated long-term changes in DNA methylation subsequent to specific environmental events or simply the passage of time. Old people have different average patterns of DNA methylation from young people, and these and related epigenetic changes can be used to define a biological, as distinct from chronological age, which turns out to be a good predictor of age of death. Identical twins – who have identical DNA – have very similar **epigenomes** when they are born, but these can diverge later. One question is how far these changes are just random and how far they constitute a cumulative epigenetic memory of past events. A second question is how far the observed changes in DNA methylation are the cause of changes in the behaviour of individuals and cells, and how far they are just *consequences* of such changes.

Metabolic programming: fetal origins of adult disease

In the 1990s the Southampton-based physician David Barker noticed a correlation between the birthweight of a baby and its risk of obesity, heart

disease and type-2 diabetes in later life. The risk of these outcomes was greater for low-birthweight babies. These observations were explained by the 'Barker hypothesis', that the nutritional status of a fetus somehow 'sets' the **metabolism** of the person throughout later life. Fetuses suffering sub-optimal nutrition in the womb were programmed to a permanent 'thrifty phenotype' that hoovered up whatever nutrients were available. In conditions of post-natal affluence and abundant food this led to obesity and type-2 diabetes (PMID:11809615)[1].

Much additional evidence supports Barker's original observations. For example, in 1944 the Nazis blockaded western Holland, reducing the entire population to semi-starvation for several months. Babies conceived during this Dutch 'hunger winter' grew into adults with elevated levels of obesity. Numerous experiments with laboratory mice have clearly demonstrated specific long-term responses in pups to the nutritional status or environment of their mothers during pregnancy. The Barker hypothesis of fetal origins of adult disease (FOAD) is now widely – but not universally – accepted. The molecular mechanisms are not well understood, but it is assumed the metabolic programming could not be genetic since conventional mutations could not produce so pervasive an effect in so short a time, so epigenetic mechanisms have been proposed.

Maternal nutrition and offspring obesity

Paradoxically, empirical observation has confirmed that both insufficient and excessive nutrition prior to and in pregnancy give rise to offspring with a tendency towards obesity and type-2 diabetes (see figure overleaf).

This observation has been underscored by a study looking at genome-wide methylation in over a thousand individuals from the Avon Longitudinal Study of Parents and Children (ALSPAC, also known as Children of the 90s, a world-leading birth **cohort** study, charting the health of 14,500 families in the Bristol area) (PMID:25855720): methylation of DNA in newborns of both underweight and obese mothers was different from methylation in babies of

---

[1]the PMID number is the identifying number for a publication in the PubMed database. See https://www.ncbi.nlm.nih.gov/pubmed.

normal weight mothers; methylation was unsurprisingly also different *between* the newborn offspring of underweight and obese mothers, but *both* patterns were associated with a tendency to obesity in the offspring. A further analysis of several large studies (PMID:29016858) looked at pre-pregnancy maternal obesity and methylation at over 450,000

Figure: There is a U-shaped relationship between maternal obesity or starvation to the offspring's predisposition to obesity.
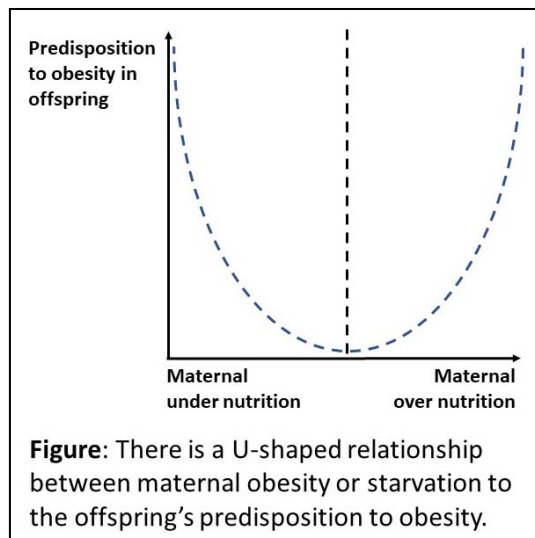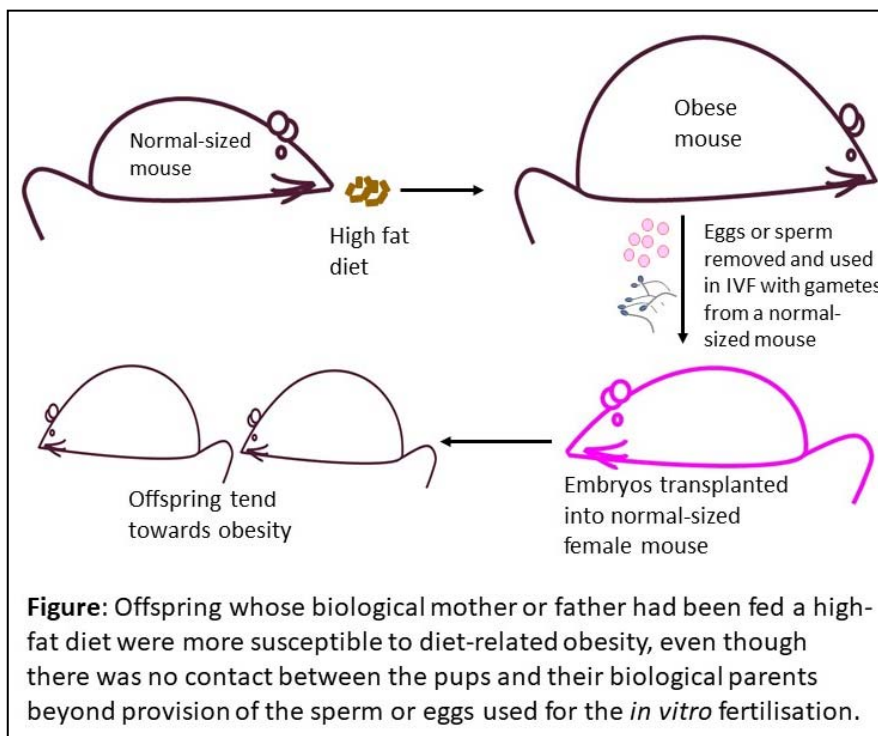
sites in newborn blood DNA in 9,340 mother-newborn pairs. The analysis found evidence that the correlation between maternal obesity and altered newborn DNA methylation persisted into adolescence.

