Harry van der Hulst

Gene Basics

Introduction

Each cell is an extremely complex structure, a rich 'world' that contains numerous molecules that each have some kind of function. Here we focus on a molecular structure that consists (among other things) of DNA, which stands for *Deoxyribose Nucleic Acid*, which is contained in the cell nucleus.[1] There are four 'bases': ***A**denine,* ***G**uanine* (together called *purines*) and ***C**ytosine* and ***T**hymine* (*pyrimidines*).[2] These molecules form 23 long strands that we call *chromosomes*, such as (1) (but then much, much longer, into the millions; see below):

(1) …AGACTATG…

Chromosomes come in pairs, so the total number is 46 (see below). The number of chromosomes differs for different species. Each chromosome contains *two* DNA strands that coil around each other forming a 'twisted ladder' which is called *the double helix*. The 'steps' of this ladder are formed by fixed pairs (called *base pairs*) of DNA letters like:

(2) …AGACTATG…
| | | | | | | |
TCTGATAC

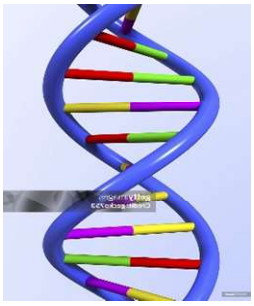


Figure 1: The double helix[3]

[1] A small amount of DNA is contained in the mitochondria which occur in the fluid around the nucleus. Note also that while white blood cells have a nucleus with DNA, mature red blood cells and platelets don't. (Otherwise, it would not be possible to use blood samples to extract DNA.)

[2] Adenine, Guanine, Cytosine and Thymine are not themselves nucleic acids, but are four different "bases" that together with other molecules (sugar and phosphate groups) make up a *DNA polymer acid*. I will refrain from further technical details.

[3] Image credit: https://www.gettyimages.nl/detail/foto/dna-royalty-free-beeld/182674440?searchscope=image%2Cfilm&adppopup=true

Each DNA 'letter' A pairs with a T and each C pairs with a G: [A-T] and [G-C].[4] This means that the two sides of the ladder are what we call *antiparallel*. The fact that a chromosome contains two antiparallel strings of DNA letters, provides, as we will see below, the basis for copying the genome when *cell division* takes place and it also explains *heredity*, i.e., how genes can pass on from one generation to the next.[5] The molecular structures that the DNA 'letters' stand for are surrounded by a host of other molecules that keep the whole ladder tightly 'packed', so that it does not break into pieces. As mentioned, each cell contains *two* sets of chromosomes, which total 46 chromosomes.[6] One set comes from your dad, the other from your mom. A pair of chromosomes is called a *homologous pair*. Homologous means that the two sets of 23 contain the same genes (although possibly in slightly different variants, called *alleles*; see below), except for the so-called sex chromosomes: females have two X chromosomes, while males have an XY pairing.

The first cell that eventually became you, was formed by combining the mother's egg cell and a father's sperm cell. Unlike 'regular cells', egg and sperm cells have only one set of 23 chromosomes (a haploid set), which adds up to you having 46 chromosomes when they are combined (forming a diploid set).[7] (Every egg cell has the X chromosome, while sperm cells can have the X or Y chromosome.)

The structure of the double helix was discovered in 1953 by two scientists, James Watson (b. 1928) and Francis Crick (1916–2004), who revealed their discovery in a two-page paper in the journal *Nature* that ended with the following sentence:[8]

> It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Here they were referring to the paring in (2). This modest sentence has been called the biggest understatement in biology. Both scientists were awarded the Nobel prize for their discovery in 1962 (along with Maurice Wilkins who had contributed to the discovery). Another scholar whose work was essential for Crick and Watson to make their breakthrough discovery, Rosalind Franklin (1920–1958), could not be included in this honor because she had passed away before the price was awarded (those are the Nobel prize rules), likely due to exposure to radiation during her crucial experiments.[9]

---

[4] The bases are divided in two groups: purines (A, G) and pyrimidines (T, C). In each case, one of the purines (A, G) is complemented by one of the pyrimidines (T, C).

[5] Here I ignore the important view that what the offspring inherits from the parent can come from other biological sources than the genes. See: Jablonka, Eva. 2001. The systems of inheritance. In Oyama, Susan, Paul E. Griffiths and Russell D. Gray. Eds. *Cycles of contingencies: Developmental systems and evolution*. Cambridge, MA: The MIT Press. 99–116.

[6] The set of 46 chromosomes form the so-called *diploid* set, which contains two *haploid* sets of 23. The number of chromosomes can be different in some humans in certain types of pathology (such as Down syndrome, Turner syndrome or Klinefelter syndrome). Organisms which reproduce asexually are *haploid*, not *diploid* like humans and most other species.

[7] You will frequently read that each set of 23 chromosomes contains approximately 3 billion DNA letters, but the number is actually 6 billion because each DNA letter is paired with another one forming the steps in the twisted ladder. And because there are two versions of each chromosome, the total number of nucleotides in a cell is thus 12 billion, an astonishing number, also given everything else that is contained in cells.

[8] Watson, James D. and Francis H. C. Crick. 1953. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* 171: 737–738.

[9] A recent assessment of the extremely significant and underappreciated role that Franklin played is: Cobb, Matthew and Nathaniel Comfort. 2023. What Rosalind Franklin truly contributed to the discovery of DNA's structure. *Nature* 616: 657–660.

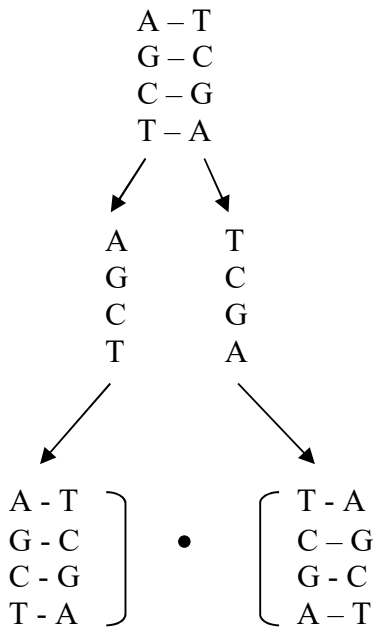
Mitosis and meiosis

Human life begins when a sperm cell enters an egg cell, which is called *fertilization*. As mentioned, sperm cells and egg cells contain only one set of chromosomes (that are called *haploid gametes*). (Gametes are also called reproductive cells or sex cells.) These two cells merge and produce a *diploid zygote*, which is a cell that contains two versions of each chromosome: one is the paternal version, the other is the maternal version. Due to a process called *meiosis*, which involves the division of egg or sperm cells into four different haploid gametes *before* fertilization occurs, the gametes that merge into a zygote do not actually mirror the genes of either one of your parents precisely because the process of meiosis makes them unique by 'switching' stretches of DNA between the four haploid gametes.

