

INTERNATIONAL ADVANCED LEVEL

BIOLOGY

TOPIC GUIDE: EPIGENETICS

Pearson Edexcel International Advanced Subsidiary in Biology (XBI11)

Pearson Edexcel International Advanced Level in Biology (YBI11)

First teaching September 2018

First examination from January 2019

First certification from August 2019 (International Advanced Subsidiary) and
August 2020 (International Advanced Level)



Contents

Introduction	3
The genome	4
DNA sequencing	4
Using genome sequence to define species	5
Analysing evolutionary patterns	6
How the amino acid sequence of proteins is determined	7
Links to genetic disorders	8
Introns, exons and splicing	10
Factors affecting gene expression	11
Promoters	12
Enhancers	12
Transcription factors	12
DNA methyltransferases	13
Histone modifying enzymes and chromatin remodelling complexes	14
Regulatory RNAs	15
Epigenetic memory	16
Stem cells	17

Introduction

This guide is intended to provide additional teaching support material and background information for the following aspects of the new Pearson Edexcel IAL Biology 2018 qualification.

2.14 - (i) understand how errors in DNA replication can give rise to mutations (substitution, insertion and deletion of bases)

7.22 - understand how genes can be switched on and off by DNA transcription factors

3.19 - understand how one gene can give rise to more than one protein through post-transcriptional changes to messenger RNA (mRNA)

3.20 (ii) know how epigenetic modification, including DNA methylation and histone modification, can alter the activation of certain genes

(iii) - understand how epigenetic modifications can be passed on following cell division

3.11 (i) understand what is meant by the terms *stem cell*, *pluripotent* and *totipotent*, *morula* and *blastocyst*

(ii) be able to discuss the ways in which society uses scientific knowledge to make decisions about the use of stem cells in medical therapies

6.19 - understand how DNA profiling is used for identification and determining genetic relationships between organisms (plants and animals)

It assumes you are already familiar with the structure of DNA and RNA (including 5' and 3' ends) and the basics of gene transcription, translation and the genetic code.

The same material is also found in Pearson Edexcel GCE A level Biology A, Topic 3 (The voice of the genome).

The genome

Your genome is the totality of your DNA – not just the protein-coding genes, but all the non-coding DNA within (introns) and between the protein-coding genes. It does not include all the various RNA species present in cells.

One of the surprising features of the human genome is how little of it is protein-coding – only about 1.2%. The same is true of the genomes of other higher organisms. About half of the rest is repetitive, comprising huge numbers of copies of certain short sequences whose function, if any, is mostly unknown. Much of the non-repetitive DNA is involved in regulating expression of the protein-coding sequences. Gene regulation is the subject matter of epigenetics.

DNA sequencing

The standard technique for identifying the sequence of nucleotides in a piece of DNA was developed by Dr Fred Sanger in Cambridge in the 1970s. It earned him a share of the 1980 Nobel Prize for Chemistry. It works by using a DNA polymerase enzyme to make copies of the DNA to be sequenced, but spiking the pool of individual nucleotides with a small amount of a chemically modified nucleotide (a dideoxy nucleotide) that will terminate growth of any copy in which it gets incorporated (Figure 1).