To underline the connection between obesity and the epigenome, a major 2017 study of 10,000 individuals showed that obesity was associated with epigenome-wide changes in DNA methylation (PMID:28002404). The differentially methylated areas included genes involved in lipid (fat) metabolism and the data could be used to predict the likelihood of future type-2 diabetes. Again, we must be very cautious in interpreting these data, as a correlation of two phenomena does not prove that either one causes the other.

Whilst Dr Barker's original observations concerned poor maternal nutrition during pregnancy, subsequent studies have gathered more evidence pointing to the long-term outcomes of other influences, with individuals exposed during fetal or early life to poor environment such as stress, low mood or trauma in their mothers showing later effects on their health and personality. The theory rests on the existence of specific developmental phases when a human organism is 'plastic' or 'sensitive' to stressors that ensures its adaptation and best chance of survival in its environment (PMID:21684471).

Transgenerational effects

Although the Barker hypothesis concerns the effects of the intrauterine environment, experiments with laboratory mice have suggested that there may be additional ways that parental obesity might influence the offspring. In one typical experiment (PMID:26974008), mice were fed a high-fat diet so that they became obese, insulin resistant and glucose intolerant. Sperm or eggs from these obese mice were used for *in vitro* fertilisation, and the embryos implanted in surrogate mothers that were fed a normal diet. The resulting pups were not obese – but when they were challenged with a high-fat diet, pups whose parents had been fed a high-fat diet were more susceptible to diet-related obesity than those whose parents had been fed a normal diet (see figure below). This was true whether it was the fathers or the mothers that had been fed the high-fat diet, although the effect was most pronounced in female pups whose biological mothers had received the high-fat diet. The pups had an entirely normal intrauterine environment and had

Figure: Offspring whose biological mother or father had been fed a high-fat diet were more susceptible to diet-related obesity, even though there was no contact between the pups and their biological parents beyond provision of the sperm or eggs used for the *in vitro* fertilisation.

no contact with their biological parents beyond provision of the sperm or eggs used for the *in vitro* fertilisation. Whatever caused the effect must have been present in the sperm or eggs.

Other experiments (PMID:26721680) suggested that the effect in sperm was due to the presence of RNA fragments from breakdown of **transfer RNAs**. How these fragments exert their effect remains unexplained. Alternative explanations centre around the way the DNA is packaged in sperm.

It is intriguing that such experiments involve obesity and type-2 diabetes – the very characters that are currently increasing so alarmingly across the world. It is as though mice and humans have a specific built-in programme leading to these problems that can be activated in various ways. Maybe the current epidemic is not solely the result of the poor lifestyle of those directly affected? Over-eating and under-activity are undoubtedly strongly influential factors, but might the extreme obesity that is now so common on our streets be an exaggerated response to a poor lifestyle that was triggered by something in the sperm or eggs of moderately obese parents?

The cases described above show how either the intrauterine environment, or something in the gametes, can affect offspring. A fetus *in utero* already has the primordial **germ cells** that will make up the next generation. Thus, no extra theories are needed to explain how a mother's environment might affect her grandchildren. Reports of grand*paternal* effects are more challenging. For example, in the Överkalix region of Sweden, the risk of cardiovascular and diabetes-related death of individuals correlates with increased food supply during the prepubertal growth period of their grandfathers (PMID:16391557). There are other reports of similar grand*paternal* effects in humans and in mice (PMID:23435955).

