Describe the organisation of the human genome. How has knowledge of the genome sequence assisted our understanding of this organisation?

Since Avery, MacLeod and McCarty were able to prove that DNA was the genetic blueprint to life in 1944, research into the human genome has been one of the fastest evolving areas of science as scientists sequence and compare different sections of DNA of animals. All of this is with the purpose of finding new evidence for links between the genome and diseases, evolutionary ancestors, purposes of specific proteins and just out of scientific curiosity. The largest of these projects was the Human Genome Project which, completed in April of 2003, had successfully sequenced 99% of the euchromatic human genome with 99.99% accuracy [1]. This was a huge step in understanding the genome and its organisation for the purpose of gene expression and control. In this essay I will discuss the different sections of the human genome which have been sequenced, with the intent to explain our current understanding of the genome's sequence and therefore its organisation.

BASIC STRUCTURE

The human genome is made up of DNA (deoxyribose nucleic acid) which is a double stranded molecule made up of repeating units of deoxyribose (a sugar), a phosphate group, and one of four nitrogenous bases which can be adenine (A), thymine (T), guanine (G) or cytosine (C) making what is known as a nucleotide. This polymer forms a string of over 3 billion letters which are the basis for everything that our bodies are capable of doing due to different arrangements of these same nucleotides [2]. This mass of a polymer is contained within the nucleus of every cell within the body, which can be categorised into 22 autosomes (chromosomes organised numerically by their length) and a pair of sex-determining chromosomes.

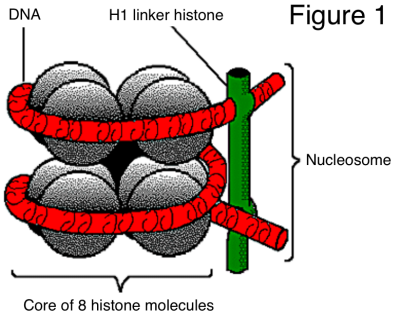

Figure 1

Due to the sheer amount of genetic information each of our cells contain, DNA must be compressed in order to fit it all into the body. To do this, the DNA is wrapped around proteins known as histones, reducing the size of the DNA strands so that the primary physiological state of DNA as chromatin is thousands of times shorter than an unpacked molecule would be. In humans, there are 4 different core types of histone protein (H2A, H2B, H3 and H4) [3] which each make up 2 of the 8 histone proteins in each nucleosome. In addition to those, 2 types of linker histones (H1 and H5) are found on each end of a nucleosome to keep the DNA in place. Interaction between these nucleosomes leads to further condensing from chromatin into chromosomes during cell division. These proteins are highly basic in order to facilitate association with DNA which, as its name suggests, is acidic [4]. This association is one of the ways in which gene expression is controlled. The more closely associated a gene is with its histones, the more repressed it is. Also, the histones are capable of undergoing post translational modifications in order to alter the interactions between the histones and either nuclear proteins or the DNA itself. These modifications include methylation, acetylation, phosphorylation, and ubiquitination amongst others which each have separate effects on the expression of, and activity on the DNA sequence [5]. This organisation of the genome around the histones is therefore vital in the compression and expression of DNA within each of our cells.

PROTEIN CODING REGIONS

With the completion of the Human Genome Project in 2003 came the realisation that the genome contains only about 20,000 protein-coding genes [1], similar to the number and function of the genes in much simpler organisms such as fruit flies and roundworms. This comes to only 2% of the entire genome. This was surprising as it had long been assumed that our complexity as organisms was as a direct result of our genes. What we understand now is that humans are able to generate a greater number of protein variants from one gene than simpler organisms. It has been found that over 60% of human genes are capable of these variants due to splicing [6], a process in which sections of pre-mRNA are removed from the molecule so that only the sections which code for a desired protein are left as mRNA. Due to our ability to manipulate the genome in this way, the sequence is highly efficient as the same section can code for different proteins. Splicing is also used to remove non-protein-coding sections of DNA, known as introns. All of our genes are made up of these introns interspersed by exons, regions that do code for proteins. The introns themselves are also very useful sections of DNA as they perform a range of functions such as coding for non-protein-coding functional RNA and regulating gene expression. Our current knowledge of the way in which genes are composed of both introns and exons has helped us understand the function of these sections, and therefore the importance of how gene expression is regulated by both introns and exons.

'JUNK' DNA

As previously mentioned, the introns and exons only make up about 2% of the entire genome. The other 98%, colloquially referred to as 'junk DNA', is constantly being discovered (using high-throughput sequencing techniques for forensics, medicine and anthropology) [7] to have a variety of different functions which are just as vital in our cells as the protein coding regions. As there is so much, its roles are diverse from structural (acting as anchor points between chromosomes during cell division, or as protective structures which prevent the DNA from unravelling and getting damaged), to many other coding roles such as non-protein-coding RNA or sections which code for viral RNA.