Mitosis

Before we discuss meiosis, let us take a closer (but necessarily brief and non-technical) look at mitosis. Before a cell divides into two identical cells, the 'ladders' of all 46 chromosomes are unwound and the two strands of DNA are separated (see step a below). This separation allows each strand to serve as a template to guide the construction or synthesis of another strand (step b). Recall that '§' represents the two strands of DNA that form the double helix. The superscripts 'n' represents the chromosome number.

Duplication of chromosomes



$§^{n(P)}$ (i.e., a paternal chromosome number 'n')

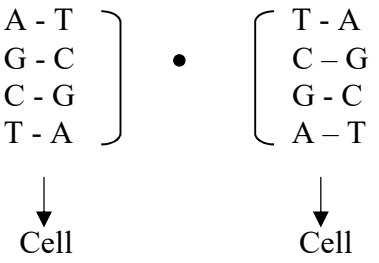
a. Separation of strands of base pairs

b. Matching: making each side complete which results into two identical chromosomes: $§^{n(P)}•§^{n(P)}$

The two identical duplicates of each chromosome, which are called *sister chromatids*, are connected by the so-called *centromere* (here symbolized as **'•'**). Because we start with a total of

46 chromosomes, this means that after duplication, we have 46 chromatid pairs, each member of such chromatid pairs being a complete chromosome (having the double two-stranded helix structure). This makes the next step possible, which is the separation of chromatid sisters into two separate cells.

Separation of chromatid sister into two cells (the actual cell division)



In mitosis, all of the 46 chromosomes are duplicated in this way and the two duplicates are distributed over two cells. This completes mitosis, the result of which is two genetically identical daughter cells, each containing 23 homologous pairs.[10]

Meiosis

While mitosis is the process responsible for generating all the diploid body cells of an individual, there is another process called *meiosis* which is responsible for generating sperm and egg cells that contain only one pair of chromosomes, i.e. 23 chromosomes. Such cells are called *haploid gametes*.[11]

Meiosis is a process by which the diploid sperm or egg cell (which both contain 23 homologue pairs for each chromosome) divides into four gamete cells that contain only one version of the 23 chromosomes. Meiosis begins and ends in the same way as mitosis. First, there is duplication of the DNA and meiosis ends with separating chromatid sister over two cells. What makes meiosis different from mitosis is that in between these two steps, there is an additional step of cell division which involves two crucial processes called *recombination* (or cross-over) and *separate alignment*.

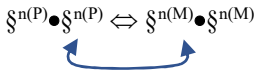
[10] In the process of matching a mistake can occur, which means that one of the cells now has a *mutation*, i.e., a DNA letter in a spot that does not match that spot in the original cell. Setting that aside, following mitosis there are two cells that contain the same DNA sequence and all 46 chromosomes but that are not necessarily identical in function. For example, *stem cells* give rise to one daughter cell that is another stem cell (like itself) and another daughter cell that is more specialized for a particular cell lineage or tissue type. The genomes have the same sequence, but epigenetic factors are now different, resulting in different cells.

[11] Gametes have to be haploid cells because when they merge they form one diploid cell with two pairs of 23 chromosomes. If sperm and egg cells were diploid, the fertilized egg would contain 92 chromosomes and would then double the number of chromosomes with each new each generation.

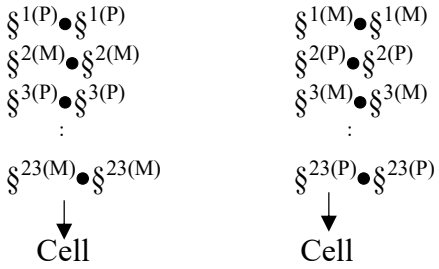
Meiosis

Phase 1: Duplication of chromosomes (see above)

Phase 2: The crucial step
Each chromosome, which contains two chromatid sisters (due to duplication), 'bonds together' with its homologue counterpart, one being the maternal and the other the paternal version of that chromosome. This bonding is here symbolized by the double shafted arrow '⇔'. Bonding then allows *recombination* (or cross-over), which means that genes from paternal chromatids can swap with the corresponding maternal genes, here symbolized with ↔:



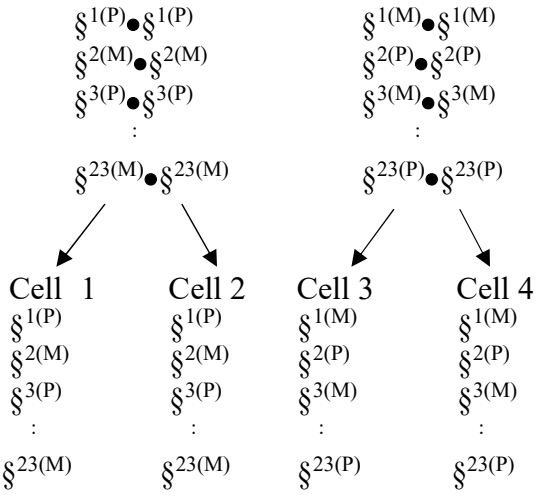
Due to recombination the two chromatid sisters of each chromosome that started out as identical duplicates may now differ in their DNA sequences. Subsequently, the 46 chromosomes (still consisting of chromatid sisters) are separated into two sides of the cell, with, and this is important, on each side a mix of paternal and maternal chromosomes. This is called *independent assortment*. These two sides then split to form two diploid cells, each having 23 chromosomes (i.e. chromatid pairs) that are a mix of paternal and maternal chromatid pairs:



The step that concludes meiosis is that for both cells the chromatid sisters are divided over two cells (which is the same step that concludes mitosis).

Phase 3: Separation of chromatid sister over two cells

The chromatid sisters of each of the cells containing 23 chromosomes are now separated, which then produces four gamete cells from the original sperm or egg cell, each containing 23 chromosomes (which are no longer chromatid pairs). (Meiosis II is thus identical to the only cell division that we see in mitosis.)



The end result of meiosis is four haploid cells that each contain only one set of 23 chromosomes. You might think that Cell 1 and 2 are the same, and Cell 3 and 4 are the same too, but this is not the case. Due to the process of cross-over the DNA sequences in each haploid cell are different from each other. This means that even though you inherited DNA from your parents, the specific DNA sequences that make up the chromosomes that come together during fertilization are not exact copies of the DNA of your parents. They are quite unique.[12]

As a linguist, I am inclined to regard meiosis and mitosis as small rule systems which differ in that one has certain rules that the other is missing:

	Mitosis	Meiosis
Rule 1: DNA duplication	√	√
Rule 2: Recombination	-	√
Rule 3: Independent alignment	-	√
Rule 4: Cell division	-	√
Rule 5: Separation of chromatic sister	√	√
Rule 6: Cell division	√	√

In the process called meiosis, rules 2-4 are interspersed between rule 1 and 5. Rules 1, and 6 are the same for both mitosis and meiosis. The critical rule for meiosis is rule 4 which introduces an

[12] Independent assortment combined with crossing-over along the length of the chromosome yields seemingly limitless possibilities for the resulting genome in the gamete, which makes every person genetically unique.

extra step of cell division which is how you can divide one cell into four. Mitosis has only one cell division rule and thus get you two cells from one cell.

