Genes and Genetic Codes

3.1 Genes

Genes are the basis of inheritance and are transferred from parents to offspring, and traits are inherited. Genes are arranged within chromosomes which has the single, long DNA molecule, a portion of which corresponds to a single gene. A gene could be as short as a few hundred base pairs or as long as many thousands. The term gene was introduced in 1909 by Danish geneticist, Wilhelm Johannsen. It was derived from the ancient Greek word, γόνος, *gonos*, which means offspring and procreation.

In a total genome, the genes can be classified as functional and non-functional genes. In humans, functional genes contribute to only 3% of the total genome and the remaining portion of 97% consists of "junk DNA." Some of this junk DNA consist of pseudogenes, which becomes non-functional genes. The non-coding DNA consists of dispersed or clustered repeated sequences of varying length, from one base pair (bp) to thousands of bases (kilobases, kb) in length.

- Dispersed repeated sequences—They occur as copies which spread across the genome and are categorized as long or short interspersed nuclear elements, long terminal repeats, and DNA transposons.
- Clustered repeated sequences—The repeated sequences occur in tandem copies and depending on the length of the repeat unit, they are called as satellites, minisatellites, and microsatellites. Repeated elements can constitute up to 40% of the genome.

3.1.1 Activities of Genes

- A gene can be replicated and thus genetic information can be passed from generation to generation unchanged.

- The sequences of bases in the RNA depends directly on the sequences of bases in the gene. Most of these RNAs serve as templates for making protein molecules. Thus, most genes are essentially blueprints for making proteins. The production of protein from a DNA blueprint is called gene expression.
- A gene can accept occasional changes or mutations.

3.1.2   Structure of Gene

A gene is a small portion in a DNA which is held as a code. This code specifies the amino acid sequence of a protein. The sequence information for a single gene will not be continuous along the DNA, but will be interspersed with pieces of non-coding sequence. The coding parts of a gene sequence are exons, and the non-coding parts are introns. Genes are DNA strands made up of the nucleotide chain. The chemical structure of a gene comprises of nucleotides. The nucleotide of one strand binds with the nucleotide of the opposite strand by hydrogen bonding, whereas adjacent nucleotides bind with each other by phosphodiester bonds. Gene structure consists of core elements and regulatory elements. The core elements like exons are involved in protein formation, and the regulatory elements such as promoters, enhancers, and silencers maintain gene expression. Another type of element, the maintenance element contains information for DNA repair, DNA modification, and DNA replication. The functional structure of a gene comprises of introns, exons, promoters, enhancer, and untranslated region. In a given gene, after splicing, the coding parts are joined and are called exons, whereas the non-coding sequences that are removed from the final transcript are called introns. The elements located at the end of a gene are called regulatory elements (Fig. 3.1).

The non-coding sequences that facilitate binding sites for enzymes and transcriptional factors are called promoters. The promoter is located near 5′ end, and it consists of TATA box and CCAAT sequences for enzyme binding. The promoter is made up of

- Core promoter—helps in RNA polymerase bindings to initiate transcription
- Proximal promoter sequence—provides bindings for transcriptional factors

The induction and repression of transcription are by enhancer and silencer, respectively, and are located away from exon to regulate gene expression. The 3′ untranslated regions are non-coding regions of gene helping in aborting the process of transcription and forming the final transcript. Once the RNA polymerase reaches the untranslated region, it stops synthesizing RNA and detached from the strand.