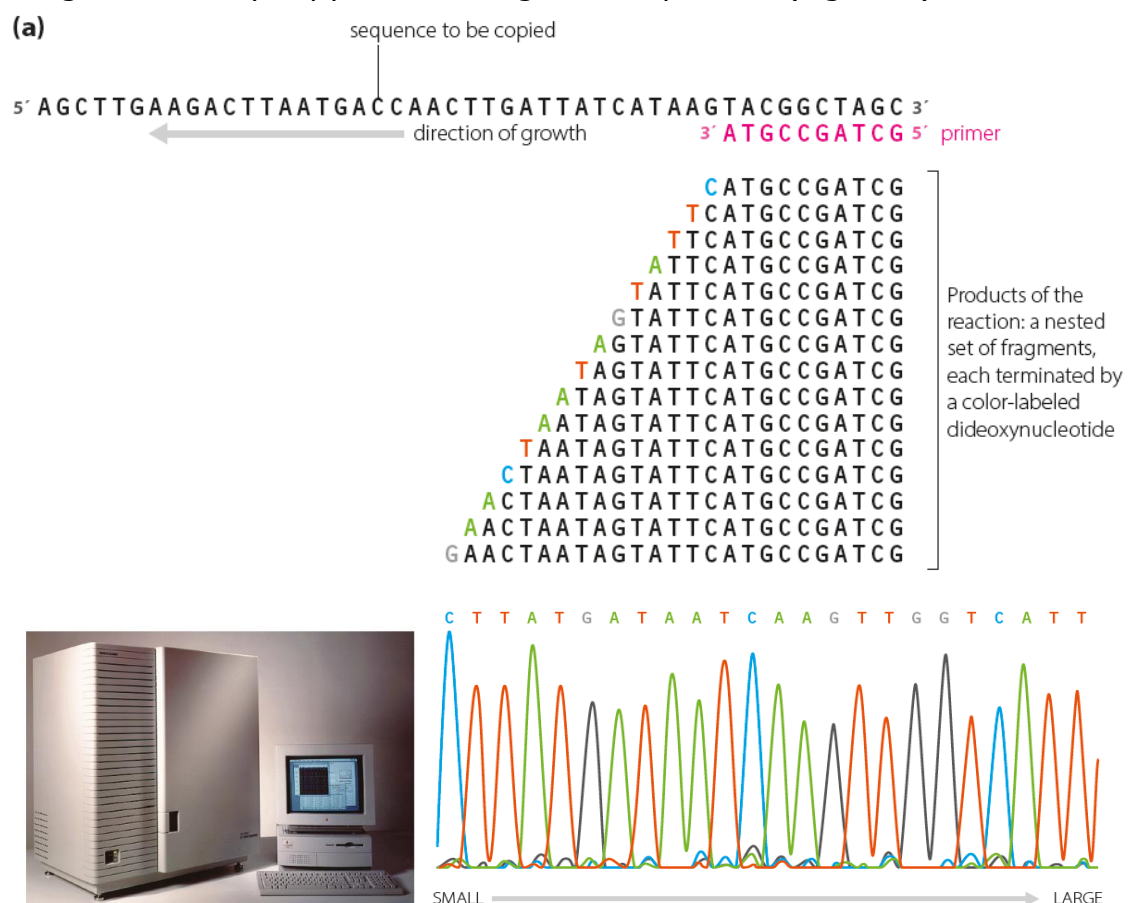


Figure 1: The principle of dideoxy (Sanger) sequencing of DNA. (a) DNA polymerase makes many copies of the test sequence by extending a specially designed primer oligonucleotide. Whenever by chance it incorporates a dideoxy nucleotide instead of the corresponding normal deoxy nucleotide, the chain terminates. Each dideoxy nucleotide is tagged with a different coloured molecule. (b) An automated sequencing machine uses electrophoresis to separate the reaction products by size. (c) It reads the colours and shows the sequence as a series of coloured peaks. From *New Clinical Genetics*, Read & Donnai, Scion Publishing 2015.

Sanger's method can give very accurate sequence of a DNA fragment up to around 800 base pairs in length. The Human Genome Project used Sanger sequencing (on banks of automated sequencing machines); it was necessary to piece together millions of short sequences in the computer to produce the overall 3200 million base pair human genome sequence. It took 15 years and cost around 3 billion dollars.

Starting around the year 2005, a number of revolutionary new DNA sequencing technologies became available. Different competing companies produced different methods, but all the so-called 'Next-Generation Sequencing' methods have in common that they sequence millions of random DNA fragments in parallel. Depending on the technology, the fragments may be fixed on nanobeads in arrays of tiny wells; they may be anchored in arrays to a solid surface, or they may be in arrays of nanopores in a membrane. Sequencing works by synthesis, like Sanger sequencing. In different technologies each nucleotide added generates a light signal or a pulse of hydrogen ions. Whatever the detailed technology, use of these methods has vastly increased the amount of DNA a lab can sequence, to the point that it is now possible to sequence an individual's whole genome in a week for around £1,000. We are only beginning to see the impact of this new capability on the National Health Service.

Using genome sequence to define species

Classically, species are defined as groups of individuals able to interbreed and produce fertile offspring. That requires observation of their behaviour, and maybe experimental crosses. An alternative approach is to consider their genome sequence. This is not completely straightforward, because genome sequences vary between individuals of the same species – reflecting the fact that we are all different. But just as we can readily appreciate that all humans, despite their individual differences, are more similar to each other than to chimpanzees, so we can see from the DNA sequence that humans are one species and chimpanzees another.

An interesting example of defining species based on DNA sequence concerns the Wood White butterfly, *Leptidea synopsis* (Figure 2).

This butterfly, rare in Britain though less so in Ireland, looks fairly similar to the common Small White (*Pieris rapae*), but be readily distinguished by a trained eye. However, it has turned out that 'Wood Whites' actually comprise three species, *L. synopsis*, *L. reali* and *L. juvernica* that can only distinguished reliably by sequencing their DNA (Dincă *et al.*, 2011).



Figure 2: Wood White butterfly
©Davidtomlinson
photos.co.uk

can

be

Analysing evolutionary patterns

When genome sequences of related species are compared, the degree of difference between each pair can be used to construct an evolutionary tree. One might use the DNA sequences of one or a few selected genes that are present in each species. Alternatively, the gene sequences can be translated to give the amino acid sequences of the proteins they encode. This approach is preferred for more distantly related species, because it ignores changes that simply convert one codon for an amino acid into another for the same amino acid (see below). Constructing a tree for real uses computer programs that apply elaborate statistical arguments (there is an example in the Dincă *et al*/ paper mentioned above).

Figure 3 shows a simple example.

