Information sheet 1: what is a gene?

Our bodies are made from thousands of different building blocks called proteins. They may have a structural role – such as a type of protein called 'collagen' – or an important role in controlling metabolism – such as proteins called 'enzymes'. Each individual protein building block ('molecule') is manufactured on-site within a cell using a set of instructions called a *gene* (which we inherit from our parents) and a set of ingredients called amino acids (which we absorb from our food). There are over 20,000 genes in each cell of our body; scientists have recently predicted that it is possible to create somewhere between 0.6 – 6 million different proteins using these 20,000 genes.[1]

By contrast, our genome looks fairly modest when compared to the Marbled Lungfish (*Protopterus aethiopicus*), which has the biggest animal genome ever found at 133 billion base pairs. This is beaten by the Japanese plant Paris Japonica, at 149 billion bases, which at present appears to be the largest genome in nature. The prize for nature's smallest genome currently goes to a parasitic bug that lives inside insects (*Carsonella ruddii*) at just 159,662 bases.

All genes in nature are written using a simple code made from a chemical called DNA. There are only 4 'letters' in the DNA code, known as 'bases' or 'nucleotides': A, T, C or G. These are abbreviations for four chemicals adenine, thymine, cytosine and guanine. The bases are joined together in enormously long chains called chromosomes. Along each chromosome (there are normally 46 in each cell), the genes are separated by long stretches of 'packing' DNA whose function is only now becoming clear. Interestingly, only 1.5% of the DNA in our cells makes up genes: the rest is packing. Overall, we inherit around 3 billion bases of DNA from each parent, so only 45 million bases actually make up genes.

A typical gene (Figure 1) has a switch at the beginning that regulates how much protein is being manufactured. The 'coding sequence' – which is the part of the gene that tells the cell how to make a protein – is usually arranged in different sections called 'exons', separated by sections of more packing DNA called 'introns'. Within the introns scientists have discovered all sorts of detail that suggests that introns often have a function, rather than just being bland packaging material. For example,

---

[1] Ponomarenko and colleagues. The Size of the Human Proteome: The Width and Depth. International Journal of Analytical Chemistry 2016, Article ID 7436849, http://dx.doi.org/10.1155/2016/7436849 accessed 6th July 2018.

they sometimes contain 'fine tuning' instructions, or even instructions that control other genes nearby.
Some genes do not encode proteins but have different functions within the cell, usually the fine-tuning of protein synthesis. The smallest human gene is just 76 bases long and encodes a molecule called 'transfer RNA'. This is not a protein, but has a critical function in the synthesis of proteins.

By contrast, the largest human gene is 2,220,390 bases long. It encodes a protein called dystrophin, which consists of 3,684 amino acids. This is not, however, the largest protein. That accolade goes to the *TTN gene*, which is 'only' 80,781 bases long but encodes at least 13 different forms of a giant protein called titin. The largest form of titin is a huge 35,991 amino acids long (depending on the specific form).

DNA is read in blocks of three bases called 'codons'. Each codon corresponds to either a 'stop' or 'start' signal or tells the cell to add a particular amino acid to the protein being manufactured.

Table 1: examples of codons and their meanings

Codon	**Amino acid**
TTC	phenylalanine
AAT	asparagine
GCC	alanine
CGA	arginine
TGA	STOP

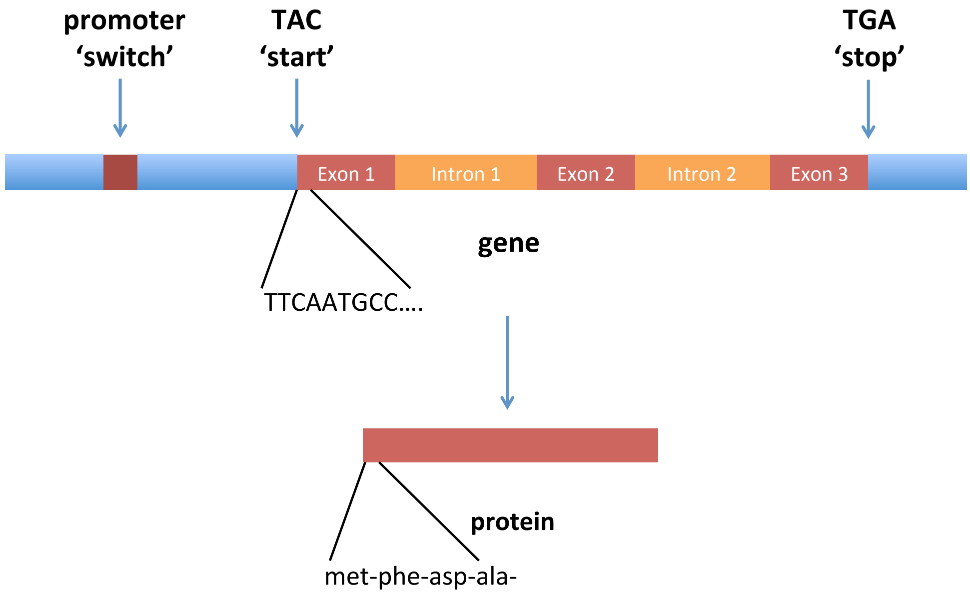
Importantly, too, there are usually a number of different ways to encode an amino acid. For example, GCT, GCC, GCA or GCG all encode the amino acid alanine[2]. This means that DNA code can vary considerably between different people without it affecting the meaning. This is important when we consider DNA testing: first, we must understand 'normal'.

Only exon code is required to create a template for protein synthesis. When a cell needs to make a protein, a copy of the gene is made using a similar chemical to DNA called RNA. The intron code is removed ('spliced out') from the RNA copy. This is a bit like taking a photocopy of a recipe and cutting out any unnecessary information with scissors.

---

[2] For more detail, see https://en.wikipedia.org/wiki/DNA_codon_table

Figure 1: typical structure of a gene



NB: codons and amino acids are those cited in Table 1.