These, known as endogenous retroviruses, are estimated to make up between 1% and 8% of the genome and can sometimes still make their proteins or reproduce by making the cell replicate its DNA section. They have been inherited through successive generations and vary in their effects. Most of these are inactive relics as they are suppressed by the body which has developed defence mechanisms such as coating the DNA with molecules which will suppress these viral genes [8]. A few of the endogenous retroviruses have been repurposed by the human genome to be beneficial to human health and development. An example of this is the gene for a viral envelope which is now used in pregnancy for the development of the multinucleate layer which keeps the mother and foetus' blood separate [9]. Other examples of endogenous retroviruses are however harmful to the body and have been linked to cancers and autoimmune deficiency due to certain cases of them activating proto-oncogenes and some having been shown to encode immunosuppressive proteins [10]. As a result of our comprehension of the different roles and origins of these sections of code, we can appreciate their random arrangement within the genome.

Other parts of this junk DNA are specially located to perform their function due to their structural importance. For example, when condensed into chromosomes, repeat sequences

of the bases TTAGGG can be found on either end of these large structures [12]. This comes as a result of the telomerase mechanism in each of our cells which is made up of a template RNA molecule for the base sequence known as hTR [14] and hTERT, the telomerase itself [13] which attaches the new base sequence to the end of the chromosomes to lengthen them. This forms a region known as a telomere. During DNA replication, the mechanism does not allow for the chromosome to be entirely replicated, resulting in a shortening of the length of the chromosome after each replication. If there were no telomeres, the chromosome itself would deteriorate, and important genetic information could be lost. The telomeres are therefore disposable sections which protect the important genetic information. This sequence of bases being repeatedly added to the ends to lengthen the telomeres increases the length of cell life as they can continue dividing for many cell generations without triggering apoptosis due to genetic deterioration. This is particularly useful in stem cells which must continue for many generations to produce replacement cells throughout the body and so are especially enriched for telomerase activity. Thus, knowledge of the sequence of DNA which codes for both components of the telomerase mechanism has assisted our understanding of how chromosomes are arranged in their condensed state, and why this organisation is so important for the protection of genetic information so that it is safely passed onto daughter cells after cell division.

Also located on each chromosome is a centromere - an area which links sister chromatids during DNA replication and the area of spindle fibre attachment during cell division. The centromere is identifiable by the 171 repeating base pairs forming an alpha satellite DNA unit which are associated with histones much alike the rest of DNA, however with the major H3 histone having been replaced by a CENPA histone [12] which facilitates the attachment of various proteins forming a disc-shaped structure known as a kinetochore which the microtubules of spindle fibres will attach to during metaphase and anaphase [15]. These different histones also make the region more rigid which makes the chromosomes easier to pull apart. The base pair sequence itself is important as it dictates the specific locus on the chromosome where the separation of sister chromatids occurs. Despite the role of the centromere being the same in all chromosomes, its location can vary from metacentric (in the centre) to telocentric (on the very end) [16]. In this way, our knowledge of the sequence of base pairs which make up the centromere and the histones they associate with, has helped us to understand which proteins are capable of attaching, and therefore the organisation and role of this section of the genome.

Mitochondrial DNA

Mitochondrial DNA, referred to as mtDNA, is the only part of the human genome which is not located inside the cell's nucleus. As its name suggests it is found in the mitochondrial matrix, a structure in the cytoplasm of the cell which is responsible for producing adenosine triphosphate (ATP), the energy currency of cells. In humans as well as most complex organisms, the mtDNA is a circular structure which is covalently closed, made up of 16,569 base pairs coding for only 37 genes [17]. The two strands of DNA are separately classified as heavy and light depending on their ratios of purines (A/G) and

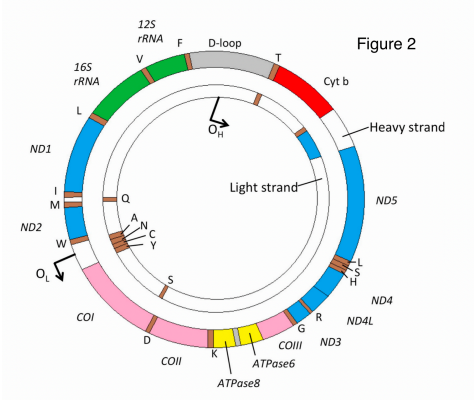

Figure 2

pyrimidines (T/C) (purines are heavier than pyrimidines as they have an extra ring). This circular shape is widely accepted to be due to the endosymbiotic hypothesis of origin which suggests that mitochondria would have originally been prokaryotic cells (thus their genome is circular) which began living inside a eukaryotic cell as this supplied the eukaryote with an energy supply. This shows how the sequencing of the mitochondrial DNA in humans has helped our understanding of why the mitochondrion's genome itself is circular, and how mitochondria came to inhabit our cells in the first place, making it the only part of the human genome not confined to the nucleus.