So, there is evidence that some effects of the environment *can* be passed down the generations (transgenerational effects). Thus, in some circumstances an organism apparently *can* pass on to its offspring characteristics that it has acquired during its lifetime.

The question whether epigenetic effects can operate across generations more broadly is a matter of great debate. Epigenetic modifications are remembered through mitosis, by definition, but not normally through

meiosis (consider X-inactivation, for example). The epigenetic marks on the DNA of an individual are mostly stripped off in their early gametes before gaining new egg- or sperm-specific marks. Then, nearly all these new marks are erased again in the early embryo so that it can establish the different epigenetic programmes needed to make all the different tissues of the fetus (see page 23). So, any transgenerational epigenetic change would have to survive two rounds of genome-wide erasure and re-establishment, indicating it would be extremely hard for epigenetic marks to be transmissible. Instead, there may be other routes for non-DNA inheritance, maybe through molecules in the cytoplasm. We know, however, that during the embryonic epigenetic reprogramming, not all such marks are entirely removed (eg: imprints), so there is scope for exceptions. In summary, transgenerational epigenetic effects are a confusing topic. Such effects undoubtedly occur, but whether they are isolated oddities with essentially trivial individual explanations, or whether they point to something fundamental – a whole major under-appreciated substratum of inheritance – is hard to say. Watch this space!

Common diseases and changes to the epigenome – cause or effect?

Many common disorders (eg: asthma, diabetes and mental health conditions) result from environmental triggers acting on a genetically susceptible person to produce pathogenic changes in gene expression. Although the study of the epigenome in common disease is a burgeoning field of research, much of the underlying biology is still a matter for speculation, and so most scientists exercise great caution in their interpretation of the data. It is yet unclear whether the changes in DNA methylation seen in many common conditions are causative or simply the result of a pathological process; for example, when a transcription factor binds to DNA it inhibits methylation of the DNA and when a transcription factor is removed, methylation can be acquired afterwards. The hope is that, regardless of whether DNA methylation changes are cause or effect, they might be useful diagnostic and prognostic indicators that will assist doctors in their analysis and treatment of their patients. In addition to the work on obesity, many other common illnesses have been subjected to DNA methylation studies, including neurodevelopmental conditions (eg: autism spectrum disorders, ASD), **atopy** (eg: hay-fever, allergic asthma and eczema),
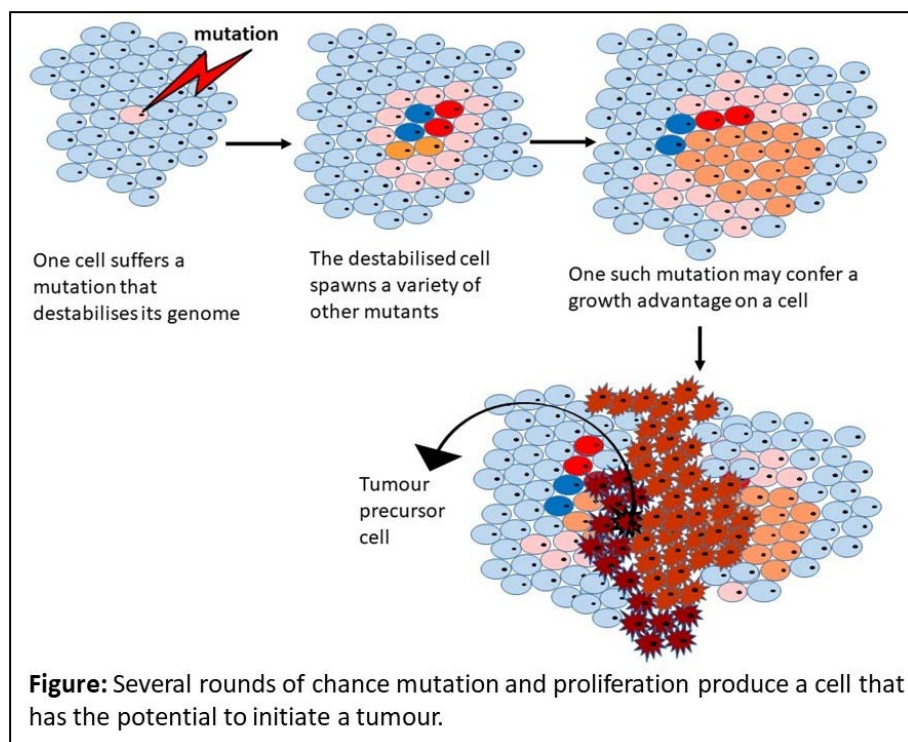
autoimmunity (when the body launches an immune attack against its own cells) and psychiatric disorders. Examples include:

- A study of families with one autistic child showed that the recurrence of signs of autism in a subsequent child at 12 months was correlated with differential DNA methylation of neurodevelopmental genes in sperm from their fathers (PMID:25878217).