However, it is even more critical for meiosis that the extra cell division is fed by the processes of recombination and independent alignment. Without those, meiosis would simple lead to four identical cells and thus not be different from applying mitosis twice.

It is interesting to think about what meiosis would produce if it had rule 2 (recombination) but not rule 3 (independent alignment) or vice versa. With only independent alignment, meiosis would give you four cells, but two of those would be identical to the other two. As we noted above, due to recombination we get four different gametes. If we just had independent alignment, we would again get four cells, but again, two of those would be identical to the other two.

Both recombination and independent alignment on their own would deliver two different gametes that have genomic structures that are different from the paternal and maternal genome. Given the logical ordering of recombination (rule 2) and independent alignment (rule 3), one might hypothesize that rule 2 evolved first and that rule 3 evolved later to ensure an even greater diversity within the species than rule 2 would deliver on its own.

Interim Summary

All cells of the *embryo* descend from the first, original zygote through the process of cell *division* which is called *mitosis*. Before a cell divides into two identical cells, the 'ladders' of all 46 chromosomes are unwound and the two strands of DNA are separated. This separation allows each strand to serve as a template to guide the construction or synthesis of another strand. Then the cells, which now contain two sets of 46 chromosome pairs, divide in two, each cell getting a complete set of 46 chromosomes. The chromosomes in both 'daughter' cells are supposed to be completely identical, but in mitosis errors can occur, which are called *mutations*. (Mutations can also occur during meiosis.) A cell with mutations can be 'killed' by the immune system, but if not, it can transmit its mutation to daughter cells and that can have bad consequences for the developing organism. It can also be harmless because large stretches of DNA seem to be *non-functional* (see below) which means that mutations in this DNA have no (apparent) effect at all.

Mitosis is thus responsible for the development of a complete organism from the first fertilized egg, the zygote. The first step is that the zygote undergoes cell division to form two daughter cells. Each of these is then divided which produces four cells. This process of division keeps repeating itself, and eventually several dozen daughter cells form a structure known as the *blastocyst*. This blastocyst will implant itself into the wall of the mother's uterus. Once attached, the cells will continue to divide, eventually forming an *embryo*.

Each cell contains the complete set of chromosomes with all their genes, but as the organism develops, cells get *specialized* which means that for each cell type only a subset of the genes will be expressed. This means that skin cells only express genes that are relevant for proteins needed for the skin, while, for example, heart cells only express genes for proteins needed in the heart, and so on.

Identical (*monozygotic*) twins result when the first cell divides into two cells that each start their own process of development. So-called *dizygotic* ('fraternal') twins result when two egg cells are fertilized by two sperm cells. In the first case we get two individuals that are genetically virtually identical, either both boys or both girls (with very rare exceptions). In the latter case, the

sharing of genetic materials is as it is between 'regular' siblings and the twins can have the same or opposite sex.

Alleles

It was mentioned that even though homologous chromosomes contain the same genes, most genes have slight variants that are called *alleles* that can differ mildly in their DNA letters. This means that the haploid sets that comes from Mom and Dad are not necessarily identical; they have the same genes, but not the same gene variants (alleles). In fact, the chance that this would be so is nil. What is important is that if the parents have different alleles of some gene, either one could end up in the fertilized egg. This means that the first cell that became you could have had two variants of many genes. When the gene variants differ, usually one will be *dominant*. The fact that genes come in variants (*alleles*) is perhaps the most basic property of the genome that ensures diversity in all species that reproduce sexually. This diversity is further enhanced due to the fact the gametes that are produced during meiosis end up being completely unique due to the ingenious copying or exchange processes that we already mentioned.

On top of all this, there is the additional factor, also mentioned, that adds to diversity: mutation. As noted, in the duplication phases of both mitosis and meiosis copying errors can occur. This can produce new alleles.[13] If this happens in the non-reproductive cells in our bodies, mutations can cause illness. If mutations occur in the formation of gametes (during meiosis), it may happen that a new allele leads to a fertilized egg that will not develop in the normal way, and depending on the mutation, the life that develops from this egg could be cut short at any time: soon after fertilization, at any time during the pregnancy or after birth. But it is also possible that the person that now carries a specific mutation can live a long life, albeit one with a subtle or non-noticeable (and non-lethal) deviation that may even prove to be advantageous. This person can then pass on the mutation to the next generation and even spread in a species. This is how evolution works!

Variation that is due to alleles has its limits. Humans have many genetically based properties that cannot vary or are very unlikely to vary. If we ignore random mutations, all members of each species share a *basic body plan* which does not vary which means that the genes that underlie this body plan do not have alleles. Hence, notwithstanding the fact that there are perfectly healthy people who have six toes on each foot, all humans have the same bodily design and, one would think, the same overall brain design as well, both allowing for variation that is due to environmental circumstances (which brings in the role of epigenetics). Of course, unlike the rest of the body, the brain is very special in two ways. Firstly, the brain, with all its neural circuits, has an astonishing complexity that has thus far defied a complete analysis. Secondly, unlike other organs in the human body, the brain gives rise to the mind. As we will learn in this chapter, while the routes from genes to brain structure are extremely difficult to map out, how brain activity produces the mind remains one of the big puzzles of modern science.

[13] In humans, given the size of the world's population, even the rarest mutations are likely to be recurrent (appearing more than once somewhere among individuals of our species).

What is a gene?

There are several different definitions of the concept 'gene'. I will here adopt a very common definition, which is that a gene is a stretch of DNA that 'codes for' a specific protein. Genes differ in how many DNA letters they contain: from a few hundred to more than two million. Proteins, which occur inside cells, are the building blocks of life. They are complex molecules that are made up of *amino acids*. There are only 20 amino acids that occur in living organisms, but they can be combined and are strung together into *polypeptide chains*, and it is these polypeptide chains that fold to form 3d structures of proteins. Proteins can contain hundreds of thousands of amino acids. Each protein-encoding gene can be expressed in different ways (see below) and deliver different versions of the protein, which can then be further modified in different ways to increase protein diversity. The route from DNA to proteins is mediated by a process that links triplets of DNA letters (called *codons*) to specific amino acids. This linking follows a universal mapping table that is called the *genetic code* which is universal across all life forms:[14]

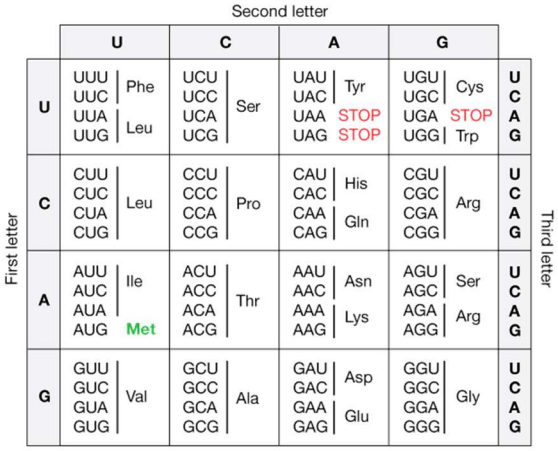


Figure 2: The genetic code[15]