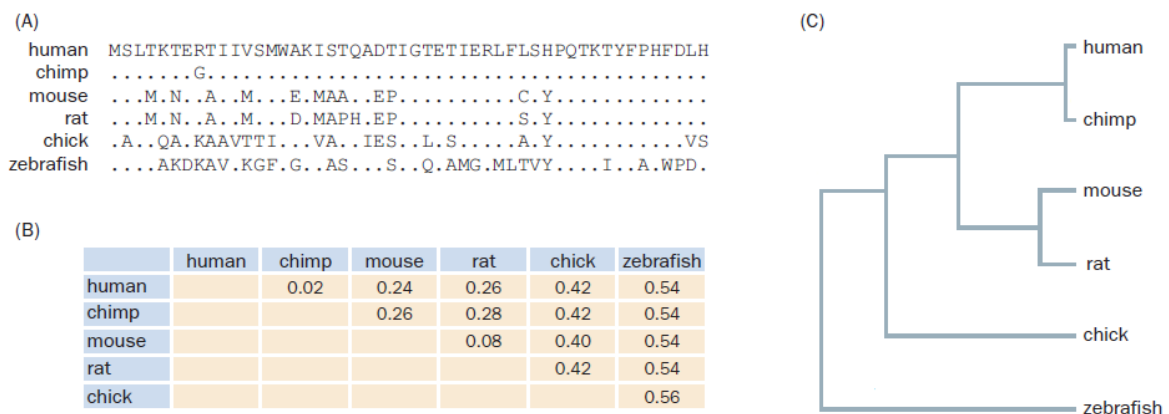


Figure 3: Comparison of the last 50 amino acids of the zeta-globin protein in six species. (a) the raw sequences, using 1-letter codes for the amino acids (see below). Dots show unchanged amino acids. (b) tabulation of pairwise differences. For example, humans and chimps differ at 1 position out of 50, so the difference is 0.02. (c) tree constructed from the data. You can see how human/chimp and mouse/rat form close couples; then chick is about equidistant from both, and zebrafish equidistant from all five. The distances can be used to estimate the time of divergence, but to do that properly requires heavy statistics and computing. From *Human Molecular Genetics* Strachan & Read, Garland 2011.

Possible teaching approach

Class could construct the table and tree from the data. Other examples can be found on the Web – some possibilities include:

<http://www-tc.pbs.org/wgbh/evolution/educators/teachstuds/pdf/unit3.pdf>

http://evolution.berkeley.edu/evolibrary/article/0_0_0/phylogenetics_01

http://serc.carleton.edu/sp/process_of_science/examples/73104.html

How the amino acid sequence of proteins is determined

When a protein-coding gene is expressed, the enzyme RNA polymerase synthesises an RNA molecule (messenger RNA) that is complementary to the sequence of one strand of the DNA (the *template strand*) and identical to the sequence of the other strand (the *sense strand*). Databases and publications always cite the sequence of the sense strand, written in the 5' – 3' direction (Figure 4).

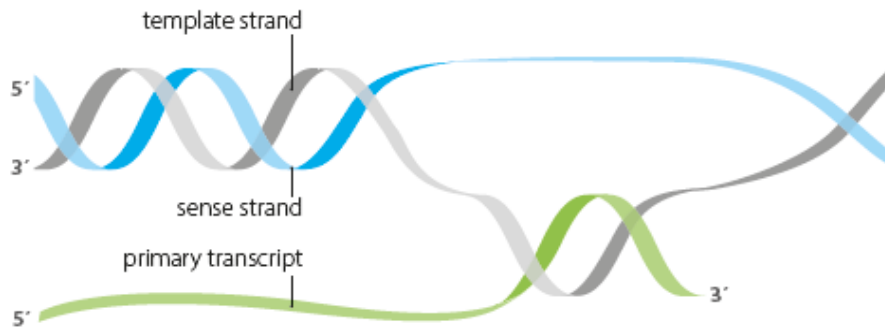


Figure 4: From *New Clinical Genetics*, Read & Donnai, Scion Publishing 2015.

Possible teaching approach

Give the class a DNA sequence as conventionally written (you can get any number of real examples from http://www.ensembl.org/Homo_sapiens/Info/Index).

Ask them to write the complementary strand, in the conventional 5' – 3' direction. Then ask them to translate each strand using the table of the genetic code. The results are completely different, making the point about the sense strand and template strand.

An alternative would be to give them a sequence of the bases on a template strand and get them to predict the sense strand, the mRNA, the tRNA and the amino acid sequence. Then they should do it backwards to prove it produces a completely different amino acid sequence.

The messenger RNA (after splicing out any introns, see below) is 'read' by ribosomes. A ribosome attaches at the 5' end of the mRNA and slides along until it encounters a start signal: the triplet AUG embedded in a suitable consensus sequence (known as the Kozak sequence). It then starts assembling a polypeptide chain, the choice of amino acid at each position being determined by a triplet of three consecutive nucleotides in the mRNA.

Individual amino acids are covalently attached to specific small RNA molecules, they transfer RNAs, by amino acid-activating enzymes that are specific for each type of transfer RNA. Three nucleotides on the transfer RNA base-pair with three nucleotides of the mRNA within a special pocket of the ribosome. When the ribosome encounters a stop codon it falls off the mRNA and releases the polypeptide it has been making.

The genetic code (Figure 5) consists of unpunctuated non-overlapping triplets of nucleotides.

UUU	Phe (F)		CUU	Leu (L)		AUU	Ile (I)		GUU	Val (V)
UUC			CUC			AUC			GUC	
UUA			CUA			AUA			GUA	
UUG			CUG			AUG	Met (M)		GUG	
UCU	Ser (S)		CCU	Pro (P)		ACU	Thr (T)		GCU	Ala (A)
UCC			CCC			ACC			GCC	
UCA			CCA			ACA			GCA	
UCG			CCG			ACG			GCG	
UAU	Tyr (Y)		CAU	His (H)		AAU	Asn (N)		GAU	Asp (D)
UAC			CAC			AAC			GAC	
UAA	STOP		CAA	Gln (Q)		AAA	Lys (K)		GAA	Glu [E]
UAG			CAG			AAG			GAG	
UGU	Cys(C)		CGU	Arg [R]		AGU	Ser (S)		GGU	Gly (G)
UGC			CGC			AGC			GGC	
UGA	STOP		CGA			AGA	Arg [R]		GGA	
UGG	Trp (W)		CGG			AGG			GGG	

Figure 5: The genetic code as mRNA codons. The corresponding DNA sequence in the sense strand has the complementary bases (so A would be T) and T instead of U. By writing out the nucleotide sequence of a protein-coding gene, you can predict the amino acid sequence of the protein it encodes. Amino acids have a standard three letter abbreviation (eg. Arg = Arginine, Leu =Leucine, but to save space in Fig. 6 they have been given a one letter code in the above.