- Evidence for a link of high immunoglobulin E (IgE) concentration – an antibody associated with hay fever and allergic asthma – with differential methylation around many genes, including several implicated in the allergic response (PMID:25707804; 29692868).

- Experiments with rats showing that pups groomed by their mothers grew up more resistant to stress, with reduction in DNA methylation of specific genes in part of the brain (PMID:9287218, 15220929).

- A study showing a relationship between methylation of certain genes in the cord blood of newborn babies and behavioural problems later in childhood (PMID:28595673).

- Evidence that the gene encoding reelin, a protein that is important in the development of neurones and the creation of synapses, was more methylated (with reduced expression) in the brains of patients with schizophrenia (PMID:15961543).

- Wide-ranging studies identifying differential epigenetic profiles in autoimmune diseases (such as in type-1 diabetes and rheumatoid arthritis), characterised by the body launching an immune attack against its own cells (PMID:29574040).

Many of these studies are either small and/or preliminary, and it must be emphasised that, even when changes in DNA methylation have been documented, it is difficult to know whether these are causes or consequences of changes in gene expression. Thus, these studies provide only intriguing hints, not firm proof. However, they do suggest that DNA methylation and other epigenetic effects could be part of long-term responses to adverse events, which in turn suggests they might act as useful **biomarkers** for particular conditions or contribute to possible novel therapeutic interventions.

Cancer

Cancer is not a single disease although all cancers are essentially diseases of growth dysregulation. A multistep process of genetic and epigenetic changes allows tumour cells to replicate when they should not, forming a tumour (see figure below) and eventually invading other tissues to form the secondary tumours that are the usual cause of a patient's death .

Figure: Several rounds of chance mutation and proliferation produce a cell that has the potential to initiate a tumour.

Early research efforts have been dedicated to distinguishing changes that are the primary causes of cancer development (so-called driver mutations) from those that are incidental consequences of the generally chaotic and unstable nature of cancer cell genomes (often referred to as passenger mutations). But it is clear that the genetic instability seen in tumour cells enables the occurrence of further mutations that promote cancer cell survival, and thus the so-called passenger mutations are also likely to be contributing to the characteristics of the cancer cell. Changes in DNA methylation, particularly loss

of methylation (hypomethylation), occur in several cancers and may contribute to the chromosome instability that is a feature of most cancers.

Several studies have looked at DNA methylation across the genome in various cancers. Because changes in DNA methylation and other epigenetic modifications are coupled with altered transcription in cancer, it can be difficult to decide which events directly cause the initiation of cancer. Genetic mutations in components of chromatin remodelling complexes, in DNA modification enzymes (DNA methyltransferase) and in writers, erasers and readers of histone modifications have been observed in several cancers, and it is therefore likely that many of the epigenetic changes that mediate cancer progression are a consequence of genetic changes (PMID:26972587).

In most cases it remains unclear whether epigenetic abnormalities *independent* of genetic mutations can cause cancer. Paediatric malignancies seem to be the exceptions. The most common childhood cancers are the leukaemias (resulting in high numbers of abnormal white blood cells), followed by tumours of the nervous system. Here DNA mutations are far fewer compared to adult cancers and where they do occur they are usually in genes encoding all classes of epigenetic regulators (writers, readers, erasers and chromatin remodellers) (PMID:25950259). There is also the predisposition to cancer in syndromes such as Beckwith Wiedemann syndrome associated with a loss of genomic imprinting. The example most commonly cited is increased methylation of part of the *H19* gene which is associated with a high risk of Wilms tumour.

Environmental factors such as chronic infection and hypoxia affect epigenetic modifications, especially DNA methylation. In many cases they must act as drivers of tumorigenesis. It has been found that some genomic regions are specifically susceptible to age-related and environmental signals. Studies of methylation in smokers might provide an example. Some points on the *AHRR* gene – coding for a protein (the aryl hydrocarbon receptor repressor) that mediates the effect of specific toxins and is involved in the regulation of cell growth and differentiation – are known to be methylated very differently in smokers and non-smokers. Several recent large studies (PMID:26667048; 27632354; 28100713) of blood samples from smokers have shown that hypomethylation of *AHRR* is associated with *future* risk of lung cancer, raising

the exciting prospect of using methylation status to provide clinically useful information that directs possible early intervention.

The observation that there is an age-related general decrease of DNA methylation – along with local hypermethylation – over the whole genome, supports the link between cancer and methylation, since it is well-known that we have an increased risk of cancer as we age. Furthermore, we know that there is an increase – over the last 50 years, or so – in the **incidence** of certain types of cancer (eg: breast and colon cancers), with age and several additional factors playing a role, including the rise in obesity and growth in alcohol consumption. This suggests the action of multiple epigenetic-mediated processes that act at the interface between the environment and key genes in cancer. Andrew Feinberg and colleagues (PMID:29617578) have proposed a model whereby environmental factors and ageing produce an epigenetically disrupted pool of **progenitor** cells (the cells that are predecessors of the fully specialised cells). They further argue that the epigenetic variation in the progenitor pool occurs prior to genetic mutation and can drive the variation in tumour cell characteristics during cancer evolution.