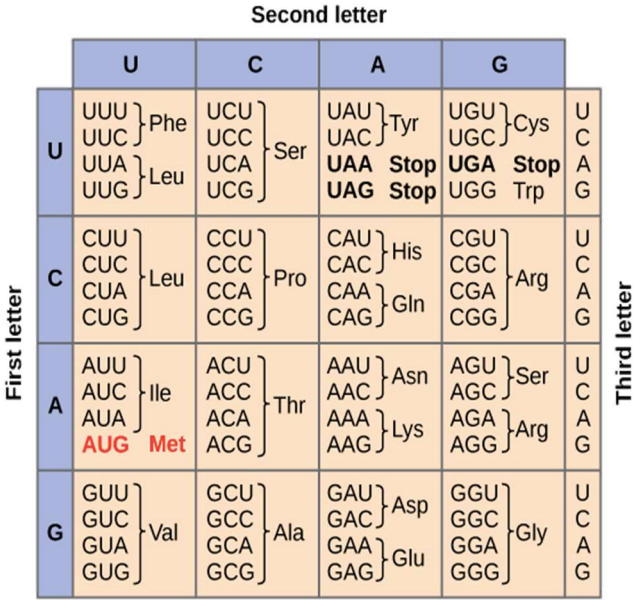
We should note that while there are 64 possible combinations of three DNA letters, these need to code for only 20 amino acids. As a result, in many cases multiple codons code for the same amino acid.[16] Also note that some codons indicate STOP, which means 'end of the protein'.[17] The 'redundancy' in the table is actually a good thing because even if some mutations occur, the necessary protein can still be formed when there are several codons that can deliver the amino acids that it contains.

A current estimate is that there are about 21,000 protein-encoding genes in the human genome. Different proteins perform all sorts of functions in your body, both in internal processes (digestion, for example), in repair if there is damage, or by providing structural coherence to muscles, tissues, organs, bones, etc.

[14] In the codon table, we see U instead of a T. This is because protein building happens via an RNA-based intermediate stage that is derived from the DNA sequence, in which there are uracils instead of thymines.
[15] Image credit: https://openstax.org/books/biology/pages/15-1-the-genetic-code (CC BY 4.0).
[16] There is only a single Met codon and only a single Trp codon. The codon table highlights that (usually) Methionine is the first amino-acid of a protein, hence the green colour
[17] Not necessarily the end of the gene, which can also contain, among others, regulatory elements beyond this point.

Protein formation

Each amino acid corresponds to a three-letter DNA sequence, called a codon. There is a list that specifies the correspondence between specific codons and specific amino acids, which is called the *genetic code* and this list is universal for all life on the planet. For example:



Note that the DNA letter T (see above) here occurs as U. The formation of proteins out of amino acids takes two steps. Protein synthesis occurs in the cytoplasm of the cell. DNA is restricted to the protective confines of the nucleus. In order to get the information contained in the DNA out to the cytoplasm to be used for protein construction, a portion of the chromosome that contains the relevant gene is 'unzipped' and the DNA sequence of a gene is 'copied' into a chemically similar molecule known as *messenger-RNA* (mRNA). The RNA sequence is identical to the protein-encoding sequence in the DNA except that Ts are replaced with Us. The sugar type of the nucleotide is ribose instead of deoxyribose and, also, mRNA is single-stranded and not double-stranded. The process of making the mRNA strand is known as *transcription* (you are transcribing or 'writing' a copy of the gene with the information you want to transport to the cytoplasm of the cell where the protein machinery is found). The process of converting the information in the mRNA into corresponding protein is known as *translation*. In this process molecules called *ribosomes* play a critical role, of 'translating' the RNA message into the amino acid sequence. The sequence for the proteins is formed by attracting an amino acid (floating around in the cell) for each codon. With the three-letter codon code, we have 64 different codes, but many of those correspond to the same amino acid, which means that there are many 'redundancies' in the code.[18] These redundancies actually make the whole process more robust, because when mutations occur in the DNA sequence, this may not make a difference as long as the codon corresponds to the intended amino acid.

In the next section, we will see that not all genes code for proteins. Some genes, called regulatory genes, have the function of turning other genes 'on' or 'off'.

[18] The codons that correspond to 'STOP' indicate the end of the formation of a protein.

About mutations

Because all genes occur twice on the homologue chromosomes, a mutation that is inherited from the parents does not have to have an effect if it only occurs in one of the alleles. When the alleles are different, with potentially different phenotypic consequences, such as causing different eye colours, they cannot both get expressed in an individual, so which one will it be? As we have seen, in case of competing or contradictory alleles, there is a ranking that designates one variant as dominant and the other as recessive, but it can also be a matter of chance.

Genetic disorders of any kind can cause all sorts of problems in the organs of the body, including problems in the brain. Serious genetic disorders are often fatal and either interrupt the development of the fetus or lead to early death. A brain problem may cause some cognitive function (like language) to fail or at least to malfunction, which is manifested in the behaviour of the individual.

Mutations arise from 'copying' errors during DNA duplication when cells divide or from changes in the DNA sequence due to, for example, radiation. Once such a mutation arises is will be duplicated in descendent cells, unless the defense mechanism in the body can 'kill' the bad cell. Mutations that are not destroyed cause new alleles that can distribute themselves through the species. If the mutations lead to an organism that is not viable, the mutation will simple disappear because it has little chance of being inherited by offspring. However, mutations can also have a non-lethal, mild effect that simply adds to the variation among its members of the species.

How many genes and 'junk DNA'

Estimates concerning the number of genes for humans have changed a lot over the years. The latest number I read is that there are 21,306 coding genes.[19] Whether you think that this number is big or small depends on many considerations. When gene research first started, humans, with their alleged 'superior' mental abilities, were expected to have the largest number of genes of any species, but this is not the case. Apparently, *daphnia*, a water flea species, has some 30,000 genes, and many types of fruit (bananas, tomatoes, etc.) apparently have more genes than humans. If members of the daphnia species have the same mental capacities as humans, they have managed to hide that well. In 2003, the so-called *Human Genome Project* sequenced almost all the DNA letters that make up a specific human genome.[20] This project, a large international cooperation, started in 1990 and was completed in 2003. The costs run into the billions. Today, it has become very easy and relatively cheap to sequence the genome of human individuals, which has led to very large databases of human genomes that can now be used for research (and other less noble activities if not well protected). We will come back to such research below. Sequencing a genome (human or of other species), does not mean, however, that this sequence has been fully 'parsed'

[19] Willyard, Cassandra. 2018. New human gene tally reignites debate. *Nature* 558: 354–355.