Links to genetic disorders

Replacing one nucleotide by another, substitution, in a protein-coding gene can have one of three effects: a synonymous variant, a mis-sense variant or a nonsense variant (Figure 6).

- (a) ATG GTG CAT CTG ACT CCT GAG GAG AAG TCT GCC GTT...
 M V H L T P E E K S A V ...
- (b) ATG GTG CAT CTG ACT CCT GAG GAG AAG TCA GCC GTT...
 M V H L T P E E K S A V ...
- (c) ATG GTG CAT CTG ACT CCT GTG GAG AAG TCT GCC GTT...
 M V H L T P V E K S A V ...
- (d) ATG GTG CAT CTG ACT CCT GAG TAG AAG TCT GCC GTT...
 M V H L T P E STOP

Figure 6: (a) the coding sequence for the start of the beta-globin gene, with the amino acids encoded. (b) A substitution mutation leading to a synonymous (same-sense) change that does not affect the amino acid encoded. (c) A substitution mutation leading to a mis-sense change, replacing glutamic acid with valine (this is the sickle cell variant; as is usually the case, the initial methionine is cleaved off during post-translation processing, so the variant can be described as Glu6Val). (d) A substitution mutation leading to a nonsense change, introducing a premature stop codon. All the coding sequences are shown as they would be on the DNA sense strand.

Inserting or deleting one or more nucleotides has a more drastic effect: it alters the reading frame (a frameshift change) and so changes the entire amino acid sequence downstream of the change.

```

(a)   ATG GTG CAT CTG ACT CCT GAG GAG AAG TCT GCC GTT...
      M  V  H  L  T  P  E  E  K  S  A  V ...
(b)   ATG GTG CAA TCT GAC TCC TGA GGA GAA GTC TGC CGT T...
      M  V  Q  S  D  S  STOP
(c)   ATG GTC ATC TGA CTC CTG AGG AGA AGT CTG CCG TT...
      M  V  I  STOP

```

Figure 7: (a) the wild-type beta-globin sequence. (b) inserting a single nucleotide alters the entire message (and in this case introduces a premature stop codon). (c) deleting a single nucleotide again alters the entire message (and, again introduces a premature stop codon)

Possible teaching approach

This readily lends itself to class exercises. To access endless examples, go to http://www.ensembl.org/Homo_sapiens/Info/Index; enter a gene or condition, e.g. 'cystic fibrosis', 'Factor VIII'. From the list that appears, click on a promising looking gene; click on a transcript, then the 'cDNA' item on the top left.

Predicting the effect of a change on the protein encoded is fairly straightforward (and can be made the subject of many classroom exercises). Predicting the effect on the person carrying the variant is not at all straightforward. Some changes will have a major effect, like the sickle cell mutation. Some will slightly alter the structure or activity of the protein, maybe contributing a little to susceptibility or resistance to a common multifactorial (not monogenic) condition like diabetes or hypertension. Some will have no overt effect on the person, even if there is a very major effect on the protein – some proteins are not important, or their role can be taken over by other proteins.

Moreover, supposing a sequence change makes an important protein non-functional, we can ask whether we can get by with a single functional copy (remember, we are diploid and have two copies of each autosomal gene), or whether 50% overall function is not sufficient. In the first case the condition will be recessive: carriers of one non-functional copy will be normal. Cystic fibrosis is an example. In the second case the condition will be dominant – for example, individuals with achondroplastic dwarfism have a single malfunctioning copy of the *FGFR3* (fibroblast growth factor receptor 3) gene. Which of these alternatives happens depends on the detailed role of that particular protein in the cells where it is expressed.

The general conclusion is that without very detailed knowledge of the particular protein and its exact role in the biology of specific cells, it is impossible to predict the phenotypic effect of a DNA sequence change, however radical the effect may be on the encoded protein.

Introns, exons and splicing

In most genes in humans and other multicellular organisms, the protein-coding sequence is split into segments (exons) that are separated by non-coding sequence (introns). This arrangement was a complete surprise when first discovered in the late 1970s. Bacterial genes, which were the best understood genes at the time, do not have introns. It seems completely counter-intuitive. The number of exons in genes varies with no apparent logic (Figure 8). The average is around 8–10, but there are genes with no introns, and the record is held by the gene for the muscle protein titin, which has 362 exons.

Gene sizes also vary independently of the number of exons, because introns vary extremely widely in size, both within and between genes. Some introns are only a few dozen base pairs, some are more than 100 kilobases. In Figure 8, all the gene diagrams have been made to fit the box, but the real sizes vary widely: 1.43 kb for the insulin gene, 1.61 kb (beta-globin), 4.62 kb (*HLA-A*), 80.72 kb (phenylalanine hydroxylase) and 188.7 kb (*CFTR*, the gene mutated in cystic fibrosis).