Manipulating epigenetics
Epigenetic mechanisms as drug targets
Epigenetic mechanisms are attractive targets for drug development because epigenetic changes are reversible. Thus, a suitable drug might be able to restore normal functioning to the disordered epigenetic states that characterise many diseases. This is a particularly attractive proposition for cancer. The pathological behaviours of cancer cells are the result of a combination of genetic and epigenetic changes. However, there is no obvious general way of reversing genetic changes that change the DNA sequence. The epigenetic changes may present more tractable targets, as epigenetic drugs target the readers, writers or erasers of the epigenetic marks described earlier. These enzymes are favoured targets for drug developers because there is extensive experience in designing small (and so relatively inexpensive) molecules that inhibit specific enzymes. This is just one of many approaches being pursued to develop effective anticancer drugs; it remains to be seen how generally useful this approach will turn out to be.

There are many recognised inhibitors of DNA methylation enzymes: the chemotherapy drugs azacytidine and decitabine are the two oldest. They form an irreversible complex with the enzyme that results in its degradation, preventing it from catalysing further DNA methylation. Various less toxic alternatives have been developed and are being tested in trials, including several naturally derived plant flavonoids (like one found in green tea and others sold as anti-oxidant food supplements); some of these have been shown to inhibit DNA methylation in human cancer cell lines.

In this era of multimillion pound drug development programmes, medicines developed for one purpose are being reassessed for their action in others, including cancer. Just one example is hydralazine – used to treat hypertension and heart failure – which has also been shown in experiments to inhibit the methylation of DNA and may be most effective in combination with valproic acid. Valproic acid – a long-used antiepileptic and mood stabiliser – turns out to target histone deacetylase; probably this is not its major mechanism of anti-epileptic action, but the epigenetic effect may explain some of the unwanted side-effects and perhaps suggests new applications. Some examples of anti-depressants known as monoamine oxidase (MAO) inhibitors appear to break down a histone demethylase, which may be part of its mechanism of action as an antidepressant. There are many more examples of small molecules that inhibit one or other of the enzymes active in altering epigenetic marks.

Currently the US Food and Drugs Administration has approved four such drugs:

- Azacytidine and decitabine, inhibitors of DNA methylation.

- Vorinostat and romidepsin, inhibitors of histone deacetylase.

All are licensed for treatment of specific cancers but seem likely to have wider applications. Many other drugs that target epigenetic mechanisms are currently in clinical trials for various forms of cancer. Some (GSK126 and EPZ-5676 for example) target writers; JQ1 and UNC669 target readers, while entinostat and mocetinostat target erasers of histone marks. Some caution is in order because the targeted enzymes have very general roles in gene control, thus there is legitimate concern that any inhibitor might have unwanted side-effects. Much of the current excitement in anticancer therapies centres on a quite different idea: encouraging a person's immune system to

destroy tumours. Nevertheless, researchers are pursuing many different approaches to treating cancer, and it remains to be seen which, if any, will form the basis of effective general therapies.

Manipulating epigenetics through diet

We have seen that diet, as well as drugs, might be used to modify epigenetic effects. The main interest is in DNA methylation, which may be sensitive to levels of folic acid and other B vitamins in the diet. Biochemically, folic acid (vitamin B9) is an important part of the metabolic pathways that provide the methyl groups for methylation of DNA. It is well established that folic acid supplementation in early pregnancy greatly reduces the risk of a woman having a baby with spina bifida. Futhermore, the incidence of certain childhood cancers – many of which show fewer overall DNA mutations than in adult cancers, yet a higher frequency of mutation in epigenetic regulators – declined in the USA after the introduction of mandatory folic acid fortification of cereal grain products from 1998 (PMID:25950259).

Many studies in mice and rats have investigated the effect on DNA methylation of diets high or low in methyl-donor molecules such as folic acid. No simple picture has yet emerged from these animal investigations, or from human epidemiological studies. Further work is needed to decide whether this so-called nutri-epigenomics is a beneficial enterprise. However, one study (PMID:24074862) has clearly shown how a mutation in a gene (*MTTR*) that encodes a critical enzyme in the folic acid metabolic pathway leads to birth defects in mice. These defects were related to the **genotype** of the *maternal* grandparents – even when the grandchildren had normal genotypes – indicating a clear transgenerational effect. Furthermore, there was widespread epigenetic instability shown by altered gene expression in the placentas of the same grandchildren. This study demonstrates the importance of folic acid to the satisfactory functioning of the epigenome and how disruption of folic acid metabolism can have effects in future generations.