[20] The project "completed" in 2003 was not actually able to sequence all of the genome, there were still gaps (making up ~8%) that couldn't be determined with the technologies available at the time. The first time "all the DNA letters that make up a specific human genome" were completely sequenced was actually pretty recent, in 2022, work done by the Telomere to Telomere (T2T) Consortium.

into stretches that correspond to genes, which is a much more difficult task that is still ongoing.[21] In other words, having a full sequence of nucleotide letters of a genome doesn't tell you what the actual genes are, let alone their functions.

We have to recognize that there are genes that are said to have a *regulatory role* in the expression of other genes. These regulatory genes encode for proteins, called *transcription factors*, that influence the activities of other genes by binding to a so-called promotor region that precedes these genes, which is thus also called *the transcription site* of the gene. Regulatory genes act as 'on/off switches' for the genes that they control. Such genes can control the promotor region of more than one other gene.

In the genome of all species, the protein coding genes make up only a very small portion of all DNA. In humans, protein-coding DNA is estimated to be only 1–2% of all DNA. The rest used to be called 'junk' DNA. That so much DNA is non-coding is another reason for why the genome is robust, because mutations in non-coding areas do not seem to matter. We have to say 'seem' because based on the sequencing of human DNA in the Human Genome project, geneticists have taken a close look at the complete set of DNA letters, leading to *The Encyclopedia of DNA elements* (ENCODE),[22] which has revealed that a lot of the 'junk DNA' may not be so junky after all. It is likely that as geneticists discover and learn more, genes, including additional regulatory genes, can be identified that play a role in the development of the organism.[23] This line of research suggests that perhaps up to a quarter of human DNA may have some role to play, which then also increases the chances that certain mutations matter. These developing views lead to decreasing estimates of the amount of DNA that was initially called 'junk', which is why this term is no longer deemed meaningful in the field of genetics.

Given that humans have very large and complex brains, it was initially thought that there must be more genes in the human genome than in other species. It is not surprising that geneticists expected to find that the human genome contains many more than the approximately 21,000 genes that they eventually settled on. So, one might ask how it is possible that such an unexpectedly small number of genes can lead to enormous complexity (especially in the brain), not just in humans, but also in other species with large brains and a rich array of behaviors. The answer is that many genes have *multiple functions*, due to 'alternative splicing', which allows using different (combinations of) chunks of the same gene to be activated for the generation of different proteins. Once more, there is much more to tell about this kind of splicing, but I must try to avoid going into too much detail. Apparently, alternative splicing occurs more often in the expression of human genes than in other species. This could, at least in part, be an explanation for the greater complexity of the neural pathways in the human brain when compared to non-human brains.

So-called 'master genes'

The field called *comparative genomics* compares, as the name suggests, the genomes of different species. This has revealed that there are many genes that are shared by all species (often with small differences in parts of the DNA sequence). A related field called *evolutionary developmental*

[21] Richards, Julia E. and R. Scott Hawley. 2011 *The human genome: A user's guide*. 3rd ed. Amsterdam: Elsevier Academic Press.

[22] https://www.genome.gov/Funded-Programs-Projects/ENCODE-Project-ENCyclopedia-Of-DNA-Elements.

[23] As it stands, regulatory genes are not (and never were) part of "junk DNA", they are just protein-coding genes whose job is to regulate activities of other genes.

biology (EvoDevo) investigates how these shared genes play strikingly similar roles in species that are very distant from each other. Necessarily, despite the similarities, these master genes do have different effects during the development of members of different species, due to differences in their DNA sequences and the role of regulatory genes and of environmental factors.[24] Attention to the role of environmental factors establishes a link to epigenetics.[25]

All life forms on our planet go back to one common, very primitive life form, which is why all species share the same *genetic code* (which is, recall, the translation table from triplets of DNA, called *codons*, to the amino acids that form proteins). Master genes can be traced back to these early beginnings of life.

One example of the master genes are the so-called *Hox genes*, which is a group of regulatory genes that are essential for the development of the body plan of organisms. Hox genes are very old in an evolutionary sense, having evolved long before the diversification of the species that we know today. These genes are regulatory genes because they produce proteins (transcription factors) that control the activations of many genes that are necessary for *morphogenesis*, i.e., the development and formation of the body plan characteristics of a species. Hox genes (among many other genes) play an important role in controlling, for example, the genes that build proteins for the development of legs, arms, fingers, but also internal organs such as the heart and the brain, all of which have very similar design properties across a wide range of species. To give a simple example, assuming there are genes that are responsible for growing 'limbs', these regulatory Hox genes can then control *how many limbs* are to be formed, i.e., how often the 'limb building genes' need to be expressed, since not all animal species have the same number of limbs. Hox genes display hierarchies in the sense that hierarchically higher-level Hox genes can control lower level Hox genes. Again, this shows the startling complexity of the genetic system. If you thought that studying the complexity of language was not challenging enough, try genetics. That said, at the end of this chapter, I will briefly address the numerous articles and studies that have uncovered many interesting analogies between the genome and language.

Genome, genotype and phenotype

The organism that eventually results from the information in the genome represents, at least in part, the *nature* part of us, and of other species, here taking into account that some other biological factors (e.g., present in the mother's egg) and (epigenetic) processes that play a role in development can also be regarded as being part of our nature. We say that all members of a species have the same genome as part of their nature. However, as was already made clear, individuals can have differences in their specific DNA sequences, due to having different versions of genes that are called alleles or due to random mutations. An individual's combination of genes is referred to as one's *genotype*. We then use the term *phenotype* for the characteristics of an organism that develops *in a given environment*. In some cases, it is possible to pinpoint a direct relation between a specific gene and a specific trait of a phenotype (we then speak of *monogenic traits*), but in many cases properties of the phenotype are influenced by a small or large number of genes that are somehow working together; these traits are called *polygenic*. Also, a single gene, especially

[24] Hall, Brian K. 2012. Evolutionary developmental biology (Evo-Devo): Past, present, and future. *Evo Edu Outreach* 5: 184–193.

[25] We have seen in Chapter 4 that *morphogenesis*, i.e., the development of the morphological characteristics of species is also dependent on general, physical laws that put limits on what kinds of forms evolutionary factors can produce.

regulatory genes, can play a role in the development of many different traits, which may not have an apparent close relationship or function in the body. For example, the trait of albinism (having white skin) is caused by a gene-dependent melanin pigmentation in the skin, but this same gene also produces an effect in other bodily properties like the eyes and one's hair color.

The character and behavioral traits, as well as cognitive abilities of an individual, are a product of the structure and operation of the brain. The brain develops, just like other parts of the body, due to the expression of genes that impact on the development of brain cells and neural circuits, in many cases under the influence of environmental factors. The problem in pinning down the ultimate causes of cognitive properties is that a lot of different genes bear on the formation and development of the brain. But even if we knew how properties of the brain relate to the many genes that are relevant, we face the additional problem that the brain is so 'insanely' complex (happily leading to 'sane' brains in most cases) that we would still have a hard time pinning down the exact causes of phenotypic behavioral and character traits, including the various cognitive capacities within the brain (such as the capacity for language). As mentioned in Chapter 6, the total set of neural connections (called the connectome) has a staggering complexity which is largely uncharted. For this reason alone, there is no way of telling what the role is of genes in this highly complex structure.