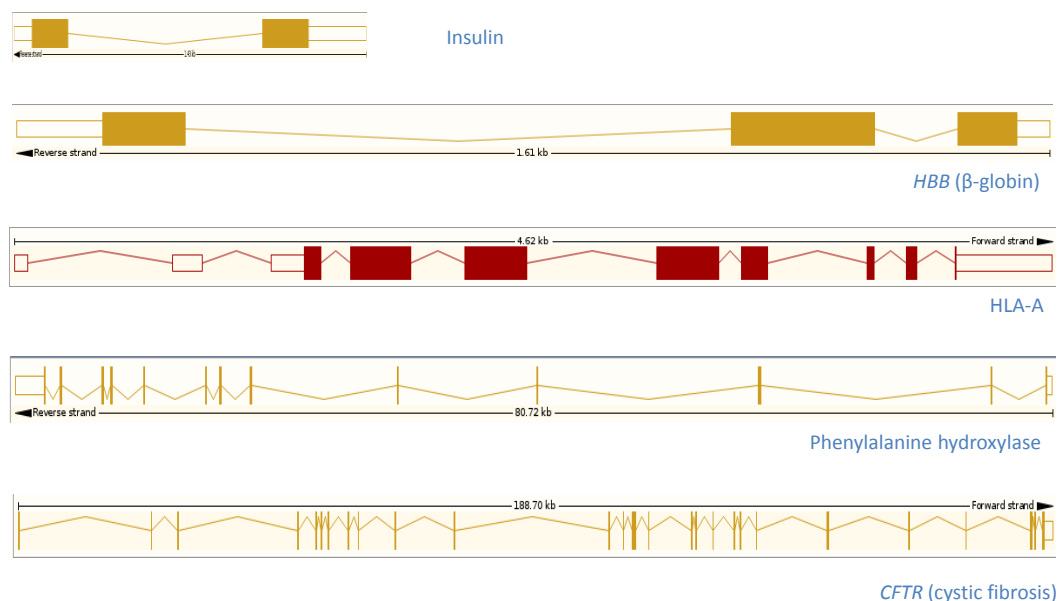


Figure 8: Data from Ensembl showing introns and exons.

When a gene is transcribed, the RNA polymerase traverses the entire sequence, exons and introns, to make the primary transcript. This is then processed, within the nucleus, by being physically cut at exon-intron boundaries; the exons are spliced together to make the mature mRNA, and the introns are discarded. The machinery that does this, the spliceosome, is exceedingly complicated, incorporating five species of small RNAs and around 170 different proteins. Many transcripts can be spliced in more than one way – certain exons may be sometimes incorporated and sometimes skipped.

Alternative splicing is often tissue-specific, and the different splice isoforms may have clearly different functions. For example, some proteins exist in either a cell-surface form or a secreted form, depending whether an exon encoding a transmembrane domain is included in the final spliced mRNA.

Alternative splicing is not a peculiar and exceptional event, it is quite normal. The average gene encodes about 5 different splice isoforms, and there are genes (neurexin

B, for example) that encode over 1,000. This forces a significant extension to the one-gene-one-enzyme hypothesis of Beadle and Tatum.

Possible teaching approaches

1. ask students to identify the parts of a gene (Figure 8)

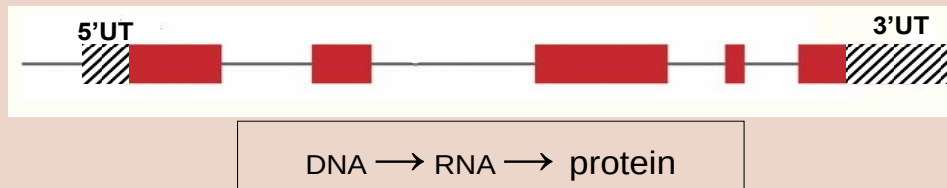


Figure 9

Where is the:

- promoter
- transcription start site
- transcription termination site
- 5' end of exon 3
- 3' end of intron 2?

What are the 5' and 3' untranslated regions?

2. Ask groups of students to access a gene in Ensembl (url as above), and to report the number of exons, the number of different transcripts and the relation between them. Suitable simple genes are *HBB* (beta-globin) or *GJB2* (connexin 26, mutated in about half of autosomal recessive profound childhood deafness). More complex genes could include *CFTR* (cystic fibrosis), *BRCA1* (familial breast cancer) and *PAX3* (mutated in the Waardenburg syndrome of hearing loss and pigmentary anomalies). The Ensembl entries include diagrams showing the exons of each transcript.

Factors affecting gene expression

Some genes are expressed in every cell of our body (so-called housekeeping genes) but most are not. Haemoglobin is made only by red cell precursors, keratins only in skin and hair; the *ADH4* alcohol dehydrogenase gene is expressed only in liver cells. Tissue-specific gene expression is the key to our complexity, compared to simpler organisms. How is it achieved?

For a gene to be expressed, two things are necessary:

- the DNA must be accessible, not buried in densely packed chromatin (a complex of highly folded DNA packaged with histone proteins)
- sequence-specific DNA-binding proteins (transcription factors) must bind to the promoter, upstream of the sequence to be transcribed, to help recruit RNA polymerase.

These depend on the interactions of a complex set of players – sequence elements (promoters and enhancers), proteins (including transcription factors, DNA methyltransferases, histone-modifying enzymes and chromatin remodelling complexes), and a battery of small RNA species.

Promoters

In order to transcribe a gene, the RNA polymerase must attach to the DNA just upstream of the transcription start site. This region is called the promoter. Binding is determined by the DNA sequence, but also by sequence-specific binding of a whole set of other proteins that together constitute the transcription initiation complex. Individual protein-DNA interactions may be quite weak, but they are cemented by protein-protein interactions. Some of those other proteins are present only in certain cells, and the many possible combinations are one route to tissue-specific gene expression.

Enhancers

Enhancers are promoter-like sequences that are located some way away from the gene they regulate. They can be upstream or downstream of the gene, and in some cases up to a million base pairs away. Like promoters, they bind a variety of proteins, many of them tissue-specific, and the DNA loops round to bring them into contact with the promoter (Figure 10). Many genes are controlled by a variety of different tissue-specific enhancers.