Epigenetic manipulation in the laboratory: induced pluripotent stem cells

Development from a fertilised egg to the cells of terminally-differentiated tissues involves successive rounds of epigenetic programming that progressively channel cells into a particular restricted identity. In normal development the

37

process is irreversible, but there are exceptions. As mentioned above, the epigenetic slate must be wiped clean (or maybe almost clean) during egg and sperm cell formation and then again in the early embryo, so that all the epigenetic programmes required during development can be applied. Embryonic stem cells, obtained from the inner cell mass of **blastocyst** embryos, are pluripotent – they can potentially develop into any differentiated cell type (see GIM3 pages 19-20). These ES cells are of great interest in medicine because they might be sources of differentiated cells to replace those lost in disease. A big problem has been the limited supply of ES cells, and the need to destroy an embryo in order to obtain them.

Great excitement ensued when it was reported in 2007-8 that ordinary differentiated cells such as fibroblasts could be reprogrammed in tissue culture to a pluripotent state by a relatively simple cocktail of transcription factors. This held out the promise of a ready supply of pluripotent cells which, moreover, could be derived from a patient needing cell therapy, so avoiding any problems of mismatch and rejection. Induced pluripotent stem cells (iPSC) are not totally identical to ES cells - their epigenomes may still carry traces of their origin - but in the laboratory they can be successfully differentiated into all manner of other cell types. As research tools they have enabled scientists to study brain cells from patients with brain disease or muscle cells from patients with muscle disease. Whether re-differentiated iPS cells will prove to be safe and effective therapeutic tools remains to be seen, but the promise is great.

Darwin versus Lamarck: a new round in the battle?

British visitors to the Père Lachaise cemetery in Paris may be surprised to see the inscription on the memorial to Jean-Baptiste Pierre Antoine de Monet, Chevalier de Lamarck (1744 –1829), better known simply as Lamarck. It hails him as 'Fondateur de la doctrine de l'evolution'. Shouldn't that be Darwin?

This is not entirely the French trying to steal British glory. Lamarck can indeed lay fair claim to being at least a main proponent of the idea of evolution, long before Darwin published *On the Origin of Species*. The argument is about mechanism.

The traditional battle ground is the giraffe's neck. How did that long neck evolve?

- On a Lamarckian view, giraffes that stretched up to feed on leaves from overhanging boughs not only lengthened their own necks by the exercise but also had offspring with longer necks as a direct consequence of the stretching. Many generations of incremental lengthening resulted in the current long neck.

- On a Darwinian view, giraffes had genetically determined variation in the natural length of their necks. In times of scarcity those with longer necks were able to access additional food, and so were more likely to survive. Their offspring inherited the genes predisposing to a longer neck. Many generations of such natural selection resulted in the current long neck.
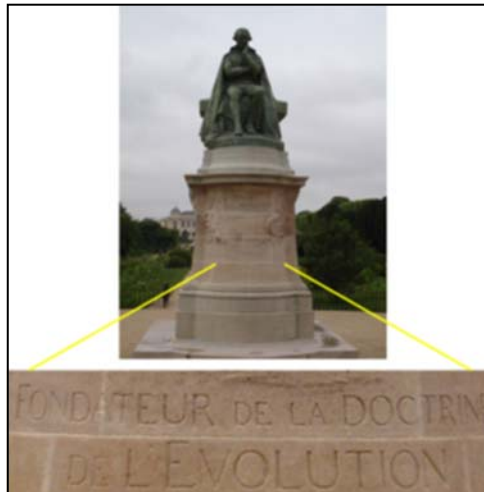
Lamarck's vision is kinder. Whole populations evolve peaceably, whereas Darwin's model relies on competition, starvation and limited numbers of survivors. But as a general explanation of evolution Lamarckism lacked a credible mechanism or experimental support. Darwin, as an outstanding

naturalist, could furnish innumerable examples to support his concept of evolution by natural selection (even if he did not have a genetic theory to explain how it could work cumulatively over many generations). During the early twentieth century the question of inheritance of acquired characters – a prerequisite for Lamarckian evolution – became exceedingly contentious. Positions on the Darwinian side became rigid and dogmatic. The Lamarckian biologist Paul Kammerer was driven to suicide by accusations that he had faked data supporting his position. Appreciation of the role of DNA appeared to settle the controversy once and for all in favour of Darwin. Environmental effects can produce *random* mutations on which selection can operate, but not the *targeted* changes in the DNA sequence that Lamarckian inheritance would require.