Mutations and biological evolution

Returning to the concept of mutation, genetic mutations can arise due to different factors. A genetic atypicality is called *congenital* if it is *inherited* from one of the parents. This means that that mutation is present in the gametes of one of the parents (or both). Such a genetic atypicality that comes from either the father or the mother may have come from *their* parents or may have arisen *de novo* in the parents' sperm or egg cells, by chance or due to specific environmental factors. Secondly, a mutation can also occur during meiosis, the process that divides egg/sperm cells in four different gametes. A genetic atypicality is called *acquired* if it arises *de novo* in the embryo's early development or in the fetal stage, due to errors in the process of cell division (mitosis). Acquired mutations can also arise during a person's whole lifetime for a variety of reasons.

While mutations are often bad news, some such mutations can be beneficial because they offer an individual an advantage of some kind, given the environmental conditions that a species faces. Beneficial mutations are thus an important driving force behind biological evolution, following the idea that the advantageous mutation may make the individual more likely to survive and thrive in life which may lead to more success in producing offspring. This offspring may then inherit this mutation and thus have the same advantage. The beneficial gene variant can then spread throughout the species and become a property of all its members.[26]

[26] The spreading (but also the disappearance) of a gene variant in a population is sometimes called a *genetic (or selective) sweep.* This concept should not be confused with *genetic drift* which is usually used to refer to gene variants which undergo random fluctuations in allelic frequency (i.e. becoming more or less frequent without any impact on phenotypic outcomes, for example after a population bottleneck).

An accumulation of beneficial mutations that reinforce each other can eventually produce a new species; this is called *speciation*.[27] Speciation also often occurs when a subpopulation of a given species moves away or is separated from the main group, living a life that meets different environmental challenges, resulting in its evolution into a new species, while the group that they came from stays the same, or changes much less and/or in different ways. In this process, we can say that the environment selects the gene variants, the alleles, that allow individual organisms, and indirectly the species, to thrive (or at least survive). This is what Charles Darwin came to understand; he called it *natural selection*. His idea was inspired by the *selective breeding* or *artificial selection* that, for example, dog breeders had practiced for a long time.

Whether or not a mutation (that is not fatal) is in fact beneficial thus depends on the environmental challenges that the members of a species face. A standard example is this. Suppose the main source of nutrition of some species consists of leaves that grow on trees and suppose that due to periods of droughts or other climatic conditions, the leaves can only be found higher up in the trees. It would then be advantageous for members of this species to be able to reach those higher leaves by having a somewhat longer neck; of course, longer legs or being able to jump well would help too. This is where chance plays a role; mutations are unpredictable. A mutation that leads an organism to have a somewhat longer neck that exceeds the variation in the length of the neck that occurs in any species that have necks, will give the 'owner' of that mutation an advantage. Hence whatever allele is responsible for these properties is likely to spread throughout the species. To put it crudely, short-necked members simply get less nutrition and may die before they have a chance to produce offspring, or they produce less offspring. Somewhat longer-necked animals will continue to benefit from their physical properties so there will be a tendency for necks to become longer and longer as the leaves continue to only grow higher up in trees. And this is how you get giraffes (who, by the way, also have very long legs)!

As we will discuss in Chapter 9, when faced with climatic changes, species are 'forced' to adapt or else go extinct. The evolutionary trajectory toward our more recent ancestor species (that walk upright and have big brains) may have been caused by dramatic shifts in climate in certain parts of Africa (specifically the so-called Great Rift Valley, an area in central-north Africa) which puts high demand on the need to continuously adapt, likely being the reason for why our own species is perhaps the most adaptable species in the natural world.[28]

Another factor that can lead to species developing certain characteristics over time is called *sexual selection*. Think of the enormous tail of the peacock. It is believed that male peacocks with bigger tails are favored by female peacocks, maybe among other reasons that we do not know, because being able to drag around such a heavy tail must mean that these males are very strong and fit, and thus good mates.[29] Certain changes are thus selected for because they increase the chance of being chosen as a mate. As we will see in Chapter 8, many animal species, especially the male members, display behavior (often called communicative) that seemingly serves that function.

Due to natural and sexual selection, gene variants that are beneficial for survival and/or reproduction will spread in the population and gene variants that do not give such advantages (or

[27] Some scholars attribute an important role to phenotypic plasticity, which precedes genetic accommodation to changes in behavior that fall within the range of variable phenotypic traits that occur within a species. See: West-Eberhard, Mary. 2003. *Developmental plasticity and evolution*. New York: Oxford University Press.

[28] Whether *Homo sapiens* (that's us) will evolve or go extinct given current climate change remains to be seen. Our species can make adaptations that are not genetic, but rather *cultural*, based on our ability to invent techniques to endure or counterbalance climate change.

[29] See: https://www.nature.com/scitable/blog/accumulating-glitches/an_introduction_to_sexual_selection/

are even disadvantageous) will tend to disappear over time. Darwin's claim to fame is the factor of natural selection, but as with so many other insights that are now commonplace, he anticipated most other factors that play a role in evolution. As mentioned, the properties that both types of selection can cause are not unlimited. For example, there are purely physical constraints on how long necks can become. In Chapter 4, we have discussed how evolutionary scientists have long recognized that evolutionary change is curtailed by constraints (which could be called 'third factors'; see Chapter 4) on the physical shape of organisms.

Epigenetics

In the field of genetics, epigenetics is generally accepted to involve changes 'above' the genome (As, Ts, Cs and Gs all remain the same but chemical groups can be added or taken away from specific Cs in the sequence or other proteins the DNA associates with). The most commonly discussed epigenetic changes involve the addition of methyl groups to Cs that are located next to a G and also adding or removing acetyl groups to histone proteins that the DNA wraps around. Various stimuli can result in altered gene expression and maintaining expression of specific genes. The idea that learning a language involves an epigenetic process is perhaps unorthodox. It is perhaps questionable to say that learning a new word, or other aspects of language involves an epigenetic process. The reasoning here is that learning involves creating or strengthening neural pathways and that this requires expression of the genes that underlie such alterations of these pathways.

Several times, in the preceding sections (and in preceding chapters), I mentioned that gene expression can be influenced by *environmental factors*. I also have mentioned that the Evolutionary Developmental (EvoDevo) approach pays attention to the role of environmental factors in the development of phenotypes. According to some researchers, the role that environmental factors play in gene expression, both pre-natal and post-natal, have long been underestimated. A field that specifically focuses on the regulatory changes that are dependent on cues from the body-internal and body-external environment on gene expression is called *epigenetics*. Epigenetics is thus not about changes in the DNA sequences. (This is not to deny that external factors such as certain extreme form of radiation can cause changes in DNA.)

To explain the term, 'epi-' means 'over, outside of, around'. The idea, and the term, was originally proposed by Conrad Waddington (1905–1975) in 1942, with specific reference to the pre-natal differentiation and specialization of cells in the formation of phenotypes.[30] He introduced the notion of *canalization* which causes a uniformity in the phenotypes of a species by imposing a limit on the effects of the environment. As just mentioned, there are, for example, physical constraints on the *morphology* (i.e. the physical form) of all species.