3.1.3   Types of Genes Based on Their Function and Position

Housekeeping genes: They are required to perform normal cell functions and generally code for transcription, translation, and replication.
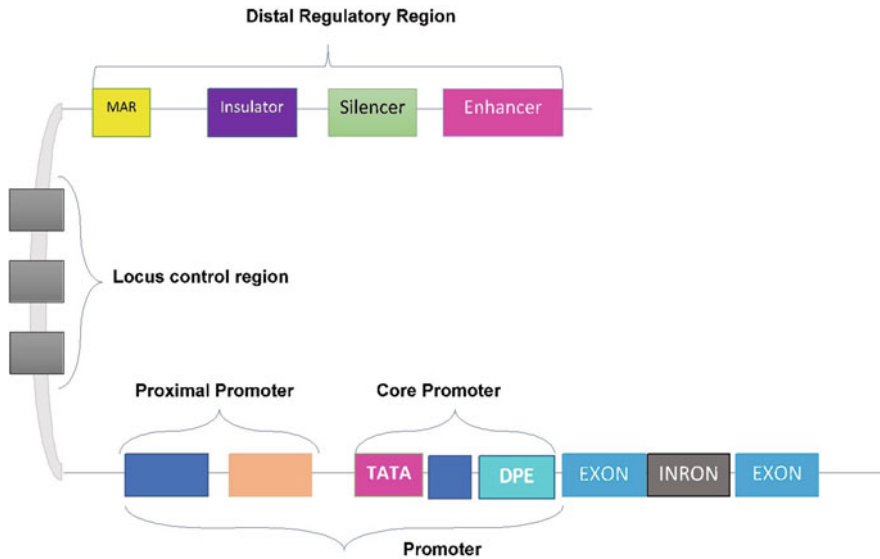
Fig. 3.1 The molecular structure of a gene

Inducible genes: They remain inactive and is expressed under the influence of extrinsic factors.

Developmental genes: They are involved in the early development of organisms.

Tissue-specific genes: They are active only in specific tissues and remain inactive in other tissues.

Homologous genes: They are inherited from a common ancestor, share a common function, and have sequence similarities and are categorized into homologous genes.

Non-homologous genes: They aren't inherited from a common ancestor; instead, it is originated due to some evolutionary forces.

Autosomal genes: They are located on autosomal chromosomes.

Sex-linked genes: They are located on sex chromosomes.

3.2 Genetic Code

The set of instructions that translates DNA into 20 amino acids are called genetic code. Codons, the three letter nucleotides are the basis of genetic code. Every codon codes only for one specific amino acid. The genetic code determines the amino acid sequence in a protein.

3.2.1   Discovery of the Genetic Code

Codon was discovered by Francis Crick and his colleagues during 1961. Later, the genetic code was deciphered by Marshall Nirenberg and his colleagues who showed that during protein synthesis, A, T, G, C form codons of different base combinations which code for all 20 amino acids. Following this, Nirenberg along with Johann Matthaei, a German scientist conducted various experiments to know about protein synthesis using synthetic RNA.

They used a "cell-free system" to add RNA strands with any of the four bases (A or G or U or C) and radioactively tagged amino acids. Using radioactive measurements, they found that when RNA containing the base U was added, molecules with one single amino acid, phenylalanine, were synthesized. Hence the triplet made up of UUU resulted in phenylalanine. Following this method, by 1963, scientists deciphered 35 codons and by 1966, they had deciphered more than 60 codons.

3.2.2   Codons and Amino Acids

In the genetic code, each codon consists of three bases which are arranged in a specific order. Each codon combination corresponds to one particular amino acid. It is known that from the four nucleotides, 64 possible permutations and combinations of three letter nucleotide sequences are possible. In total, there are 64 codons, 61 codons code for different amino acids, and three codons are considered stop codons.

3.2.2.1 START and STOP Codons

START codon begins the translation. The common start codon is AUG. In eukaryotes, AUG codes for methionine while in prokaryotes, it codes for formyl methionine. There are three STOP codons in total. They are also called as nonsense or termination codons. Codons such as UAG, UGA, and UAA are considered STOP codons and are named amber, opal, and ochre, respectively. During protein synthesis, they signal the end of the polypeptide chain. They also trigger the ribosome to release the new polypeptide chain since no tRNA anticodons complement these stop codons (Fig. 3.2).