Life events do, however, appear to influence epigenetic marks. Given the evidence for transgenerational effects that do not appear to be genetically conferred, Lamarckism now seems to have authenticated examples backed by a potentially plausible mechanism, including epigenetic ones. This does not mean that Darwinism is overthrown or even that Lamarckism is a true alternative. Evidence for evolution driven by natural selection abounds wherever one looks. Large-scale evolution is unquestionably Darwinian. It does, however, suggest that a softening of rigid stances is in order. A modern view should perhaps develop in two stages. The first stage is to become a rigid Darwinian. An understanding of evolution must be based on natural selection and rejection of Lamarckism as a general mechanism. But once the Darwinian base is established the second stage requires appreciation that things are not always quite so simple. Natural selection works on phenotypes not DNA sequences – a problem for dogmatic 'selfish gene' theories. Concepts like the evolution of adaptability blur the boundaries. Microorganisms frequently transfer genes between individuals of different species, which constitutes a clear case of inheritance of acquired characters. The same may occur with higher organisms. One review (PMID:27846492) quotes examples of non-DNA inheritance from a remarkably wide range of organisms, from microorganisms through plants, fungi and nematode worms to mammals. Unquestionably, acquired characters can sometimes be found in the next generation. Somehow there must be a

mechanism that confers this apparent heritability. Having reached this more nuanced view, there remains space for debate about just how important epigenetic mechanisms have been in potentially reversible phenotypic changes over one or two generations, and in overall evolution.

The future

Gene regulation holds the key to understanding how we function, in sickness and in health. The DNA sequencing revolution of the past decade has generated massive amounts of data on our DNA sequence and how it varies, but until recently epigenetics had scarcely entered the era of Big Data. Soon, we will be able to sequence epigenomes routinely, and new techniques and big collaborative projects are producing a flood of data on how our genomes are regulated. The potential complexity is quite frightening – we each have just two copies of our genomes, but 200 or more epigenomes. However, the rewards of a deep understanding of human genome function are immense. As we learn more about the 'epigenetic landscapes' of our genomes, we will understand more of the fundamental mechanisms that predispose a person to disease, particularly those that appear to be on the rise in the 21st century such as obesity and diabetes, heart disease, autism and atopy. Hopefully a fuller understanding of causes will also uncover more targets for epigenetic-based therapies.

Glossary

Adenine – one of the four nucleotides which make up the basic units of DNA or RNA. Represented by the letter A.

Amino acids – organic compounds that are the building blocks of proteins.

Antibody – a protein produced by the immune system in response to a foreign substance, called an antigen eg: pathogenic bacteria and viruses. Antibodies recognise and latch onto antigens to remove them from the body.

Atopy – allergic diseases such as allergic rhinitis, asthma and atopic dermatitis (eczema).

Autoimmune – antibodies or lymphocytes produced by a person's own immune system acting against substances naturally present in the body.

Base-pair – a single letter of DNA sequence, base-paired with its complement base on the opposite strand of the double helix.

Biomarker – a naturally occurring molecule, gene, or characteristic by which a disease can be identified.

Blastocyst – an early embryo which has developed to the point of having two different cell components and a fluid cavity, usually reached by day five after fertilisation.

Chromatin – a complex of DNA and proteins that forms chromosomes within the cell nucleus.

Chromosome – one of the DNA-protein packages into which the human genome is packed in the cell nucleus.

Coding DNA – DNA containing the genetic code for a protein.

Cohort – a group of people with a shared characteristic. Cohort studies are a type of medical research used to investigate the causes of disease and to establish links between risk factors and health outcomes.

Complex condition – a condition that is caused by a mixture of genetic and environmental factors.

Cytoplasm – all the material within a living cell except the cell nucleus, including all the cell's other internal sub-structures.

Cytosine – one of the four nucleotides which make up the basic units of DNA or RNA. Represented by the letter C.

DNA – deoxyribonucleic acid, the ultimate repository of genetic information.

Deletion – a missing segment of a gene or chromosome.

Diabetes – a condition of having higher than normal blood sugar levels.

Double helix – the three-dimensional structure of double-stranded DNA, in which two nucleotide strands linked by hydrogen bonds form a helical configuration. The two DNA strands are oriented in opposite directions.

Down syndrome – a genetic condition caused by the presence of all or part of an extra copy

of chromosome 21.

Embryo – the earliest stages of development in the womb up to ~9 weeks post-fertilisation.

Enzyme – a protein catalyst, acting to bring about a specific biochemical reaction.

Epigenome – the complete set of epigenetic modifications of DNA of a genome.

Expressed gene – a gene that is switched on to produce the protein it encodes.

Fetus – the stages of development in the womb from ~9 weeks post-fertilisation until birth.

Gamete – a germ cell (egg or sperm).

Gene – the unit of heredity, a sequence of DNA, transferred from a parent to child.

Genome – the totality of genes or genetic material of an individual.

Genotype – the genetic makeup of an organism with reference to a single trait or set of traits (cf phenotype).

Genetic code – the information encoded within the DNA of the genome that is translated into proteins by living cells.

Germ cell – a gamete (sperm or egg cell).

Guanine – one of the four nucleotides which make up the basic units of DNA or RNA. Represented by the letter G.

Histone – one of a number of proteins that associate with DNA in the nucleus of the cell and help condense it into chromatin.