Another scientist who pioneered epigenetic notions was the developmental psychobiologist Gilbert Gottlieb (1929–2006), who we already met in Chapter 3. Gottlieb, like Waddington, was ahead of his time, paving the way for the now widespread attention to epigenetics.[31]

---

[30] See: Waddington, Conrad H. 1942. Canalization of development and the inheritance of acquired characters. *Nature* 150 (3811): 563–565; Waddington, Conrad H. 1956. The genetic assimilation of the bithorax phenotype. *Evolution* 10: 1–13.

[31] See: Gottlieb, Gilbert. 1997. *Synthesizing nature and nurture: Prenatal roots of instinctive behavior*. New York: Lawrence Erlbaum Associates. Also: Sameroff, Arnold. 2010. A unified theory of development: A dialectic integration of nature and nurture. *Child Development* 81: 6–22.

To explain further what we mean by epigenetics, we go back to the very first cell that results from merging the sperm gamete and the egg gamete. The question is: How does the environment (both body-internal, e.g., neighboring cells and organs, and body-external, e.g., nutrition, stress) affect the development of the organism? Or more accurately: How does the environment influence the expression of genes? This question is pertinent since every cell contains all the genes in the human genome. But it is clearly not the case that all these genes are expressed in every cell. We have already mentioned that cells get to be specialized, depending on where they are located in the body. While initially the development of a fertilized egg leads to cells that are not yet very specialized, so-called *stem cells* (that are *pluripotent*), soon thereafter the cells strongly specialize, which means that in each cell most genes are 'turned off'. This is possible due to the 'on/off' switches that are part of the promotor region of genes. As was discussed, these switches are manipulated by the *transcription factors* that attach themselves to the promotor region of genes.

In early stages of development of the fetus, the attachment of transcription factors is due to neighboring cells that are part of the developing organism. This is how cell specialization takes off. I refer to such effects as involving the body-internal environment because they arise *internally* to the developing organism, and they arise *necessarily* for the organism to develop as it is supposed to develop. In short, cells can get signals from neighboring cells that tell them how to specialize in a certain way. While cell specialization that is due to internal factors falls under the heading of epigenetics, epigenetics is also crucially involved when gene expression is influenced by body-*external* factors that the mother is exposed to (i.e., what she eats, drinks, breathes in, etc.).[32] These factors are here called 'external' because they come from outside the developing organism, which may still be inside the body of the mother. After birth, the organism endures many external influences that are caused by (interaction with) the environment, including choice of nutrition, air, sunlight, that are due to the environment or to behavioral choices of the organism itself. Here we must also include psychological factors such as 'stress', feeling happy (better), doing meditation, etc.

It is clear that epigenetics is of direct relevance to the nature–nurture debate. Epigenetics teaches us that the external environment ('nurture' in a broad sense; here including both external prenatal and postnatal factors) is not a factor that independently of the genes influences the development of our bodies and minds ('nature'). Rather, the environment has a direct influence on inhibiting or accelerating gene expression. We must be cautious of one possible misunderstanding: epigenetic effects do not change the DNA of genes. The DNA sequences that make up the chromosomes remain the same. Rather, the epigenetic effects influence the expression of genes.[33] Of course, there are many limits in how the environment can impinge on gene expression. All genes that are responsible for the basic body plan of humans are not sensitive to epigenetic effects. That said, the development of the body can be significantly altered by medications that the mother takes during the pregnancy.[34]

We now see that the phenotype is the result of the intimate interaction between the genome (or rather genotype) and what is called the *epigenome*:

---

[32] I made up the distinction between 'internal' and 'external' for expository purposes. I believe it is not used in the official epigenetic literature. Likely, the difference cannot always be clearly made.

[33] A nice popular book about the relevance of epigenetics for the nature-nurture debate is: Ridley, Matt. 2003. *Nature via nurture: Genes, experience, and what makes us human*. Harper Collins Publishers.

[34] A medication that was prescribed to pregnant woman to treat morning sickness in the late 1950s turned out to cause severe birth defects in children. See: Vargesson, Neil. 2015. Thalidomide-induced teratogenesis: History and mechanisms. *Birth Defects Research, Part C: Embryo Today* 105/2: 140–156.

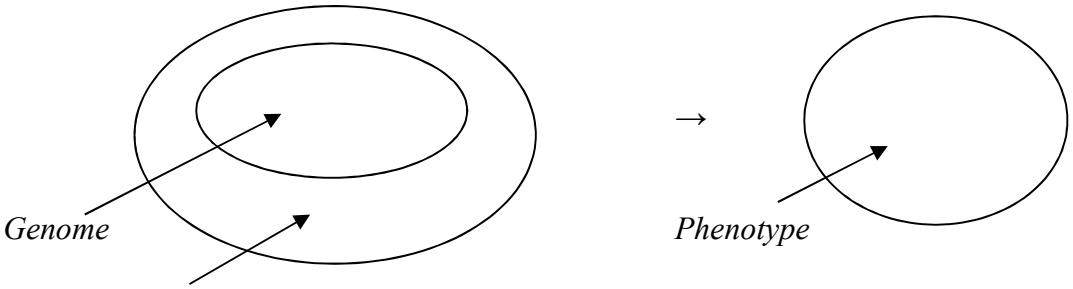


Figure 7.1: Genome, epigenome and phenotype

Thus, *nurture influences nature*, rather than both being independent forces that shape an organism. This is a rather different view from the more traditional (and still often held) view that nurture and nature each have their own *additive* contribution to the way that an organism develops. Epigenetics, which is a blossoming field in the broader context of genetics, provides a scientific perspective on the nature–nurture debate. It has of course not yet provided 'all the answers', but there is a general feeling that it will provide many new insights in the near future.[35]

A special branch of epigenetics is called *Behavioral Epigenetics*, which focuses specifically on the way in which the environment influences behavior (by controlling gene expression in the brain).

Epigenetic tags

The 'tags' are chemical groups added either to the DNA (specifically CpG dinucleotides) or Histone protein tails. The most often discussed chemical 'tags' are those of methyl groups added to the C of CpG dinucleotides in promoter sequences of a gene and also the addition or removal of acetyl chemical groups to the tails of histone proteins. The combination and quantity of these methyl and acetyl groups will impact access to the gene by transcription and regulatory proteins and therefore determine if a gene is able to be 'on' (expressed) or not. Transcription factors are the proteins needed to activate transcription. If transcription factors cannot come in contact with the gene's promoter or regulatory sequences then the gene remains 'off' (not expressed). There are several families of proteins that are responsible for switching the epigenetic marks around, most prominently: HATS (histone acetyl transferases, which add acetyl groups to histone and allow genes to be 'on'), HDACS (Histone deacetylases, which remove acetyl groups and turn genes 'off'), and DNMTs (DNA methyltransferases which add methyl groups to DNA CpG and also turn genes 'off').