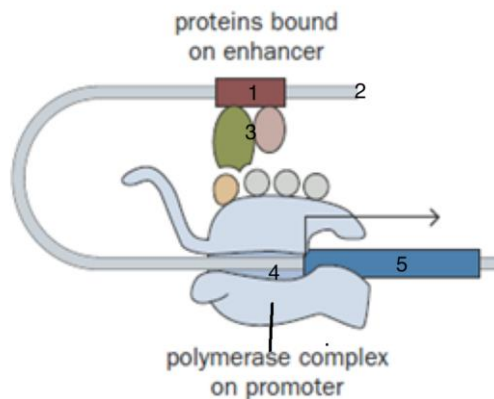


Figure 10

Gene ready to be transcribed. 1 enhancer, 2 DNA, 3 transcription activator proteins, 4 promoter, 5 gene.

Transcription factors

Transcription factors are proteins that bind to promoters and enhancers. There are general transcription factors, present in every cell and part of the basal transcription machinery, and tissue-specific factors. These in turn are produced by genes that are themselves controlled by other transcription factors, allowing a cascade of regulatory effects. Acting in a combinatorial way, around 1000 transcription factors can exert subtle control over the expression of our 20–25 000 protein-coding genes.

DNA methyltransferases

These add methyl ($-\text{CH}_3$) groups to DNA, specifically to the 5-position of cytosines that lie immediately upstream of guanines (so-called CpG dinucleotides, the p representing the phosphate joining adjacent nucleotides). 5-methyl cytosine base-pairs with guanine exactly the same as normal cytosine, but the methyl groups act as a signal to methyl DNA binding proteins, which in turn recruit other regulatory proteins.

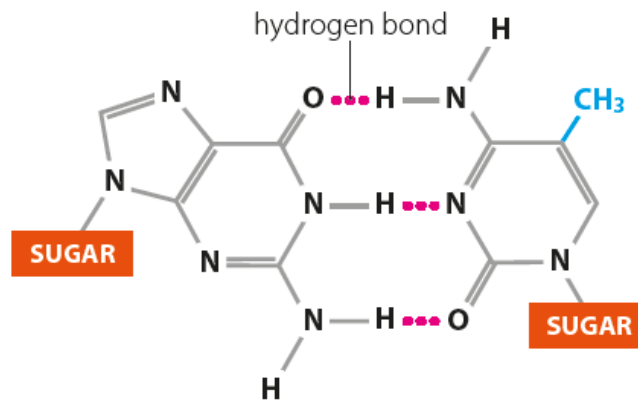


Figure 12: From *New Clinical Genetics*, Read & Donnai, Scion Publishing 2015.

Histone modifying enzymes and chromatin remodelling complexes

Every diploid human cell nucleus contains 2 metres of DNA.

Possible teaching approach

Chemically adept students could work out the molecular weight of an A-T or G-C base pair, given the formulae of nucleotides (the answer is about 550). A diploid cell contains about 6 picograms (6×10^{-12} g) of DNA. Students can use this, together with the Avogadro number (6×10^{23}), to work out the number of base pairs in a (diploid) cell. Having worked that out, or being given the figure of 6×10^9 , and knowing the spacing of base pairs is 0.34 nm (from the X-ray diffraction work of Rosalind Franklin), students can work out the length of DNA in a (diploid) cell. Given that a person consists of around 10^{13} cells, they can then work out the length of DNA in their body. If nothing else, this should give them practice in manipulating indices, and illustrate how thin the DNA double helix must be!

The DNA needs to be tightly packaged to fit into the nucleus, and the first level of packaging is into nucleosomes. A nucleosome is an octamer of histones (small basic proteins whose positive charge gives them an affinity for the negatively charged phosphate groups of DNA). Each nucleosome contains two molecules each of histones H2A, H2B, H3 and H4, with 147 base-pairs of DNA wound round it. At the basic level, DNA is organised into a string of beads, nucleosomes, separated by variable lengths of spacer DNA.

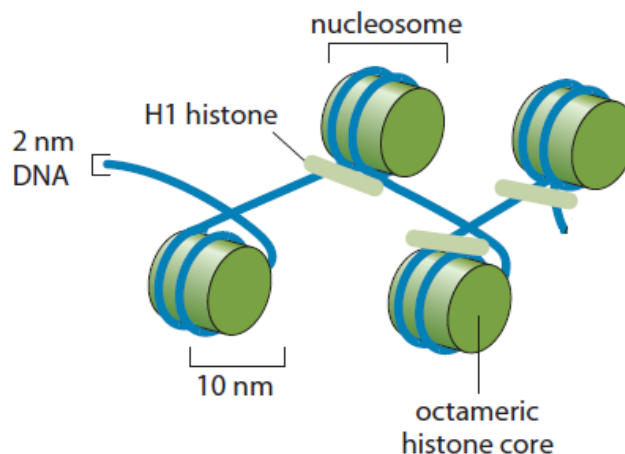


Figure 13: Nucleosomes. Histone H1 is not part of the nucleosome, but binds the immediately adjacent DNA.

If a gene is to be expressed it must be in accessible chromatin. DNA that is wrapped up in nucleosomes (and especially when the string of beads is in turn tightly coiled in higher levels of packaging) is not accessible to RNA polymerase and the other DNA-binding proteins necessary to initiate transcription. Chromatin-remodelling complexes are large ATP-driven multi-protein complexes that control the positioning of nucleosomes along the DNA so as to make specific promoters available for transcription.

In nucleosomes, the histone molecules have protruding N-terminal 'tails' that can interact with other proteins. Different proteins bind to the histone tails to stimulate or inhibit transcription. Binding is controlled by covalent modifications to the histone tails.