MtDNA, unlike nuclear DNA, is inherited solely from the mother. It has been found that this is partly due to the fact that most of the mitochondria in sperm is found in the tail as this requires the most energy, and the tail is generally lost after fertilisation; and also because there has been evidence (in mice) of specialised cellular vesicles using selective autophagy to destroy paternal mtDNA after fertilisation [18]. In very rare cases where some is inherited from the father, the embryo is generally rejected before it can develop. Due to the fact that the mtDNA is inherited uniparentally, there would be an expectation of an accumulation of mutations in the genome which could be eventually fatal. In order to not be affected by this process, human mitochondria undergo what is known as mtDNA bottleneck where the mutations in the mitochondria of the mother's oocyte (immature egg cell) results in the formation of gametes carrying varying mutant loads in their mitochondria so that the mutations are shared [19]. This explains why the number of mutations in a person and their mother's mitochondria will vary and so why they will be found in different proportions across one woman's offspring.

Ultimately, our current knowledge of the genome sequence has assisted our understanding of its organisation by revealing the function of certain sections of DNA and therefore why they are situated where they are found. This knowledge on both the structure and function of our genome will help us as we understand more about the links between various sections of DNA and medicine or evolutionary biology, as well as perhaps leading us to a further understanding of the world around us through comparisons made between the human genome and that of microorganisms, other animals or plants as many of our genes are similar in function. It is for this reason that our ability to sequence the human genome has been so vital in understanding its role and, of course, its organisation.

Word count: 2168

Bibliography:

1. En.wikipedia.org. (2019). *Human Genome Project*. [online] Available at: https://en.wikipedia.org/wiki/Human_Genome_Project#State_of_completion [Accessed 18 Feb. 2019].
2. ASM. (2015). *An Overview of Genome Organization and How We Got There: from FISH to Hi-C*. [online] Available at: https://mmbr.asm.org/content/79/3/347 [Accessed 27 Feb. 2019].
3. En.wikipedia.org. (n.d.). *Histone*. [online] Available at: https://en.wikipedia.org/wiki/Histone#Classes_and_histone_variants [Accessed 22 Feb. 2019].
4. NCBI. (2014). *Post-Translational Modifications of Histones That Influence Nucleosome Dynamics*. [online] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4375056/ [Accessed 23 Feb. 2019].

5. En.wikipedia.org. (n.d.). *Function of Histone Variants*. [online] Available at: https://en.wikipedia.org/wiki/Histone#Functions_of_histone_modifications [Accessed 2 Mar. 2019].
6. NCBI. (2015). *Introns: The Functional Benefits of Introns in Genomes*. [online] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4742320/ [Accessed 30 Feb. 2019].
7. En.wikipedia.org. (2019). *DNA sequencing*. [online] Available at: https://en.wikipedia.org/wiki/DNA_sequencing#Next-generation_methods [Accessed 30 Feb. 2019].
8. Nytimes.com. (2017). *Ancient Viruses Are Buried in Your DNA*. [online] Available at: https://www.nytimes.com/2017/10/04/science/ancient-viruses-dna-genome.html [Accessed 1 Mar. 2019].
9. NCBI. (2018). *The placenta goes viral: Retroviruses control gene expression in pregnancy*. [online] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6177113/ [Accessed 3 Mar. 2019].
10. NCBI. (2003). *Demystified . . . Human endogenous retroviruses*. [online] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1187282/ [Accessed 5 Mar. 2019].
11. Reference, G. (2014). *TERT gene*. [online] Genetics Home Reference. Available at: https://ghr.nlm.nih.gov/gene/TERT#resources [Accessed 1 Mar. 2019].
12. Carey, N. (2015). *Junk DNA*.
13. Reference, G. (2014). *TERT gene*. [online] Genetics Home Reference. Available at: https://ghr.nlm.nih.gov/gene/TERT#resources [Accessed 1 Mar. 2019].
14. Reference, G. (2014). *TERC gene*. [online] Genetics Home Reference. Available at: https://ghr.nlm.nih.gov/gene/TERC#location [Accessed 1 Mar. 2019].
15. En.wikipedia.org. (n.d.). *Kinetochore*. [online] Available at: https://en.wikipedia.org/wiki/Kinetochore [Accessed 29 Feb. 2019].
16. Nature.com. (2008). *Chromosome Segregation: The Role of Centromeres*. [online] Available at: https://www.nature.com/scitable/topicpage/chromosome-segregation-in-mitosis-the-role-of-242 [Accessed 2 Mar. 2019].
17. ScienceDirect. (1999). *The mitochondrial genome: structure, transcription, translation and replication*. [online] Available at: https://www.sciencedirect.com/science/article/pii/S0005272898001613 [Accessed 3 Mar. 2019].
18. ScienceDaily. (2014). *Inheriting Mitochondria: Where does your father's go?* [online] Available at: https://www.sciencedaily.com/releases/2014/05/140515095637.htm [Accessed 3 Mar. 2019].
19. eurekalert. (2008). *Mitochondrial 'bottleneck' cracked*. [online] Available at: https://www.eurekalert.org/pub_releases/2008-01/wt-mc012508.php [Accessed 4 Mar. 2019].

Figure 1: http://www.accessexcellence.org/RC/VL/GG/nucleosome.html
Figure 2: https://www.sciencedirect.com/science/article/pii/S0005272815000961