Transgenerational epigenetics effects: Lamarck's victory

We could ask whether Darwin's notion of natural selection may also apply to the possibility of what is called *transgenerational epigenetic inheritance*.[36] This is a question that has received a lot

[35] Armstrong, Lyle. 2013. *Epigenetics*. New York: Garland Science (Taylor and Francis Group).
[36] A key reference for this concept is: Jablonka, Eva and Marion J. Lamb. 2005. *Evolution in four dimensions: Genetic, epigenetic, behavioral and symbolic variation in the history of life*. Cambridge, MA: The MIT Press.

of attention, even among the earlier proponents of epigenetics, such as Conrad Waddington. Because, as mentioned, epigenetic processes do not alter the DNA sequences in any way, you might think that epigenetic effects are not heritable, i.e., cannot be passed on from one generation to the next. This is different from saying that a pregnant woman's body influences the development of the child that she carries, which is in a sense a next generation. But in this case the epigenetic effect occurs during the development of the embryo, potentially caused by nutrients and other substances that enter the mother's body. Smoking is bad for the unborn baby and the mother.

This raises an important question: Can epigenetic 'tags' that arise during the lifetime of a person, and are attached to their DNA sequences, be transmitted to their children? For this to happen, firstly, the tags would have to be present in the reproductive cells. Secondly, it would then have to be possible that these tags 'survive' the process of meiosis and thus are potentially present in the gametes that can form a fertilized egg. The question is whether this so-called *transgenerational epigenetic inheritance* is possible.[37]

It is important to know that epigenetic tags that are present in reproductive cells are usually wiped out before the paternal and maternal chromosomes fuse into the first cell. This means that the fertilized egg can start 'fresh' without the 'clutter' of the life experience of the parents. However, it has been shown that this erasure does not always happen completely. A subset of regulatory regions of genome sometimes might not get "reset" or wiped clean. This means that some acquired properties, i.e., properties (in the form of epigenetic tags) that a person acquires during their lifetime due to external circumstances *can* be inherited by their offspring.[38] However, examples (especially in humans) seem to be quite rare and, when they are proposed, tend to be rather controversial.

An early example of transgenerational epigenetic inheritance comes from a study of descendants of people who experienced a famine during the last year of World War II in the Netherlands. Children born during the famine were smaller than those born the year before the famine, which is not surprising because being inside the mother, they likely did not get enough nutrition because the mother did not get enough nutrition. However, it turned out that later in life they could be overweight and suffer conditions that shortened their life span. This suggests that something else is going on, although the details of how this came about are apparently still not determined. A further crucial point is that it has also been claimed that certain of these effects could be found in subsequent generations, albeit it in a limited way, perhaps not going beyond one or two generations. The phenomenon described here is sometimes referred to as *Dutch Hunger Winter Syndrome.*[39]

---

[37] Lacal, Irene and Rossella Ventura. 2018. Epigenetic inheritance: Concepts, mechanisms and perspectives. *Frontiers in Molecular Neuroscience* 11: 292. Bošković, Ana and Oliver J. Rando. 2018. Transgenerational epigenetic inheritance. *Annual Review of Genetics* 52: 21–41.

[38] Verhoeven, Koen J. F., Bridgett M. von Holdt and Victoria L. Sork. 2016. Epigenetics in ecology and evolution: what we know and what we need to know. *Molecular Ecology* 25: 1631–1638. Jablonka, Eva and Gal Raz. 2009. Transgenerational epigenetic inheritance: Prevalence, mechanisms, and implications for the study of heredity and evolution. *The Quarterly Review of Biology* 84: 131–176. Tollefsbol, Trygve O. Ed. 2019. *Transgenerational epigenetics*. 2nd edition. London: Academic Press.

[39] Painter, Rebecca C., C. Osmond, P. Gluckman, M. Hanson, D. I. Phillips and T. J. Roseboom, 2008. Transgenerational effects of prenatal exposure to the Dutch famine on neonatal adiposity and health in later life. *BJOG: An International Journal of Obstetrics and Gynaecology*. 115: 1243–1249. Another study that brought transgenerational effects to the forefront was carried out in Sweden: Kaati, Gunnar, Lars Olov Bygren and Sören Edvinsson. 2002. Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period. *European Journal of Human Genetics* 10: 682–688.

Jean-Baptiste de Lamarck (1744–1829) proposed a theory of evolution in which characteristics that a person acquires during his lifetime *can* be transmitted to the next generation. His theory was surpassed by Darwin's theory of evolution, and he was ridiculed for this claim that acquired characteristics are heritable, especially by August Weisman (1843–1914), who conducted an experiment wherein he removed the tails of rats over five generations, making the point that at no time rats were born without tails, or shorter tails for that matter. While it is obviously true that many acquired characteristics are not, or likely cannot be, passed on to next generations, it would seem that the discovery of transgenerational epigenetic inheritance validates Lamarck's idea to a very limited extent. It's quite a long way from Lamarck's vision to the limited cases mentioned here, and some would say that the similarities are superficial.

Epigenetics and imprinting

In the context of transgenerational epigenetic inheritance, *genomic imprinting* refers to the fact that paternal or maternal genes are sometimes passed on in such a way that their expression is blocked by epigenetic tags. In Chapter 3, we discussed a different notion of imprinting in the study of animal behavior, where it refers to the fact that animals sometimes learn something only if they are exposed to the right stimulus during a specific period, the so-called critical period. We referred to this phenomenon specifically as *behavioral imprinting*.

Behavioral imprinting was discovered by one of the founders of *ethology* (the study of animal behavior), Konrad Lorenz (1903–1989), who, more or less by accident, discovered that goslings will follow as their 'caregiver' whatever or whoever appears to them within a small time interval after birth, the so-called critical period.

We should ask whether behavioral imprinting, which arises during development, has a genetic basis. The answer must be yes. A specific kind of behavior is due to activity in a 'neural circuit' in the brain. To activate a neural circuit, there must be an appropriate stimulus, i.e., an external factor that *epigenetically* triggers gene expression that causes the neural activity. We can now define behavioral imprinting as the genetically-guided activation of a neural circuit that is sensitive (a) to a specific time interval in the development of the organism and (b) to a specific kind of external stimulus.

An interesting question is *how* early relevant stimuli can lead to imprinting. If the relevant environment includes the prenatal environment, right from conception, then it could be that alleged fully instinctual behavioral properties, i.e., properties that apparently need no trigger at all, *are* in fact triggered by very early cues that are already present during the development of the fetus. Indeed, Gilbert Gottlieb (see above) attributed a very significant role to environmental, including prenatal, factors in explaining imprinting and critical period effects, thus emphasizing the effects of nurturing (the environment) over innately guided behavior.

To maintain the notion of innateness, a nativist would have to assume that some properties of the brain/mind do not require *any* epigenetic factors to emerge. However, another response could be that prenatal imprinting that is part of the natural process of ontogenesis should be included in what counts as innate.