Specific enzymes tag particular amino acid residues in specific histones with methyl, acetyl and other groups to allow complex and flexible control of gene expression. There are 'writers' that apply the tags, 'readers' that bind in response to the tags, and 'erasers' that remove tags.

Regulatory RNAs

Our genomes encode a remarkable number of non-coding RNAs – that is, RNA molecules that are made by transcribing specific DNA sequences, but that are not messenger RNAs. Ribosomal RNA and transfer RNA are the best-known examples, but in recent years we have seen an explosive growth in the number of other species identified. In fact, we have more genes for non-coding RNAs than for proteins. We don't know what the function of all those RNAs is, but it is generally supposed that their primary role is, one way or another, to regulate the expression of protein-coding genes. Some have been shown to be involved in controlling chromatin structure, and hence gene expression.

You can see that controlling when and where a gene is expressed is immensely complicated and subtle. But this should not come as a surprise, given that we construct all the 200 or so different cell types of our bodies, and organise them into flexible tissues and responsive organs, using hardly more protein-coding genes than the nematode worm *Caenorhabditis elegans* uses to organise its 1000 cells into its 1 mm long body (around 22 000 in man, 19 000 in the worm).

Epigenetic memory

Epigenetics (literally 'above genetics') is about the mechanisms that allow cells to retain a memory of their particular patterns of gene expression, and to pass that memory on to daughter cells. In some cases the memory can be transmitted across generations, from parent to child, although it is quite controversial how general such transgenerational effects are in humans (they are better characterised in plants, in vernalisation for example). The epigenetic modifications themselves are the same DNA methylation and histone modifications that we have seen regulate transcription within a cell; the question is how epigenetic memory works.

The key to epigenetic memory lies in the DNA methyltransferases. Remember that these can methylate cytosines in CpG sequences – that is, cytosines immediately upstream of a guanine. In the DNA double helix, CpG will base-pair with GpC. But because the two strands are anti-parallel, reading in the standard 5' – 3' direction, opposite every CpG in one strand is a CpG in the other (Figure 13).

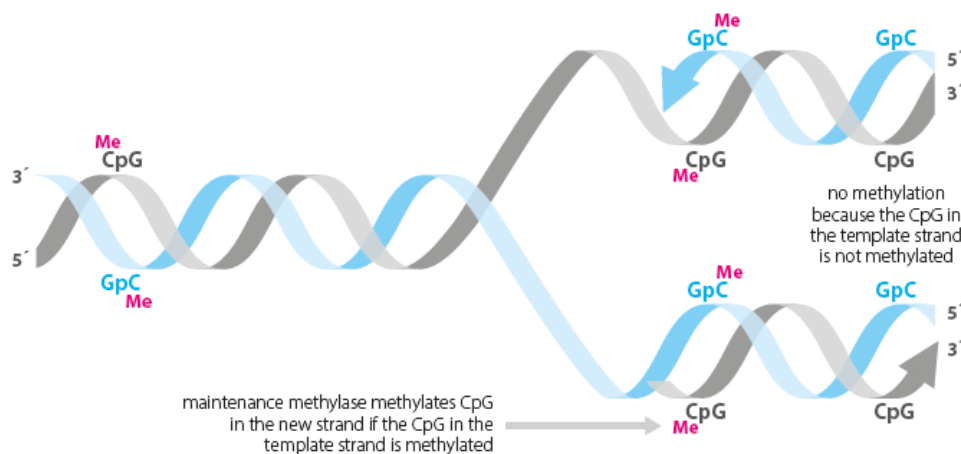


Figure 13: From *New Clinical Genetics*, Read & Donnai. Scion Publishing 2015.

We have three DNA methyltransferase enzymes. Two of them are responsible for *de novo* DNA methylation, adding methyl groups to CpG sequences that were previously unmethylated. The third, DNMT1, is the maintenance methylase. When a DNA molecule is replicated, the newly synthesised strands are initially completely unmethylated. However, DNMT1 then specifically methylates any CpG on a daughter strand that lies opposite a methylated CpG on the template strand. Thus the specific pattern of methylation is inherited from mother cell to daughter cells.

Other mechanisms besides DNA maintenance methylation may contribute to epigenetic memory, since *Drosophila* flies do not methylate their DNA, yet can clearly regulate gene expression and maintain cell differentiation. This whole area is one of active research. Perhaps the basic question is which is the primary factor – DNA methylation, histone modification or something else? It appears that the various mechanisms reinforce one another by positive feedback. Methyl DNA-binding proteins recruit histone modifying enzymes, but modified histones recruit DNA methyltransferases. It seems possible that transcription factors play the key role in all of this, and that binding transcription factors may be the primary cause, setting all the other processes in train.

Stem cells

The cells of a very early embryo are *totipotent* – that is, they can differentiate to form all cell types of the foetus and adult, including the placenta. Later, at the blastocyst stage, when the embryo consists of a hollow ball of a few hundred cells, the dozen or so cells of the inner cell mass are *pluripotent* – they can develop into all cell types of the adult body, but not into the cells of the placenta and membranes. As development proceeds, cells become more specialised. Terminally differentiated cells do not normally divide; tissues are maintained by small populations of multipotent or unipotent stem cells. Stem cells can divide symmetrically, to produce two daughter stem cells, or asymmetrically, to produce one stem cell and one cell (a transit amplifying cell) that can divide rapidly and produce the terminally differentiated cells of a tissue.