Incidence – the number of individuals who develop a disease during a particular time (cf **prevalence**).

Meiosis – the specialised type of cell division that produces gametes. During meiosis the chromosome number is halved, and maternal and paternal copies of genes are shuffled so that each gamete is genetically unique.

Mendelian – of a character, determined by a single gene.

Messenger RNA – during transcription, a single strand of DNA is decoded by RNA polymerase, and messenger RNA (mRNA) is synthesised. An mRNA molecule carries a portion of the DNA code to other parts of the cell for processing.

Metabolism – the chemical processes that occur in a living cell or organism.

Methylation – a chemical modification of a molecule, adding one or more methyl ($-CH_3$) groups. Methylation of specific bases in DNA and of histones is used as a signal to change the way the DNA is packaged and genes are expressed.

Mitosis – the normal process of cell division, which creates two genetically identical daughter cells (cf meiosis).

Mutation – a change/variant in the DNA sequence in a genome.

Non-coding DNA – DNA that does not code for protein. It may, however, still have important functions, encoding functional RNAs (like *XIST*) or controlling gene expression.

Neurone – a nerve fibre.

Nucleosome – the fundamental subunit of chromatin in the nucleus of the cell, comprising a length of DNA coiled around a core of histones.

Nucleotide – the basic unit of DNA or RNA, consisting of a base (normally adenine, guanine, cytosine or thymine in DNA; adenine, guanine, cytosine or uracil in RNA), a sugar (deoxyribose in DNA, ribose in RNA) and a phosphate.

Phenotype – the observable properties or behaviour of an organism (cf genotype).

Prevalence – the total number of individuals in a population who have a disease or health condition at a specific time (cf incidence).

Progenitor cells – cells that are more specific than a stem cell, but not fully differentiated. These cells go on to divide, becoming fully differentiated 'target' cells.

Proteins – large molecules made up of chains of amino acids that have diverse functions, eg: as structural elements in tissues, as antibodies in the immune system and as enzymes, catalysing chemical reactions in the body.

RNA – ribonucleic acid, closely related to DNA (deoxyribonucleic acid). RNA molecules are very heterogeneous and have many different functions in the cell.

Single-gene condition – a disorder that is caused by a mutation in a single gene.

Stem cell – an undifferentiated cell of a multicellular organism, capable of giving rise to cells of the same type or, by differentiation, other kinds of cell.

Synapse – the junction between two neurones, over which a neurotransmitter acts to pass an electrical signal from one neurone to another.

Syndrome – a combination of clinical features which occur together and are due to the same underlying defect or factors.

Thymine – one of the four nucleotides which make up the basic units of DNA. Replaced by uracil in RNA. Represented by the letter T.

Tissue – groups of similar cells that, together, carry out a specific function.

Transfer RNA – small RNA molecules that are part of the mechanism by which messenger RNAs specify the amino acid sequence of proteins.

Transcription – the process that involves copying genetic information from DNA to RNA.

Transcription factor – a protein that binds DNA to control the rate of transcription of genetic information from DNA to messenger RNA.

Translation – the process that converts the messenger RNA code into protein.

Tumour – an abnormal growth creating a mass of tissue; benign or malignant (cancerous).

Uracil – one of the four nucleotides which make up the basic units of RNA in place of thymine in DNA. Represented by the letter U.

X chromosome, Y chromosome – the sex chromosomes. Males have one X and one Y, females have two Xs.

Zygote – the fertilised egg cell, before the first cell division.

Further Information

(All websites accessed 14 September 2018)

Roadmap Epigenomics Consortium
www.roadmapepigenomics.org/

International Human Epigenome Consortium
www.ihec-epigenomes.org/

The Sanger Centre – educational resources
https://www.yourgenome.org/

DNA Learning Center – resources
https://www.dnalc.org/resources/

How DNA is packaged
https://www.dnalc.org/resources/3d/07-how-dna-is-packaged-basic.html

PubMed
https://www.ncbi.nlm.nih.gov/pubmed
PMID numbers are the ID numbers of journal articles in the PubMed database. Entering the number on the website will bring up the title, author list and abstract of the article, although unless the abstract is marked 'Free' you will need a personal subscription or access through a university library or similar to access the full text.

Genetics in Medicine 4. Epigenetics and Gene Regulation

This is the fourth booklet in a series on genetics in medicine, published by The Galton Institute. This booklet is aimed at non-specialists with an interest in the area and is about the burgeoning field of study of heritable and non-heritable changes in gene expression that do not alter the underlying DNA sequence.

In this booklet we describe how epigenetic processes are central to the growth and development of an individual, and how when these processes are disrupted, serious genetic disorders can result. We also look at some of the research on the role of the environment in influencing gene regulation and how this is being related to common disorders such as obesity and diabetes, including the more speculative theory of fetal origins of adult disease and the associated work on transgenerational effects.