3.2.3   Properties of the Genetic Code

1. The code is degenerate or redundant. Most of the amino acids are coded for more than one codon. Leucine, serine, and arginine have six different codons. Proline, threonine, and alanine have four. Isoleucine has three. Methionine and tryptophan have only one codon.
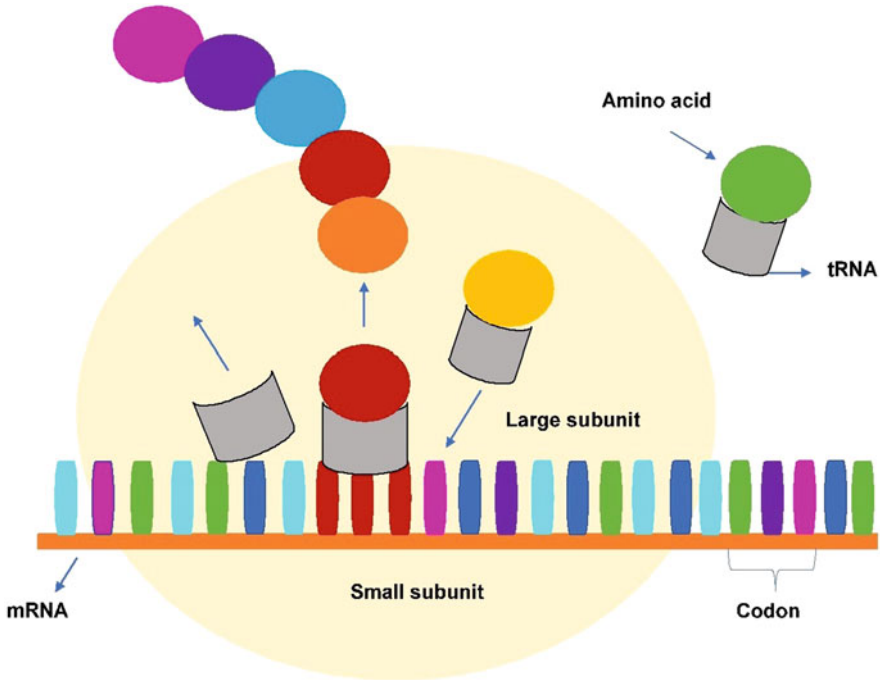2. The code is unambiguous. One codon codes for one amino acid only.

Fig. 3.2 Process of amino acid formation in ribosome from codon

3. The code is highly universal. It is the same for various different kind of organisms. However, a few exceptions to this are known as in yeast mitochondria, UGA codes for tryptophan instead of stop. In Paramecium, UAA and UAG code for glutamine instead of stop codon.

3.2.4 The Wobble Hypothesis

The property of the genetic code is that the code is unambiguous. However, sometimes, more than one codon code for one amino acid and is known as degeneracy of genetic code. In order to explain this degeneracy of codons, Francis Crick proposed "Wobble hypothesis" in 1966. The term wobble means to sway or move unsteadily. In this hypothesis, he suggested that among the three bases, the first two bases of the codon have an exact pairing with the bases of the tRNA anticodon. There may be wobbling when pairing occurs between the third base of the codon and anticodon (Fig. 3.3). This leads to the recognition of more than one codon by a single tRNA. Hence even if there are 61 codons for amino acids, there will be around 40 tRNA only due to wobbling.
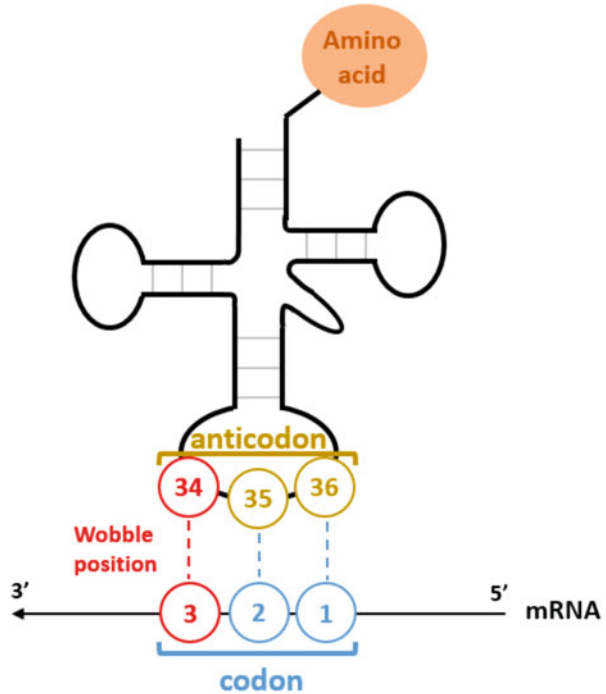
According to wobble hypothesis, the base present at 5′ end of the anticodon is not spatially confined as the other two bases and hence it forms hydrogen bonds with any of the bases located at the 3′ end of a codon. This leads to the following conclusions:

- In the codon, the first two bases have normal (canonical) hydrogen bond pairs with the second and third bases of the anticodon.
- Non-canonical pairing may occur in the remaining position. Therefore, at the third position of the codon, the wobble hypothesis proposes a flexible set of base pairing rules.
- The wobble allows the anticodon of one tRNA to pair with more than one triplet in mRNA.
- According to the wobble base pair rules, first base U can recognize A or G, first base G can recognize U or C, and first base I can recognize U, C, or A.

3.2.4.1 Wobble Base Pairs

The pairing that doesn't follow Watson—Crick base pair rules are known as wobble base pair (Fig. 3.4). The four main wobble base pairs are guanine-uracil (G-U), hypoxanthine-uracil (I-U), hypoxanthine-adenine (I-A), and hypoxanthine-cytosine (I-C). Hypoxanthine is the nucleobase of inosine; hence "I" is used for hypoxanthine.
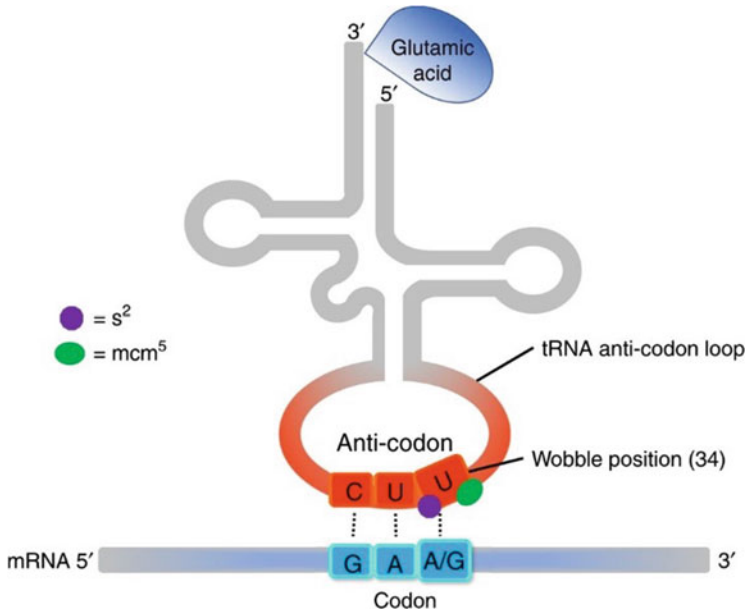
Fig. 3.4 Wobble position (reproduced from Goffena et al. 2018)

3.2.4.2 Significance of the Wobble Hypothesis

- Wobble provides broad specificity with limited amount of tRNAs.
- In RNA secondary structure, wobble base pairs are essential and are necessary for proper translation of the genetic code.
- Wobbling allows faster dissociation of tRNA from mRNA and also protein synthesis.
- The existence of wobble minimizes the damage that can be caused by a misreading of the code; for example, if the Leu codon CUU was misread CUC or CUA or CUG during transcription of mRNA, the codon would still be translated as Leu during protein synthesis.

References

Fages-Lartaud M, Hohmann-Marriott MF (2022) Overview of tRNA modifications in chloroplasts. Microorganisms 10(2):226. https://doi.org/10.3390/microorganisms10020226

Goffena J, Lefcort F, Zhang Y et al (2018) Elongator and codon bias regulate protein levels in mammalian peripheral neurons. Nat Commun 9:889. https://doi.org/10.1038/s41467-018-03221-z