All this progression is the result of successive epigenetic modification of the genome. Many years ago, long before any of this was understood, C H Waddington put forward the idea of an 'epigenetic landscape'. He conceived a model of a ball rolling down a tilted three-dimensional surface with hills and bifurcating valleys. As the ball rolls down, its options are limited to the valleys that open up from the particular valley it is currently occupying, and the further down the surface it rolls, the fewer its options are. As a model of the progressive epigenetic restriction of differentiation potency as embryonic development proceeds, it is very good.

In 2015 we can put flesh on Waddington's concept. Each valley is defined by the battery of genes a cell expresses, and this depends on the transcription factors present (Figure 14). Among those genes are some for further transcription factors, which in turn define the secondary valleys. Choices between valleys can depend on signals from the surrounding cells or medium, or they can be generated within a cell by asymmetric cell division, or simple chance. Transcription factors active in higher valleys may be actively turned off as differentiation proceeds, or they may be simply diluted out as the cells multiply. Replacing them may reverse differentiation (see below).

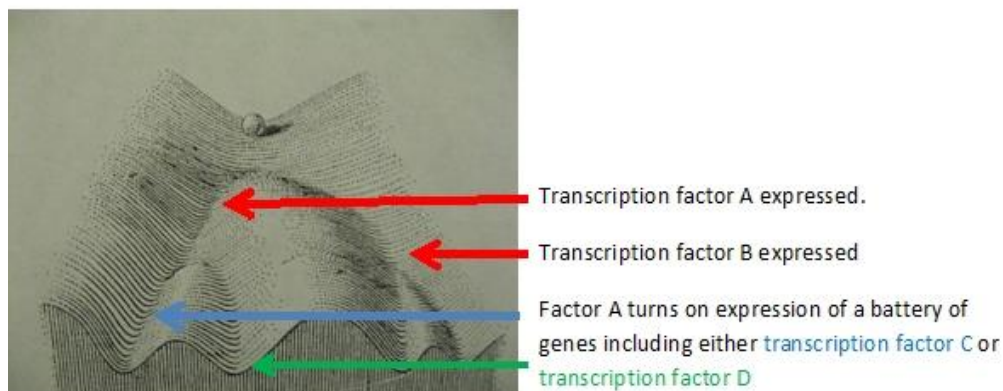


Figure 14

Possible teaching approach

All blood cell types (erythrocytes, lymphocytes, granulocytes, platelets and dendritic cells) are produced by descendants of a small population of multipotent haematopoietic stem cells in the bone marrow. This is a nice illustration of these principles (Figure 15).

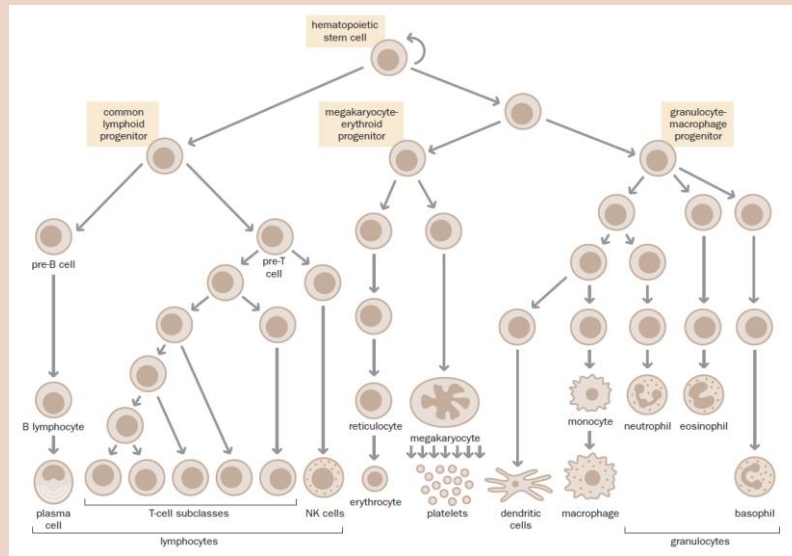


Figure 15

Pluripotent stem cells are of great medical interest because, in principle, pluripotent cells from a patient could be grown and differentiated into any body cell type, and then used to replace damaged cells or tissues of the patient without any of the problems of rejection that complicate normal transplants.

The first human pluripotent stem cells were embryonic stem (ES) cells, obtained in the late 1990s by delicate and difficult manipulation of cells from the inner cell mass of blastocysts. These proved quite controversial, because in order to obtain them a human embryo had to be destroyed. The embryos used were spare ones from *in vitro* fertilisation clinics – the procedure normally produces more embryos than would be re-implanted, and the couple concerned might agree to donate the surplus for research.

Ideally, to avoid rejection, a patient should receive ES cells derived from his own cells. This gave rise to the idea of *therapeutic cloning*, where a donated fertilised egg was enucleated and the nucleus replaced by one from a somatic cell of the patient (the procedure that created Dolly the sheep). The egg would then be grown to the blastocyst stage and patient-specific ES cells obtained.

Because of the many practical and ethical difficulties, all this remained rather theoretical, until the discovery that differentiation could be reversed. If normal, differentiated, somatic cells are treated with a special cocktail of transcription factors, some of them revert to pluripotency. With appropriate culture conditions, the pluripotent cells can be multiplied in culture and then induced to differentiate into any desired cell type.

Development of these iPS (induced pluripotent stem) cells has opened the door to a new world of clinical possibilities. Patient-specific cells of any type might now be produced in

the laboratory – neurons for a patient with Parkinson disease, blood cells for a patient with bone marrow failure, and so on, without any of the problems surrounding ES cells. Producing iPS cells is a highly skilled and uncertain business, and questions remain about the safety of introducing the derived cells into a patient – might some of them develop into tumours? Thus many questions remain, but the future looks exceedingly promising.